Data Interpretation Based on Embedded Data Representation Models

Analytical Models for Effective Online Marketing in the Fashion Industry

Embedded Data Representation Model に基づくデータ解釈に関する研究

ファッション業界を事例とした
効果的なオンラインマーケティングのための分析モデル

July, 2023

Ryotaro SHIMIZU
清水　良太郎

Data Interpretation Based on Embedded Data Representation Models

Analytical Models for Effective Online Marketing in the Fashion Industry

Embedded Data Representation Model に基づくデータ解釈に関する研究

ファッション業界を事例とした
効果的なオンラインマーケティングのための分析モデル

July, 2023

Waseda University
Graduate School of Creative Science and Engineering

Department of Industrial and Management Systems Engineering,
Research on Applied Information Science

Ryotaro SHIMIZU
清水　良太郎

Contents

Chapter 1

Introduction

This chapter introduces the background and aim of this study. In addition, the content of each chapter is explained herein.

1.1 Research Background

Online marketing has become increasingly prevalent, with the widespread use of digital devices to cater to customers and facilitate global sales. Consumers use these devices to browse e-commerce sites, social media, blogs, and video streaming applications. Furthermore, they prefer obtaining product information and completing purchases through e-commerce sites. Consequently, it is crucial in the realm of online marketing to prioritize the provision of a seamless and user-friendly experience across all stages leading to a digital purchase.

The significant characteristics of online marketing include the following.

- The ability to conduct marketing activities worldwide, regardless of the area or location of residence.
- The impossibility of directly seeing the customers.

Therefore, online sites have significantly more potential for reaching and engaging users in marketing compared with traditional brick-and-mortar stores. Hence, the accumulation and utilization of data are essential in this context. Based on these characteristics, in recent years, various companies have strived to enhance their websites' and applications' functionality, usability, and overall user satisfaction by harnessing the wealth of data stored in their databases.

One of the challenges faced in online marketing is the absence of in-person store clerks who can assist users in making purchasing decisions. This issue is particularly relevant in the fashion industry, which is the focus of the current research. When browsing fashion e-commerce websites, users often encounter questions such as “what items can I combine?,” “which item should I purchase?,” “where can I find the specific item I’m looking for?,” and similar queries. Moreover, the fashion industry is known for its inherent ambiguity, making it even more difficult for users to independently make purchasing decisions, especially when it comes to expensive items or trying out new fashion trends through online platforms.

By solving these user questions and difficulties, companies engaged in online marketing aim to realize a comfortable service for users. Therefore, these companies aim to develop a data-driven business that enhances usability and user satisfaction by acquiring essential knowledge from data and leveraging it in the following cycle.

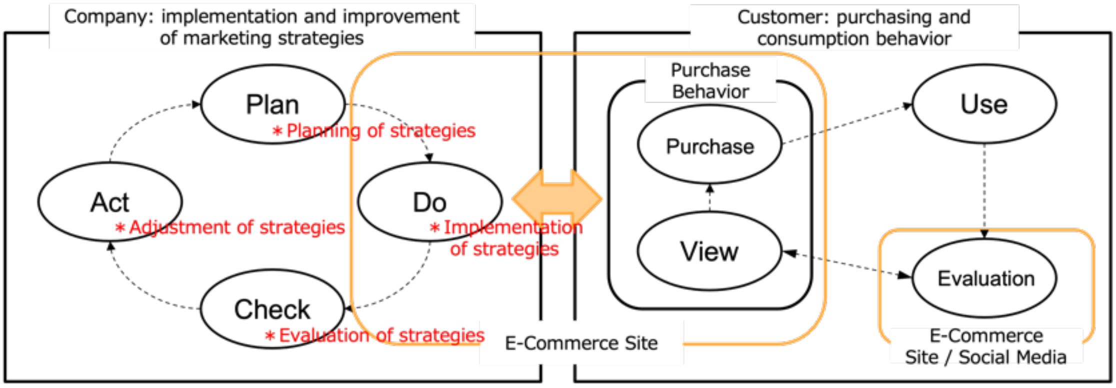

Company: implementation and improvement of marketing strategies
Plan
∗Planning of strategies
Act
∗Adjustment of strategies
Do
∗Implementation of strategies
Check
∗Evaluation of strategies
E-Commerce Site
Customer: purchasing and consumption behavior
Purchase Behavior
Purchase
View
Use
Evaluation
E-Commerce Site / Social Media

Figure 1.1: Overall view of the PDCA cycle for implementation of marketing strategies within a company and customer action cycle in online marketing

Furthermore, users are recommended several items by e-commerce sites through recommender systems. In this case, the users have to supplement the information as to “why this item is good (why it is recommended to him/her)” on their own, making it difficult for non-experts to make purchasing decisions, particularly for expensive items.

In addition, it is common on e-commerce sites and social media for full-body clothing images to be posted in the following manner through a function that adds tags to each image [1, 2, 3, 4, 5].

Multiple tags are assigned to each image as attribute information by contributors. These tags encompass both concrete and straightforward expressions (e.g., “denim,” “skirt,” “t-shirt,” etc.) as well as ambiguous expressions (e.g., “spring outfit,” “formal,” “casual,” “office-casual,” etc.). Once a specific tag is assigned, it is considered correct regardless of the contributor’s sensitivity. However, ambiguous tags are characterized by their dependence on the individual contributor’s

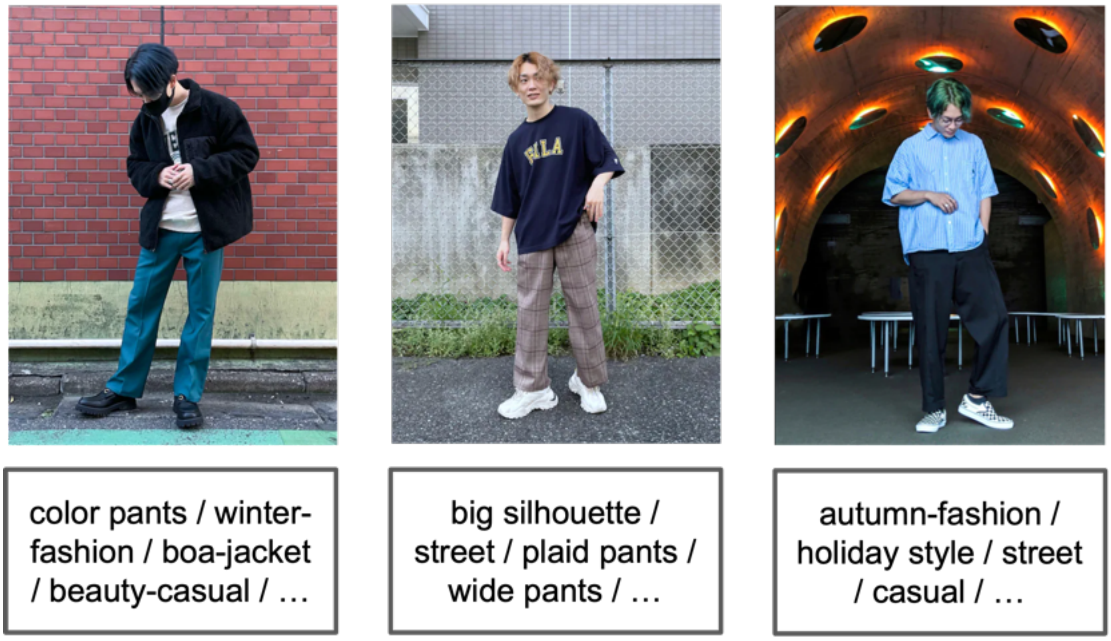

color pants / winter-fashion / boa-jacket / beauty-casual / …
big silhouette / street / plaid pants / wide pants / …
autumn-fashion / holiday style / street / casual / …

Figure 1.2: An example of how fashion images are described on e-commerce sites and social media [6]

sensibilities and may or may not be assigned. For instance, if contributor A perceives image A as entirely "casual," it is appropriate to assign the "casual" tag. Conversely, if contributor B considers image A as partially casual, they may choose not to assign the "casual" tag. Moreover, if contributor C deems the expression "adult-casual" more suitable than "casual," they would assign "adult-casual" instead. The requirement for users, particularly non-experts, to interpret these ambiguous expressions themselves is a primary reason for the challenges encountered in the fashion domain.

1.2 Research Purpose

Despite the previously mentioned challenges in online marketing, users actively desire a seamless online purchasing experience. Specifically, general consumers' recognition and evaluation of fashion items can be ambiguous, making online user support vital for addressing user queries effectively.

Therefore, this research employs machine learning techniques to develop a system that assists users in comprehending and interpreting fashion items. The primary objective is to enhance online usability and improve user satisfaction. Consequently, the present study aims to alleviate online marketing challenges and meet the evolving needs of users.

This study presents two main approaches as proposals.

- Approaching the contact point between the company and users [7].
- Approaching the steps for evaluating and specifying the fashion item during the users' purchasing processes [8, 9, 10].

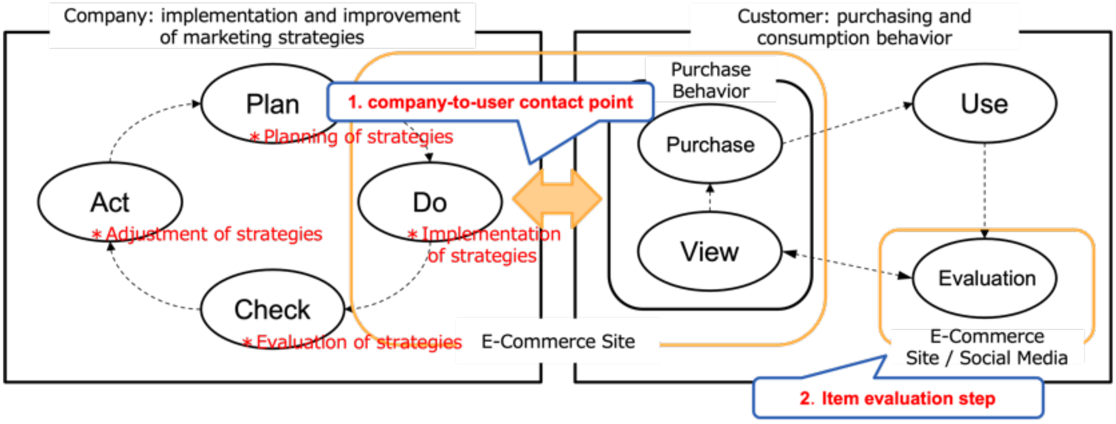

Company: implementation and improvement of marketing strategies
Customer: purchasing and consumption behavior
Plan
＊Planning of strategies
1. company-to-user contact point
Act
＊Adjustment of strategies
Do
＊Implementation of strategies
Check
＊Evaluation of strategies
E-Commerce Site
Purchase Behavior
Purchase
View
Use
Evaluation
E-Commerce Site / Social Media
2. Item evaluation step

Figure 1.3: Positioning of each approach proposed in this study

The point of contact between a company and its customers has been the subject of several studies, particularly in the context of recommender and retrieval systems. In Chapter 3, the focus shifts to the technology of "explainable recommendation," which aims to enhance the explanatory capabilities of recommender systems. The objective is to develop this technique and propose a model (referred to as model 1) that efficiently learns and utilizes vast amounts of diverse side information for explanation purposes. The proposed model 1 is based on a graph neural network model that leverages a knowledge graph and employs the self-attention mechanism to facilitate intrinsic explanations within the model. The aim is to enable users to comprehend the strengths and reasons behind the recommendation, ultimately contributing to the realization of an effective recommender system.

During the users' purchasing process evaluation stage, they assess items across various platforms such as e-commerce sites, social media, blogs, and video streaming applications. Particularly in recent years, users have found it easier to refer to other people's outfits through social media and similar services during this evaluation phase. In Chapters 4-7, a new technology called the "fashion intelligence system" is proposed. This system aims to automatically interpret ambiguous fashion images, supporting users in understanding fashion and assisting them with various fashion-related decisions, including item purchases and styling choices, through different applications.

The fashion intelligence system is based on the visual-semantic embedding (VSE) model technology. We propose three types of VSE models.

Proposed model 2 VSE can map a massive amount of full-body outfit images with abundant tags containing various ambiguous expressions into the same space using foreground-centered learning, background regularization, and other schemes [8].

Proposed model 3 Partial VSE (PVSE) that enables sensitive learning of each part [9].

Proposed model 4 Dual Gaussian VSE (DGVSE) enables the analysis of the meaning and diversity of mapped elements, such as outfits, items, and ambiguous expressions [10].

Chapter 4 focuses on proposing a "fashion intelligence system" based on the VSE (model 2). The mapping mechanism of the VSE is achieved through foreground-centered learning, background regularization, and other approaches. Additionally, various applications of the proposed VSE are presented. Despite its relatively simple structure, model 2 offers a range of applications such as image retrieval, re-ordering, and Attribute Activation Map (AAM) creation. Furthermore, the effectiveness of each application is demonstrated through multifaceted evaluation experiments utilizing real-world service datasets.

In Chapter 5, the PVSE (model 3) is proposed, enabling the extraction of features for individual components of a full-body outfit, including the hairstyle, face, jacket, t-shirt, pants, and shoes. This contrasts model 2, which learns from full-body images as a whole. The objective of Chapter 5 is to cater to the specific needs of users by focusing on designated parts, which is not feasible with model 2, and address more detailed aspects of fashion.

Chapter 6 introduces the DGVSE model (model 4), which maps each element onto the projective space as a distribution rather than a single point. This approach enables a detailed analysis of the meaning of mapped elements and the diversity of their applications, which is impossible with models 2 and 3. This chapter aims to enhance users' understanding of ambiguous fashion images by providing a more comprehensive view.

1.3 Structure of Dissertation

This study consists of eight chapters, and the overviews of each chapter are as follows.

Chapter 2 of this study focuses on the target problem and related studies. It provides an overview of a wide range of machine learning techniques employed for user support in online marketing, while also clarifying the position of each proposal within this research.

Chapter 3 introduces a model for interpreting the rationale behind recommendations, utilizing an explainable recommendation model that leverages diverse side information. The proposed model efficiently learns and leverages various types of side information accumulated on e-commerce sites to quantify the contribution of each piece of side information to a user's purchase decision. Through extensive evaluation experiments and real-world data analysis, the chapter demonstrates that the proposed model maintains high recommendation accuracy while reducing computational complexity. Moreover, the potential and practical usefulnesses of the proposed approach are showcased by exploring its applications in various scenarios.

In Chapter 4, a novel technology and research area called the "fashion intelligence system" is presented to support users in fashion-related decisions. The system employs automatic interpretation of ambiguous fashion expressions, enabling users to obtain answers to challenging and intricate questions. By conducting multifaceted evaluation and analysis experiments using real-world data, the chapter effectively demonstrates how the proposed system facilitates users in selecting and acting upon various aspects of fashion, including clothing choices and item purchases.

Chapter 5 introduces a fashion intelligence system based on PVSE to enable fine-grained learning for individual parts of fashion outfits. This concept emerged from the observation that traditional VSE-based fashion intelligence systems learn full-body images as a whole, making it challenging to develop applications focusing on specific parts. Through comprehensive evaluation and analysis experiments with actual data, the proposed system in this chapter successfully realizes applications that prioritize specific parts, thereby catering to the more detailed needs of users.

In Chapter 6, a fashion image analysis model based on DGVSE is proposed. While conventional embedded data representation models like VSE and PVSE map each component as a single point in the destination projective space, ambiguous expressions such as "casual" and "formal" encompass various images for different users, and a single expression may exhibit a wide range of meanings and variations. The chapter demonstrates that the proposed model, which considers this aspect, enables a detailed analysis of the distribution and interpretation of meanings associated with each component (representation and image).

Lastly, in Chapter 7, an overall discussion of the study's findings is presented, accompanied by a description of how the four proposed models can be effectively applied in real-world applications.

Finally, in Chapter 8, we report the conclusion, summarizing the results and findings of this study and discussing further prospects.

Chapter 2

Conventional Research

This chapter describes conventional research, from the application of artificial intelligence (AI) and detailed technologies being mainly studied in the fashion industry. Additionally, the relationship between this study and this conventional research is described.

2.1 Application of Artificial Intelligence in the Fashion Industry

Numerous studies have proposed various AI methods for application in the fashion and apparel industries [11]. For example, several methods have been proposed to improve the efficiency of the production process and supply chain and to increase sales [12], especially in fabric selection and evaluation [13, 14], AI utilization in the manufacturing process and distribution [15, 16, 17], and recommendation systems on online shopping sites [18]. In addition, explainable recommendation and image retrieval studies aim to improve sales. Existing studies and technologies have in common that they pertain to technologies utilized in the manufacturing-to-sales flow (supply chain). In other words, they are intended to support business decision-making, improve business efficiency, and replace the role of experts. In a broad sense, these studies are included in the framework of business intelligence [19].

2.2 User Support with Recommender System in the Fashion Domain

Recommendation systems in the fashion industry play the role of supporting users by recommending fashion items and outfits. Many studies consider the interaction information between the user and the item and the image information [20, 21].

For example, He et al. [22] proposed a simple recommendation model, Bayesian Personalized Ranking (BPR), based on the interaction information between the user and the item in the purchase history data. Visual BPR (VBPR) [21] is a new recommendation model that combines the image information of items with the BPR model. He et al. [23] further improved the accuracy by combining the hierarchical structure of product categories with VBPR. These are recommendation systems based on purchase history data. These systems are expected to improve user satisfaction by enabling users to reach the items they are interested in purchasing quickly.

However, various studies recommend combinations of fashionable items based on the user's past purchases [24, 25]. In addition, some studies suggest what combination of items would be best based on the contents of the user's closet [26]. Research on recommendation systems also considers which items should be combined with existing item sets. These studies recommend "which item sets are fashionable" by considering interactions and compatibility among items. By receiving recommendations on how to dress, users are expected to help solve questions about how to combine items they already own and what new items to purchase to add to their existing set of items to become more fashionable.

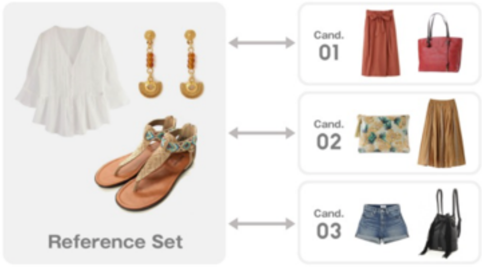


Figure 2.1: Image of outfit recommender system (1) [24]

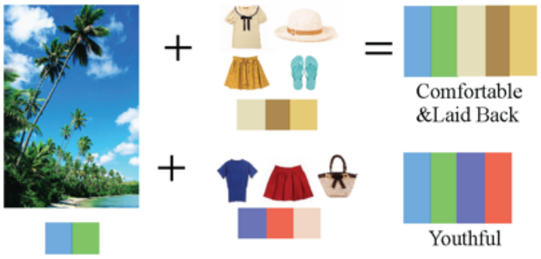


Figure 2.2: Image of outfit recommendation system (2) [27]

One of the unique studies also recommends a photo-worthy outfit according to the travel destination [27]. This research can also be considered a recommendation system that combines information on travel destinations with consideration of interaction and compatibility of each item. This method is expected to alleviate the user's concern of "what should I wear at my trip destination?" This method is expected to mitigate the user's problem of "what should I wear at the destination?"

2.3 User Support with Fashion Image Retrieval

Fashion image retrieval is one of the most active research areas in image processing. Several studies [28, 29, 30] provide complete content-based search techniques, searching from query images for other similar images. Moreover, a study exists to cross-domain fashion image retrieval, e.g., to search the professional's photo (used in an online store) from the user's photo of the item [31, 32, 33]. Furthermore, a study was conducted on image retrieval techniques to search the daily (realway) clothes similar to the query image of a person on the runway [34]. Other studies exist that learn words (attributes) as side information and utilize them for improving image retrieval accuracy [35]. Dong et al. [36] proposed an image retrieval method that learns attributes (e.g., "lapel design," "neckline design," "collar design") as side information and focuses on these specific areas. Furthermore, research on techniques for searching for individual items that match fashion items from the query images of the individual items have been actively conducted [24, 37].

Moreover, there are studies on techniques for searching images by manipulating images and attributes [38, 39, 40]. The contributions of these studies have resulted in, for example, if the word "short-sleeve" is added to an image of a long-sleeved blue shirt, an image of a short-sleeved blue shirt will be obtained as a search result. This is a useful technique for improving online image retrieval efficiency and contributes to improving user satisfaction.

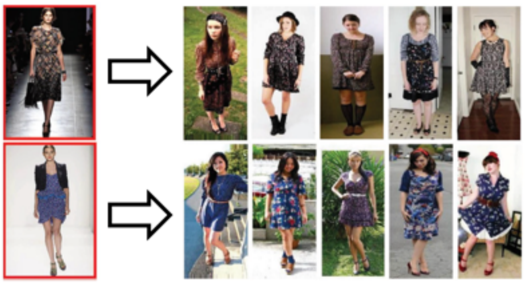


Figure 2.3: Image of fashion image retrieval (1) [34]

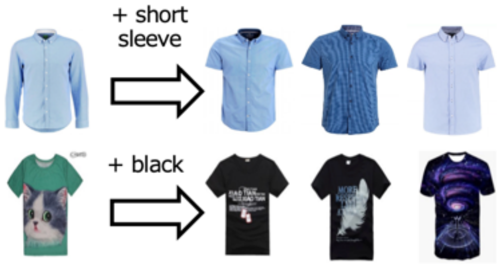


Figure 2.4: Image of fashion image retrieval (2) [38]

2.4 User Support with Explainable Recommendation

2.4.1 Explainable Recommendation

Explainable recommender systems aim to improve understanding of the reasons for recommendations output by machine learning models, which might be expressed as the capacity to answer the question, “why was this item recommended to this user?” Explainable recommendations are expected to help users make better decisions [41] and improve reliability, effectiveness, persuasiveness, transparency, and user satisfaction [42, 43, 44, 45, 46]. Various companies have published related research [47, 48].

The explainable recommendation approach can be broadly divided into two categories [42].

1. Model-agnostic approaches (post-hoc approaches) train a model to explain (interpret) the reason for recommendations separately from the recommendation model.
2. Model-intrinsic approaches train a transparent and directly interpretable recommendation model by various means.

2.4.1.1 Model-Agnostic Approach

In the model-agnostic approach, a recommendation model is trained, and then an explainable model is trained separately to explain the reasons for the provided recommendations. This approach has the advantage that the complexity of the recommendation (decision-making) model itself is irrelevant. Moreover, the model-agnostic approach imitates the common human decision-making mechanism of making intuitive decisions first and considering the reasons for the decision later.

As a model-agnostic approach, LIME is among the most well-known methods for interpreting machine learning model decision-making [49]; many studies have used LIME to perform recommendation [50, 51, 52]. In this method, the local structure of a part of the output gained by a complex model used for decision-making is learned ex post facto using a simple and highly interpretable model. LIME has the advantage of being highly explainable and applicable to any complex model. In contrast, Peake et al. [53] proposed an approach that interpreted the output of the recommendation model based on matrix factorization [54] by association analysis [55].

Furthermore, although reinforcement learning is still considered a black box, it is deemed beneficial to acquiring scientific insight into the internal behavior of decision-making models [56]. Utilizing this feature, various studies have also been conducted on model-agnostic approaches that utilize reinforcement learning for explainable recommendations [41, 48].

However, in the model-agnostic approach, the reasons are not directly obtained from the recommendation model because the reasons for recommendation are interpreted via post-hoc learning with another model. Therefore, there is no assurance that the decision-making model can be accurately explained or that the reason can be accurately expressed. This challenge is widely understood as a significant problem in the model-agnostic approach.

2.4.1.2 Model-Intrinsic Approach

In contrast to the model-agnostic approach, the model-intrinsic approach can obtain the decision-making rationale directly from the recommendation model. This approach aims to recreate a situation where decisions are made from the beginning for coherent reasons. Specifically, the procedures used to decide whether to purchase a certain item by considering various aspects (brand, price, etc.) are important, and the model reflects this by quantifying the importance of each factor. In other words, while the model-agnostic approach adopts an intuitive decision-making flow, the model-intrinsic approach uses a rational decision-making flow. In this situation, the answer to the question, "why did a given user purchase this specific item?," must not be a retroactively constructed reason, but rather the actual reason the user purchased a certain item.

For example, Abdollahi et al. proposed model-intrinsic approaches to learn the objective function by adding a value expressing interpretability to the loss function of the recommended model based on matrix factorization [57] and restricted Boltzmann machines [58]. Furthermore, Seo et al. [59] and Chen et al. [60] proposed models to acquire interpretability using a neural network with an attention mechanism that predicted the rating given to an item by the user using text review data.

The overwhelming advantage of this model-intrinsic approach is that interpretations can be obtained directly from the recommendation model. In contrast, the strength of the model-agnostic approach is that the recommendation model is learned independently, and the recommendation accuracy is not reduced for interpretability. In other words, the model-intrinsic approach can be regarded as superior if an interpretation of the desired type of information can be obtained from a recommendation model without compromising accuracy.

2.4.2 Explainable Recommendation Using Side Information

Many studies use side information to obtain information that cannot be obtained only from the interaction data between users and items, and to realize highly accurate recommendations, such as [61]. Side information is used not only to make highly accurate recommendations but also to make explainable recommendations. Essentially, an explainable recommendation uses a wealth of side information about users and items to interpret the reasons for a recommendation, because checking how much the side information used relates to the recommendation result with a model trained on a massive volume of side information leads to strong interpretability.

For example, if the image information is included in interpretable model learning, an interpretation of which points in the image contribute to the recommendation can be acquired [45, 62]. If social information is included, an interpretation of which friends' purchases or favoriting of items contributed to the recommendation can be acquired [63, 64]. If text review data are included, an interpretation of which words in the review text contributed to the recommendation can be acquired [65, 66, 67].

However, using massive volumes of side information increases the computation required in both approaches. Moreover, the difficulty with the model-intrinsic approach using such a large volume of side information is that a high recommendation accuracy must be achieved while ensuring interpretability. Training models to learn all the side information available without deep consideration reduces their accuracy and increases computational time. In other words, the side information must be examined carefully, particularly in the model-intrinsic approach. Xian et al. [68] considered the volume problem of side information in real-world scenarios and proposed an interpretable recommendation model using an efficient route search based on reinforcement learning. However, their goal was to achieve both interpretability and recommendation accuracy, and it remained unclear whether the computational cost was realistic. Owing to these difficulties, it has been reported that most existing studies using side information related to explainable recommendations adopt the model-agnostic approach [69].

2.4.3 User Support with Explainable Recommendations in the Fashion Domain

Explainable recommendations have been proposed and are currently being investigated as a means to aid online user decision-making [42]. In this research field, beyond the recommendation system's decision to "recommend this item to this user," the reasons "why this item is recommended to this user" are also given by the system. These studies are based on the argument that by showing the reason for recommendations, the opacity in online purchasing with recommender systems is eliminated, and user satisfaction is improved [42, 44, 45]. Various innovative technologies have been proposed specifically for the fashion field; for example, graph neural network-based methods have been proposed that output the reasons for a recommendation and its strength based on large amounts of peripheral information [70]. Moreover, convolutional neural network (CNN)-based methods exist that indicate where the user is likely to be interested in the image, using item image information and textual information [45, 71, 72].

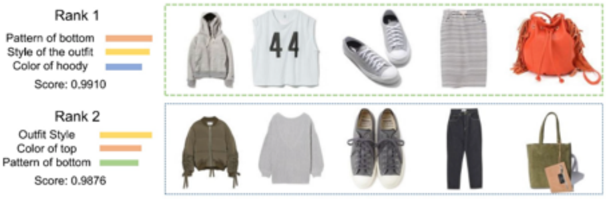


Figure 2.5: Image of fashion explainable recommendation (1) [70]

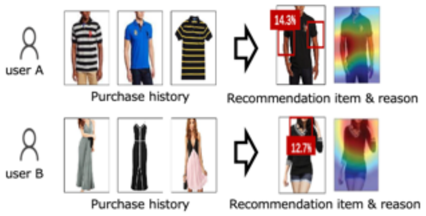


Figure 2.6: Image of fashion explainable recommendation (2) [71]

2.5 Visual-Semantic Embedding

VSE models for image retrieval [73, 74, 75], visual question-answering [76], hashing task [77, 78], zero-shot learning [79, 80], person re-identification [81], and image descriptions generation [82] have been widely researched.

In the fashion domain, VSE is used for text-based and individual clothes image retrieval [36, 83, 84], and learning outfit compatibility (individual outfit item matching) [85]. VSE included in [86] is a method of embedding a fashion item image and a specific word in the item description in the same projective space. As a result, VSE in [86] makes it possible to search for fashion images by calculating similarities between individual item images and simple words and to find specific points on the target item image that are highly related to the word (thereby creating an attribute activation map). These multiple functions constitute the advantage of VSE.

2.6 Relationships between Conventional Research and Proposed Methods

The relationships and positions of the conventional and proposed methods in this study are arranged based on the above.

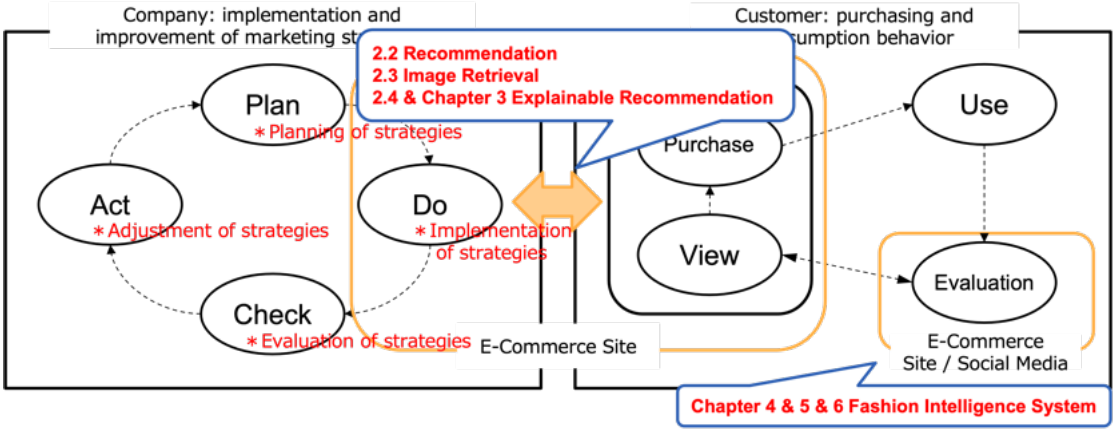

Company: implementation and improvement of marketing st
Customer: purchasing and umption behavior
2.2 Recommendation
2.3 Image Retrieval
2.4 & Chapter 3 Explainable Recommendation
Plan
∗Planning of strategies
Act
∗Adjustment of strategies
Do
∗Implementation of strategies
Check
∗Evaluation of strategies
E-Commerce Site
Purchase
View
Use
Evaluation
E-Commerce Site / Social Media
Chapter 4 & 5 & 6 Fashion Intelligence System

Figure 2.7: Relationship between conventional research and proposed methods

The recommendation, image retrieval, and explainable recommendation technologies can be positioned as approaches to the contact point between the company and users. These technologies enable users to quickly find items they want to purchase (sometimes with a recommendation reason). Specifically, recommendation technology is effective for users who prefer recommendations for their favorite items automatically. In addition, image retrieval technology is effective for users who want to find the item they are looking for quickly. Additionally, we believe that explainable recommendation technology will be helpful for users who wish, “I don’t know what to buy” or “I want a reason to make a decision.”

In other words, these technologies can improve online services’ convenience. In Chapter 3, we propose a framework for an explainable recommendation that enables us to utilize a significant amount of side information.

In Chapter 4, we define the fashion intelligence system that can be positioned as approaching to the steps of evaluating and specifying the fashion items during the users’ purchasing processes. In Chapters 4, 5, and 6, we propose specific models to realize the fashion intelligence system. The approach to this step aims to support understanding the user’s fashion image. In addition, (of course, although it is possible to contribute significantly depending on how it is used) improving business indicators such as sales and profits are not set as premised goals.

Chapter 3

An Explainable Recommendation Framework with Massive Volumes of Side Information

This chapter describes the proposed explainable recommendation framework that enables the use of massive volumes of side information effectively. A detailed explanation of the method, various evaluation experiments using real-world service data, and practical application of the method are included.

3.1 Purpose of this Chapter

In recent years, the application of machine learning to actual marketing problems on e-commerce sites has been widely accepted. Among these, machine learning-based recommender systems have greatly increased e-commerce site sales. However, most successful machine learning models for recommender systems lack explainability (interpretability) in terms of the reasons for each recommendation [42]. In other words, the learning process of the model is self-directed (a black box), and the reasons for the effectiveness of the recommendations provided are not evident. Thus, methodologies that enable understanding of the reasons for recommendations are referred to as explainable recommendation [42] and have been actively studied in recent years. The explainable recommendation applies explainable (interpretable) artificial intelligence [87, 88] (XAI) to recommender systems, and this approach improves the reliability of recommendation models and user satisfaction [42, 43, 44, 45, 46]. Thus, in recent years, several companies have published various research results on the performance of explainable recommendation [47, 48].

Explainable recommendation utilizes a wealth of side information to explain the reasoning behind recommending a particular item to a particular user. Examples of side information that can be accumulated on e-commerce sites include brands, sellers, price ranges, types of items, age and gender of users, and data on their bookmarked (favorite) brands and shops. Companies are highly interested in using such side information to improve the accuracy and explainability of their recommendations. By obtaining detailed insights into why a user favors a particular item, an interpretable explanation recommendation model, supported by abundant side information, holds the potential to inform marketing strategies such as acquiring new users, planning new items, and offering transparent recommendations.

Wang et al. [89] proposed the knowledge graph attention network (KGAT) model to realize high recommendation accuracy using the available side information of recommendable items. Based on a self-attention mechanism and a deep learning architecture, KGAT provides high-accuracy recommendations considering higher-order relationships. Although KGAT was originally proposed as a highly accurate recommendation model utilizing the side information of items, in this study, we mainly consider using KGAT for explainable recommendation. In the KGAT model, the reasons for recommendations can be obtained directly from the model by interpreting the attention weights and graph structures. Therefore, it can be regarded as an explainable recommendation model following the model-intrinsic approach [42]. This approach enables the interpretation of the reasons directly from a complex recommendation model. In addition, KGAT can be applied to various marketing strategies, such as planning new items and acquiring new users. Therefore, KGAT can potentially enhance marketing strategies in the EC business.

In explainable recommendation, the explanation is usually enriched by the number of connections (variables) of the learned side information. Thus, the obvious goal is to maximally train numerous side information variables to improve the explainability. However, learning a substantial volume of variables raises the computational complexity. Specifically, while learning side information containing numerous variables leads to rich interpretability, learning numerous variables is unrealistic as it increases the computational cost and degrades the accuracy of recommendation models. This challenge is also prominent in conventional KGAT. The quantity of variables included in the side information stored by companies managing the major services is characteristically massive. Realizing richer explainability by utilizing as many side information variables as possible is considered an important goal. Therefore, it is necessary to clarify problems such as calculation costs for actual applications.

In this chapter, we present a new knowledge-based explainable recommendation framework learning model based on an improved KGAT model. While we successfully decreased computational cost, the model maintains a high accuracy and exhibits increased interpretability. The algorithm of the proposed improved KGAT model enables it to learn a knowledge graph, including edges with soft probability obtained by compressing a massive volume of side information. Consequently, the computational cost issues of using a massive volume of side information are significantly mitigated, facilitating direct and visual interpretation of recommendation reasons in detail. Moreover, while maintaining recommendation accuracy, the proposed framework enables the model to learn types of side information that were difficult to learn with conventional KGAT. our approach enhances the number of nodes and edges that can be used to explain the reasons for each recommendation. The results demonstrate the considerably improved practicality of the improved KGAT model as a model-intrinsic approach for real-world services.

To demonstrate the effectiveness of the proposed framework, we conducted an experiment to test its feasibility on real-world data collected through ZOZOTOWN [4], the largest fashion e-commerce site in Japan. We demonstrate the value of utilizing the proposed framework for real-world data by conducting evaluation experiments and multifaceted analysis of the obtained results. The results showed that the computational time was reduced by approximately 80% without a decrement in the accuracy, thereby dramatically increasing the practicality of the proposed approach for real-world application. By adding a new type of side information to the learning model and enabling it to learn types of side information, we demonstrate that the proposed framework exhibits richer interpretability compared to the conventional model. In summary, in this chapter, a multifaceted analysis suggests that the proposed framework not only realizes an explainable recommendation model but is also a powerful tool for planning various marketing strategies.

The main contributions of this chapter can be summarized as follows. Firstly, we introduce a novel knowledge-based explainable recommendation framework learning model, which builds upon an enhanced KGAT model. Secondly, we conduct empirical comparisons between the proposed framework, the conventional KGAT, and other baseline models. The experimental results demonstrate that the proposed framework effectively reduces computational costs while maintaining high accuracy and improved interpretability. Thirdly, we provide a comprehensive analysis and consider the application of knowledge graph embedding in marketing strategy planning. Moreover, we highlight the advantages of explainable recommendation beyond just explainability.

3.2 Knowledge Graph Attention Network

The methods of learning the graph network structure as an embedded representation have attracted considerable research attention in recent years, and their application to recommendation [90, 91] and other applications [92] are actively studied. In particular, various attractive methods such as recommendation [93] that learn the structure including time series have been proposed in the applications to the recommendation systems.

Among the approaches based on side information, a method has been developed that directly interprets the reasons for decision-making by treating side information as a knowledge graph while maintaining high recommendation accuracy using attention weights. By tracing the nodes and edges included in the knowledge graph structure given as input data, this method directly and visually obtains the reason for each recommendation (model-intrinsic). Based on this method, analysts can understand the reasoning behind the model’ s recommendation of a given item to a certain user by dividing the recommendation into several factors and the number of contributions. The several factors are expressed by the elements included in the information used as input data (users, items, and their side information), and the number of contributions part is expressed by the attention weight.

The conventional recommendation model KGAT is based on graph attention networks (GATs) [94], which learn the graph structure, including item attribute information, while considering which connections to emphasize using attention weights. A key feature of KGAT is that higher order relationships, including the side information of items, can be modeled by an end-to-end learning method based on the framework of deep learning. Moreover, through the obtained attention weight, KGAT exhibits interpretability as a model-intrinsic approach. Furthermore, as reported in certain studies on context-aware recommendations [95, 96, 97], KGAT improved the recommendation accuracy and addressed cold-start problems using side information.

However, similar to other methodologies of model-intrinsic approaches, challenges remain when considering an application to real-world data, particularly in terms of the computational time required. Furthermore, because Wang et al., who proposed KGAT, did not originally design it with XAI as a primary purpose, KGAT only supports limited side information (only of items). Only the nodes and edges considered in learning can be used for interpretation when a graph-based model is used as an XAI. Essentially, enriching various types of side information results in a richer interpretability.

In this study, we improved the existing conventional KGAT and developed a framework with an improved KGAT model designed for application to real-world services. With our proposed framework, the computational time can be reduced significantly while maintaining and improving the recommendation accuracy without significant negative effects on the KGAT learning algorithm, which can obtain a direct interpretation from the recommendation model using the existing knowledge graph. To the best of our knowledge, no prior works reported in the relevant literature have focused on improving the computational time and interpretability of the model-intrinsic approach with GATs (at least KGAT) by devising a method to optimally handle the knowledge graph. In this respect, the proposed approach is original and novel.

3.3 Proposed Model

We present the proposed framework enabling a massive volume of side information based on the proposed improved KGAT model.

As a premise, in KGAT, each node is called an "entity" $e \in \mathcal{E}$, and each edge is called a "relation" $r \in \mathcal{R}$. Moreover, "triplet" $\{(h, r, t) | h, t \in \mathcal{E}, r \in \mathcal{R}\}$ is a set of three elements: the first entity (called head entity h), the relation r, and the last entity (called tail entity t). In addition, the bipartite graph of users and items $G_{\mathrm{CF}} = \{(h, r, t) | h, t \in \mathcal{E}_{\mathrm{CF}}, r \in \mathcal{R}_{\mathrm{CF}}\}$ and the knowledge graph comprising the side information of items and users $G_{\mathrm{KG}} = \{(h, r, t) | h, t \in \mathcal{E}_{\mathrm{KG}}, r \in \mathcal{R}_{\mathrm{KG}}\}$ are collectively called the "collaborative knowledge graph (CKG)" $G_{\mathrm{CKG}} = \{(h, r, t) | h, t \in \mathcal{E}_{\mathrm{CKG}}, r \in \mathcal{R}_{\mathrm{CKG}}\}$. Here, $\mathcal{E}_{\mathrm{CF}}$ is a set of entities included in the bipartite graph, and $\mathcal{R}_{\mathrm{CF}}$ is a set of relations included in the bipartite graph (equivalent to "buy" or "bought"). $\mathcal{E}_{\mathrm{KG}}$ and $\mathcal{R}_{\mathrm{KG}}$ are sets in the knowledge graph, and $\mathcal{E}_{\mathrm{CKG}}$ and $\mathcal{R}_{\mathrm{CKG}}$ are the sets in the collaborative knowledge graph.

In addition, the input/output of the proposed model and the target loss function do not differ significantly different from the basic mechanism of the conventional KGAT. The input data are the CKG structure, and the output is the probability $\hat{y}(u, i)$ that user $u \in \mathcal{U}$ will purchase item $i \in \mathcal{I}$. In addition, training is performed to minimize the loss functions L_{KG} (pairwise ranking loss: a loss function designed to measure whether the positional relationship of entities in the projective space conforms to the structure of the input data) and L_{CF} (Bayesian personalized ranking (BPR) [22] loss: a loss function designed to measure whether the purchase probability calculated from embedded representations can simultaneously reproduce the user's actual behavior.

The proposed framework efficiently learns a massive volume of side information and reduces computational time without reducing interpretability. To realize both reduced computational time and high interpretability as an explainable recommendation model, the proposed framework includes the following two improvements over the approach using the conventional KGAT.

1. It allows for probabilistic relation learning, which realizes rapid calculations with a massive volume of side information by incorporating the following three improvements:
   (a) compression of many-to-many relationship side information based on a latent class model,
   (b) considering the probabilistic strength of relations when calculating the attention weight and loss for the CKG structure, and
   (c) being given a prior distribution of the probability of sampling during learning to prevent biased and inefficient learning.
2. Moreover, it allows the side information of users to be learned alongside those of items to realize greater interpretability.

Including the aforementioned improvements, the learning algorithm of the proposed framework consists of five steps, as outlined hereafter, and the details of each step are explained in the following sections.

1. Compression by Latent Class Model: A large volume of side information with a many-to-many relationship is compressed using a latent class model.
2. CKG Embedding Layer: An embedded vector representation preserving the structure of the input graph data is acquired.
3. Attentive Embedding Propagation Layer: An embedded vector representation designed to calculate the purchase probability is acquired while considering the relationship with neighbors' entities.
4. Prediction Layer: The purchase probability is calculated using the embedded vector representation acquired in the previous layer.

5. Recombining Compressed Information: Relations with soft probabilities compressed by the latent class model are converted into hard relations by calculating the inner product.

3.3.1 Compressing by Latent Class Model

A massive volume of side information is compressed. Particularly, we focus on many-to-many side information in the knowledge graph as a compression target.

Age and gender are two examples of one-to-many user-side information. Similarly, item brands and the sellers from which they were purchased are examples of one-on-one item side information. Because only one of these one-to-many side information types exists for each user or item, the one-to-many side information does not cause the total volume of the side information to become extremely large.

In contrast, as an example of many-to-many user side information in the case of the fashion e-commerce site, information about brands or shops registered as favorites by each user can be considered (called “favorite brands” and “favorite shops” in this study). Moreover, item descriptions and review texts are also side information in which items and words are connected in a many-to-many relationship. Because one or more many-to-many side information for each user or item is given in most cases, many-to-many side information results in massive volumes of total side information. An average of 25 pieces of records on “favorite brands” for each user were included in the target dataset. Therefore, this is clearly the cause of the large volume of side information. Essentially, while many-to-many side information enhances interpretability, it may cause an increase in the amount of computation required and a corresponding decrease in the recommendation accuracy.

Furthermore, in the proposed framework, a latent class model is used to compress the many-to-many side information. Compression of high-dimensional data with a latent class model is a common and powerful methodology that has yielded successful outcomes in several studies [98, 99]. This approach allows for capturing the complex structure underlying side information data and acquiring potential information. For example, by compressing the information of “favorite brands” with the latent class model, the latent classes and the class membership probability, which expresses the strength of each user and each brand belonging to each class, are given a value in the range [0,1]. The latent class acquired here is treated as an entity, and the class membership probability is used as the strength of the relation between “users (head entities) and latent classes" or "brands (tail entities) and latent classes" in the knowledge graph.

In this study, a set of target relations for compression is expressed as $\mathcal{R}^{c}_{\text{CKG}}$. In addition, when relation r in the triplet (h, r, t) is included in $\mathcal{R}^{c}_{\text{CKG}}$, a latent class $\mathcal{Z} = \{z_k : 1 \leq k \leq K\}$ is assumed between head entity h and tail entity t. The probabilistic model is formulated as follows.

$$P(h, t) = \sum_{k=1}^{K} P(z_k)P(h|z_k)P(t|z_k), \tag{3.1}$$

where z_k is introduced between the connection of "h and t" and is decomposed into "h and z_k" and "t and z_k" connected by the probability $P(h|z_k)$ and $P(t|z_k)$. Finally, the compressed triplet (h, r, t) is excluded from CKG, and the triples "h and (r = "belong") z_k" and "t and (r = "belong") z_k" are added to the graph. Herein, the set of these compressed relations is expressed as $\mathcal{R}^{c'}_{\text{CKG}} \subseteq \mathcal{R}_{\text{CKG}}$. Moreover, the probabilities $P(h|z_k)$ and $P(t|z_k)$ are used in the following steps as the strength of the relation in the triplet ϵ^{lca}.

Several triplets were attributed to the number of variables in the side information before compression was reduced to (the number of latent classes K) × (the number of head entities) + K × (the number of tail entities). By setting a smaller value of K according to criteria such as AIC [100] and BIC [101], we significantly reduced the number of side information triplets. Notably, the latent class model significantly reduces computational costs by compressing many-to-many relationship side information in the training data.

Although it is not the subject of this research, as a side note, it is possible to group components with similar tendencies behind the data structures. In addition, interpretations can be given to those obtained classes from post-hoc analysis. For example, by compressing the information of "favorite brands" using the latent class model, the latent classes are acquired with post-hoc interpretations such as "luxury brands," "sports brands," and "fast fashion." Because many studies have used this interpretation for marketing purposes [102], the latent class model is a powerful method for extracting necessary information while compressing complex data.

3.3.2 Collaborative Knowledge Graph Embedding Layer

Embedded representations (for each entity and relation) that retain the structure of a CKG are acquired in this layer. During training, each triplet is vectorized using a graph-embedding technique called TransR [103], and the parameters are updated to maximize the difference between the triplets that exist on the graph and those that do not by minimizing pairwise ranking

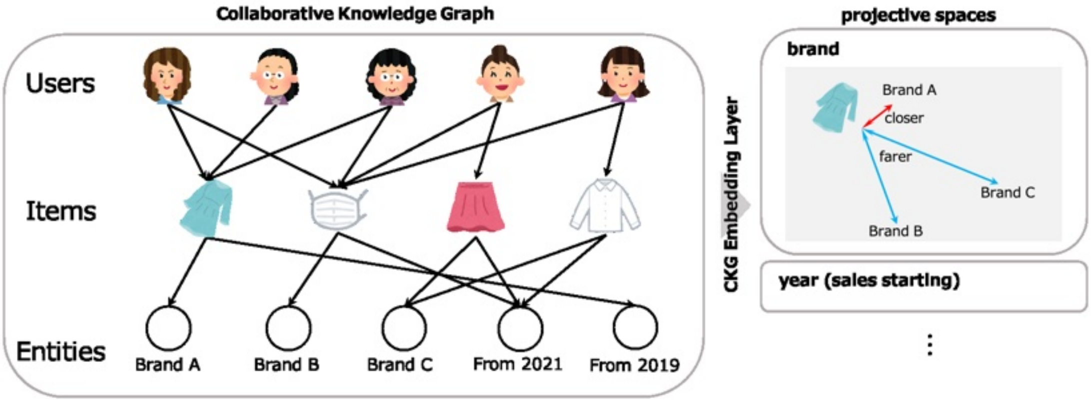


Figure 3.1: Illustration of the CKG embedding layer

loss L_{KG}, which is a common settlement for materializing preference learning [104] expressed as follows.

$$L_{\mathrm{KG}} = \sum_{(h,r,t,t')\in\mathcal{T}} -\ln(\epsilon_{(h,r,t)}\sigma(g(h,r,t') - g(h,r,t))), \tag{3.2}$$

where $\sigma(\cdot)$ is a sigmoid function. Here, the positive triplet (h, r, t) and the negative triplet (h, r, t') are sampled from $\mathcal{T} = \{(h, r, t, t')|(h, r, t) \in \mathcal{G}_{\mathcal{CKG}}, (h, r, t') \notin \mathcal{G}_{\mathcal{CKG}}\}$ based on the prior probabilities (stated below). Moreover, $g(h, r, t)$ is the plausibility score, which expresses the likelihood that a triplet exists on the graph, formulated as follows.

$$g(h,r,t) = \|\mathbf{W}_r\mathbf{e}_h + \mathbf{e}_r - \mathbf{W}_r\mathbf{e}_t\|_2^2, \tag{3.3}$$

where $\mathbf{W}_r \in \mathbb{R}^{m\times d}$ is the matrix used for transformation entities from the d-dimensional entity space into the m-dimensional relation r space. Here, $\mathbf{e}_h, \mathbf{e}_t \in \mathbb{R}^d$ are embedded representations in the projective spaces of entities h, t, and $\mathbf{e}_r \in \mathbb{R}^m$ is the embedded representation in the projective space of relation r.

Here, the conventional KGAT could not learn the relation given stochastically. In contrast, the proposed improved KGAT model allows probabilistic relation learning. Two improvements were included in the CKG embedding layer. The probability $\epsilon_{(h,r,t)}$ is given to each triplet contained in Eq. (3.2). $\epsilon_{(h,r,t)}$ represents the strength of the relation r that connects the head entity h and tail entity t, and it is expressed as follows.

$$\epsilon_{(h,r,t)} = \begin{cases} 1.0 & (r \notin \mathcal{R}^{c'}_{\mathrm{CKG}}), \\ \epsilon^{\mathrm{lca}}_{(h,r,t)} & (r \in \mathcal{R}^{c'}_{\mathrm{CKG}}). \end{cases} \tag{3.4}$$

In this study, a probability of 1.0 is assigned to all one-to-many type relations. In contrast, the triplet (h, r, t) of compressed many-to-many side information is given the probability obtained from the latent class model $\epsilon^{\mathrm{lca}}_{(h,r,t)}$. $\epsilon_{(h,r,t)}$ plays a role in making it possible to consider the loss appropriately for the strongly connected relations when calculating L_{KG}.

Then, the prior probabilities for positive sampling are given by calculating L_{KG}. Because the conventional KGAT does not give prior probabilities, the problem of a massive volume of side information, which is the type including many triplets (= many-to-many), is frequently selected and over-learned. To solve this problem in the proposed model, triplets containing strongly connected relations are sampled relatively more by assigning $\epsilon_{(h,r,t)}$ as prior probabilities. In other words, triplets containing strongly connected relations are intensively learned. This second improvement prevents inefficient learning owing to sampling numerous less important triplets, which are weakly connected. Therefore, the proposed model improves the computational efficiency of the learning process for graph data, including probabilistic relations.

Consequently, the entities “user” and “item purchased by user,” “user” and “user side information,” and “item” and “item side information” are arranged close to one another in the projective space through the CKG embedding layer.

3.3.3 Attentive Embedding Propagation Layer

The importance of each triplet is calculated in this layer. Moreover, while considering which relationship is emphasized using this importance value, the embedded representation is used to calculate the purchase probability in the next prediction layer for each item and the user. To enable learning considering probabilistic relations, this layer also includes an improvement over the conventional KGAT. Specifically, when calculating the attention weight $\pi(h, r, t)$ of each triplet based on the distance between e_h and e_t in the projected space of r, the probability $\epsilon_{(h,r,t)}$ given to the relation contained in each triplet is reflected.

$$\pi(h, r, t) = \epsilon_{(h,r,t)}(\mathbf{W}_r\mathbf{e}_t)^{\mathsf{T}} \tanh(\mathbf{W}_r\mathbf{e}_h + \mathbf{e}_r). \tag{3.5}$$

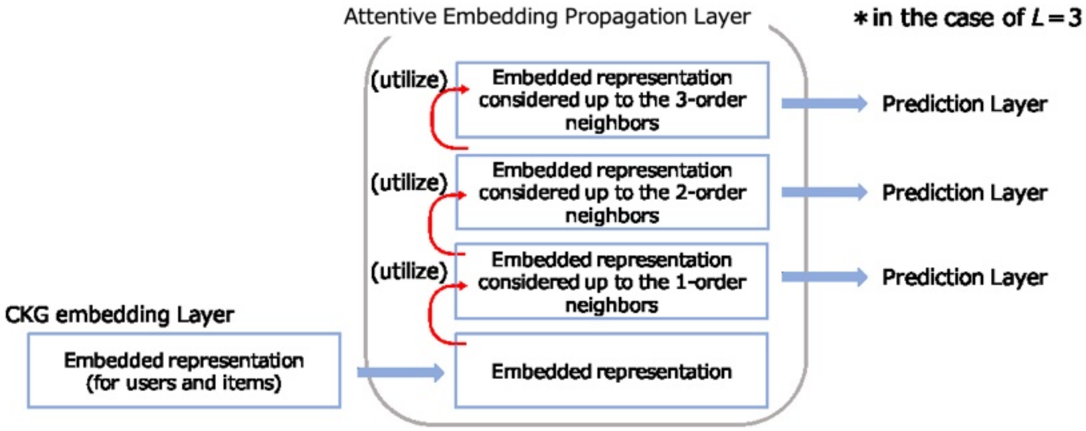

Attentive Embedding Propagation Layer
∗in the case of $L=3$
(utilize)
Embedded representation considered up to the 3-order neighbors
Prediction Layer
(utilize)
Embedded representation considered up to the 2-order neighbors
Prediction Layer
(utilize)
Embedded representation considered up to the 1-order neighbors
Prediction Layer
CKG embedding Layer
Embedded representation (for users and items)
Embedded representation

Figure 3.2: Illustration of attentive embedding propagation layer

Here, tanh is a nonlinear activation function. In addition, a set of all triplets that are connected to the head entity h is stated as $N_h = \{(h, r, t) | (h, r, t) \in \mathcal{G}_{\text{CKG}}\}$. The attention weight $\pi(h, r, t)$ for triplet (h, r, t) is normalized across N_h as follows.

$$\pi_{\text{norm}}(h, r, t) = \frac{\exp(\pi(h, r, t))}{\sum_{(h,r',t') \in N_h} \exp(\pi(h, r', t'))}. \tag{3.6}$$

Furthermore, this layer has an L-times overlapped structure, which facilitates consideration of the relationship up to the L-th neighborhood. The embedded representation that considers the entities up to the l-th order neighborhood for head entity h is calculated as follows.

$$\mathbf{e}_h^{(l)} = f(\mathbf{e}_h^{(l-1)}, \mathbf{e}_{N_h}^{(l-1)}). \tag{3.7}$$

Herein, as the function $f(\cdot)$, we select the bi-interaction aggregator defined as follows.

$$\begin{aligned} f(\mathbf{e}_h, \mathbf{e}_{N_h}) \; = \; & \text{LeakyReLU}(\mathbf{W}_1(\mathbf{e}_h + \mathbf{e}_{N_h})) \\ + \; & \text{LeakyReLU}(\mathbf{W}_2(\mathbf{e}_h \odot \mathbf{e}_{N_h})), \end{aligned} \tag{3.8}$$

where $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{d' \times d}$ are the parameters and are used as weight matrices, d' is the size of the transformation, LeakyReLU$(\cdot)$ is leaky rectified linear unit, and $\odot$ denotes the element-wise product. Here, $\mathbf{e}_{N_h}$ is the embedded representation considering a set of all triplets around h and is formulated as follows.

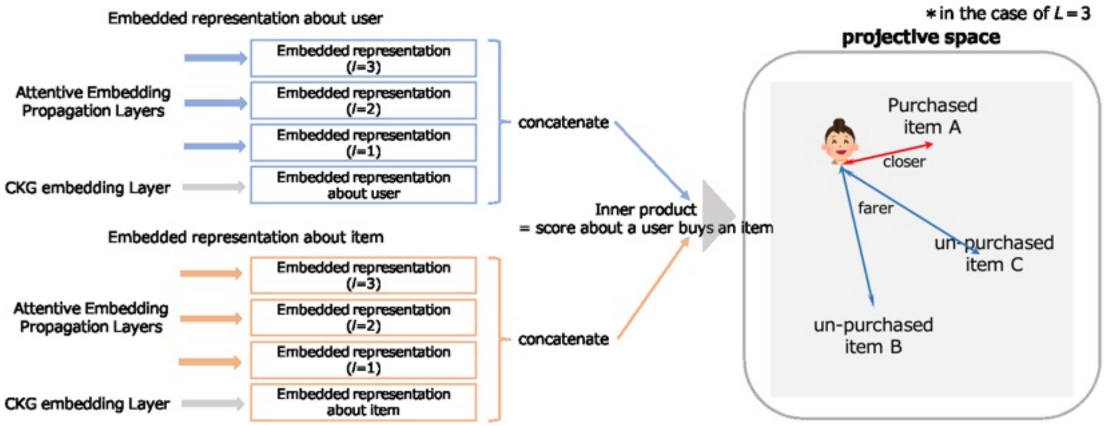

Embedded representation about user
Attentive Embedding Propagation Layers
CKG embedding Layer
Embedded representation (l=3)
Embedded representation (l=2)
Embedded representation (l=1)
Embedded representation about user
concatenate
Embedded representation about item
Attentive Embedding Propagation Layers
CKG embedding Layer
Embedded representation (l=3)
Embedded representation (l=2)
Embedded representation (l=1)
Embedded representation about item
concatenate
Inner product
= score about a user buys an item
＊in the case of L=3
projective space
Purchased item A
closer
farer
un-purchased item C
un-purchased item B

Figure 3.3: Illustration of prediction layer

$$\mathbf{e}_{N_h} = \sum_{(h,r,t)\in N_h} \pi_{\text{norm}}(h,r,t)\mathbf{e}_t. \tag{3.9}$$

Thus, a new embedded representation is obtained by aggregating the characteristics of the peripheral entity. In particular, the weighted average value of the embedded representation of the peripheral entity weighted by the importance of each triplet is calculated. In addition, the new embedded representation for each item and user obtained from this layer is used to calculate the purchase probability in the next prediction layer. Essentially, a more sophisticated embedded representation for calculating the purchase probabilities is acquired according to the L-th neighborhood relationships considered important in the recommendation.

3.3.4 Prediction Layer

Based on the embedded representation for each user and each item obtained from the attentive embedding propagation layer, the purchase probability is calculated for each user and item pair. The probability of user u buying item i is calculated as follows.

$$\hat{y}(u,i) = {\mathbf{e}_u^*}^{\mathsf{T}} \mathbf{e}_i^*, \tag{3.10}$$

where $\mathbf{e}_u^*$ and $\mathbf{e}_i^*$ are the embedded representations considering up to the L-th neighbor entities formulated by the concatenation operation $\|$, as given below.

$$\mathbf{e}_u^* = \mathbf{e}_u^{(0)}\|\cdots\|\mathbf{e}_u^{(L)}, \mathbf{e}_i^* = \mathbf{e}_i^{(0)}\|\ldots\|\mathbf{e}_i^{(L)}. \tag{3.11}$$

The learning process involves calculating the difference between the scores (purchase probabilities) for pairs of users and items that exist in actual purchase data and the scores for non-existing pairs. The loss function used to represent purchase behavior in this context is known as BPR loss, which can be formulated as follows:

$$L_{\mathrm{CF}} = \sum_{(u,i,j)\in O} - \ln \sigma(\hat{y}(u,i) - \hat{y}(u,j)). \tag{3.12}$$

where $O = \{(u,i,j)|(u,i) \in \mathcal{G}_{\mathrm{CF}}, (u,j) \notin \mathcal{G}_{\mathrm{CF}}\}$ represents the training set for calculating the BPR loss. Moreover, the positive sample with pair (u,i) and a negative sample of pair (u,j) are sampled randomly for each step.

Finally, the entire model is optimized using the combination loss function of L_{KG} and L_{CF}.

$$\mathrm{argmin}_{\Theta} L_{\mathrm{KGAT}} = L_{\mathrm{KG}} + L_{\mathrm{CF}} + \lambda \|\Theta\|_2^2, \tag{3.13}$$

where $\Theta = \{\mathbf{E}, \mathbf{W}_r, \forall r \in \mathcal{R}_{\mathrm{CKG}}, \mathbf{W}_1^{(l)}, \mathbf{W}_2^{(l)}, \forall l \in \{1, \ldots, L\}\}$ is a set of parameters that are estimated by minimizing Eq. (3.13). Here, $\mathbf{E}$ is a set of all embedded representations.

3.3.5 Recombining Compressed Information

The compressed many-to-many side information is restored to its original state using the following formula.

$$\pi(h,r,t) = \sum_{k=1}^{K} \pi_{\mathrm{norm}}(h,r,z_k) \times \pi_{\mathrm{norm}}(z_k,r,t). \tag{3.14}$$

Thus, the attention weight between head entity h and tail entity t is restored by calculating the inner product of the attention weight of the triplet, which is separated into each head entity h and latent class z_k and the latent class z_k and each tail entity t. This restoration enables the model to obtain a direct interpretation of the same connection as the original data, rather than an ambiguous interpretation of the connection with the latent class. In other words, this restoration prevents a reduction in the interpretability.

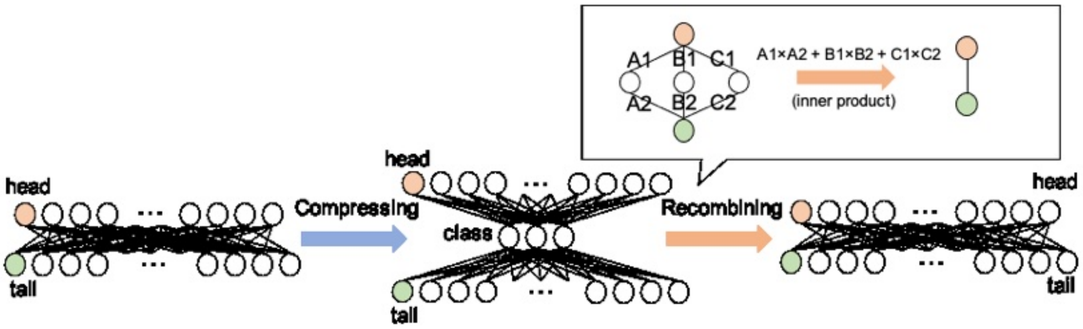


Figure 3.4: Illustration of compressing and recombining

3.4 Experiments and Evaluation

To validate its effectiveness, the proposed approach was applied to the actual purchase history data and the side information of each user and item stored in ZOZOTOWN. The recommendation's rationale is visualized and explained based on the obtained results. Additionally, an evaluation experiment was conducted to assess the proposed framework's recommendation accuracy.

3.4.1 Experimental Settings

To prepare the experimental data, random sampling was conducted from the purchase history data of users who purchased items more than five and less than 60 times during the year from February 2020 to January 2021. Particularly, 260,881 purchase history data related to 25,757 users were acquired. The number of items included was 11,741. In addition, the dataset included 10 types of side information such as the items' brands, sellers, categories, and price ranges, as well as users' ages, and genders, including many-to-many information (favorite brand/shop) and 790,083 items in total (one-to-many:many-to-many=149,992:608,131). The total data were divided into a training dataset and a testing dataset in a time series at a ratio of 8:2. The parameters were set by referring to the experimental conditions in [89]. The number of dimensions used for the embedded vector representation in each entity and relation was set to 64. The attentive embedding propagation layer was set to 3 ([64, 32, 16]) layers. Moreover, the top-20 and top-60 accuracies (Recall, NDCG) were used as the evaluation index. In addition, we adopted probabilistic latent class analysis [105] (pLSA) and latent Dirichlet allocation [106] (LDA) as latent class models. Furthermore, for each side information to be compressed, with the judgment of the

AIC criterion, the number of latent classes was uniformly set as six, and five and seven classes were also adopted for comparison. Finally, a GPU (NVIDIA Tesla T4×1,4 vCPUs, 15 GB) was used for the entire computation.

A summary of the abbreviation notation for each method used in the experiment is presented in Table 3.1.

Table 3.1: Notation, name, and role for each model

Notation	Model name & role
BPRMF	Bayesian Personalized Ranking Matrix Factorization [22]; comparison model without side information. Not explainable by this model itself.
CFKG	Collaborative Filtering over Knowledge Graph [107]; comparison model learning a knowledge graph structure.
KGAT	Knowledge Graph Attention Network [89] (KGAT); conventional model.
KGAT+(pLSA-k)	Proposed framework with an improved KGAT model using pLSA with k classes; proposed framework.
KGAT+(LDA-k)	Proposed framework with an improved KGAT model using LDA with k classes; proposed framework.

3.4.2 Visualization of Recommendation Reasons

In this section, we present the results of visualizations performed using the results obtained from KGAT+(pLSA-6). Figure 3.5 and Figure 3.6 show two examples that visualize and explain the recommendation reason obtained from the proposed framework. In the map, each entity indicates a user or item included in the purchase history data or an entity included in the side information. Moreover, each relation has a value (attention weight) that indicates the importance of the relationship for deciding on a recommended item. By observing these entities and attention weights, the reason each item is recommended to each user is interpreted quantitatively.

In Figure 3.5 and Figure 3.6, we characterize the reasoning behind recommending item i_5 to user u_1. Moreover, Table 3.2 presents more detailed information about attribute-based reasons, which can be observed in Figure 3.5. By checking each path, as shown in Figure 3.5 (Table 3.2), we reveale that the reason why item i_5 is recommended to user u_1 is most influenced by the fact that item i_5 is sold by brand e_1, which is favored by user u_1. Here, this information is the type of side information that cannot be learned easily by the conventional model because it causes an increase in the volume of side information. Moreover, as validated in Figure 3.6, item i_5 being recommended for user u_1 was influenced by the fact that item i_5 was purchased by some users related to item i_1, i_2, i_3, and i_4, which were purchased by u_1 in the past.

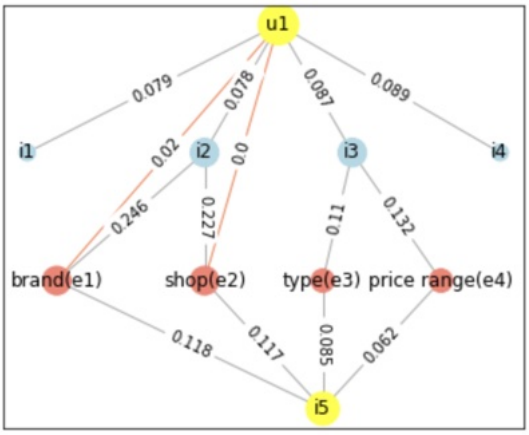


Figure 3.5: An example of a map visualizing the attribute-based reason for recommendation

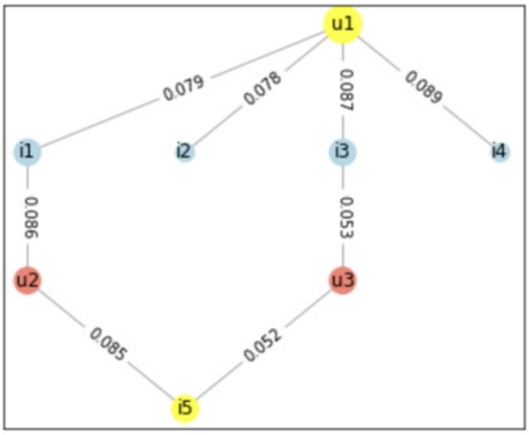


Figure 3.6: An example of a map visualizing the behavior-based reason for recommendation

Table 3.2: Recommendation (attribute-based) reasons and total attention weight by observing each path

No	Path	Weight	Reason
1	$u_1 \to e_1 \to i_5$	2.36×10^3	user u_1 favors brand e_1.
2	$u_1 \to i_2 \to e_1 \to i_5$	2.23×10^3	item i_5 is sold by the same brand e_1 as i_2 purchased by user u_1 in the past.
3	$u_1 \to i_2 \to e_2 \to i_5$	2.07×10^3	item i_5 is sold by the same shop e_2 as i_2 purchased by user u_1 in the past.
4	$u_1 \to i_3 \to e_3 \to i_5$	0.81×10^3	item i_5 is the type category e_3 as i_3 purchased by user u_1 in the past.
5	$u_1 \to i_3 \to e_4 \to i_5$	0.01×10^3	item i_5 is sold in the same price range e_4 as i_3 purchased by user u_1 in the past.
6	$u_1 \to e_2 \to i_5$	0.00×10^3	user u_1 favors shop e_2.

For example, by utilizing the results shown in Figure 3.5, the strategies such as emphasizing that item i_5 is an item of brand e_1 when recommending item i_5 to user u_1 may be easily interpreted through the visual display. Moreover, in addition to the recommended items, a simple strategy that displays the reason for the recommendation and the score is also possible, as shown in Table 3.2. However, by utilizing the results in Figure 3.6, actions such as displaying to the user “this item i_5 is preferred by other users who have purchased item i_1 that you purchased in the past” can be performed.

As a result, users may enjoy shopping more, perhaps feeling that they can better understand the reasons why they purchased various items, noting that they were recommended for specific, given reasons. Users might also find similar insight for items they felt less likely to purchase by noting the reasons those items were recommended.

3.4.3 Recommendation Performance

Table 3.3 shows the summary of the evaluation experiment results. The models other than the proposed framework were not superior to the conventional KGAT model for any accuracies. In particular, the result that KGAT and (all) KGAT+ were superior to BPRMF, which used the same BPR loss as KGAT in learning purchase behavior, suggests the effectiveness of recommendation using side information. Furthermore, the comparison result obtained with CFKG, which learns the same knowledge graph structure, suggests the difficulty in learning the knowledge graph structure and the KGAT's ability to learn it well. Overall, these results suggest the excellent functionality of KGAT itself as a recommended model.

As the most important result, Table 3.3 shows that the proposed framework reduced the computational time by approximately 63% to 75% compared to the conventional model. This result is a benefit of compressing a part of a massive volume of side information by the latent class model. Because the number of relations whose probability becomes 0.0 differs due to compression, the number of side information differs between pLSA and LDA even if the number of classes is the same. This result suggests that the proposed framework greatly improves upon the limitations of the original method, overcoming the key challenges in the real-world application of the conventional KGAT model.

Table 3.3: Summary of evaluation experiment results

Model	Recall		NDCG		# Side	Time
	@20	@60	@20	@60	info	(/epoch)
BPRMF	0.0647	0.1160	0.0482	0.0659	0	-
CFKG	0.0369	0.0759	0.0279	0.0416	785,122	-
KGAT	0.0748	0.1369	0.0537	0.0742	785,122	142.92 s
KGAT+(pLSA-5)	**0.0762**	**0.1409**	0.0543	0.0756	238,387	**35.28 s**
KGAT+(pLSA-6)	0.0761	**0.1409**	**0.0548**	**0.0760**	246,662	36.36 s
KGAT+(pLSA-7)	0.0758	0.1399	0.0546	0.0758	257,147	37.08 s
KGAT+(LDA-5)	0.0755	0.1395	0.0542	0.0752	313,746	44.88 s
KGAT+(LDA-6)	0.0740	0.1392	0.0533	0.0747	346,461	49.68 s
KGAT+(LDA-7)	0.0739	0.1370	0.0528	0.0737	441,740	53.76 s

In addition, the proposed KGAT+(pLSA-6) framework achieved the best accuracy for most indicators. Furthermore, all KGAT+ frameworks using pLSA showed a higher accuracy compared to conventional KGAT and other comparison methods. However, a part of KGAT+ frameworks using LDA did not have better accuracy than conventional KGAT. However, it showed almost the same degree of accuracy, while considerably shortening the computational time. These results suggest that equal or higher accuracy than the conventional model can be expected by appropriately setting the number of latent classes based on criteria such as AIC in the proposed framework. In the following, the results obtained from the most accurate KGAT+(pLSA-6) were analyzed to represent the results from the proposed frameworks.

In Figure 3.7 and Figure 3.8, the vertical axis (bar graph) shows the ratio of the users located in the target category. The vertical axis (line graph) shows the Recall@20 and NDCG@20 transition. In addition, the horizontal axis shows the users' categories grouped by the total number of purchased items for each user included in the training data. For example, "< 5" expresses a group of users with fewer than five purchases included in the training data.

Figure 3.7 and Figure 3.8 shows that the proposed framework was always the most accurate for all users' groups. This result indicates that robust recommendations can be made to the group of users who have not yet accumulated significant purchase history data in the training data compared to other models. Thus, the experimental results suggest that the proposed framework can handle the cold-start problem better than other models.

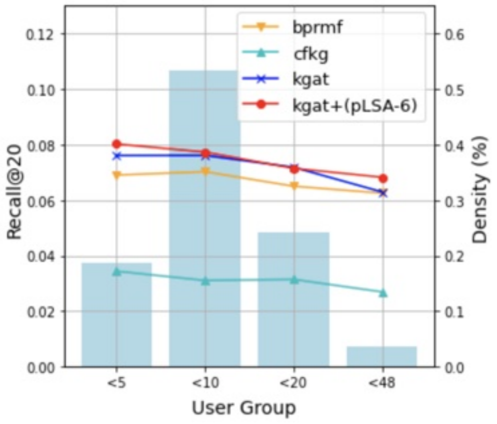


Figure 3.7: Result for cold-start problem (Recall)

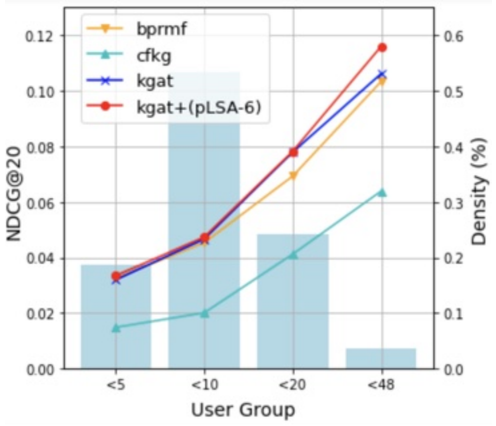


Figure 3.8: Result for cold-start problem (NDCG)

3.4.4 Contribution to Accuracy of Each Proposed Element

In the proposed KGAT model in the proposed framework, the following three improvements were included to learn probabilistic relations:

1. compression of many-to-many relationship side information by the latent class model,
2. consideration of the probabilistic strength of relation when calculating the attention weight and loss for CKG structure, and
3. being given the prior distribution to the probability of sampling during learning.

Table 3.4 shows the results recorded while observing how each element contributed to the results of recommendation accuracy. In this table, the model with all the improvements showed the best accuracy. This result suggests that each improvement contributes to the successful learning of massive volumes of side information caused by the many-to-many type.

Table 3.4: Contribution to accuracy for each proposed element

Excluded	Explanation	Recall		NDCG		Time
		@20	@60	@20	@60	(/epoch)
-	KGAT+	**0.0761**	**0.1409**	**0.0548**	**0.0760**	36.36 s
1(, 2, 3)	Without side information compressed by latent class model (= conventional KGAT)	0.0748	0.1369	0.0537	0.0742	142.92 s
2(, 3)	Without considering probabilistic relations, and defining to belong to only the class with the highest probability (= compression, but hard clustering)	0.0759	0.1393	0.0539	0.0749	**30.60 s**
3	Without prior distribution given to positive sampling in training	0.0735	0.1398	0.0527	0.0743	37.68 s

3.4.5 Impact of the Number of Latent Classes

In the proposed framework, changing parameter K of the latent class model (pLSA) changes the number of side information after compression, which affects the computational time and accuracy. Figs.3.9–3.11 show the comparisons of the changes in computational time and accuracy when the number of classes K is changed with the baseline (conventional KGAT).

Figure 3.9 indicates that the computational time can be shortened by approximately 65% to 80% compared to the conventional KGAT by changing the number of classes. The reduction

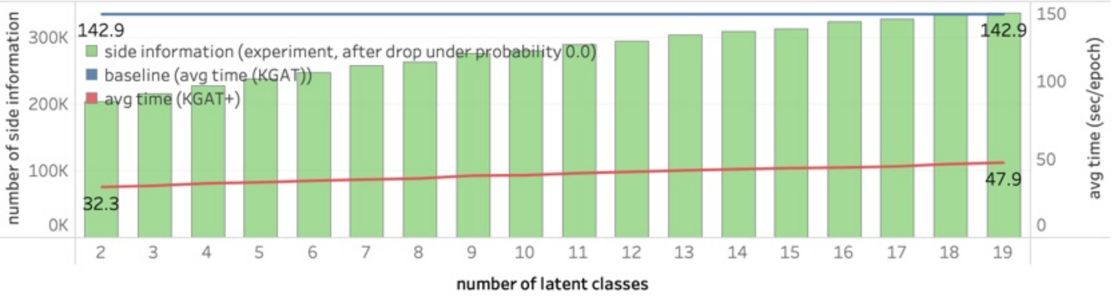


Figure 3.9: Transition of computational time when the number of latent classes K is changed

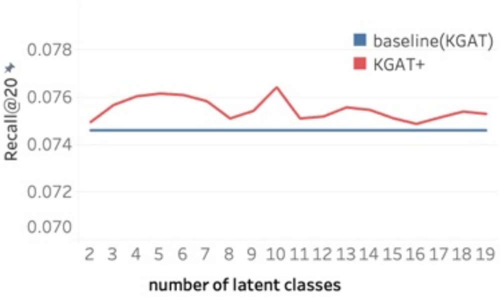


Figure 3.10: Transition of Recall@20 when the number of latent classes K is changed

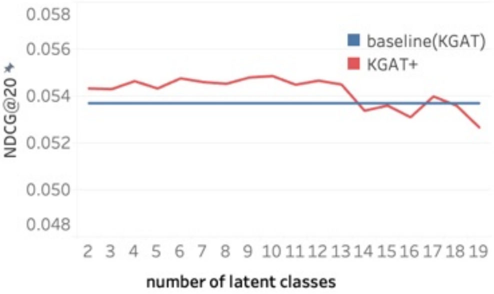


Figure 3.11: Transition of NDCG@20 when the number of latent classes K is changed

in the volume of side information contributes to this phenomenal result. If the volumes of side information can be reduced, then the computational load will reduce as well. Thus, the proposed framework has a stable contribution. In addition, the volume of side information that can be compressed by the proposed framework are discussed theoretically in the “Discussion” section.

Furthermore, as illustrated in Figure 3.10 and Figure 3.11, if the number of classes is not extremely large, the accuracy tends to be higher than that of the conventional KGAT. The latent class model performs to extract characteristic information from the original data. Therefore, if the number of latent classes is considerably increased, the information contained in the original side information is excessively acquired. These results suggest that in the proposed framework, if the number of classes is set appropriately (not set excessively large), useful information for recommendation can be extracted from side information, and highly accurate recommendation can be achieved.

3.5 Additional Analysis

Analyzing these results is also beneficial for planning marketing strategies because an embedded representation can be obtained for each entity (item, user, and side information) and relation. In this section, we analyze the embedded representation for each entity and the attention weight for each triplet obtained by applying the proposed framework to real-world data. In addition, we suggest the usefulness of applying the proposed framework to actual services.

3.5.1 Utilization of Obtained Embedded Representation

Numerous mappings can be drawn using the embedded representation for each entity obtained from the proposed framework. Among these mappings, Figure 3.12 and Figure 3.13 are enlarged views of the whole brand mapping around a part of certain brands. In these mappings, each embedding is compressed to two dimensions with t-distributed stochastic neighbor embedding [108]. These mappings allow for grasping the positional relationship between all entities included in the CKG on the projective space based on the users' purchasing behavior.

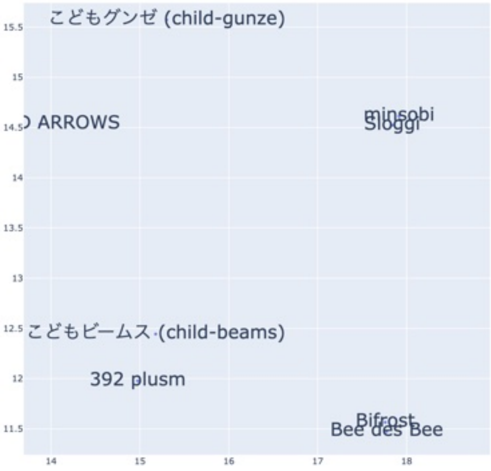


Figure 3.12: Diagram mapping embedded representations around brands for mother and child

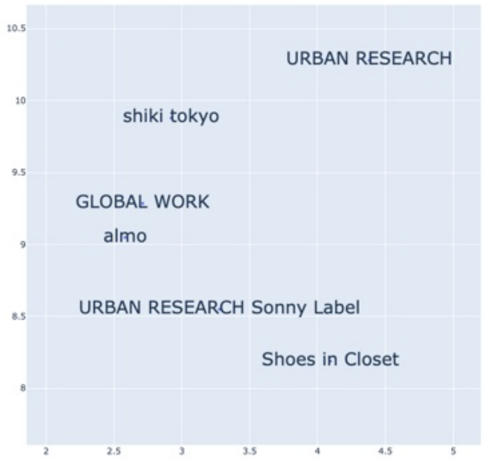


Figure 3.13: Diagram mapping embedded representations around URBAN RESEARCH Sony Label

To elaborate, in the Figure 3.12, famous brands such as "Bee des Bee" [109], "392 plusm" [110], "child-gunze" [111] and "child-beams" [112] that mainly handle items for children are gathered. This suggests the validity of the experimental results. Moreover, when enlarging the area around the URBAN RESEARCH Sony Label [113] (Figure 3.13), URBAN RESEARCH [114] existed nearby. Furthermore, if this situation is suboptimal, it is possible to contemplate the design of items and the type of items being showcased in order to align more closely with other desirable brands. Consequently, the positional relationship of each entity can be visualized by taking into account the user's on-site behavior.

3.5.2 Utilization of Obtained Attention Weights

By utilizing the attention weights assigned to each triplet through the proposed framework, it becomes feasible to analyze the significance of individual relationships within the CKG with regards to users' purchasing behavior. Attention weights are assigned to all triplets, and in this case, our focus is on the relationships between users and two specific brands: "BEAMS" [115], a relatively high-priced brand irrespective of gender, and "earth music&ecology" [116], a brand targeted towards young women. By examining the attention weights associated with each brand-user relationship, it is possible to verify the importance of each brand in influencing the purchasing behavior of individual users.

Figure 3.14-3.17 shows the average attention weight of the relations connected from each user to these two brands by age and gender.

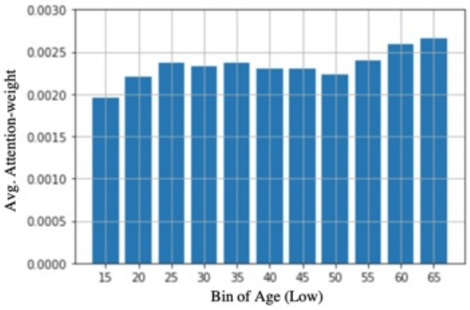


Figure 3.14: Distribution of the mean attention weight given to user-connected relations with BEAMS (for men)

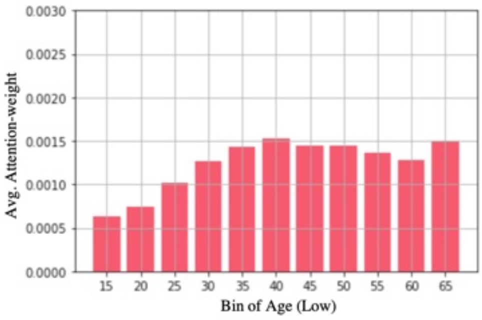


Figure 3.15: Distribution of the mean attention weight given to user-connected relations with BEAMS (for women)

In the brand that was high-priced regardless of gender (BEAMS), the attention weight associated with men was generally higher; however, the weights for both men and women tended to be relatively high. Particularly, a tendency for relatively older generations to place more importance on the brand was notable. This result can be considered convincing because BEAMS is a high-priced brand, and men's items are handled much more frequently than women's items at

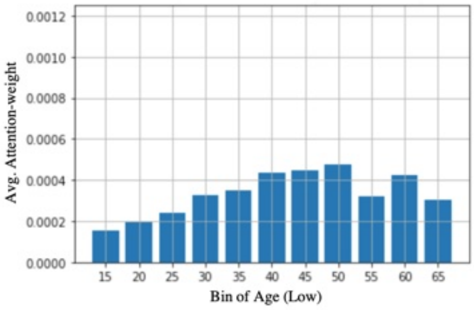


Figure 3.16: Distribution of the mean attention weight given to user-connected relations with earth music&ecology (for men)

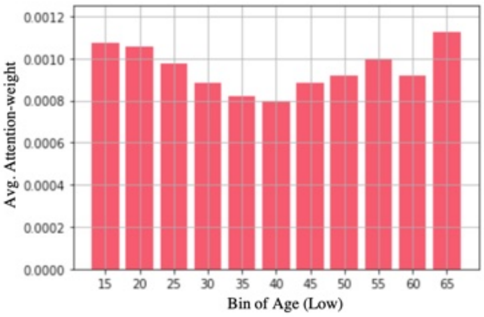


Figure 3.17: Distribution of the mean attention weight given to user-connected relations with earth music&ecology (for women)

ZOZOTOWN.

In addition, the brand for young women (earth music&ecology) was indeed assigned considerable importance by women, especially by younger females. This result suggests the validity of the obtained results. Furthermore, users of older age groups liked earth music&ecology to some extent can probably be attributed to the fact that they used the same user account in their families. The rationale may be found in the fact that relatively high importance was gained from relatively older male users.

In addition, a list of the top few users who placed particular importance on both brands can also be created.

Table 3.5: Top eight users for BEAMS and earth music&ecology

Rank	For BEAMS	For earth music&ecology
1	15504 / Woman in late 30s	24061 / Woman in early 20s
2	24681 / Man in late 40s	20641 / Woman in early 20s
3	19760 / Woman in early 30s	20078 / Woman in early 40s
4	17128 / Man in early 30s	16353 / Woman in early 20s
5	5836 / Woman in early 30s	10470 / Woman in early 30s
6	11102 / Woman in early 30s	18714 / Woman in early 20s
7	22647 / Man in late 40s	23891 / Woman in late 10s
8	13978 / Woman in late 40s	17451 / Woman in early 20s

The users listed in Table 3.5 displayed varying degrees of significance attributed to each brand in relation to their purchasing behavior on the platform. While individuals in their 30s and 40s, both men and women, demonstrated a balanced ranking in their valuation of BEAMS, young women predominantly occupied the top positions, mirroring the pattern observed for earth music&ecology. Thus, Table 3.5 serves as a valuable resource for implementing diverse strategies, including brand-focused recommendations and targeted email advertisements. In essence, the attention weight results obtained from the proposed framework represent a highly important compilation for devising effective marketing strategies.

3.6 Discussion

3.6.1 On the Experimental Results and Additional Analysis

Summarizing the above experimental results, we can point out the following suggestions.

- The proposed method maintains the interpretability based on the knowledge graph for item side information. In addition, interpretability can be further reinforced using the user side information.
- Compared with the comparison models, the proposed framework is effective in terms of recommendation accuracy.
- The proposed framework can deal with cold-start problems effectively using various types of side information.
- In the proposed framework, the computational load can be significantly reduced according to the volumes of side information after compression.
- In the proposed framework, if the number of classes is set appropriately, useful information for recommendation will be extracted from side information and recommendation with accuracy equal to or higher than the baseline will be realized.
- Each of the improvement techniques for learning probabilistic relationships introduced to realize efficient learning helps improve the recommendation accuracy and computational cost.

Furthermore, in the additional analysis for utilizing the obtained results, we clarified that the following analyses can be achieved by applying the proposed framework to actual business data.

- Visualizing the obtained embedded representation allows for a better understanding of the positional relationship between entities in the projective space.

- By observing the obtained attention weights, it is possible to discover users who have each entity as an important motivation for their purchasing behavior.

These analyses confirmed that the proposed framework renders the model interpretable in terms of the reasons for recommendations and facilitates the utilization of multifaceted results. Furthermore, we devised actual marketing strategies from the analysis results and considered how they could be used for actual services. These multifaceted utilization possibilities can be regarded as strengths, along with the explainability and high accuracy of the proposed framework. In addition, this highlights the advantages of the proposed framework compared with other explainable recommendation models.

3.6.2 Comparison with Conventional Recommender Systems

Conventional recommender systems aim to increase sales within the framework of recommending any of the items to existing customers. However, by clarifying the recommendation reasons, for example, the following can be achieved.

- Identifying the attributes of the users who strongly demand an itemcan help cultivate a new customer base.

- Items matching customer attributes can be recommended to new users.

- Understanding the attributes of the recommended items can facilitate new product planning.

The ability to leverage these diverse applications demonstrates that the proposed framework is a powerful marketing tool that surpasses the confines of a typical recommendation model.

3.6.3 On the Reduction Amount of Computational Time

In the proposed framework, the computational time that can be reduced depends on the extent to which the many-to-many relationship can be compressed. In our experiment, triplets whose class membership probabilities became 0.0 because of compression are dropped and not counted in the side information. Therefore, the number of side information after compression and dropping cannot be pre-determined before the latent class model is learned by a training data set. However, in theory, the number of many-to-many side information after compression (and before dropping) $N^{c'}_{\text{side}}$ can be computed easily using the following formula.

$$N^{c'}_{\text{side}} = \sum_{d=1}^{D}(K_d N_{h^c_d} + K_d N_{t^c_d}), \tag{3.15}$$

where the number of relations contained in the side information to be compressed is expressed as D and each type of side information is compressed by an independent latent class model with the number of classes $K_1, K_2, ..., K_D$. The numbers of head entities and tail entities included in the side information about the d-th compression target relation are expressed as $N_{h^c_d}$ and $N_{t^c_d}$ respectively.

While comparing the theoretical number of side information after compression computed by Eq. (3.15) with the number after dropping the side information whose class membership probability became 0.0, the transition of the computational time in the experiment can be confirmed in Figure 3.18. In the experiment of section 4.5, pLSA is used for compression. If the number of latent classes is set to twenty or more, the number of side information after compression will theoretically increase from the number of original side information before compression, thus the upper number of latent classes is set to nineteen.

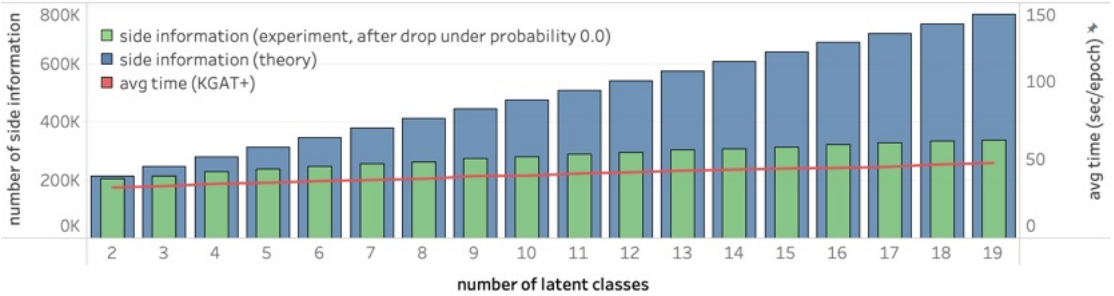


Figure 3.18: Transition of the number of side information (theoretical and experimental) and transition of computational time (experimental) when the number of latent classes K is changed

Figure 3.18 shows that as the number of classes increases, the proportion of dropped side information due to a class membership probability of 0.0 after compression deviates from the theoretical number of classes. Additionally, the reduction in computational time corresponds to the number of side information after compression and dropping. Moreover, the computational cost of training a latent class model is significantly lower compared to the training cost of KGAT, and the increased computational cost resulting from the number of classes is negligible.

Therefore, when determining the number of latent classes, it is advisable to search for the number that achieves the highest accuracy or utilize criteria such as AIC and BIC while narrowing down the range based on the theoretical number of side information after compression. Furthermore, in our experiment, the number of classes is uniformly set to simplify the model, regardless of the relation of the compression target. However, if complexity can be accommodated, adjusting the number of classes for each relation can further reduce computational time and improve accuracy.

3.6.4 On the Loss Function

In the improved KGAT included in the proposed framework, the loss L_{KG} and L_{CF} are minimized simultaneously during training, as expressed in Eq. (3.13). However, a problem that often occurs in such a loss function including multiple types of loss is that each loss cannot be completely converged. In contrast, in our experiments, as shown in Figure 3.19, both losses converge stably as learning progresses.

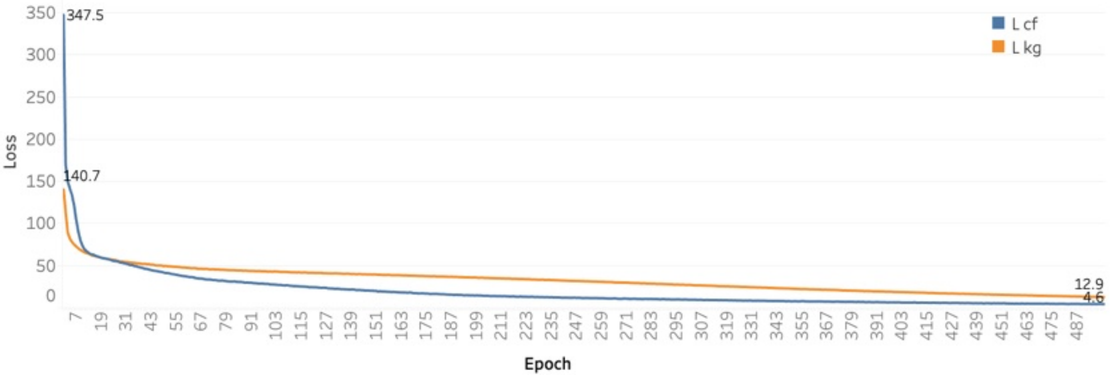


Figure 3.19: Result of the convergence analysis experiment of each loss (KGAT+(pLSA-6))

In the improved KGAT, as in the conventional KGAT, both losses were trained equally without weighting. In our experiments, we did not observe a significant difference in the magnitudes of the two losses, indicating that the equal weighting method was appropriate. However, in cases where one loss dominates the other in magnitude, it may be necessary to balance the losses by assigning appropriate weights. This is a common practice in many state-of-the-art studies [61, 92, 117] to ensure stable training and convergence.

3.6.5 On the Harmful Effects of Compression

In the proposed approach, many-to-many side information is compressed and learned, and the original relationship is restored by calculating the inner product of attention weights divided for each latent class. However, this approach may not completely reproduce the original side information present in the original data before compression. Specifically, the final calculated attention weight may become a positive value for a triplet that does not exist in the original data and is used for interpretation. Conversely, the attention weight value of the existing triplet in the original data may become 0.0, rendering it unusable for interpretation. Essentially, the new triplets are used for interpretation, and the existing important triplets are judged not to be excluded based on the similarity of the side information. Thus, these phenomena are not necessarily problematic; they can be considered beneficial depending on how these results are used. In addition, when performing recommendations with the explanation system in practice, for a triplet that does not exist in the original data, for example, a sentence such as "this item is recommended to you because this item is sold as the brand preferred by users who have similar preferences to you in terms of their favorite brands" can be presented. This may lead to a better presentation of the reason for the recommendation. Given that in the conventional approach, only triplets that existed in the original data could be used for interpretation, it may be considered that the proposed approach improves the richness of the interpretability of the model.

3.6.6 Applicability to Other Side Information

In this study, we focused on many-to-many type side information and compressed it; however, there is various other high-dimensional side information that can be included in learning the model of the model-intrinsic approach. For example, image data (e.g., item image) and text data (e.g., reviews and descriptions of items) are generally high-dimensional. They are typical examples in which compression is likely to be effective. In applying the proposed framework, it is necessary to devise compression and restoration methods based on each data. However, the concept of the proposed framework for compression and restoration alleviates the challenge of computational time and the challenge of a trade-off between interpretability and accuracy. Therefore, the proposed framework opens up possibilities for utilizing various types of (previously challenging) high-dimensional side information in the knowledge graph-based model-intrinsic approach.

3.6.7 Applicability to Other Models

Furthermore, this study has improved the KGAT model to allow edges with probabilistic relations to be considered during learning, resulting in a significant reduction in the amount of calculation. Moreover, the proposed framework can be applied to other similar knowledge graph-based model-intrinsic approaches. By considering the probabilities at specific points during internal calculations, the reduction in computational complexity can be achieved relatively easily by introducing additional steps, such as incorporating prior probabilities into sampling. Hence, the proposed model is also valuable because of its high versatility.

3.7 Conclusion of this Chapter

In this chapter, we have proposed an explainable recommendation framework with an improved KGAT model. In the experiments conducted, the proposed framework significantly improved the performance of prior methods, such as the conventional KGAT model, in terms of computational time. Therefore, the proposed approach is expected to dramatically increase the likelihood that such methods will be applied in practice by overcoming the limitations of existing approaches. In addition, while the proposed framework retained the capability of explanation, the results show that equal or higher accuracy can be expected compared to the conventional KGAT. Notably, in the proposed framework, the users' side information that was not considered in the conventional model for explanation can be utilized to construct the explanation. Effectively learning the side information that previously involved substantial difficulties in high computational time enabled our model to exhibit a higher interpretability. These results were obtained by employing the proposed framework's capabilities of compressing and recombining data structures, which enabled the KGAT-based machine learning models to consider soft edges. Moreover, this idea of learning soft edges can be extended to other models that employ similar knowledge graph structures for explainable artificial intelligence, resulting in significant reductions in computational time for various tasks. Therefore, the proposed framework has demonstrated its potential for enabling the practical application of knowledge graph-based explainable recommendation models to real-world services and data.

Chapter 4

Automatic Fashion Image Interpretation via Fashion Intelligence System

This chapter defines a novel technology and research field named “fashion intelligence system” that supports consumers in understanding fashion-specific ambiguous verbal expressions. A new method that utilizes the vision and language model as the starting point of the fashion intelligence system is proposed in this chapter.

4.1 Purpose of this Chapter

In recent years, it has become common for consumers to be acquainted with the fashion outfits of others through social network services (social media) and e-commerce sites and engage in their fashion-item-purchasing activities. In addition to multi-topic-type social media and e-commerce sites, such as Instagram [1], Facebook [2], Twitter [3], and Amazon [118], fashion-specific services, such as ZOZOTOWN [4], WEAR [5], and Snap Fashion [119], are also used extensively. Users’ searches for full-body outfit images on these services’ applications (apps) or e-commerce sites are crucial activities for refining their fashion-based shopping experiences concerning fashion and making purchasing decisions. Therefore, making fashion-based browsing (internet surfing in the fashion domain) comfortable is important because it develops relations in the fashion industry.

However, fashion is a fuzzy and complex domain that contains many abstract elements, and it can be challenging for consumers to understand and interpret fashion, especially for non-experts, because abstract expressions such as “casual,” “formal,” and “cute” are normally used when explaining fashion. For example, questions such as “what factor makes this outfit casual?,” “how

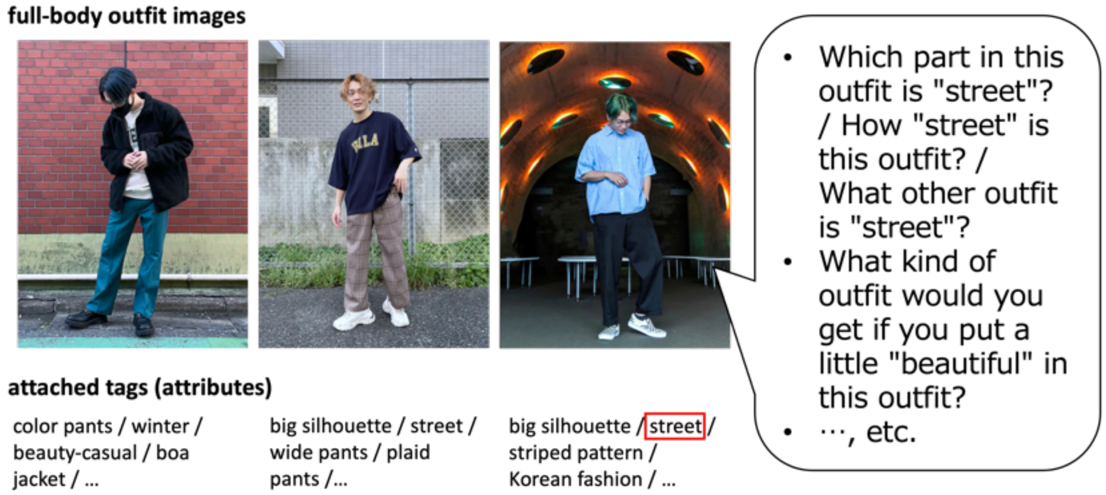

full-body outfit images
attached tags (attributes)
color pants / winter / beauty-casual / boa jacket / ...
big silhouette / street / wide pants / plaid pants /...
big silhouette / street / striped pattern / Korean fashion / ...
Which part in this outfit is "street"? / How "street" is this outfit? / What other outfit is "street"?
What kind of outfit would you get if you put a little "beautiful" in this outfit?
···, etc.

Figure 4.1: Image of fashion abstract problem

casual is this outfit?," and "what kind of outfit would it be if this outfit was made a little more formal?" are difficult to answer, especially for non-experts (and not easy for experts either). This difficulty in interpreting fashion is often encountered on the internet because users cannot rely on human experts. Thus, the ambiguity and complexity of online shopping can make it difficult for many users to engage with new lines of outfits and hinder their deep interest in the fashion industry. Therefore, the automatic discovery of answers to such questions is expected to greatly contribute to the fashion industry by making users' online full-body-outfit-image-retrieval and fashion-item-purchasing more comfortable. Simultaneously, it is expected to broaden users' perception and help interpret fashion outfits and encourage their interests, not always for business purposes. These aspects indicate a pressing need for a "fashion intelligence system" that will help to broaden the perceptions of and interests in fashion, rather than a "business intelligence system" that aims to plan a marketing strategy.

For this purpose, this study proposes a fashion intelligence system for online outfit coordination based on a visual-semantic embedding method for automatically learning and interpreting fashion and obtaining answers to users' questions. Our proposed method can embed the abundant abstract tag information in the same projective space as the outfit images. Calculating similarities between images and tags in a mapped space helps search outfit images using fashion-specific abstract words. Furthermore, visually estimating the degree of relevance between images and tags makes it possible to interpret abstract words. In this study, we focus on full-body outfit images, which include an extensive range of backgrounds even after detecting a person in a rectangular form due to the shape of the object (person). Therefore, we introduced foreground-centered learning and background regularization terms that utilized grid weight maps obtained through semantic segmentation. Moreover, the proposed system incorporates sensitive negative sampling based on a latent class model to address the complexity and interpretational challenges arising from abstract expressions, such as "casual," "formal," and "cute," present in the target datasets. These abstract expressions significantly contribute to the intricacy and difficulty of interpreting the fashion domain.

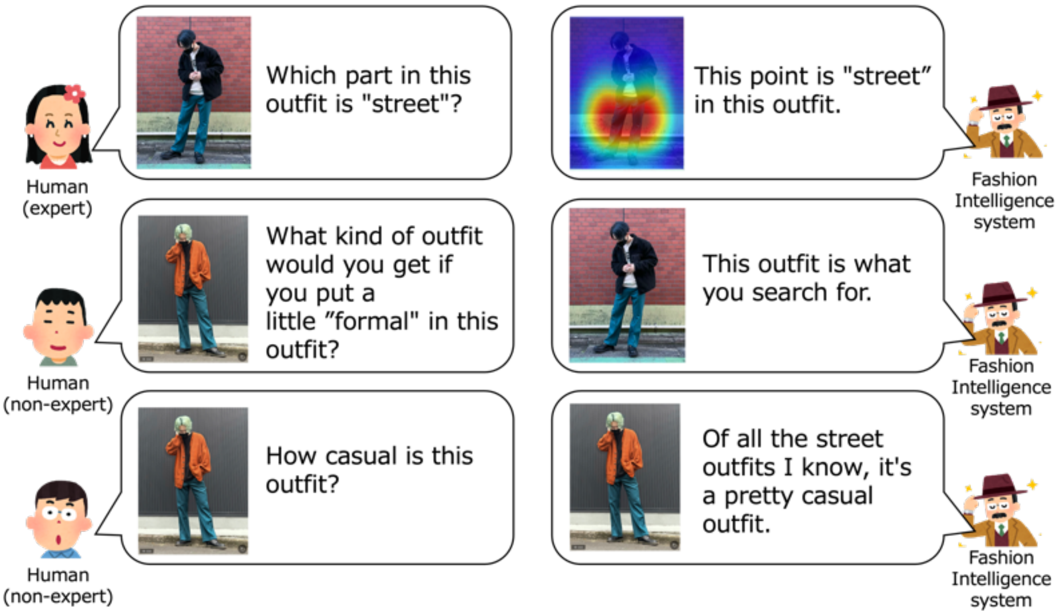

Human (expert)
Which part in this outfit is "street"?
This point is "street" in this outfit.
Fashion Intelligence system
Human (non-expert)
What kind of outfit would you get if you put a little "formal" in this outfit?
This outfit is what you search for.
Fashion Intelligence system
Human (non-expert)
How casual is this outfit?
Of all the street outfits I know, it's a pretty casual outfit.
Fashion Intelligence system

Figure 4.2: Image of fashion intelligence system

Furthermore, results from multifaceted evaluation experiments and analyses using accumulated data from real-world fashion services suggest that the proposed system has valuable applications in the industry. Thus, this research helps decrease fashion-specific ambiguities and complexities (including the dependence on personal judgments) and supports users (experts and non-experts alike) in fashion-related marketing activities and choices.

The main contributions of this chapter are summarized as follows. 1) We define a novel technology and research area called "fashion intelligence." 2) We propose a fashion intelligence system (based on a visual-semantic embedding method) enabling the embedding of full-body outfit images and rich fashion-specific abstract tags in the same projective space. 3) The proposed system is applied to multiple datasets accumulated from actual fashion-related services, and its efficacy is verified based on the results of multifaceted evaluation experiments and analyses. 4) Based on the multifaceted experimental results and analyses, we introduce the application of the proposed system to fashion marketing strategy planning.

4.2 Differences from Related Works

4.2.1 Application of Artificial Intelligence in the Fashion Industry

As mentioned in Chapter 2, numerous studies have proposed various artificial intelligence (AI) methods for application in the fashion and apparel industries [11]. In addition, in a broad sense, these studies are included in the framework of business intelligence [19] and do not extend this concept.

In contrast, we propose "fashion intelligence," defined as intelligence about fashion. More specifically, this concept involves diverse knowledge obtained by analyzing information for various fashion-related decisions and a mechanism to obtain such knowledge. Whereas conventional studies have not gone beyond the business intelligence framework, our definition of fashion intelligence envisions the generation and discovery of new knowledge by targeting fashion, which is evaluated and imaged differently depending on people's preferences, values, and cultural backgrounds. Thus, we focused on the question of "how to handle ambiguity in human evaluations and judgments," and the handling of this ambiguity is also our primary focus. In this area, we define a "fashion intelligence system" as a mechanism that facilitates the discovery of new knowledge and the creation of new values about fashion by automatically interpreting fashion and collaborating with people (users) through dialogues.

Clearly, the concepts underlying the fashion intelligence system proposed in this study differ fundamentally from those associated with conventional business intelligence systems, such as those used by companies to make business decisions. Fashion intelligence systems are primarily used by ordinary consumers and experts who support consumers (although they can be used meaningfully for corporate activities) and provide functions that support the discovery of knowledge and values about fashion. Thus, a fashion intelligence system is not an AI that is introduced considering a perspective that is easily evaluated objectively, such as profit or efficiency, or an AI that replaces the role of experts. In contrast, it aims to strongly support end-users who desire to revel in the fashion world by acquiring knowledge and making new discoveries.

4.2.2 User Support with Explainable Recommendation in the Fashion Domain

As mentioned in Chapter 2, the concept of explainable recommendations has been proposed and is currently actively being researched to support online user decision-making [42]. In this research field, beyond the recommendation system's decision to "recommend this item to this user," the reasons "why this item is recommended to this user" are also given by the system.

In contrast, the desired decision support for users, especially non-experts, is to interpret the abstract and difficult aspects of the fashion domain, They seek answers to questions like "where does this fashion item (outfit) fall on the casual (formal) spectrum?" This interpretation of fashion is crucial, especially in online settings, where users, both experts and non-experts, rely on their own experiences to interpret fashion independently. It allows them to determine what to wear, what is considered fashionable, and what is not. While easy access to answers would enhance users' understanding of fashion, their motivation to explore clothing options, and their desire to make purchases, these questions pose significant challenges for non-experts (and even for experts). However, addressing this support requirement is not the primary focus of explainable recommendation technology.

Furthermore, achieving complex multimodal learning with item image information, user information, and sometimes textual information within a single model has led to intricate model structures. However, the practical implementation of these models is challenging due to learning complexity and computational costs. To ensure robust and practical support applications, it is preferable to adopt a simple and easily understandable model structure whenever possible.

4.2.3 User Support with Fashion Image Retrieval

As mentioned in Chapter 2, fashion image retrieval is a highly active research area in image processing. However, most existing studies focus on clothes retrieval rather than outfit (full-body) retrieval. Full-body images, including single images with multiple products, have an extensive range and complex background, and targeting them is a challenging task [120].

Moreover, in previous studies, the words added or removed from images were specific (non-abstract) words (such as "short-sleeve" and "black"). Although this improves the search efficiency, it does not contribute to users' understanding and interpretation of fashion. For instance, in Figure 2.4, if the word "short-sleeve" is added to a long-sleeved blue shirt, it is easy, even for non-expert users, to answer "short-sleeved blue shirt." Conversely, because of its difficulty, what users (especially non-experts) truly want to be supported on is to obtain the answers to abstract and somewhat complex questions, such as “what kind of fashion outfit would you recommend if I want to make this fashion outfit a little more casual (formal)?” Obtaining answers to these questions is extremely helpful in deciding what clothes users will wear on the day and what items they will purchase. In [121], item search is realized by considering only eight types of selectable abstract category information, such as “party,” “outdoor,” and “summer.” However, this previous study only dealt with a limited number (eight) of information categories.

In this study, we focus on a system that includes a search function for full-body outfit images by utilizing more than 1,000 types of tags and a visual interpretation function via an attribute activation map (AAM). As we mentioned, it is extremely difficult for non-experts to play this role independently (and it is not easy for experts either).

4.2.4 Fashion Concept Discovery

VSE, included in [86] as a discovery concept, is a method of embedding in the same projective space a fashion item image and a specific word contained in the item description. Thus, VSE in [86] allows searching for fashion images by calculating similarities between individual item images and simple words and to locate specific points on the target item image that are highly related to the word, thereby creating an AAM. These multiple functions constitute the advantage of VSE. The main task of [86] is the automatic discovery of concepts such as “size,” “color,” and “shoulder shape.” For example, words such as maxi and mini belong to the concept of “size,” and words such as “one-shoulder” and “no-sleeves” refer to “shoulder shape.” By utilizing the embedded representations of images and words acquired in this concept discovery process, it is possible to calculate similarities between images and words and discover parts of images with high relevance to related words. Compared to other attribute-based individual clothes image retrieval methods, the strength of VSE in [86] is that it not only executes image-retrieval tasks but also creates an AAM due to the simplicity of its model structure.

This fashion concept discovery process consists of three processes: “visual-semantic embedding training,” “spatially-aware concept discovery,” and “concept subspace learning,” with image and word embedding being mainly performed in the “visual-semantic embedding training” step. The strength of this research includes the results of searching and extraction of attention parts but also the relatively simple structure of the VSE model. Specifically, only 1) the embedded representation of each word, 2) the transform matrix for converting the output obtained from a CNN into the embedded representation of the image, and 3) the parameters included in the CNN itself are targeted for update. The simplicity of the model contributes to the ease of implementation and management, the ease of updating parameters, and the realization of short computational times. Hence, this aspect enables the realization of model application in actual business scenarios.

However, the main task of this previous research was to discover concepts automatically, and the target image data was an image of an individual fashion item (similar to other studies on explainable recommendations and fashion image retrieval). Moreover, because the target image data are related to products for sale, it is usually an aesthetically pleasing image captured by an expert photographer with a simple background. Moreover, owing to the shape of the item, most rectangles, after detecting the area where the fashion item appears, contain almost no background. Furthermore, the target word data is simple (as in other studies on explainable recommendation and fashion image retrieval) and incorporates concrete words, such as “maxi,” “mini,” “one-shoulder,” and “no-sleeves.” Consequently, the image retrieval function can realize simple search tasks that even non-experts can easily answer, as in other studies. Therefore, extracting attention parts on images that are highly related to words is simple enough that even non-experts can easily answer, such as “no-sleeves in this item is here.”

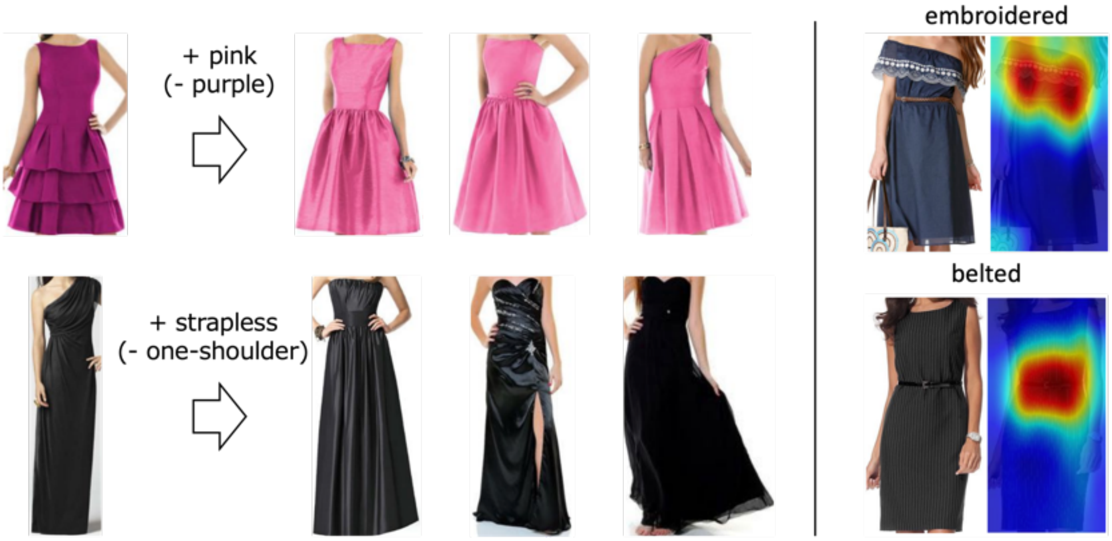
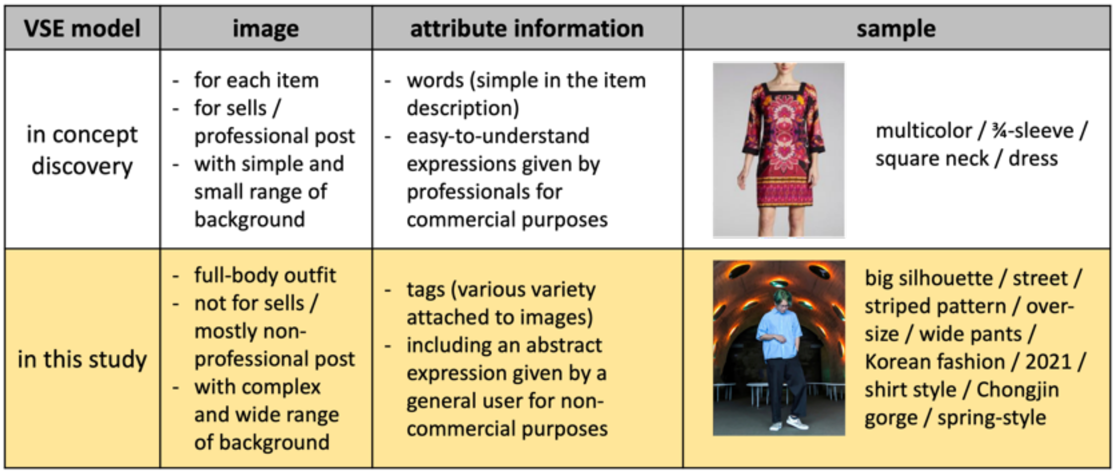


Figure 4.3: Contributions of the previous visual-semantic embedding method in fashion concept discovery [86, 122]

In contrast, the target data in this study are images that include full-body outfits. As most general users post those images on social media, there can be a variety of backgrounds. In addition, because full-body outfits are targeted, the rectangle after object detection always includes an extensive range and variety of backgrounds, depending on the pose and shape of the person. However, unlike in previous studies, tag data added for non-commercial purposes to convey images of outfits are targeted. Therefore, many abstract expressions are also included in addition to specific tags.



VSE model	image	attribute information	sample
in concept discovery	- for each item - for sells / professional post - with simple and small range of background	- words (simple in the item description) - easy-to-understand expressions given by professionals for commercial purposes	multicolor / ¾-sleeve / square neck / dress
in this study	- full-body outfit - not for sells / mostly non-professional post - with complex and wide range of background	- tags (various variety attached to images) - including an abstract expression given by a general user for non-commercial purposes	big silhouette / street / striped pattern / over-size / wide pants / Korean fashion / 2021 / shirt style / Chongjin gorge / spring-style

Figure 4.4: Comparison with previous visual-semantic embedding in fashion concept discovery [86, 122]

In fashion concept discovery, individual models (processes) were learned for each category of fashion items, such as "dress," "top," and "pants." In contrast, in our proposal, a set (full-body outfit) including all of these items is collectively learned with one model. This is greatly beneficial because applying the models to businesses reduces the number of models to be managed. However, considerable ingenuity is required to learn all the information in full-body outfit images and abstract tags simultaneously and efficiently.

4.2.5 Motivation for this Chapter

In summary, the motivations for this chapter, based on the review of related works, are given as follows:

- To accurately quantify full-body outfit images and related abundant tags, including abstract tags, in one comprehensive model.

- To achieve image retrieval that is difficult for users (experts and non-experts) by calculating full-body outfit images and tags similarities from fashion-specific abstract expressions.
- To clarify the parts in the image that are highly related to each tag (thereby creating an AAM) and make fashion-specific abstract tags visually interpretable.

Considering the aforementioned motivations, this study proposes a visual-semantic embedding model that enables the interpretation of fashion attribute information (including abstract aspects) and decreases dependencies on human experiences or knowledge in the fashion field. We present an application of the model that broadens users' perceptions and supports their online purchasing and browsing activities. We propose a novel fashion intelligence system that makes it possible to perform tasks difficult for users (experts and non-experts), which were not targeted in conventional studies on explainable recommendations and fashion image retrieval.

4.3 Proposed System

4.3.1 Problem Definition

In this study, we propose a system that supports the visual interpretation of fashion information by embedding full-body outfit images and tag information in the same projective space to quantify the information. By completing this research, users (especially non-experts) will be able to understand fashion independently, leading to increased motivation for fashion and purchasing desire. Thus, the proposed system is expected to improve the usability of all social media and e-commerce sites handling fashion items, including fashion-related posts. Furthermore, it is expected to broaden the understanding and perception of users (including experts) and aid in planning marketing strategies.

The image data targeted in this study are full-body fashion pictures of single subjects (persons) captured against various backgrounds. Even if the smallest rectangle reflecting only one person is extracted from these images, a wide and complex background is included in the rectangle to account for the problem introduced by the shape of the person. In particular, the left and right areas of the neck, between the legs, and the left and right areas of the legs are all part of the background. In addition, background patterns vary broadly (especially in social media data), and many cases exist wherein images other than the subject are not one-tone backgrounds.

Therefore, if the entire contents of a rectangle are learned in their original form, the background becomes noise, and sensitive learning cannot be realized.

Multiple tags are attached to each image as attribute information from post users. This tag information includes specific and simple tags (e.g., “denim,” “skirt,” and “t-shirt”) along with abstract tags (e.g., “spring-style,” “formal,” “casual,” and “office-casual”). Regarding their characteristics, specific tags, once attached, are always correct regardless of the sensibility of post users. Conversely, the characteristics of abstract tags, whether attached or not, are uncertain and depend on the sensibility of post users. For example, in the sensibility of post user A, if image A is completely “casual,” then the tag “casual” may be correctly attached. In contrast, if post user B feels that image A is only partially casual, the “casual” tag may not be attached by this user. For post user C, if the expression “holiday-style” seems more appropriate than “casual,” “holiday-style” will be attached rather than “casual.” Thus, a target image includes not only specific tags but also abstract tags. Abstract expressions are one of the major reasons why users (especially non-experts) find the fashion domain difficult.

4.3.2 Visual-Semantic Embedding

The proposed system allows embedding specific and abstract tags in the same space as full-body outfit images. To handle a complex and extensive background included in a rectangle and abundant tag information, including abstract expressions, the proposed system introduces certain innovations to enable sensitive learning. The entire model structure is shown in Figure 4.5. In the following sections, a systematic description of the proposed model is reported.

4.3.2.1 Person Detection

Based on a trained object detection model, the smallest rectangle reflecting only one person is extracted from an entire posted image related to a full-body outfit image. Single shot multibox detector (SSD) [123] and faster-RCNN [124] are generally used for person detection (we chose SSD, but this is not the main point of this research). The smallest rectangle, where only one person is reflected, and the maximally scraped-off background, is obtained using these models.

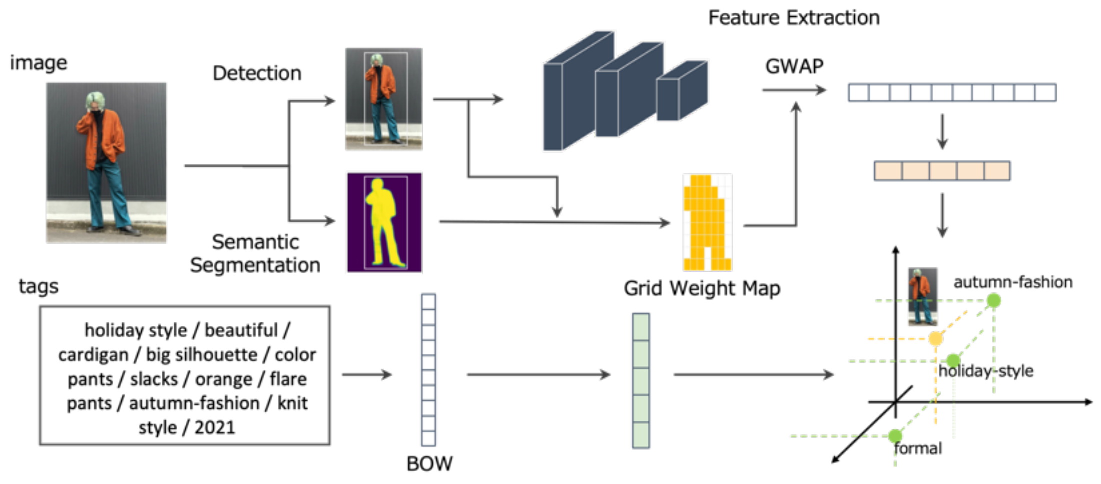


image
Detection
Semantic Segmentation
Feature Extraction
GWAP
Grid Weight Map
tags
holiday style / beautiful / cardigan / big silhouette / color pants / slacks / orange / flare pants / autumn-fashion / knit style / 2021
BOW
autumn-fashion
holiday-style
formal

Figure 4.5: Structure of a prototype of our visual-semantic embedding model proposal

4.3.2.2 Grid Weight Map Based on Semantic Segmentation

Several studies have focused on person foreground segmentation [125, 126]. From the perspective of ease of handling, in this study, semantic segmentation is performed using a trained fully convolutional network (FCN) [127]. During semantic segmentation, we determine the probability of whether or not a cell corresponds to a "person." This output is obtained for each pixel. Based on the value calculated for each pixel, an average value is determined for each grid, which refers to an area obtained when a target image is divided vertically into I and horizontally into J. If the average value of the probabilities calculated for each grid equals or exceeds a threshold α, the weight corresponding to the (i, j)-th grid is set to 1.0, and the others are set to 0.0. A grid weight map corresponding to the foreground is obtained by normalizing the weight obtained through this procedure. Moreover, a grid weight map corresponding to the background is obtained in the same step with the 1.0 and 0.0 settings swapped.

4.3.2.3 Loss Function

Upon transferring the trained CNN and re-training with a target dataset, an extractor for acquiring image features is obtained. Several types of CNN models can be used as extractors; however, the complexity of the dataset and the allowable computational cost are the determining factors. In this research, the GoogleNet Inception V3 model [128] is used as an extractor based on the original study [86].

Global weighted average pooling (GWAP) [129] is performed using the image features extracted from the CNN and the grid weight map. This implies replacing the global average pooling (GAP) layer remaining after the final convolutional layer of a general CNN with the GWAP layer. To extract high-quality image features, the idea of delicately handling the features for each grid is effective [130]. This GWAP operation significantly reduces the problem of capturing noise caused by the GAP layer, which is not considered in numerous related methods [86, 28]. Assuming K as the dimension of the embedded space, the embedded representation of image $\mathbf{x}_{\mathrm{f}} \in \mathbb{R}^K$ in the foreground can be calculated as follows:

$$\begin{aligned} \mathbf{x}_{\mathrm{f}} &= \mathbf{W}_{\mathrm{I}}\mathbf{f}_{\mathrm{f}}, \qquad (4.1) \\ \mathbf{f}_{\mathrm{f}} &= \sum_{i,j} g_{(i,j)}\mathbf{q}_{(i,j)}, \qquad (4.2) \end{aligned}$$

where $\mathbf{W}_{\mathrm{I}} \in \mathbb{R}^{K \times D}$ is the transform matrix for converting the image feature vector extracted from the CNN to image embedded representation (D is the number of dimensions of the final convolutional layer of the CNN), $\mathbf{f}_{\mathrm{f}} \in \mathbb{R}^D$ is a weighted image feature vector for the foreground acquired via the grid weight map obtained from the GWAP layer, $g_{(i,j)}$ is a grid weight for the (i, j)-th grid corresponding to the foreground, and $\mathbf{q}_{(i,j)} \in \mathbb{R}^D$ is the foreground image feature vector for the (i, j)-th grid obtained from the final convolution layer of the CNN. In contrast, we calculate the embedded representation of the image in the background $\mathbf{x}_{\mathrm{b}} \in \mathbb{R}^K$ as follows:

$$\begin{aligned} \mathbf{x}_{\mathrm{b}} &= \mathbf{W}_{\mathrm{I}}\mathbf{f}_{\mathrm{b}}, \qquad (4.3) \\ \mathbf{f}_{\mathrm{b}} &= \sum_{i,j} g'_{(i,j)}\mathbf{q}_{(i,j)}, \qquad (4.4) \end{aligned}$$

where $\mathbf{f}_{\mathrm{b}} \in \mathbb{R}^D$ is a weighted image feature vector for the background acquired via the grid weight map obtained from the GWAP layer and $g'_{(i,j)}$ is a grid weight for the (i, j)-th grid corresponding to the background.

In addition, tags attached to datasets can vary in frequency. For example, seasonal tags such as “spring-style” are frequently added to the entire set of images posted in spring. By contrast, tags such as “beret” are attached only to the images of people wearing them; therefore, they are rarely attached to a whole dataset. Typically, infrequent tags are likely to be important factors that characterize an image (i.e., differentiate from other images). To use this property in this study, we obtained the embedded representation of the tags $\mathbf{v} \in \mathbb{R}^K$ attached to a target image using Eq. (4.5) by weighting the tags such that the less frequent tags have a greater effect.

$$\begin{aligned} \mathbf{v} &= \sum_t w_t \mathbf{a}_t, && (4.5) \\ \mathbf{a}_t &= \mathbf{W}_\mathrm{T} \mathbf{e}_t, && (4.6) \\ w_t &= \frac{1/\log_e(N_t+1)}{\sum_t 1/\log_e(N_t+1)}, && (4.7) \end{aligned}$$

where $\mathbf{a}_t \in \mathbb{R}^K$ is an embedded representation for the t-th single tag among the tags attached to the target image, $\mathbf{W}_\mathrm{T} \in \mathbb{R}^{K \times C}$ is a transform matrix for converting tag feature vectors with tag embedded representations (C is the number of tags in the entire dataset), $\mathbf{e}_t \in \mathbb{R}^C$ is an one-hot vector for the t-th single tag, and N_t indicates the total attachment frequency of the t-th attached tag to the target image in the entire batch dataset.

Using these features, we realize an operation that focuses only on the foreground (i.e., foreground-centered learning) and is robust to the noise of the wide and complex background (background regularization). Thus, the loss function is defined as follows:

$$\begin{aligned} \mathcal{L}(\Theta) = &\sum \max\left(0, m - s(\mathbf{x}_\mathrm{f}^+, \mathbf{v}^+) + s(\mathbf{x}_\mathrm{f}^+, \mathbf{v}^-)\right) \\ &+ \sum \max\left(0, m - s(\mathbf{v}^+, \mathbf{x}_\mathrm{f}^+) + s(\mathbf{v}^-, \mathbf{x}_\mathrm{f}^+)\right) \\ &+ \beta \sum \max\left(0, m + s(\mathbf{x}_\mathrm{b}^+, \mathbf{v}^+)\right), \qquad (4.8) \end{aligned}$$

where $\Theta = \{\mathbf{W}_\mathrm{I}, \mathbf{W}_\mathrm{T}, \mathbf{V}\}$ is a set of parameters to be optimized, $\mathbf{V}$ is a parameter set contained in CNN, $s(\mathbf{x}, \mathbf{y})$ indicates the cosine similarity between vectors $\mathbf{x}$ and $\mathbf{y}$, and $\beta > 0$ is a positive hyperparameter to adjust the importance of the background regularization term. Furthermore, the superscript sign + of A^+ indicates that A is a variable related to the positive sample, and − of A^- indicates that A is a variable related to the negative sample.

By updating each parameter to optimize the loss function in Eq. (4.8), we obtain a transform matrix for mapping the embedded representation corresponding to each tag and the image features obtained from the CNN in the same space as the tag.

4.3.2.4 Negative Sampling Based on the Latent Class Model

Among the optimization steps, for each set of an image and tags (positive samples), a set of an image and tags (negative) is sampled (negative sampling). As in many other studies, if sampling is performed with equal probability (i.e., completely random) from all the sets of images and tags, it becomes difficult to realize efficient and sensitive optimization. Devising negative sampling to fit the target problem facilitates efficient and effective learning [131, 132]. For example, many studies in areas such as natural language processing [133, 134] and graph convolution network [135, 136] have devised negative sampling mainly with the aim to reduce the computation amount. In this study, we additionally improve negative sampling for learning abstract tags, in terms of effectiveness, efficiency, and sensitivity.

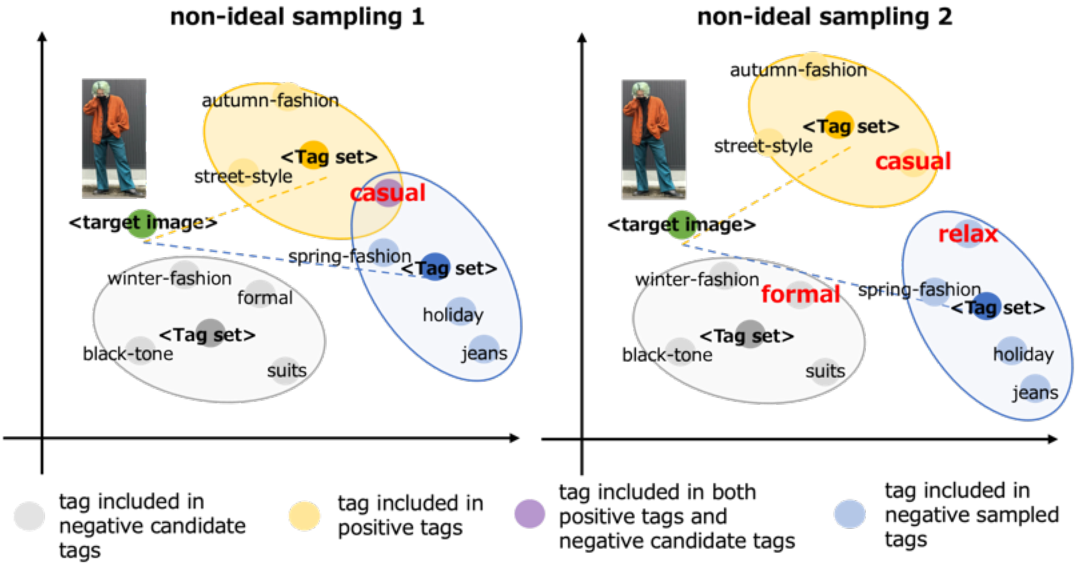


Figure 4.6: Image of a negative sampling

In this study, the first problem is that, as shown on the left in Figure 4.6, candidates including duplicated tags with positive tags and candidates not including them are treated equally. For instance, if the tag set containing "autumn-fashion," "street-style," and "casual" is a positive sample, then the other set, set 1 = ("winter-fashion," "formal," "black-tone," and "suits") and set 2 = ("casual," "spring-fashion," "holiday," and "jeans"), is a candidate for negative sampling. In this case, set 1 should be sampled because there is a duplication of tags between a positive sample and set 2. Therefore, in the proposed system, the tag sets that include a duplicated tag with the positive sample tags are excluded from being candidates for negative sampling.

As shown on the right in Figure 4.6, the set containing "autumn-fashion," "street-style," and "casual" is a positive sample, and set 1 = ("winter-fashion," "formal," "black-tone," and "suits") and set 2 = ("relax," "spring-fashion," "holiday," and "jeans") are the candidates for negative sampling. Herein, both sets have no duplication of tags with the positive tag set. However, as set 2 contains similar tags (e.g., “casual” and “relax”), set 1 should be sampled compared to set 2. This problem often happens when target tags data include abstract tags (in this case, the post user in set 2 chose the tag “casual” instead of “relax”). Therefore, in the proposed model, data related to tags are applied to a latent class model [105, 106], which is a powerful method for extracting necessary information while compressing complex data [102, 137], and the topic distribution related to the tags is acquired for each candidate. Moreover, negative sampling is performed based on the distance of distributions (Kullback-Leibler divergence) between the positive sample and all negative sample candidates.

The following sample probability p_i is given to the i-th sampling candidate by the two proposed methods. Eq. (4.9) is a definition formula of probability $p_i^{(1)}$ that excludes tag sets duplicate with a positive sample, and Eq. (4.10) is a definition formula of probability $p_i^{(2)}$ that considers the distribution of tags similarity between a positive sample and each candidate for negative sampling.

$$p_i^{(1)} = \begin{cases} 0.0 & (\text{if } \mathbf{e}^{+\mathsf{T}}\mathbf{e}_i > 0), \\ 1.0 & (\text{if } \mathbf{e}^{+\mathsf{T}}\mathbf{e}_i = 0), \end{cases} \tag{4.9}$$

$$p_i^{(2)} = \begin{cases} 0.0 & (\text{if } \mathbf{e}^{+\mathsf{T}}\mathbf{e}_i > 0), \\ \mathrm{D}_{\mathrm{KL}}\left(q_{\mathrm{lda}}(Z|T^{+}), q_{\mathrm{lda}}(Z|T_i)\right) & (\text{if } \mathbf{e}^{+\mathsf{T}}\mathbf{e}_i = 0), \end{cases} \tag{4.10}$$

where $\mathbf{e}^{+} \in \mathbb{R}^{C}$ is a bag-of-words representation of the positive tag set, $\mathbf{e}_i \in \mathbb{R}^{C}$ is a bag-of-words representation for the tag set of the i-th candidate, $\mathrm{D}_{\mathrm{KL}}(A, B)$ expresses Kullback-Leibler divergence between A and B, $q_{\mathrm{lda}}(Z|T^{+})$ is a topic distribution for the positive sample, and $q_{\mathrm{lda}}(Z|T_i)$ is a topic distribution for the i-th candidate obtained from latent Dirichlet allocation (LDA) based on the attached tags T_i. Furthermore, the probability p_i is normalized and used for sampling.

Thus, the tag set that does not have duplicate tags with the positive tags and has a distant meaning from the attached tags is sampled for the negative tag set. Hence, the tag information is effectively used to realize more efficient and sensitive optimization.

4.3.3 Creating an Attribute Activation Map

The degree of relevance between the arbitrary single tag embedded representation $\mathbf{a} \in \mathbb{R}^K$ and the arbitrary image embedded representation $\mathbf{x} \in \mathbb{R}^K$ is expressed as follows:

$$\begin{aligned} s(\mathbf{a}, \mathbf{x}) &= \sum_k a_k x_k \\ &= \sum_k a_k \sum_d W_{I_{kd}} f_{f_d} \\ &= \sum_k a_k \sum_d W_{I_{kd}} \sum_{i,j} g_{(i,j)} q_{(i,j)_d} \\ &= \sum_{i,j} g_{(i,j)} \sum_k a_k \sum_d W_{I_{kd}} q_{(i,j)_d}, \end{aligned} \tag{4.11}$$

where a_k and x_k are the values in the k-th dimension of tag embedded representation $\mathbf{a}$ and the image embedded representation $\mathbf{x}$, respectively. In addition, $W_{I_{kd}}$, f_{f_d} and $q_{(i,j)_d}$ are the values in the d-th dimension of the k-th vector $\mathbf{W}_{\mathrm{I}_k} \in \mathbb{R}^D$ in the transform matrix $\mathbf{W}_{\mathrm{I}}$ for the image, the output of the GWAP layer $\mathbf{f}_{\mathrm{f}} \in \mathbb{R}^D$ for the foreground, and the (i, j)-th grid obtained from the final convolution layer of the CNN $\mathbf{q}_{(i,j)} \in \mathbb{R}^D$, respectively.

Based on Eq. (4.11), the degree of relevance between the tag and the (i, j)-th grid on the image is expressed as follows:

$$M(i, j) = \sum_k a_k \sum_d W_{I_{kd}} q_{(i,j)_d}, \tag{4.12}$$

and a set of $M(i, j)$ for all grids is called AAM between the image and the tag.

AAM allows users to visually understand where the target tag is related to the target image. For example, if a user wants to know “what is the casual point on a full-body outfit image A?,” it can be visually grasped by measuring the degree of relevance between image A and the tag “casual.”

4.3.4 Image Retrieval

Image retrieval is realized by adding (positive) and subtracting (negative) a tag to the query image, and is given by the following Eq. (4.13).

$$\mathbf{x_o} = \operatorname*{argmax}_{\mathbf{x}} s\left(\sum_{i,j} \delta_{q(i,j)} \left(\mathbf{W}_{\mathrm{I}}\mathbf{q}_{\mathrm{q}(i,j)} + \mathbf{a}_{\mathrm{p}} - \mathbf{a}_{\mathrm{n}}\right), \mathbf{x}\right), \tag{4.13}$$

$$\delta_{q(i,j)} = \begin{cases} 1.0 & (\text{if } M_n^q(i,j) g_{(i,j)} > 0), \\ 0.0 & (\text{if } M_n^q(i,j) g_{(i,j)} \leq 0), \end{cases} \tag{4.14}$$

where $\mathbf{x_o} \in \mathbb{R}^K$ is the image embedded representation of the search result, $\delta_{q(i,j)}$ is a binary output indicator of whether the (i, j)-th grid in the query image is the computed operation target, $\mathbf{q}_{\mathrm{q}(i,j)} \in \mathbb{R}^D$ is the output for the (i, j)-th grid of the query image obtained from the final convolution layer of the CNN, $\mathbf{a}_{\mathrm{p}} \in \mathbb{R}^K$ is the embedded representation of the positive tag, $\mathbf{a}_{\mathrm{n}} \in \mathbb{R}^K$ is the embedded representation of the negative tag, $M_n^q(i, j)$ represents the degree of relevance between the (i, j)-th grid on query image and the negative tag, and $g_{q(i,j)}$ is the weight of the (i, j)-th grid in the query image.

Based on this calculation, a negative tag is subtracted only in an area highly related to the specific grid on the query image, and a positive tag is added to that area instead of a negative tag. The image of this operation is shown in Figure 4.7.

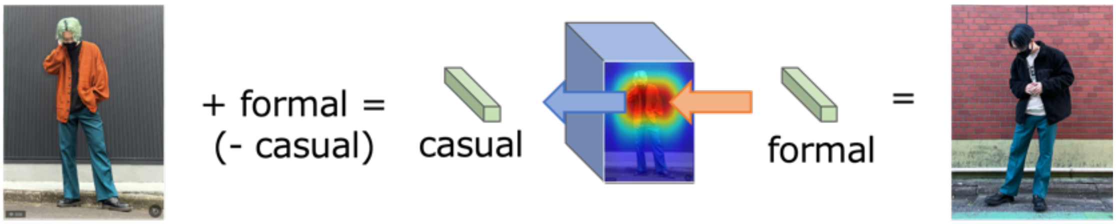


Figure 4.7: Image of calculation between tag and image for retrieval

With this operation focusing on the individual grid, an image can be acquired in which changes have been made only to the parts that should be changed on the image. This makes it possible to perform calculations that do not excessively change the entire image in the full-body outfit image of a person coordinating multiple clothes.

4.4 Experimental Evaluation

To confirm its effectiveness, the proposed system was applied to actual posted full-body outfit image data and the tags information attached to each image accumulated in WEAR [5][†]. An evaluation experiment was conducted to confirm the effectiveness of the proposed system in terms of 1) loss transition, 2) similarity between the embedded representation of the image and the attached tag, 3) relevance score between foreground/background and each tag (AAM), and 4) evaluation questionnaire for image retrieval results for both experts and non-experts.

4.4.1 Experimental Settings

Table 4.1 summarizes the two types of datasets used in the experiment.

Table 4.1: Summary of dataset features

	dataset-1	dataset-2
number of snaps	18,050	15,740
number of unique tag	1,753	1,104
gender of the subject	women	women
background	Contains considerable noise	Contains some noise
situation assumption	social media	social media / e-commerce site
sample		

With dataset-1, we verified the robustness of learning on data with a highly complex and extensive background. With dataset-2, we assessed the effectiveness of the proposed system on data with a relatively simple background but an extensive range. Dataset-1 comprises images posted by general users, while dataset-2 consists of images shared by professional users. In both datasets, each image features a single person, accompanied by an extensive background.

The embedded representation dimension included in the VSE model is set to 64, the learning rate is 0.001, the number of epochs is 50, the batch size is 32, threshold α for the grid weight map is 0.1, the parameters for regularization terms β is 0.1, and margin m is set to 0.2. Furthermore, VGG16 [138] (a CNN with a relatively large number of parameters) and GoogleNet Inception

[†] WEAR is the largest fashion outfit-sharing application in Japan. On WEAR, a user can search for the fashion items that interest him/her from over 13 million outfits (as of March 2023).

V3 (a CNN with a relatively small number of parameters) pre-trained on ImageNet [139] are respectively used as extractor for assessing the impact of CNN. As preprocessing, SSD (extractor: MobileNet V2 [140]) trained on Open Images OpenImages2 is used for object detection, and FCN (extractor: ResNet) trained on MS-COCO [141] is used for semantic segmentation. The number of LDA topics used for negative sampling was set to eight for dataset-1 and seven for dataset-2 through prior experiments.

4.4.2 Loss Transition

By checking the loss transition for each epoch, we validated the effectiveness of 1) foreground-centered learning by the grid weight map and 2) negative sampling according to the duplicated relation and the prior distribution that reflects the co-occurrence of the attached tags. These two methods are commonly proposed for efficient and sensitive learning. To confirm the individual improvement effects in detail, a comparison of the previous method [86] and proposed methods that changed only the grid weight map (foreground-centered learning) is shown in Figure 4.8-4.11. GAP represents the previous method, and GWAP(threshold) represents the completely proposed method. Furthermore, as a comparison method, GWAP(original) excludes the threshold part of the grid weight map and uses each grid's probability of the foreground softly. GWAP(center) learns only the 3/4 range of the center all images are applied.

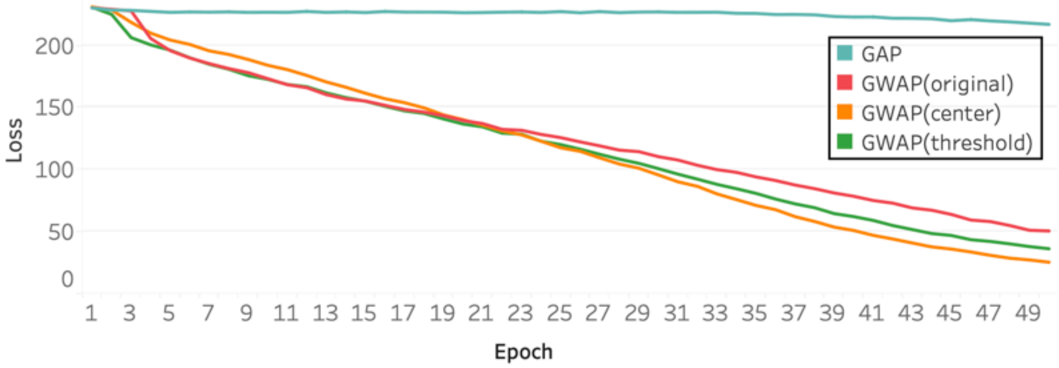


Figure 4.8: Transition of loss for each epoch (about grid weight map / dataset-1, VGG16)

From Figure 4.8 and Figure 4.9, for datasets with complex and wide range backgrounds (dataset-1), learning hardly progresses with the previous method (VGG16). In contrast, with the proposed method and the comparison methods (including foreground-centered learning), the learning progresses relatively smoothly for each epoch. Moreover, in the case of Inception V3,

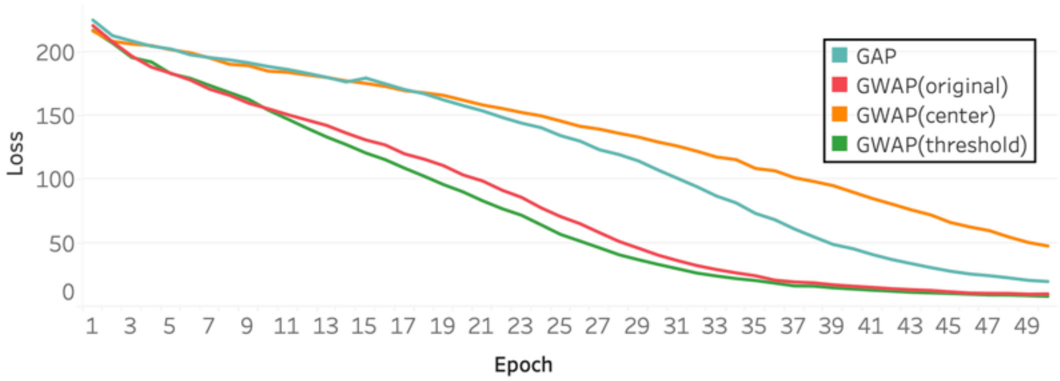


Figure 4.9: Transition of loss for each epoch (about grid weight map / dataset-1, Inception V3)

Additionally, in the case of Inception V3, the proposed method and the GWAP (original) exhibit more efficient learning progress compared to the previous method, especially when foreground-centered learning is employed and the position of the foreground is accurately provided. These results indicate that the proposed approach, which employs foreground-centered learning with a grid weight map, is a robust method for addressing background noise. However, it is worth noting that GWAP (center) did not progress smoothly in dataset-1, which can be attributed to the presence of various human poses in the dataset. In many images, the foreground does not necessarily fall within the 3/4 range of the center.

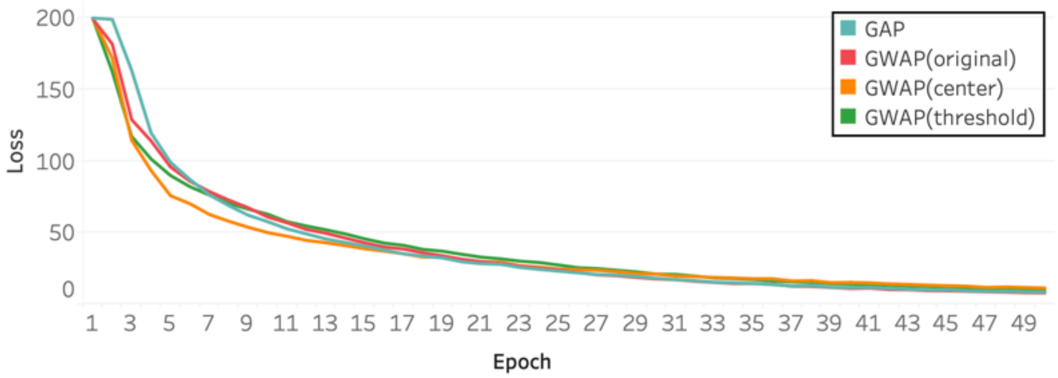


Figure 4.10: Transition of loss for each epoch (about grid weight map / dataset-2, VGG16)

As demonstrated in Figure 4.10 and Figure 4.11, we could not confirm the remarkable improvement effect of the loss transition based on the grid weight map for the dataset with a relatively simple background. Therefore, as presented in the next section, we will validate the proposal from another perspective by confirming the acquired embedded representation in more detail with respect to dataset-2. All the following experiments will also show the experimental

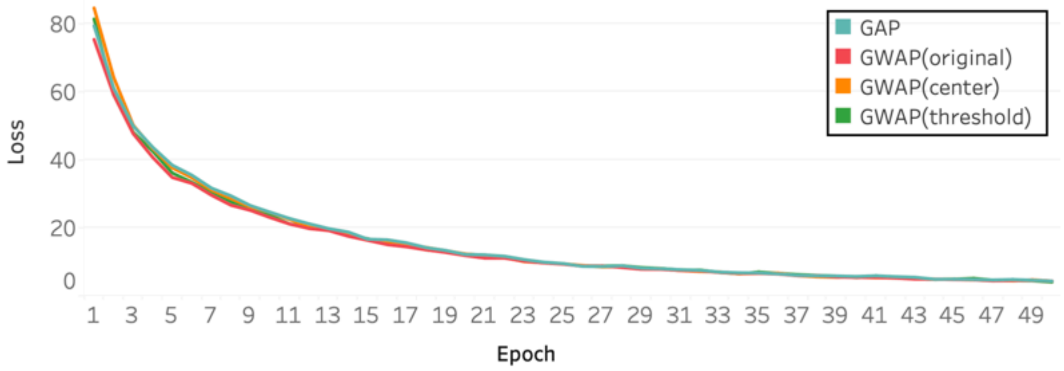


Figure 4.11: Transition of loss for each epoch (about grid weight map / dataset-2, Inception V3)

results of dataset-2 and Inception V3, which are relatively proper to learn, and the effectiveness could not be confirmed only by the loss transition.

Thereafter, to validate the effectiveness of the proposed negative sampling, the result of the method in which only the negative sampling part is changed (no grid weight map) from the previous method is shown in Figure 4.12. Herein, random represents a method of completely randomly selecting negative sampling (previous method), p1 represents a method for performing negative sampling using Eq. (4.9), and p2 represents a method using Eq. (4.10).

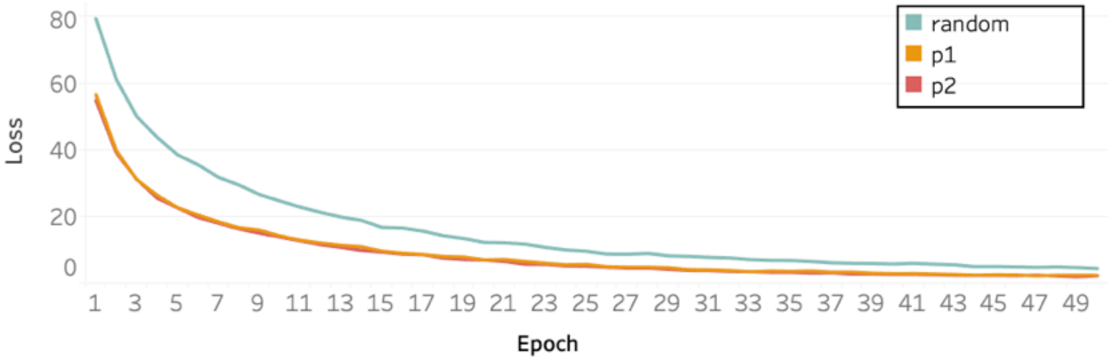


Figure 4.12: Transition of loss for each epoch (about negative sampling / dataset-2, Inception V3)

Figure 4.12 provides clear evidence that efficient learning is achieved through the additional changes made to the negative sampling technique. This outcome strongly indicates that the proposed approach is capable of selecting appropriate negative samples in comparison to the previous method. Consequently, effective and precise learning becomes crucial for embedding the full-body outfit image and tags, including abstract tags as the focus of this study. In this regard, the rules for excluding duplicate tags and the topic distribution of attached tags by LDA make valuable contributions.

4.4.3 Detailed Verification of the Effectiveness of the Grid Weight Map

As illustrated in Figure 4.10 and Figure 4.11, for a dataset with a relatively simple background, the effect of foreground-centered learning by the grid weight map could not be confirmed only by the viewpoint for the transition of loss. Therefore, we confirm the validity of the proposal through two types of detailed analyses of the obtained embedded representation while considering another research perspective.

4.4.3.1 Similarity between Images and Attached Tags

The effectiveness of the proposed method (the effect of foreground-centered learning) using the similarity of embedded representation for the image and the attached tags obtained from the proposed method is verified. If it is accurately embedded in the space of the mapped destination, the image would be embedded around the attached tag. Based on this assumption, in this experiment, the images attached to each tag are defined as the ground truth, and the negative images are obtained randomly from other images. The number of negative images is 10 times the number of ground truth. Under this circumstance, top-$K = \{5, 10, 15\}$ images that are particularly similar to the embedded representation of each tag are acquired and evaluated on how many ground truths are included in them. The evaluation indicators are Precision and NDCG. The experiment is repeated 30 times, and the average of the results is shown in Table 4.2. Moreover, the t-test was performed between the results of the GAP and other methods, and the significance level was set to 5%.

Table 4.2: Summary of evaluation values for top-K images selected using similarity for each tag

	Precision			NDCG		
top-K	5	10	15	5	10	15
GAP	0.581	0.571	0.555	0.532	0.551	0.549
GWAP (original)	0.602**	0.607**	0.603**	0.554**	0.581**	0.587**
GWAP (center)	**0.700****	**0.663****	**0.653****	**0.637****	**0.645****	**0.646****
GWAP (threshold)	0.659**	0.631**	0.621**	0.604**	0.614**	0.615**

Table 4.2 reports that GWAP exceeds GAP in terms of all indicators, and learning by focusing on the foreground is an effective method for accurately embedding the foreground and tags in the same space. In particular, GWAP (center) has the highest accuracy. In dataset-2, a foreground is included in the central 3/4 area of most images (the background is not included at all). This outcome suggests that learning progresses smoothly even when completely disregarding the arms, legs, and clothing edges. Consequently, this result further implies that achieving accurate learning involves entirely ignoring the background rather than selectively considering clothing edges while incorporating some background. However, since it is undesirable to disregard the clothing edge area for fashion interpretation, GWAP (threshold) emerges as the most effective approach. Subsequently, GWAP (threshold) is comprehensively employed in the subsequent experiments.

4.4.3.2 Attribute Activation Map

Based on the results, the effectiveness of foreground-centered learning using the grid weight map was observed. In this section, the effectiveness of background regularization, which is the third main change proposed in this study, is verified. Unlike the previous two approaches (foreground-centered learning/negative sampling), background regularization is a method proposed to reduce the relevance of the background and the tag. By reducing the degree of association between the background and the tag, it is anticipated that the likelihood of the background being colored in the colored AAM will decrease, leading to enhanced user satisfaction.

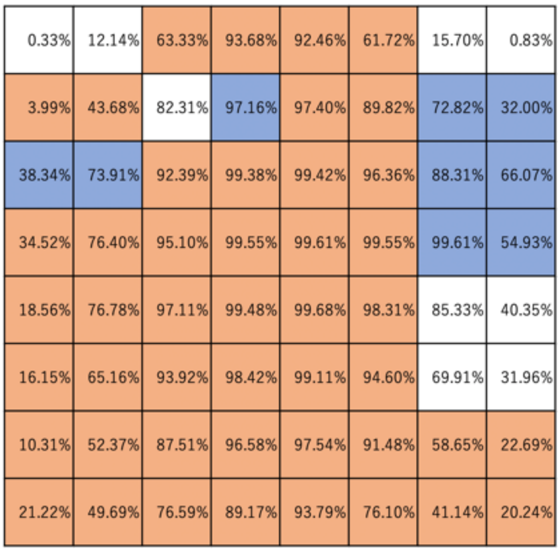
Specifically, in this experiment, the degree of relevance between all the images in the data and all the individual tags attached to them is calculated. Subsequently, the degrees of relevance for each foreground and the background area are examined individually. As the background of dataset-2 is relatively simple, there must always be no relevance between the tag and the background. Essentially, if the embedded representations of the image and the tag are learned properly, the acquired embedded representations should exhibit low relevance between the background and the tag, while demonstrating high relevance between the foreground and the tag. Table 4.3 shows the result of the average of the relevance between each grid of foreground/background and the attached tags. Herein, GAP represents a method that does not use the grid weight map at all, GWAP represents a method in which only foreground-centered learning is performed, and GWAP+reg represents a method in which background regularization is combined with foreground-centered learning. Moreover, the t-test was performed for the results of foreground/background, respectively, between the results of GAP and other methods, and the significance level was set to 5%.

Table 4.3: Overall average of the relationship between tags and foreground, background

	foreground	background
GAP	0.060	-0.127
GWAP(threshold)	0.063**	-0.131**
GWAP(threshold)+reg	**0.089****	**-0.186****

As reported in Table 4.3, unless background regularization is added, even if learning focusing on the foreground is performed, the degree of relevance will be as high as that of the previous method. In contrast, with background regularization, a higher relevance is observed between the grids of the foreground and the attached tags, and a lower relevance is observed between the background and the attached tags. This result suggests that the contribution of the regularization of the background with the grid weight map makes it possible to prioritize the foreground over the background and embed the image and the attached tag closer to each other.

Figure 4.13 (Figure 4.14) presents the outcomes obtained by calculating the degree of relevance between the tag and each grid encompassed within the foreground (background), and subsequently comparing GWAP+reg with GWAP. Each cell within the figure represents a grid, with the numbers denoting the proportion of the grid belonging to the foreground (background). Cells are highlighted in orange if GWAP+reg yields significant results at the 5% level, in blue if GWAP demonstrates significant results at the 5% level, and in white if there is no statistically significant difference at the 5% level. It is expected that each grid within the foreground exhibits a strong correlation with the attached tags, while each grid within the background displays a weak correlation with the attached tags.

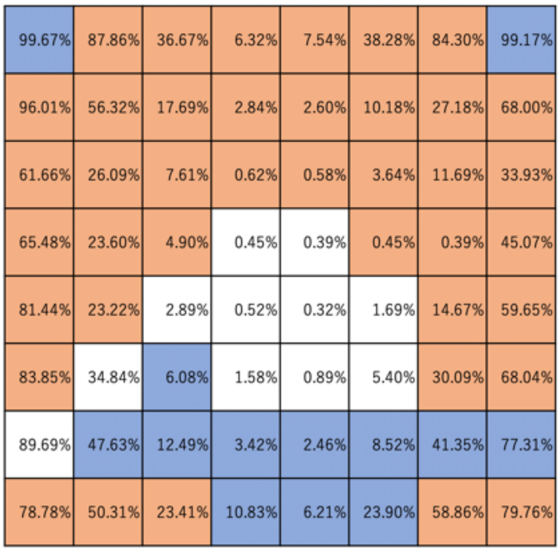


0.33%	12.14%	63.33%	93.68%	92.46%	61.72%	15.70%	0.83%
3.99%	43.68%	82.31%	97.16%	97.40%	89.82%	72.82%	32.00%
38.34%	73.91%	92.39%	99.38%	99.42%	96.36%	88.31%	66.07%
34.52%	76.40%	95.10%	99.55%	99.61%	99.55%	99.61%	54.93%
18.56%	76.78%	97.11%	99.48%	99.68%	98.31%	85.33%	40.35%
16.15%	65.16%	93.92%	98.42%	99.11%	94.60%	69.91%	31.96%
10.31%	52.37%	87.51%	96.58%	97.54%	91.48%	58.65%	22.69%
21.22%	49.69%	76.59%	89.17%	93.79%	76.10%	41.14%	20.24%

Figure 4.13: T-test results with and without background regularization for individual grid average of the relevance between tags and foreground

Based on Figure 4.13 and Figure 4.14, it can be observed that background regularization yields improved results for most grids in both the foreground and background. Although there were a few grids that produced inferior outcomes, no grids led to worse results simultaneously for both the foreground and background across all grids. Consequently, the incorporation of background regularization ensures that the embedded representations of tags are learned to be distinct from the background, thereby enabling closer alignment with the foreground. These results provide clear confirmation that foreground-centered learning and background regularization utilizing the grid weight map contribute to emphasizing the foreground and de-emphasizing the background during the learning process.

Considering that the colored background in the AAM is presented to users in real-world applications, it consistently triggers user skepticism. In the case of GWAP+reg, the relevance between the background grids and the tag decreases significantly. This finding suggests that the proposed method, which reduces the likelihood of coloring the background, is more effec-



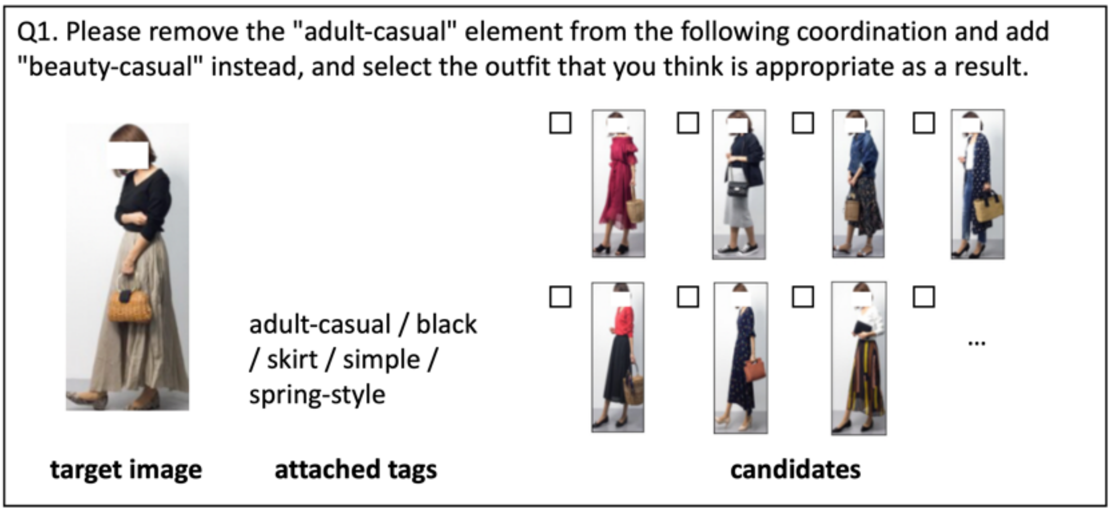
99.67%	87.86%	36.67%	6.32%	7.54%	38.28%	84.30%	99.17%
96.01%	56.32%	17.69%	2.84%	2.60%	10.18%	27.18%	68.00%
61.66%	26.09%	7.61%	0.62%	0.58%	3.64%	11.69%	33.93%
65.48%	23.60%	4.90%	0.45%	0.39%	0.45%	0.39%	45.07%
81.44%	23.22%	2.89%	0.52%	0.32%	1.69%	14.67%	59.65%
83.85%	34.84%	6.08%	1.58%	0.89%	5.40%	30.09%	68.04%
89.69%	47.63%	12.49%	3.42%	2.46%	8.52%	41.35%	77.31%
78.78%	50.31%	23.41%	10.83%	6.21%	23.90%	58.86%	79.76%

Figure 4.14: T-test results with and without background regularization for individual grid average of the relevance between tags and background

tive than the approach without background regularization when considering the application of real-world services from the perspective of AAM. Furthermore, with the availability of the semantic segmentation results obtained during the preprocessing stage of our proposed system, it becomes possible to set the degree of relevance to the background area as 0 before displaying the AAM results to users. In our experiments, we set the weight parameter β associated with the background regularization term to 0.1 by prioritizing the minimization of loss. However, by increasing this parameter, it is conceivable that the superiority of the proposed system can be extended to a wider range.

4.4.4 Image Retrieval Evaluation

Using the proposed system (including foreground-centered learning, background regularization, and negative sampling with Eq. (4.10)), certain tags are added (or subtracted) to evaluation images, and a user questionnaire verifies this. For evaluation purposes, 21 randomly sampled full-body outfit images were used for image retrieval, and a questionnaire was conducted. One abstract tag attached to the target image was subtracted, and another abstract tag was added instead. This tag was chosen arbitrarily. This questionnaire confirms whether the proposed system can accurately perform image retrieval operations (via image and tag calculation) on images containing a wide range of backgrounds and tags, including abstract tags, compared to the comparison method [86]. A sample of the questionnaire is shown in Figure 4.15.



Q1. Please remove the "adult-casual" element from the following coordination and add "beauty-casual" instead, and select the outfit that you think is appropriate as a result.

adult-casual / black / skirt / simple / spring-style

Figure 4.15: A sample evaluation questionnaire

Specifically, for each question, we selected the five most prominent search result images from both the comparison method and the proposed system. Out of the ten images presented, the images that the participants considered as correct search results were chosen (allowing for duplication). The average number of selected images was calculated and compared between the two methods. A perfect score would be 5.0, indicating that the participants selected all five images. We categorized the participants into four types of experts, as presented in Table 4.4. However, it is important to note that some participants possessed multiple areas of expertise, leading to their inclusion in multiple expert groups. Considering these conditions, we provide the summarized survey results from a total of 58 participants in Table 4.4.

Table 4.4: Result of questionnaire on outfit image retrieval

	number	comparison	proposal
experts: subjects have an experience working in fashion shops or companies	10	0.919	1.880
experts: subjects have a confident in fashion and fashion knowledge	13	0.915	1.695
experts: subjects have an experience posting fashion items on social media or blogs	17	1.103	1.885
experts: subjects sometimes called fashionable from around them	24	1.009	1.781
non-experts	21	1.199	1.927

Table 4.4 reports that the proposed system achieves more appropriate image retrieval results regardless of the group. Compared to other groups, subjects belonging to a group confident in fashion tend not to judge several images as correct answers; however, the proposed system still showed better results compared to the comparison method.

Furthermore, examples of image retrieval results are shown in Figure 4.16.

The aforementioned two examples illustrate the utilization of the image retrieval function with specific tags. For example, if the query image, which features a person wearing jeans and a khaki shirt, has the “khaki” and “casual” tags attached, removing the “khaki” tag and adding a “yellow” tag would retrieve images of individuals wearing jeans and a yellow shirt, while still maintaining the casual style. Conversely, removing the “white” tag from a query image with a “white” and “formal” tag (depicting a person wearing a white shirt and black trousers) and adding a “yellow” tag would retrieve images of individuals wearing black trousers and a yellow shirt, while preserving the formal ambiance. These examples demonstrate that operations involving specific tags can be reasonably executed to a certain extent.

In addition to the aforementioned examples, the results of image retrieval using abstract tags, which constitute one of the primary objectives of this research, are presented. For instance, removing the “casual” tag from the left query image and adding a “formal” tag would retrieve images featuring black trousers and a green shirt, thereby maintaining a formal atmosphere. Furthermore, if the “adult-casual” tag is added, the green shirt would be replaced with a longer green item, resulting in the retrieval of full-body outfit images with a more mature overall ap-

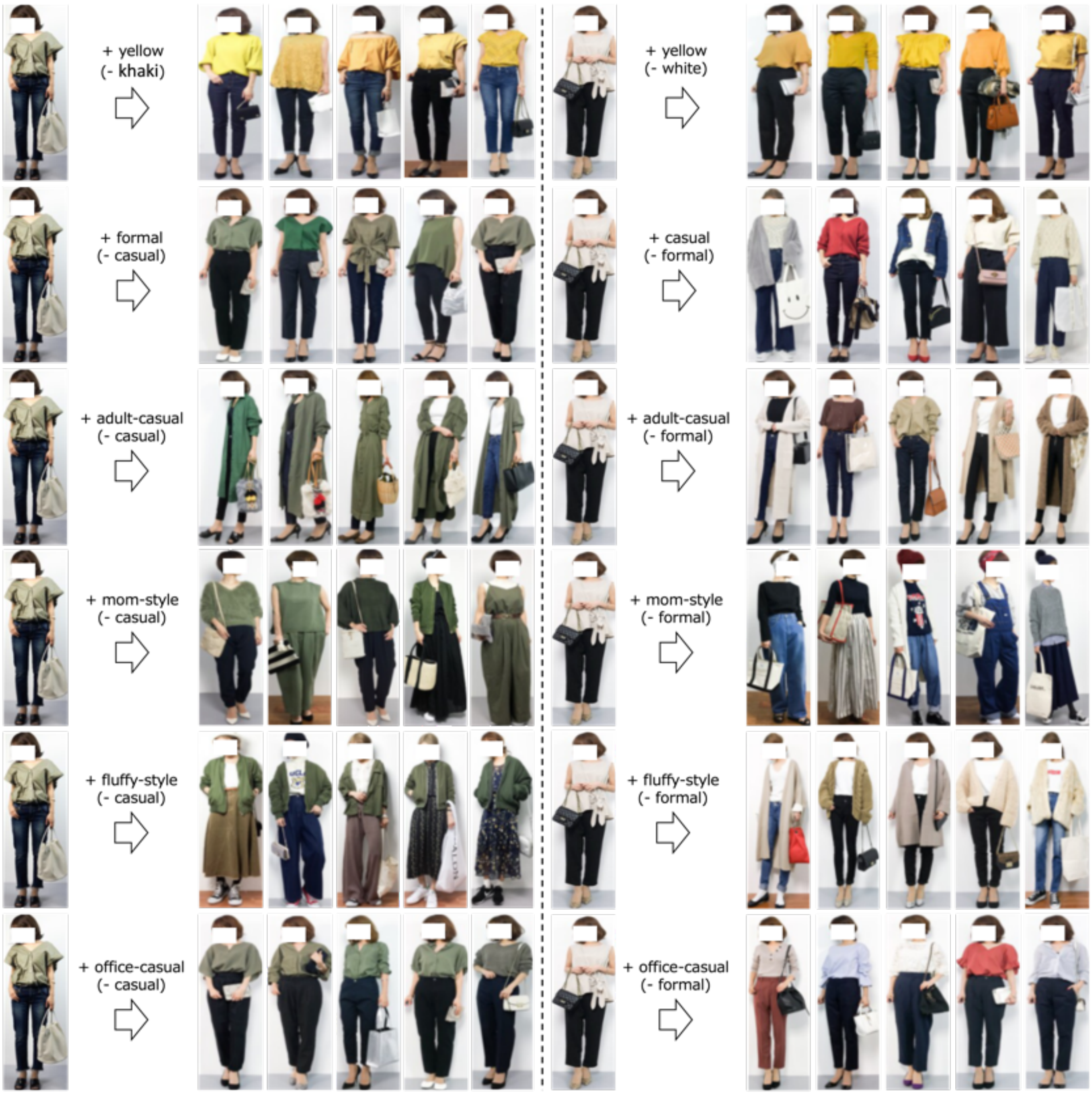


Figure 4.16: An example of image retrieval

pearance. The outcomes illustrated in Figure 4.16 indicate that the proposed system generates appropriate results for image retrieval.

By utilizing such an image retrieval function, even non-expert users can interpret fashion-specific abstract tags. This interpretability is expected to enhance users' interest in and motivation for fashion, potentially influencing their purchasing behavior. Furthermore, experts are anticipated to make new discoveries and expand their perceptions through the use of this system.

4.5 Additional Analysis

Because the proposed model can acquire embedded representations for each image and tag in the same projective space, it is possible to perform multifaceted analyses other than analysis related to image retrieval. Moreover, analyzing these results is also useful for the interpretation/understanding of fashion and planning marketing strategies. This section analyzes the embedded representations obtained by applying the proposed system to actual data in more detail. Moreover, we clarify the usefulness of applying the proposed model to real-world services.

4.5.1 Ranking-Based Image Retrieval Using the Relevance Score

In addition to the aforementioned image retrieval function using images and words, searching for images with a particularly high (low) relevance to the tag is also possible. An example is shown in Figure 4.17.

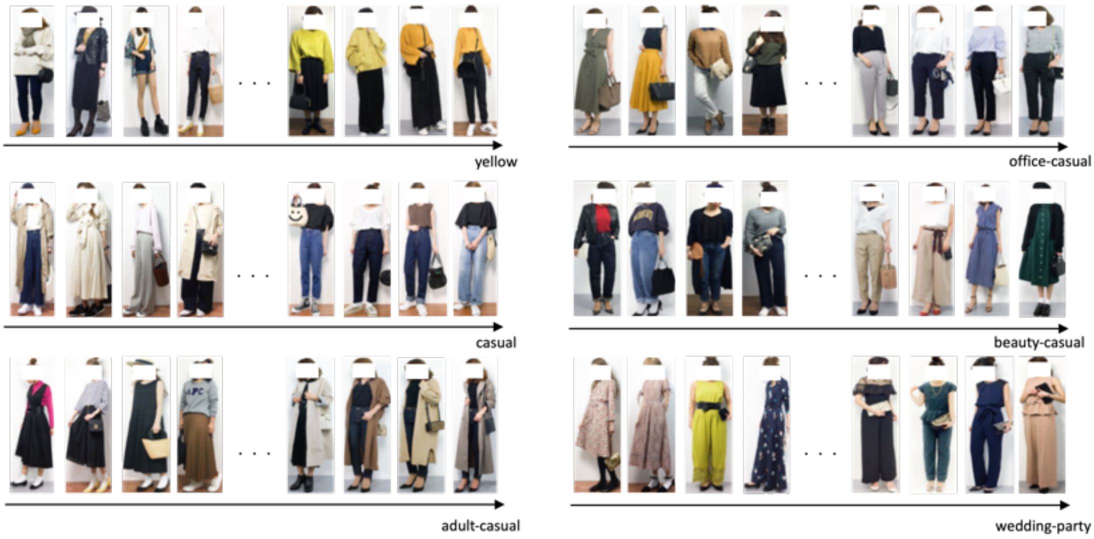


Figure 4.17: An example of ranking-based image retrieval using relevance score

For example, outfits with a higher relevance score with the “yellow” tag tend to have a higher proportion of the yellow part in the entire image. On the contrary, if the score is low, outfits including only a few parts of yellow appear. In the previous system, it was only possible to show all images with the tag “yellow” in a batch; however, using this function, it is possible to search for clothes based on the user’s purpose, such as “I want to find a yellow atmosphere as a whole,” or “I want to incorporate yellow only for one point.” Images can also be searched by the degree of relevance to abstract expressions, such as “casual,” “office-casual,” “beauty-casual,” and “adult-casual,” and scenes, such as “wedding-party.” For example, denim tends to be included especially for casual outfits, and light-colored long coats and long skirts tend to be included for less casual outfits. Outfits that are highly relevant to the wedding also tend to include one-tone dresses that are not extremely bright and outfits with low relevance tend to include patterned dresses and dresses that are excessively bright in color. By utilizing this result, the user can determine outfits that should be worn, especially for the wedding party, and identify the more suitable outfits for office.

4.5.2 Visual Interpretation of Abstract Tags by AAM

AAM allows for understanding which region on each image is relevant to a tag. This function supports the visual interpretation of the meaning of abstract fashion-specific tags. In this section, certain tags, including abstract tags, are visually interpreted using AAM. An example of the result is shown in Figure 4.18.

As shown in Figure 4.18, it is possible to visually interpret which area on the image the tag attached to each full-body outfit image has high relevance. For example, regarding the AAM between the specific tag “yellow” and each image, the importance of actual yellow regions on the image becomes more prominent; thus, appropriate results can be obtained. In addition, for “office-casual” clothing, shoes seem important for all images. Moreover, the rightmost image is particularly “office-casual” for the full-body image.

Thus, it has become possible to visually interpret abstract fashion terms that are difficult for users to interpret and understand their relevance to the full-body outfit image. This functionality is expected to improve the understandability of coordinated images, generate interest in fashion, and encourage users to engage in activities such as studying fashion and making clothing purchases in real-world services.

4.5.3 Checking the Average of AAM

In the previous section, AAM was computed for individual combinations of images and tags. By calculating the AAM for all images with individual tags and averaging the results, it becomes possible to identify which areas of an image are more likely to be associated with a particular tag. Figure 4.19 presents several examples of the average AAM.

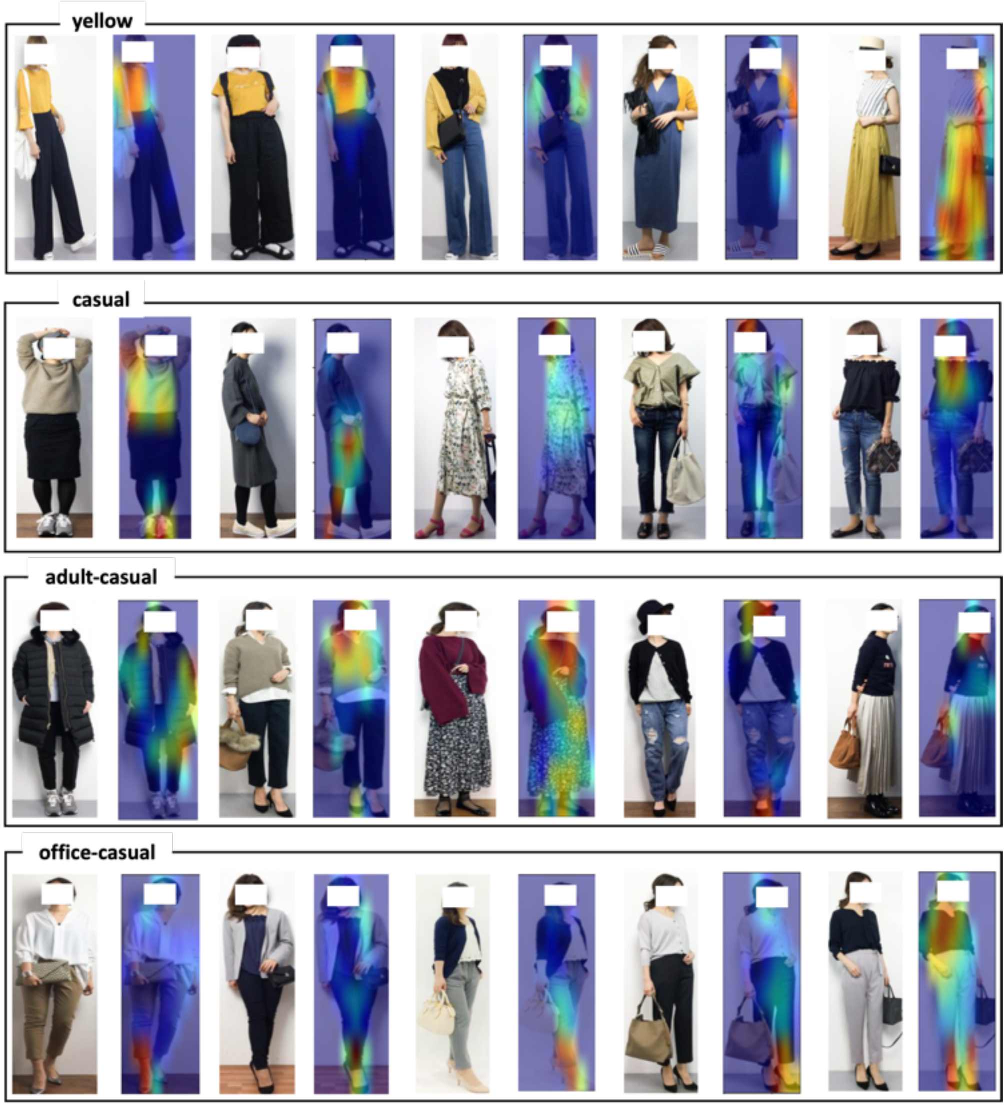


Figure 4.18: An example of visual interpretation of tags using AAM

From Figure 4.19, it can be observed that specific tags such as sneakers, denim, and dresses tend to exhibit higher relevance in the corresponding areas of the image. This result suggests that, at least for specific tags, the embedded representations can be accurately obtained by considering the relevance of the image features. Conversely, it is challenging to determine from this figure whether abstract tags have been learned effectively. However, for instance, it can be observed that the feet are significant for the tag "office-casual." Based on this finding, when examining images associated with the "office-casual" tag, a considerable number of them were found to depict high-heeled shoes. Thus, the average AAM can be utilized to gain an overview of the features across all images associated with a specific tag and facilitate the interpretation of

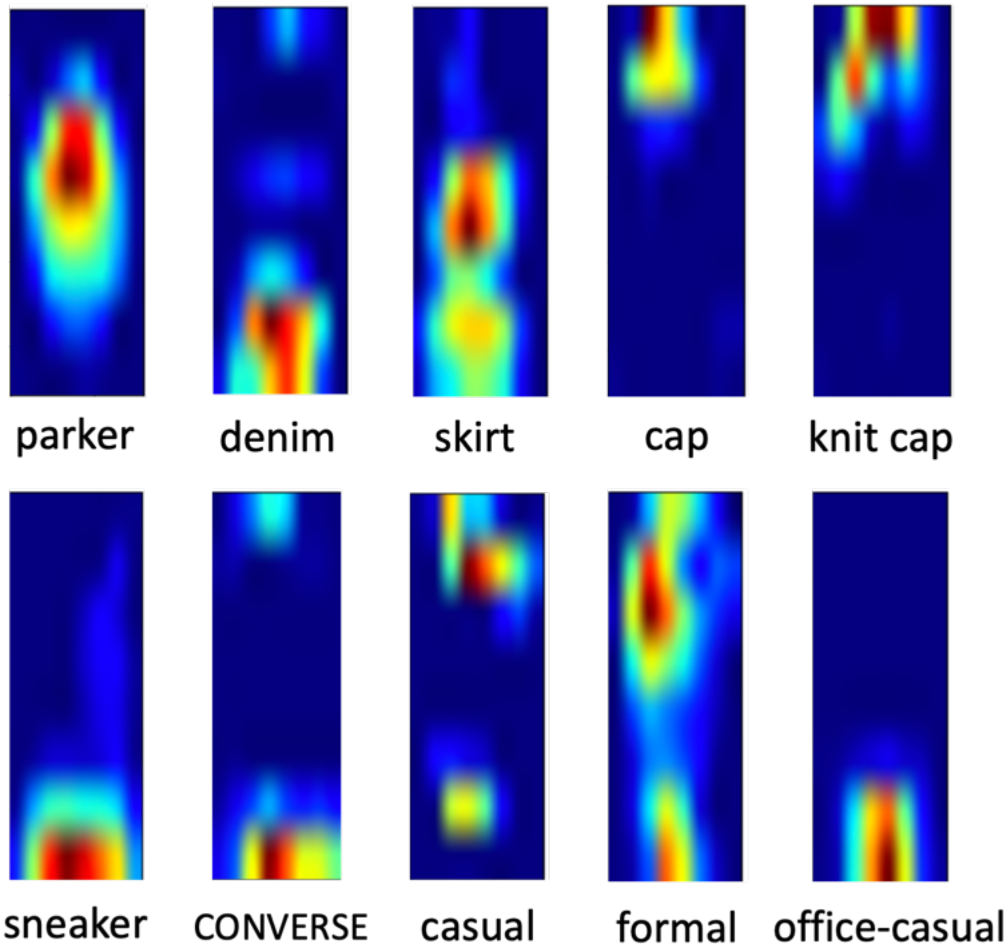


Figure 4.19: An example of average AAM for each tag

individual images. This enables users to anticipate the importance of paying attention to their feet when they have questions such as "what does office casual mean?"

Furthermore, the usefulness of the average AAM may be questionable for expressions like "casual" where the significance varies significantly depending on the combination of multiple clothing items. Although this result may not have direct marketing implications, it serves as a testament to the success of the learning process.

4.6 Discussion

4.6.1 Considerations Based on Experimental Results and Additional Analyses

Considering the aforementioned experimental results, we can provide the following summary.

- The proposed system can properly learn even for complicated datasets (the full-body outfit images, a wide range and noisy backgrounds, and various tags, including abstract tags, compared to the conventional method.
- Each major improvement point (foreground-centered learning, background regularization, and negative sampling) contributes to appropriate learning.
- The proposed system realized a more accurate image search by observing the image retrieval results using the calculation operation of obtained embedded representations.

Furthermore, in the additional analysis for the obtained results, we clarified that the following analysis could be achieved by applying the proposed system to real-world service data.

- By visualizing the degree of relevance between each image (area) and each tag, it is possible to interpret which area corresponds to the tag, even for abstract tags.
- By visualizing the average value of the degree of relevance for each tag and all the images that the tag attached, the proposed system can embed tags and images in the same space appropriately to some extent.

These analyses confirmed that the proposed system successfully learned fashion-specific knowledge contained in complex datasets. Moreover, the results suggest the potential to obtain valuable insights for users in real-world services by analyzing the obtained results from various angles. These multifaceted features are expected to provide online support to users, thereby achieving the purpose of this research.

4.6.2 Comparison with Conventional Studies

As mentioned in the related research section, the reason for recommending "the user seems to prefer this area in this item" can be obtained by the explainable recommendation methods using image information. However, it is unclear whether such information is what the users really desire. In contrast, the system proposed in this study makes it possible for users to obtain the information they truly desire to obtain by themselves because of fashion-specific abstract expressions. Specifically, it is possible to obtain information such as "the casual point of this outfit is this area" or "if you make this outfit a little more casual, the answer is this outfit." This capability of the system truly assists users in understanding fashion, motivating them to make purchases, and aiding them in making informed decisions. In addition, the main task of the conventional image retrieval task was to search for similar items from the single clothes images dataset or for similar outfits from the outfit images dataset. These are easy tasks even for non-experts to understand. Moreover, there were also tasks such as Figure 4.16 that search for the single item image by adding an image and specific words (such as “short-sleeves” and “black”); however, this is also an easy task even for a non-expert. In contrast, in this study, full-body outfit image retrieval is realized by adding a tag containing abstract expressions peculiar to the fashion domain. It also provides other functions, such as a visually understanding support function for fashion-specific expressions using AAM. Compared to conventional studies in the fashion clothes retrieval field, our study offers multiple functions that can genuinely assist users based on their specific needs.

Thus, this study proposes a novel task of “automatically interpreting the fashion domain,” which is completely different from the conventional explainable recommendation and image retrieval task. This unique aspect represents a major contribution of our research, making it valuable in this field.

4.6.3 Model Structure Considerations

As a major feature of the proposed model structure, the person area detection and semantic segmentation need to be performed as preprocessing steps, which are then used as inputs. While person area detection is generally applied as a preprocessing step in several studies, foreground and background classification can be performed using methods such as learning semantic segmentation tasks with the same model, detecting regions for each clothes item part in advance and learning them individually, or introducing a self-attention mechanism to distinguish foreground and the background within the model. However, to ensure practical application, this study focuses on the simplicity of the model. Furthermore, in the target data, the foreground is a person. If the grid in which the person is reflected can be fortunately detected, then the foreground/background separately learning becomes possible. As it is easy to use a public pre-trained model for semantic segmentation recently, it is possible to eliminate the trouble of preparing a separate dataset for semantic segmentation task. Furthermore, suppose accurate foreground and background information are obtained in advance. In that case, it is better to be able to focus on the main task of “fashion interpretation” rather than learning multiple tasks at the same time. The accuracy of the preprocessing was confirmed only visually this time. However, unlike the image where a person is small and, in the corner, only one person is shown in the center, so semantic segmentation in pre-progressing is not difficult. Therefore, this task's accuracy will not be a problem in this study.

Commonly, the exclusion of background during learning is one of the major issues in image processing. Foreground-centered learning using the grid weight map and background regularization proposed in this study can be applied to all models using CNN (including the GAP layer). Therefore, the proposal of this versatile method is one of the major contributions of this research.

Furthermore, the advantage of utilizing the separately obtained grid weight map extends beyond the ability to focus on the primary objective of interpreting fashion. It also allows for the adjustment of the degree of relevance between a tag and the background in the AAM, even if it becomes substantial. This advantage becomes particularly crucial when presenting the AAM to users in a real-world online service. If the importance of the background area is emphasized excessively, it can create a sense of distrust among users. In essence, employing a grid weight map for AAM provides a significant practical advantage in ensuring the appropriate balance between foreground and background elements.

4.6.4 Loss Function and Regularization Term Considerations

In the experiment conducted in this study, the hyperparameters α and β were set to 0.1 and 0.1. The transition of the loss term and the regularization term at that time is shown in Figure 4.20.

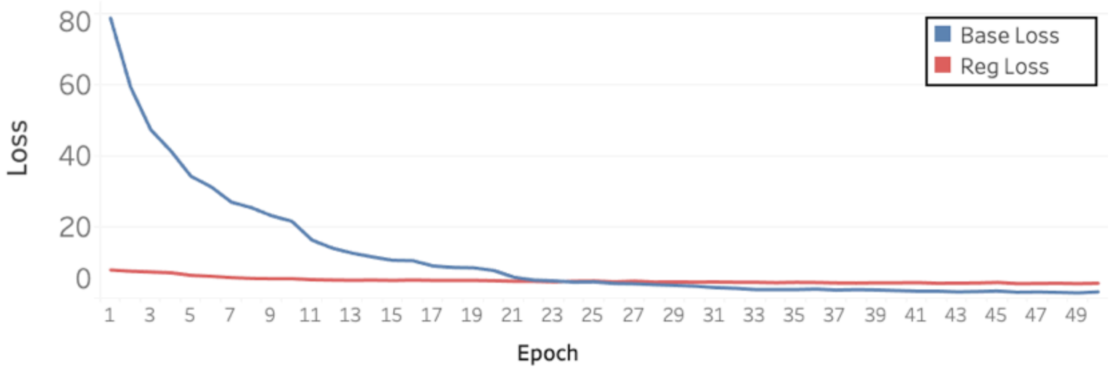


Figure 4.20: Variation of loss terms and regularization items with number of epochs

Figure 4.20 reveals that the regularization term has almost no adverse effect on the whole loss function because both steadily decrease in this condition and setting. If α is set as an extremely large value, the area wherein people are shown will also be treated as the background. Therefore, the threshold should be set by comparing the obtained grid weight map with the corresponding image. Similarly, setting β to an extremely high value would lead to interference between the subtask of maintaining distance between the background and the tag and the main task of bringing the foreground and the tag closer (see Figure 4.21). Considering a potential risk of adversely affecting the entire loss function due to the background regularization term, careful consideration is necessary while monitoring the transition of the overall loss and the background regularization term.

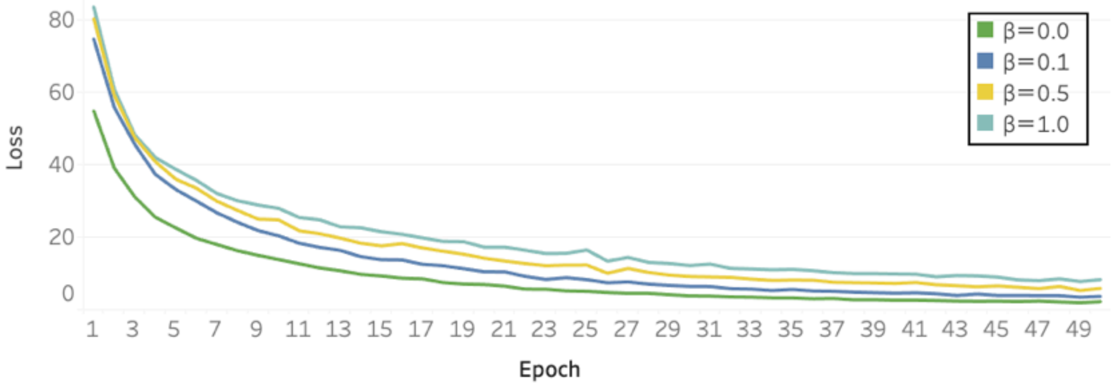


Figure 4.21: Variation of loss terms with number of epochs and hyperparameter β

4.6.5 Image Retrieval Improvement Considerations

To realize more sensitive fashion interpretation learning, it is conceivable to embed images for each clothing part (collar, sleeves, torso, and legs) individually or to perform search calculations for each part individually. However, in the target dataset, each tag is attached to a combination of full-body clothes, making it inappropriate to learn for each individual item. Although the problem can be avoided if all parts are comprehensively trained with one model, it results in a complex model. Particularly, challenges arise when it comes to effectively combining the features of each part. On the other hand, the multifaceted evaluation experiments and analyses conducted in this study indicate that the proposed system may have some degree of usefulness, suggesting that it can serve as a reasonable starting point in the field of "fashion intelligence." Future efforts to improve accuracy should consider this trade-off between enhancing sensitive learning and managing model complexity.

To combine self-attention mechanism for embedding attributes [8, 142] and hashing (tag complementation) algorithm [143, 144, 145, 146] simultaneously as learning a model for visual-semantic embedding can also be considered. However, because each factor is a research area that has been deeply tackled as an independent field, comprehensive learning of these technologies with one model leads to the complexity of the model. For example, for hashing, it is necessary to properly learn visual-semantic embedding while ensuring the accuracy of tag completion. In brief, it is necessary to consider the trade-off between accuracy and model complexity.

4.6.6 Considerations Regarding Actual Service Application

Regarding the image retrieval experiments on dataset-2 and interpretation with AAM, the results obtained were definitive. However, in another experiment (not published in this paper) on dataset-1, it was difficult to obtain an intuitive image retrieval result and AAM. Thus, even if the learning progresses well against ground truth tags, the excessive variety of attached tag noise, background, and pose makes learning extremely difficult. Accurate interpretation for datasets that contain such large amounts of noise is reserved for future research. On the contrary, it is now evident that accurate interpretations and useful results can be obtained by carefully selecting a suitable dataset in advance, as demonstrated by dataset-2. This implies that the proposed model can effectively support e-commerce site users dealing with images with minimal noise. Moreover, it can also be utilized in the context of social media by selectively choosing the data, as demonstrated by dataset-2. Therefore, the proposed system holds great potential for practical applications.

4.7 Conclusion of this Chapter

In this chapter, we introduced a novel research area called “fashion intelligence” and proposed a fashion intelligence system that enables the interpretation of abstract fashion attribute information. The proposed system is based on a visual-semantic embedding method, incorporating foreground-centered learning, background regularization, and negative sampling based on duplicated relations and prior distributions that reflect the co-occurrence of attached tags. The system utilizes various methods such as image retrieval using rich tags (including abstract tags), similarity calculation for full-body outfit images, and visual interpretation support through AAM. The effectiveness of the proposed system was demonstrated by applying it to real-world service datasets and conducting evaluation experiments. The experimental results analysis highlights the proposed system's potential in reducing fashion-related ambiguity and complexity, including the reliance on subjective judgments, thereby supporting users in understanding and interpreting fashion items. Furthermore, the proposed system has promising applications in real-world scenarios, supporting fashion studies and facilitating online user purchasing activities.

Chapter 5

Fashion Intelligence System Considering Parts by Partial Visual-Semantic Embedding

This chapter proposes a new model acquiring embedded representations corresponding to each part in a full-body clothing image. Through various experiments, we will explain that the proposed model realizes a more powerful fashion intelligence system.

5.1 Purpose of this Chapter

Unlike offline shopping, when browsing fashion items online, users must interpret fashion images to resolve difficult questions that arise in their minds without help from the shop attendants. The shopper typically asks the following questions: 1) “what would this outfit look like if it were more formal?”, 2) “how office-casual is this outfit?”, and 3) “what makes this outfit street?”, Even experts find it difficult to answer these questions. This ambiguity inherent in the fashion field may hinder the users from pursuing their interest in the fashion industry, making it difficult for them to explore new genres of clothing. Therefore, automatically obtaining answers to these questions is expected to arouse interest among users and broaden their perceptions, helping them interpret fashion clothing.

In this regard, as mentioned in Chapter 4, we proposed a “fashion intelligence system” to aid the interpretation of these terms in various applications by employing a VSE model [8]. This system helps users obtain answers to ambiguous questions by clarifying the relationships between the full-body outfit images and various verbal expressions, including ambiguous expressions specific to the fashion domain. Answers to ambiguous questions from users (such as Questions 1—3 above) can be obtained by embedding a full-body outfit image and tag information attached to the image in the same projective space and using the embedded representation of the image and tags in this projective space. By enabling users to obtain answers, the ambiguity inherent in fashion can be reduced to support users in their fashion-related decisions and actions, e.g., what to wear and which items to purchase.

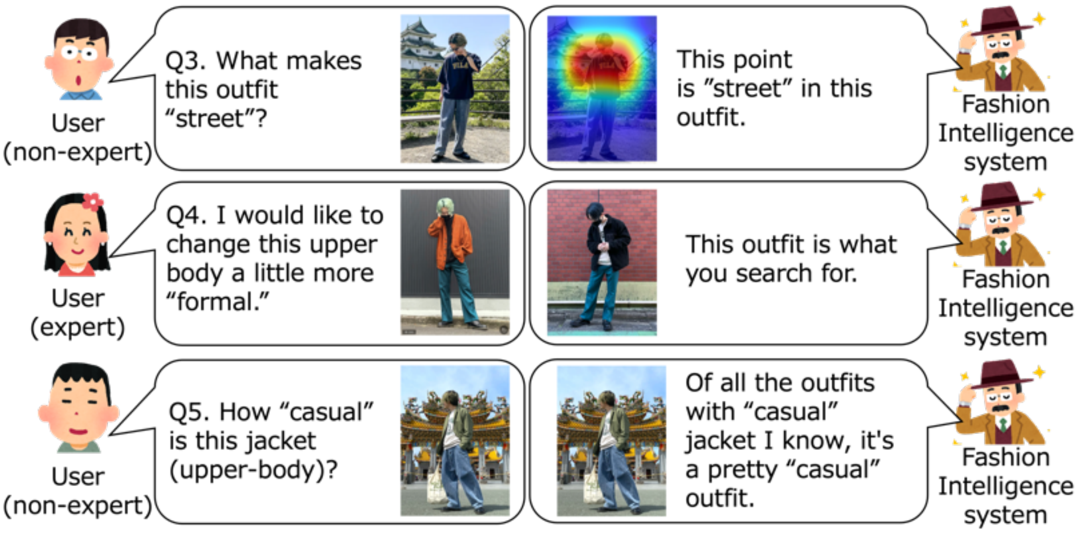

User
(non-expert)
Q3. What makes this outfit "street"?
This point is "street" in this outfit.
Fashion Intelligence system
User
(expert)
Q4. I would like to change this upper body a little more "formal."
This outfit is what you search for.
Fashion Intelligence system
User
(non-expert)
Q5. How "casual" is this jacket (upper-body)?
Of all the outfits with "casual" jacket I know, it's a pretty "casual" outfit.
Fashion Intelligence system

Figure 5.1: Image of fashion intelligence system

Full-body outfit images consist of many elements (parts), such as hair, tops, trousers, skirts, and shoes. Furthermore, the pose of the subject is diverse, and fixed parts do not appear in fixed patches. However, the VSE in [8] has a simple model structure that maps the full-body outfit image to the projective space as a batch. One of its greatest limitations is obtaining the embedded representation corresponding to each part. The problem caused by this limitation is that the AI model cannot answer the questions that users truly want answered: 4) "what would the outfit look like if I changed the top to make it a little more formal?"; and 5) "how casual is this jacket?" Specifically, image-retrieval results should not show images in which changes are made to the entire body [147] because users consider how to dress based on the outfits they already have.

In this study, we propose a partial VSE (PVSE) model that enables the acquisition of embedded representations corresponding to each part of a full-body outfit image while maintaining a simple model structure and a low computational complexity. The proposed model retains previous practical application functions and enables image-retrieval tasks in which changes are made only to specified parts (answering Question 4) and image reordering that attentively focuses on the specified parts (answering Question 5). This is not possible with conventional models. We demonstrate that the proposed model has superior functionality to conventional models through multiple quantitative evaluation experiments and qualitative evaluation analyses using real-world service data.

The main contributions of this study are as follows: 1) We devised a PVSE model that can map a full-body outfit image and rich tags onto the same projective space and obtain an embedded representation corresponding to each part in the full-body outfit. Consequently, the proposed model handles the unique characteristics of full-body outfit images and realizes a novel partial fashion intelligence system. 2) As indicated by multifaceted unique evaluation experiments using real-world data, the simple structure of the proposed model contributes to mapping operations more accurately while maintaining the computational complexity of the conventional VSE model. 3) The proposed model maintains three practical application functions. These support the user's fashion interpretation (as does the conventional model). It also identifies image-retrieval and image-reordering tasks that attentively focus on specific parts (which cannot be identified by the conventional model) through multifaceted analysis experiments and examinations using real-world data. Consequently, the proposed model reduces the ambiguity and complexity inherent in fashion through various visualization methods and supports the marketing activities and fashion decisions of users.

5.2 Preliminaries

5.2.1 Problem Setting

The images targeted in this study are full-body fashion outfit photos of a single person dressed in various items (parts), taken against various backgrounds. Full-body fashion outfit images have the following unique characteristics:

- A full-body outfit can be thought of as a set consisting of multiple individual items. This set always includes the items necessary for a full-body outfit.
- Each item satisfies the item category-level compatibility condition. Specifically, the shirt included in image A can be replaced with other shirts but cannot be replaced with a hat or sneakers.

- The poses of the people in the images are diverse, i.e., fixed parts do not appear in fixed patches.

When creating a full-body fashion outfit, the relationships between each item should be considered, and it is necessary to be able to handle difficult questions such as, "how should I change the upper-body (using the replaceable alternative items)?" In addition, this study focuses on the use of full-body outfit images instead of individual item images linked by a special reference database that records compatible combinations of each item. However, some special handling is essential because the region that shows each item is different in each image. Furthermore, because data generally accumulated in the contemporary world are full-body outfit images, a specific model that handles these characteristics is indispensable. Therefore, current adaptations of existing models from other fields cannot handle the specific characteristics of these full-body outfit images.

Furthermore, multiple tags (natural language expressions) are attached by the user posting the image as attribute information. The given tag information contains a mixture of specific and ambiguous tags with the characteristics listed in Table 5.1.

Table 5.1: Characteristics of specific tags and ambiguous tags

Type	Characteristic	Examples
Specific	Once attached, it is always the correct tag regardless of the sensibility of the contributor	hat, denim, skirt, t-shirt, sneaker, etc.
Ambiguous	Its uncertainties depend on the sensibility of the contributor	formal, casual, office-casual, kawaii, spring-style, dating-style, etc.

For instance, as per the opinion of contributor A, if image A is completely "kawaii," the "kawaii" tag can be added by contributor A. Conversely, if contributor B feels that image A is only partially kawaii, the "kawaii" tag may not be added by this contributor. In addition, if the expression "cute" seems more appropriate than "kawaii" to contributor C, they would add the tag "cute" instead of "kawaii." Thus, a target full-body outfit image includes not only specific tags but also abstract tags. The ambiguous natural language expressions are one of the major reasons why users find the fashion domain difficult.

In recent years, a massive number of posts about fashion, which consist of full-body outfit images and multiple tags attached to the photos and include the realistic trends of the time, have been posted by a high number of users on social media and e-commerce sites. Thus, a model that can extract practical fashion knowledge from these data and interpret ambiguous expressions peculiar to fashion would be a powerful tool for users. However, it is impossible to implement a partial fashion intelligence system simply by applying image processing models proposed in other fields because of the abovementioned characteristics peculiar to full-body outfit images. Therefore, one of the essential contributions of this study is to handle these characteristics and propose a model that can provide powerful application functions.

5.2.2 Part-by-Part Learning of Fashion Style Images

There have been studies wherein fashion images of each item included in a full-body outfit were independently learned, and items were recommended based on the compatibility between different types of items [148, 149, 150, 151]. Furthermore, studies have derived which combination of candidate (in a wardrobe) items match [152, 153]. Moreover, studies that search for other items that match a query item [154, 155, 156] have been reported.

Although these tasks are concerned with the fashionable combinations of items, our study focused on the following: “what happens if this casual outfit becomes more casual?” and “how casual is this outfit?” Thus, the focus of this study is fundamentally different. Furthermore, that each item is independently applied to a backbone model, such as a convolutional neural network and vision transformer (ViT) [157], is a significantly different aspect of our study, which focuses on directly learning full-body outfit images. These studies are also remarkably different because the computational cost of the backbone model increases with the number of items, and experts are needed to perform large-scale labeling to determine the items that match with other items. This is a significant hurdle when considering an application to serve in the fashion industry, where trends change rapidly.

In social media data, it is almost impossible for a full-body outfit image (one post), where real trends in fashion have accumulated most recently, to be linked to the data on individual items. The problem caused by this limitation is that the model cannot answer the questions that users truly want answered. Acquiring the features of each part from a single full-body image, as envisioned in this study, is particularly desirable when e.g., analyzing social media data (which are useful for e-commerce site data). Studies in person re-identification have used segmentation-based methods to extract the features of each part of single full-body images [158, 159]. However, such studies are different from this study because they include operations to mix the features of each part because ordering the parts to correspond to each dimension of the final required features is unnecessary. Furthermore, another method extracts a bounding box for each body part from a single image by detection and performs learning for each part [160, 161]. However, this detection-based (patch-based) method is less sensitive than the segmentation-based method, and the computational complexity necessary for applying each part to a large network remains challenging.

Conversely, several fashion image generation approaches have been studied based on shape features obtained by semantic segmentation and capturing features for each part [147, 162, 163, 164]. Particularly, [147] performed excellent work based on the claim that "making minor changes to fashion is important to become fashionable." However, these approaches are based on image generation (i.e., items that do not strictly exist are generated), and the research does not focus on attributes (expressions) but on the question, "to become fashionable what minor changes should be made to the target full-body outfit?" In contrast, this study is different in that the approach is to search for fashion images consisting only of existing items. This approach is more useful for real-world applications, where the end-user is the target because it avoids the problem of presenting items that do not exist. In addition, the flexibility to answer various questions, such as "how do I make it casual?," "how do I make it office-casual?," and "how do I make it adult-casual?"

5.3 Methodology

5.3.1 Model Architecture

The structure of the entire model is shown in Figure 5.2.

The key feature of the proposed model is the inclusion of an architecture that considers a full-body outfit image as a collection of fashion items (parts) and acquires the embedded representation corresponding to each part. A simple but effective architecture is obtained by extending foreground-centered learning [8] and combining the grid weight map corresponding to each part obtained from the semantic segmentation model and the embedded image feature with global weighted average pooling (GWAP) [129]. With this method, regardless of the number of parts into which the full-body outfit image is divided, the number of times that the backbone model is applied to each image per epoch can be limited to only one. This avoids increasing the computational complexity.

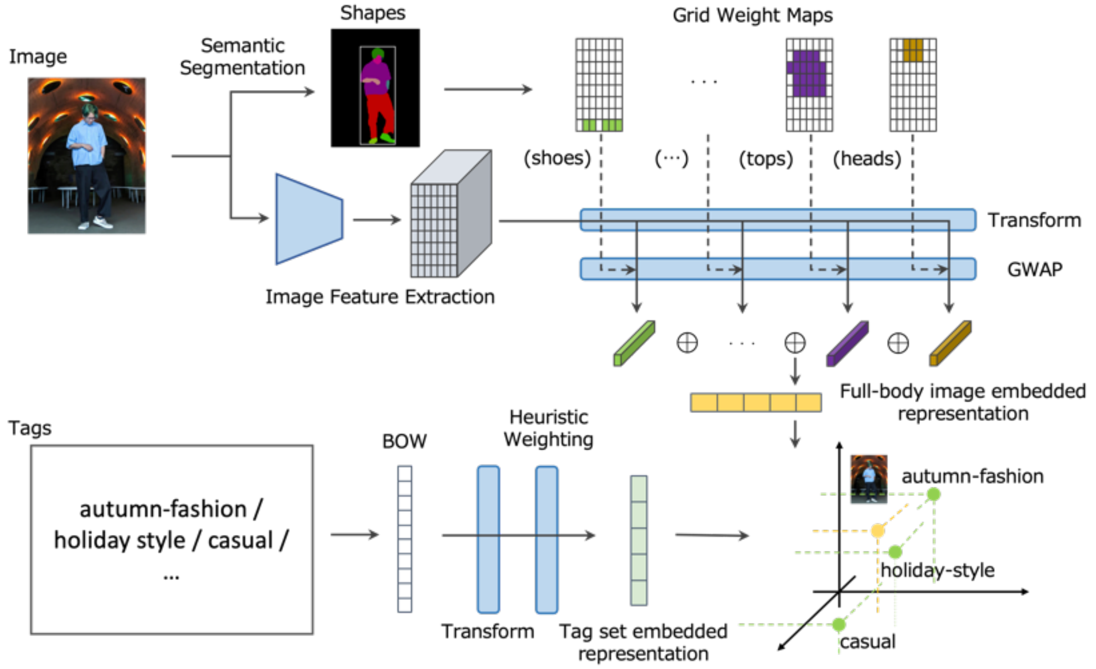


Figure 5.2: Structure of a prototype of our partial visual-semantic embedding model proposal

5.3.2 Part-by-Part Grid Weight Map Acquisition

The semantic segmentation in this study entails calculating the probability of the fashion item appearing in each pixel. The part with the highest probability for each pixel is considered part of that pixel. Here, the grid refers to the area where the target image is divided vertically into I and horizontally into J. The grid weight map for the l-th part is defined as $G_l = \{g_{(1,1),l}, ..., g_{(i,j),l}, ..., g_{(I,J),l}\}$ when the number of all parts is defined as L. $N_{(i,j),l}$ is defined as the count of the l-th part of the pixels contained in the (i, j)-th grid and $g_{(i,j),l} = N_{(i,j),l} / \sum_{i=1}^{I} \sum_{j=1}^{J} N_{(i,j),l}$. Differently expressed, $g_{(i,j),l}$ is the percentage of pixels in the (i, j)-th grid out of all the pixels in the l-th part. The grid weight map G_l achieves precise learning for each part.

5.3.3 Parameter Optimization

The dataset used in this study consisted of a single full-body outfit image to which multiple tags were assigned. The image was embedded using the image features obtained from the backbone model and a grid weight map. Eq. (5.1) was used to obtain a concatenated embedded representation of the features for each part of each image.

$$\mathbf{x} = [\hat{\mathbf{x}}_1; \cdots; \hat{\mathbf{x}}_l; \cdots; \hat{\mathbf{x}}_L], \tag{5.1}$$

$$\hat{\mathbf{x}}_l = \sum_{i=1}^{I}\sum_{j=1}^{J} g_{(i,j),l}\mathbf{W}_{\mathrm{I},l}\mathbf{f}_{(i,j)}, \tag{5.2}$$

where $[\mathbf{a}; \mathbf{b}]$ is the concatenate operation between vectors $\mathbf{a}$ and $\mathbf{b}$, $\mathbf{x} \in \mathbb{R}^{KL}$ is the embedded representation (vertical vector) of the full-body outfit image, and $\hat{\mathbf{x}}_l \in \mathbb{R}^K$ is the embedded representation (vertical vector) of the l-th fashion item part in the image. K is the number of dimensions of the embedded representation for each part. $\mathrm{W}_\mathrm{I} = \{\mathbf{W}_{\mathrm{I},1}, \cdots, \mathbf{W}_{\mathrm{I},l}, \cdots, \mathbf{W}_{\mathrm{I},L} | \mathbf{W}_{\mathrm{I},l} \in \mathbb{R}^{D\times K}\}$ is a set of transformation matrices for mapping image features (vertical vector) of the (i, j)-th grid $\mathbf{f}_{(i,j)} \in \mathbb{R}^D$ obtained from a backbone model into the projection space, where D is the number of dimensions of the obtained image feature from the backbone model. All the vectors defined in this study are vertical vectors unless specified otherwise. This operation makes it possible to proceed with subsequent learning based on the understanding of the parts that correspond to each dimension of the embedded representation to be acquired. Specifically, the embedded representation of the full-body outfit image $\mathbf{x}$ can be conceived as a concatenation of the embedded representations of L parts $\hat{\mathbf{x}}_1, \cdots, \hat{\mathbf{x}}_l, \cdots, \hat{\mathbf{x}}_L$ by Eq. (5.1). Thus, it is clear which part each element of $\mathbf{x}$ refers to. Therefore, an operation such as changing only a specific part while leaving other parts unchanged is possible by changing only the l-th part of embedded representation $\hat{\mathbf{x}}_l$ in the full-body outfit image embed representation $\mathbf{x}$. Thereby, it realizes an embedded representation model that is extremely easy to handle.

The embedded representation of the tag set assigned to an image is heuristically weighted to generate an embedded representation considering the bias in the frequency with which each tag is assigned to the entire dataset. The heuristic weighting rule is based on the assertion that “tags appearing infrequently in the overall dataset are more likely to be important elements that characterize the image (differentiate it from other images).”

$$\mathbf{v} = \sum_{t=1}^{T} w_t\hat{\mathbf{v}}_t, \tag{5.3}$$

$$w_t = \frac{1/\log(N_t + 1)}{\sum_{t=1}^{T} 1/\log(N_t + 1)}, \tag{5.4}$$

where $\mathbf{v} \in \mathbb{R}^{KL}$ is the embedded representation of the tag set, $\hat{\mathbf{v}}_t \in \mathbb{R}^{KL}$ is the embedded representation of the t-th single tag, N_t indicates the total attachment frequency of the t-th attached tag to the target image in the entire mini-batch, and T is the total number of tags included in the target image.

By optimizing Eq. (5.5), which includes the aforementioned features, the full-body outfit image, and the attached tags are mapped into the same projective space, and the embedded representation for each part corresponding to the full-body outfit image and the embedded representation for the tags are obtained.

$$\begin{aligned} l_{\text{npair\&ang}}(\mathrm{O}) &= l_{\text{npair}}(\mathrm{O}) + \lambda\Bigg(\frac{1}{2N}\sum_{n=1}^{N}\log\Big(1+\sum_{m\neq n}\exp\{f_{\text{ang}}(\mathbf{x}_n,\mathbf{v}_n,\mathbf{v}_m)\}\Big) \\ &\quad + \frac{1}{2N}\sum_{n=1}^{N}\log\Big(1+\sum_{m\neq n}\exp\{f_{\text{ang}}(\mathbf{v}_n,\mathbf{x}_n,\mathbf{x}_m)\}\Big)\Bigg), \end{aligned} \tag{5.5}$$

$$\begin{aligned} l_{\text{npair}}(\mathrm{O}) &= \frac{1}{2N}\sum_{n=1}^{N}\log\Big(1+\sum_{m\neq n}\exp\{\mathbf{x}_n{}^{\top}\mathbf{v}_m - \mathbf{x}_n{}^{\top}\mathbf{v}_n\}\Big) \\ &\quad + \frac{1}{2N}\sum_{n=1}^{N}\log\Big(1+\sum_{m\neq n}\exp\{\mathbf{v}_n{}^{\top}\mathbf{x}_m - \mathbf{v}_n{}^{\top}\mathbf{x}_n\}\Big), \end{aligned} \tag{5.6}$$

where $\mathrm{O} = \{\mathrm{V}, \mathrm{W}_{\mathrm{I}}, \mathbf{W}_{\mathrm{T}}\}$ is a set of target parameters to be optimized, V is a parameter set contained in the backbone model, $\mathbf{W}_{\mathrm{T}} \in \mathbb{R}^{H\times KL}$ is the transform matrix from a bag-of-words representation to the t-th tag-embedded representation $\hat{\mathbf{v}}_t$, and H is the number of unique tags in the entire dataset. Additionally, N is the number of positive samples in the batch data. Furthermore, λ is a positive hyperparameter that compensates for the N-pair loss [165] and angular loss [166], and α is the angular loss margin (angle). Each embedded representation is normalized when calculating the loss. The detailed operation of $f_{\text{ang}}(\cdot)$ is described in Eq. (5.8) after the derivation process. In addition, note that the number T in Eqs. (5.3)–(5.4) is variable for each target image when calculating $\mathbf{v}_n$ and $\mathbf{v}_m$. For example, Eq. (5.3) can be expressed as $\mathbf{v}_n = \sum_{t=1}^{T_n} w_t\hat{\mathbf{v}}_t$ strictly in the case of the n-th tags embedded representation $\mathbf{v}_n$ where T_n is the total number of tags attached to the n-th image.

The loss function is defined by combining N-pair loss and angular loss, which is more stable than the triplet loss [167] employed in many VSE models. The loss is calculated, as shown in Figure 5.3.

In the N-pair loss, all positive samples in the mini-batch, except for the positive sample corre-

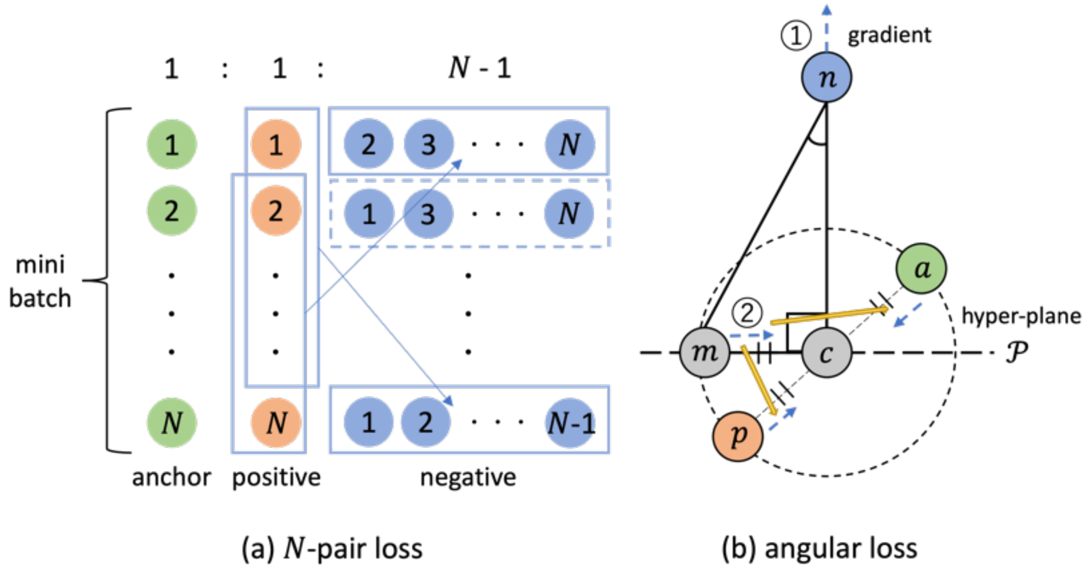


Figure 5.3: Images of N-pair loss & angular loss

sponding to the target anchor sample, are treated as negative samples and trained to move away from the anchor sample. This system allows us to use numerous samples in a single training session without increasing the computational complexity, thereby achieving stable learning.

Angular loss considers the relative positional relationship (angle) between the anchor and positive and negative samples to achieve stable learning. As shown in Figure 5.3(b), triangle $\triangle cmn$ is structured by 1) midpoint c (coordinate vector $\mathbf{e}_c$) between anchor point a (coordinate vector $\mathbf{e}_{\text{anc}}$) and positive point p (coordinate vector $\mathbf{e}_{\text{pos}}$); 2) negative point n (coordinate vector $\mathbf{e}_{\text{neg}}$); and 3) point m (coordinate vector $\mathbf{e}_m$) on hyperplane $\mathcal{P}$ perpendicular to edge nc and on the circumference of the circle of radius $ac(cp)$ centered at point c and is used to achieve learning by considering the relative positions of the anchor, positive, and negative. The basic concept is that by making the angle $\angle cnm$ smaller than the margin α, the gradient works in two directions (1 and 2 in Figure 5.3). The negative sample moves away from the anchor sample and the positive sample moves closer. This concept is expressed through trigonometric functions, as expressed by Eq. (5.7).

$$\tan \angle cnm = \frac{||\mathbf{e}_m - \mathbf{e}_c||}{||\mathbf{e}_{\text{neg}} - \mathbf{e}_c||} = \frac{||\mathbf{e}_{\text{anc}} - \mathbf{e}_{\text{pos}}||}{2||\mathbf{e}_{\text{neg}} - \mathbf{e}_c||} \leq \tan \alpha, \tag{5.7}$$

where $||\mathbf{e}_m - \mathbf{e}_c|| = \frac{||\mathbf{e}_{\text{anc}} - \mathbf{e}_{\text{pos}}||}{2}$ is established because the edge cm is half of the diameter ap.

Eq. (5.7) is expanded in Eq. (5.8).

$$
\begin{aligned}
& f_{\text{ang}}(\mathbf{e}_{\text{anc}}, \mathbf{e}_{\text{pos}}, \mathbf{e}_{\text{neg}}) \\
& \quad = ||\mathbf{e}_{\text{anc}} - \mathbf{e}_{\text{pos}}||^2 - 4||\mathbf{e}_{\text{neg}} - \mathbf{e}_{\text{c}}||^2 \tan^2 \alpha \\
& \quad = 4(\mathbf{e}_{\text{anc}} + \mathbf{e}_{\text{pos}})^{\top} \mathbf{e}_{\text{neg}} \tan^2 \alpha - 2\mathbf{e}_{\text{anc}}{}^{\top} \mathbf{e}_{\text{pos}}(1 + \tan^2 \alpha),
\end{aligned}
\tag{5.8}
$$

where the coordinates of point c are expressed as $\mathbf{e}_c = \frac{||\mathbf{e}_{\text{anc}} + \mathbf{e}_{\text{pos}}||}{2}$, and the constant terms that depend on the value of $\mathbf{e}$ are dropped in the process of unfolding. Eq. (5.5) was derived by extending this angular loss to N pairs (batch angular loss) and combining it with the N-pair loss.

5.3.4 Image Retrieval

Images can be retrieved using image- and tag-adding or subtracting operations because the proposed model maps tags and images into the same projective space. Basic image retrieval is accomplished by adding (positive) and subtracting (negative) tags to the query image and is expressed as Eq. (5.9).

$$
\mathbf{x}_{\text{o}} = \underset{\mathbf{x}}{\operatorname{argmax}}\, s\left(\mathbf{x}_{\text{q}} + \mathbf{v}_{\text{pos}} - \mathbf{v}_{\text{neg}}, \mathbf{x}\right), \tag{5.9}
$$

where $\mathbf{x}_{\text{o}}, \mathbf{x}_{\text{q}} \in \mathbb{R}^{KL}$ denote the embedded representation of the output and query images respectively, $\mathbf{v}_{\text{pos}}, \mathbf{v}_{\text{neg}} \in \mathbb{R}^{KL}$ are the embedded representation of the positive and negative tags respectively, and $s(\mathbf{x}, \mathbf{y})$ indicates the cosine similarity between vectors $\mathbf{x}$ and $\mathbf{y}$. This operation enables, for example, an image search for “I want to know the coordination of office casual by subtracting the casual element from the target coordination.”

However, image retrieval based on the above calculation is a function that is also provided in the conventional VSE model in [8] and cannot meet the detailed needs of users who want to make minor changes only to the tops. In contrast, the proposed PVSE model allows the user to know the parts to which each dimension in the embedded representation of images and tags corresponds. Using this advantage, delicate image retrieval, which makes changes only to the parts specified by the user, is achieved by adding or subtracting the embedded representation of the tag, as expressed in Eq. (5.10).

$$\tilde{v}_{\mathrm{pos},k} = \begin{cases} v_{\mathrm{pos},k} & (\text{if. } k \in \mathrm{K_q}), \\ 0.0 & (\text{otherwise}), \end{cases} \tag{5.10}$$

where $v_{\mathrm{pos},k}$ denotes the k-th element of $\mathbf{v}_{\mathrm{pos}}$, and $\mathrm{K_q}$ is the set of dimensions corresponding to the query parts (target parts to be modified) specified by the user. Additionally, $\tilde{\mathbf{v}}_{\mathrm{pos}} \in \mathbb{R}^{KL}$ constructed by each element $\tilde{v}_{\mathrm{pos},k}$ is used instead of $\mathbf{v}_{\mathrm{pos}}$ in Eq. (5.9). In addition, the negative tag $\tilde{\mathbf{v}}_{\mathrm{neg}} \in \mathbb{R}^{KL}$ is also calculated by the same operation. Therefore, this simple operation Eq. (5.10) realizes image retrieval by focusing on a specific part.

Furthermore, a positive tag and its corresponding negative tag must be specified to maintain the overall atmosphere of the query image in the conventional image and tag computation for retrieval using the VSE model. However, the overall atmosphere can be maintained by the embedding representation of dimensions corresponding to parts other than the specified parts with the embedding representation obtained from the proposed PVSE model. Therefore, even without selecting a negative tag, Eqs. (5.10)–(5.11) and (5.12) enable image retrieval with minor changes made only to the specified part.

$$\mathbf{x}_{\mathrm{o}} = \underset{\mathbf{x}}{\mathrm{argmax}}\, s\left(\tilde{\mathbf{x}}_{\mathrm{q}} + \tilde{\mathbf{v}}_{\mathrm{pos}}, \mathbf{x}\right), \tag{5.11}$$

$$\tilde{x}_{\mathrm{q},k} = \begin{cases} 0.0 & (\text{if. } k \in \mathrm{K_q}), \\ x_{\mathrm{q},k} & (\text{otherwise}), \end{cases} \tag{5.12}$$

where $x_{\mathrm{q},k}$ denotes the k-th element of $\mathbf{x}_{\mathrm{q}}$, and $\tilde{\mathbf{x}}_{\mathrm{q}} \in \mathbb{R}^{KL}$ constructed by each element $\tilde{x}_{\mathrm{q},k}$ is used as the query image in Eq. (5.11). Additionally, if multiple tags are used to create $\tilde{\mathbf{v}}_{\mathrm{pos}}$ (e.g., "casual" and "khaki-colored" upper clothes), the average of those tags is obtained and applied to Eq. (5.10) above.

5.3.5 Image Reordering

Because the proposed model maps words and images into the same projective space, the similarities (relevance scores) of all the images (to which the target tag is attached) to the target tag can be calculated, and the images are sorted in order of the scores. This function is also possible with the conventional VSE model. However, in this study, image reordering by focusing on a specific part can be obtained by calculating the relevance score of the images and target tags in only the features in the dimensions corresponding to the target part. This feature can be used, for instance, to respond to the natural desire of the user to "look up a coordinated outfit with particularly (or not particularly) casual upper garments."

5.3.6 Attribute Activation Map Creation

An AAM can be obtained using the VSE model by creating a heatmap of the relevance scores between the embedded representation corresponding to each grid and the specified tag.

Because each grid contains either single or multiple parts, a weighting calculation using grid weight map $G'_{(i,j)} = \{g'_{(i,j),1}, \cdots, g'_{(i,j),l}, \cdots, g'_{(i,j),L}\}$ is used to calculate the embedded representation corresponding to each grid, where $g'_{(i,j),l} = N_{(i,j),l}/N_{(i,j)}$ and $N_{(i,j)}$ denote the number of pixels included in the (i, j)-th grid. Therefore, $g'_{(i,j),l}$ is the fraction of pixels that contain the l-th part in all pixels in the (i, j)-th grid.

Eq. (5.13) expresses the embedded representation corresponding to the (i, j)-th grid $\mathbf{x}_{(i,j)} \in \mathbb{R}^K$, considering the (single or) multiple parts included in the grid.

$$\mathbf{x}_{(i,j)} = \sum_{l=1}^{L} g'_{(i,j),l} \mathbf{W}_{\mathrm{I},l} \mathbf{f}_{(i,j)}. \tag{5.13}$$

The embedded representation of the tag to be compared with the (i, j)-th grid in the image when calculating the relevance score $\mathbf{v}_{(i,j)} \in \mathbb{R}^K$ is determined using Eq. (5.14).

$$\mathbf{v}_{(i,j)} = \sum_{l=1}^{L} g'_{(i,j),l} \mathbf{v}_{\mathrm{q},l}, \tag{5.14}$$

where $\mathbf{v}_{\mathrm{q},l} \in \mathbb{R}^K$ denotes the embedded representation of the l-th part of the query tag.

The relevance score in the (i, j)-th grid between the image and tag is obtained by calculating the similarity between $\mathbf{x}_{(i,j)}$ and $\mathbf{v}_{(i,j)}$. An AAM can be created while considering the ratio of each part reflected in each grid using this score.

5.4 Experimental Evaluation

Two types of evaluation experiments were conducted to quantitatively evaluate the effectiveness of the proposed PVSE model. Specifically, the experiments were conducted in terms of 1) the similarity between the embedded representation of the image and the attached tag and

2) the relevance scores between parts and tags. Although the first experiment has already been proposed as an evaluation experiment method using the VSE space [8], the second experiment is a new fashion-specific, evaluation method using parts.

5.4.1 Experimental Settings

5.4.1.1 Dataset Details

The experiments used data collected from posts on the fashion outfit-sharing application WEAR [5]. From a large number of user posts, we extracted posts that contain clear full-body images (from head to toe) with relatively little noise in the background to construct an experimental dataset. The number of full-body outfit images in the experimental data was 15,740, and 1,104 unique tags were attached to these images. All target images were of women. An example of a full-body outfit image and its tags are shown in Figure 5.4.

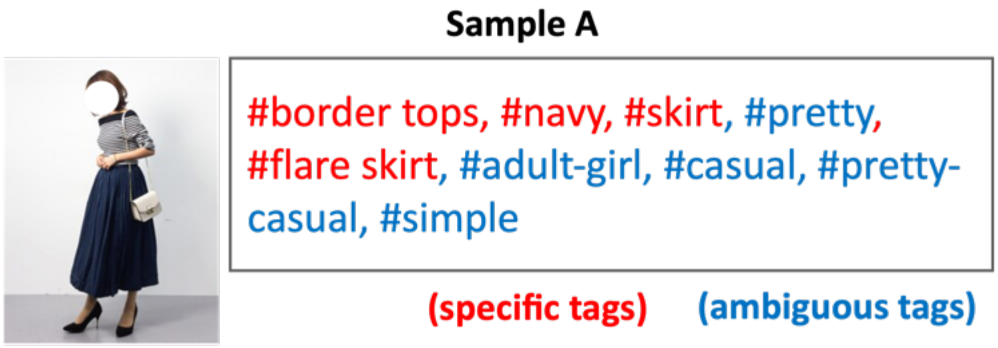

Sample A
#border tops, #navy, #skirt, #pretty, #flare skirt, #adult-girl, #casual, #pretty-casual, #simple
(specific tags)
(ambiguous tags)

Figure 5.4: Example of samples in the target dataset [5]

The image data used in this study are full-body outfit images of a single subject (a person). Each image is assigned several tags as attribute information by the user who posted the image. In addition, the tag information includes not only concrete and simple tags, such as “border tops,” “navy,” “skirt,” and “flare skirt,” but also ambiguous tags, such as “pretty,” “adult-girl,” “casual,” “pretty-casual,” and “simple.”

5.4.1.2 Parameter Settings and Preprocessings

The dimensions of the embedded representation included in the PVSE model were set as $KL = 128$. A stochastic gradient descent optimizer was used with an initial learning rate of 0.01

and was halved every five epochs. The overall number of epochs was 50, the batch size was 32, and the margin α was set to 36°. Based on preliminary experiments, we used GoogleNet Inception V3 [128] pre-trained on ImageNet [139] comparing several backbone models [138, 168], including ViT and bidirectional encoder representation from image transformers (BEiT) [169], which have achieved high accuracy in many recent studies [170, 171, 172, 173, 174, 175].

Furthermore, the unique characteristics of the fashion field mean that each part included in a full-body outfit must be detected, and many datasets and trained models have been released to realize this task. Among the many open-source pre-trained models, the context embedding with edge perceiving (CE2P) framework [176] (backbone model: ResNet-101 [168]) with the self-correction for human parsing (SCHP) strategy [177] pre-trained on the LIP [178] was used for the preprocessing (semantic segmentation) step in this study [179].

5.4.2 Evaluation Experiment 1: Position of Each Component in the Projection Space

The validity of the obtained embedding representation is verified by whether the image and tag that are nearby are closely mapped together. For example, an image with a “casual” tag is quantitatively evaluated based on the requirement of it being mapped near a “casual” tag. An image dataset was created by combining images that had the target tag and ten times as many images without the tag. Subsequently, we retrieved the top-M={5, 10, 15} images from the embedded representations of the images in the created image dataset that were closely mapped to the embedded representation of the target tag. It was determined whether the target tag had been assigned to each of these retrieved images. Precision and normalized documented cumulative gain (NDCG) [180] were used as the accuracy measures (P@M and N@M, respectively). For comparison, we tested the eight methods listed in Table 5.2. Positional embedding was removed from transformer-VSE (TVSE) [181] because text data were not the focus of this study, and H in TVSE-H represents the number of heads in the MHA layer. To check the change in the proposed model’s accuracy depending on the division of the parts, we tested the following three models: 1) PVSE-4, which is the proposed model that divided the full-body outfit image into four parts {Head, Upper-body, Lower-body, and Shoes} and was trained; 2) PVSE-8, which is the proposed model with the eight-part setting {Head, Upper-body, Dress, Coat, Lower-body, Arm, Leg, and Shoes}; 3) PVSE-16, which is the proposed model with the sixteen-part setting

{Hat, Hair, Glove, Sunglasses, Upper-body, Dress, Coat, Socks, Trousers, Jumpsuits, Skirt, Face, Arm, Leg, Left-shoe, and Right-shoe}. The experiment was repeated three times, and the mean and standard deviation values were calculated.

Table 5.2: Summary of model-type evaluation values from evaluation experiment 1

	P@5	P@10	P@15	N@5	N@10	N@15
Random	0.092	0.086	0.091	0.083	0.079	0.088
	±.005	±.003	±.004	±.005	±.006	±.006
VSE [86]	0.412	0.373	0.355	0.377	0.378	0.368
	±.010	±.002	±.004	±.010	±.005	±.005
VSE+ [8]	0.541	0.485	0.454	0.488	0.489	0.473
	±.005	±.009	±.011	±.004	±.006	±.009
GVSE [182]	0.423	0.399	0.397	0.392	0.395	0.387
	±.010	±.013	±.012	±.013	±.010	±.010
DGVSE	0.456	0.431	0.412	0.420	0.423	0.416
	±.029	±.025	±.023	±.026	±.027	±.012
TVSE-4	0.273	0.263	0.256	0.252	0.254	0.254
	±.017	±.012	±.010	±.015	±.012	±.012
TVSE-8	0.266	0.263	0.255	0.250	0.253	0.252
	±.008	±.008	±.008	±.003	±.007	±.006
TVSE-16	0.269	0.262	0.249	0.249	0.252	0.249
	±.013	±.009	±.012	±.003	±.009	±.006
PVSE-4	**0.927**	**0.831**	**0.765**	**0.826**	**0.831**	**0.797**
	±.001	±.002	±.002	±.002	±.001	±.001
PVSE-8	0.912	0.820	0.757	0.815	0.821	0.788
	±.011	±.008	±.010	±.012	±.010	±.010
PVSE-16	0.886	0.799	0.740	0.796	0.799	0.770
	±.004	±.006	±.007	±.005	±.005	±.006

Table 5.2 indicates that the proposed models are more accurate than the comparison models, including conventional VSE models, regardless of how the parts are divided. The low accuracy of TVSE suggests that the transformer may have low compatibility with the target problem of mapping a tag set containing both ambiguous and specific expressions (and the number of tags in each image is not uniform) and images into the same space. Furthermore, the high accuracy of PVSE suggests that the heuristic weighting works well, at least for the target problem. Additionally, the accuracy varies depending on how the parts are divided. In this experimental case, the rule that divides the parts into {Head, Upper-body, Lower-body, Shoes} shows the best accuracy. The proposed model map is more effective and sensitive compared to the conventional

method because all the dimensions of the conventional methods have semantic representations of all parts. This result indicates the effectiveness of the proposed model by mapping each dimension of the embedded representation to a single part. It is intuitively clear that excessively dividing parts does produce higher accuracy. Excessively fine partitioning yields an embedded representation of images for situations in which many parts are missing (e.g., dress and coat not worn together). It is unclear whether this is desirable. More parts and the fewer dimensions are included that represent the information of the important parts, when the number of dimensions KL is fixed. Therefore, setting an appropriate number of parts and dividing the information obtained from the full-body outfit image is critical to ensure that important information is not lost in the mapping process.

5.4.3 Evaluation Experiment 2: Attention to Appropriate Parts

The regions in an image and tag that are highly relevant can be determined by calculating the relevance score for each grid and tag using the VSE model. We checked whether the relevance score is high for an appropriate region. When the relevance scores are calculated among tags such as "t-shirt," "jeans," and "sneakers," along with the images to which these tags are attached, the relevance scores should be higher for the regions containing "Upper-body," "Lower-body," and "Shoes," respectively. Then, we compared the relevance scores calculated between a specific tag and each image attached to the tag to obtain the top M=5 regions. The regions' true labels were based on semantic segmentation results, while precision and NDCG were used as accuracy indices. The comparison models similar to those of the target models were applied in Section 5.4.2. As specific tags in the experiments, five tags that clearly corresponded to each of the four categories of {Head: (beret, glasses, hair bun, bob hair, and knit hat), Upper-body: (one-piece dress, blouse, cardigan, t-shirt, and outer), Lower-body: (denim, wide-trousers, skirt, trousers, and black skinny), and Shoes (ballet shoes, Converse, sneakers, sandals, and loafers)} were selected in order of their attached frequency in the entire dataset. The experiment was repeated three times, and the mean and standard deviation values were calculated.

This evaluation method, which checks the quality of the representations deeply, is unique to fashion data. This evaluation method, which thoroughly evaluates the quality of the representations, is unique to fashion data. Moreover, even if previous studies have suggested the AAM, evaluating it quantitatively has been challenging and limited because only a portion of the results were observed. By contrast, this evaluation is a novel methodology for evaluating whether the AAM can be created accurately and quantitatively. In addition, it plays a role in assessing the quality of the AAM itself and in evaluating the validity of the meaning aggregated in each dimension of the embedded representation. Namely, an evaluation of the validity of the detailed meaning possessed by each dimension of the obtained representations is possible. This is a task ignored in previous embedded representation-related studies because of its difficulty.

Table 5.3: Summary of model type evaluation values from evaluation experiment 2

	Head		Upper-body		Lower-body		Shoes	
	P@5	N@5	P@5	N@5	P@5	N@5	P@5	N@5
Random	0.151	0.136	0.539	0.487	0.343	0.307	0.093	0.083
	±.007	±.006	±.008	±.008	±.005	±.004	±.005	±.005
VSE	0.398	0.363	0.715	0.646	0.892	0.808	0.187	0.173
	±.165	±.151	±.054	±.051	±.002	±.002	±.024	±.024
VSE+	0.538	0.493	0.795	0.715	0.868	0.786	0.219	0.213
	±.091	±.080	±.004	±.005	±.032	±.027	±.119	±.120
GVSE	0.328	0.300	0.763	0.683	0.847	0.716	0.183	0.173
	±.072	±.063	±.059	±.053	±.064	±.042	±.019	±.019
DGVSE	0.460	0.417	**0.944**	**0.853**	0.888	0.801	0.161	0.154
	±.033	±.031	±.015	±.012	±.008	±.011	±.098	±.094
TVSE-4	0.524	0.479	0.760	0.682	0.676	0.617	0.023	0.020
	±.064	±.061	±.090	±.086	±.113	±.111	±.010	±.009
TVSE-8	0.568	0.523	0.760	0.679	0.633	0.578	0.031	0.027
	±.062	±.048	±.122	±.116	±.172	±.157	±.015	±.014
TVSE-16	0.560	0.510	0.745	0.669	0.715	0.653	0.031	0.027
	±.049	±.043	±.131	±.120	±.068	±.064	±.014	±.013
PVSE-4	0.752	0.689	0.848	0.768	**0.977**	**0.884**	0.485	0.465
	±.062	±.062	±.015	±.015	±.006	±.004	±.048	±.047
PVSE-8	0.708	0.664	0.871	0.793	**0.977**	**0.884**	0.517	0.503
	±.160	±.149	±.040	±.035	±.002	±.001	±.098	±.095
PVSE-16	**0.877**	**0.807**	0.803	0.731	0.972	0.880	**0.646**	**0.622**
	±.019	±.015	±.094	±.087	±.003	±.004	±.015	±.012

The results presented in Table 5.3 indicate that the proposed model exhibits better accuracy for most of the indices compared with the comparison models, including the conventional VSE models. The DGVSE indices are higher only for the upper-body, while the other parts do not yield higher accuracy. This phenomenon, which is that the accuracies for upper-body (and lower-body) are high whereas the accuracies for shoes (and head) are quite low, was obtained for all models other than PVSE. This suggests that the embedded representations overly reflect the upper and lower body features, which occupy large regions in the image, and that small but essential parts, such as the head and shoes, are not properly learned. Thus, PVSE is not a suitable learning method for handling all parts with equal accuracy. However, the proposed model is universally accurate in all cases. This result can be attributed to the fact that each part is learned, suggesting that the proposed model can learn in a way that is sufficient to accurately map each component.

5.5 Additional Analysis

The experimental evaluation shows that the proposed model can map a full-body outfit image and the attached tags into the same projective space more reasonably than the comparison models. This section demonstrates the effectiveness of the PVSE model as a fashion intelligence system with multiple types of practical applications based on the results obtained from the proposed model. The analysis conditions are the same as those in Section 5.4.1. This section includes the qualitative evaluation.

5.5.1 Image Retrieval and Reordering

Examples of image retrieval obtained by image and tag operations are shown in Figure 5.5.

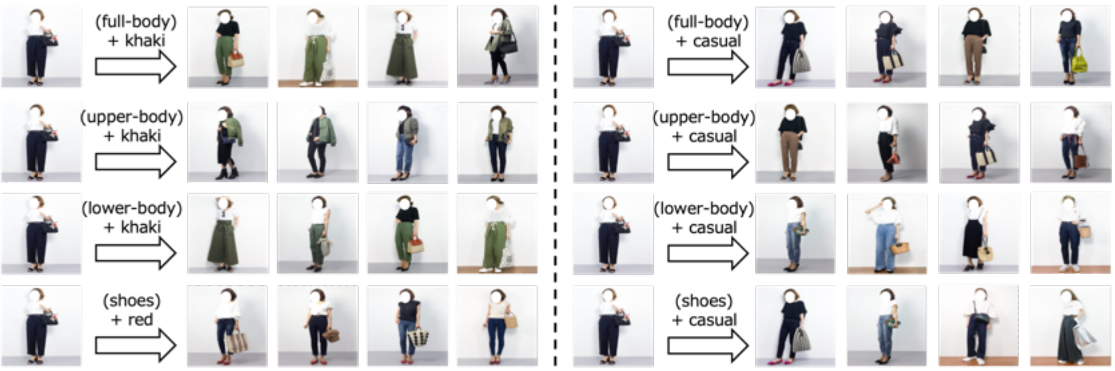


Figure 5.5: Example of image retrieval focusing on the specified part

We checked the validity of the results by observing the search results with specific tags. When the embedded representation of the “khaki” tag is added to the entire embedded representation of the query image (casual adult attire consisting of a white t-shirt and navy trousers), an outfit in which the upper or lower body turns khaki in color is returned. In contrast, specifying the changed parts returned outfits in which those parts had been changed (i.e., minor changes were made). These results indicate that the proposed model can acquire an embedded representation for each part and retrieves the image by making minor changes around the specified part. Furthermore, the dressing method that renders a query attire casual can be grasped by using ambiguous tags. For example, to change a full-body outfit, the top could be made black with a colorful item added for it to be more casual. To make a minor change to the lower body to create an overall casual atmosphere, or from navy skinny to jeans or loose skirts. Thereby, the results obtained from the proposed model can be used to answer ambiguous questions unique to fashion that are difficult for non-experts (and not easy for experts).

The results of image reordering obtained by image and tag operations are shown in Figure 5.6.

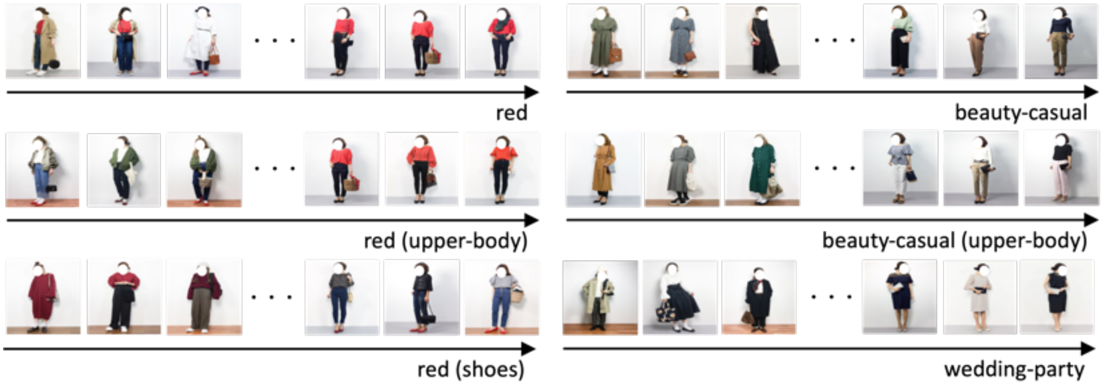


Figure 5.6: Example of image reordering focusing on the specified part

We checked the validity of the results by observing the results sorted by specific tags. When sorted using the “red” tag, which was limited to the upper-body and shoes, the images with red items in the specified parts ranked higher. Therefore, the results can be confirmed as reasonable. Additionally, ambiguous tags can be used for sorting. For example, “beauty-casual” clothes with a thin silhouette are more “beauty-casual” than those with loose silhouettes. Furthermore, dresses are the usual clothes for weddings and those that are not dresses are unusual. If a user wants to go with a unique outfit, it is preferable to choose an outfit with a low relevance score with the “wedding-party” tag. On the other hand, if the user wants to go with a typical outfit, it is preferable to choose an outfit with a high relevance score. Thereby, images can be rearranged by specifying attention parts to meet the detailed needs of the user to discover typical full-body outfits indicated by ambiguous tags and suitable clothing for each situation.

Ambiguous tags can be used to interpret fashion-specific ambiguous expressions even without specifying parts. For example, other results of image retrievals without specifying parts are shown in Figures 5.7-5.8.

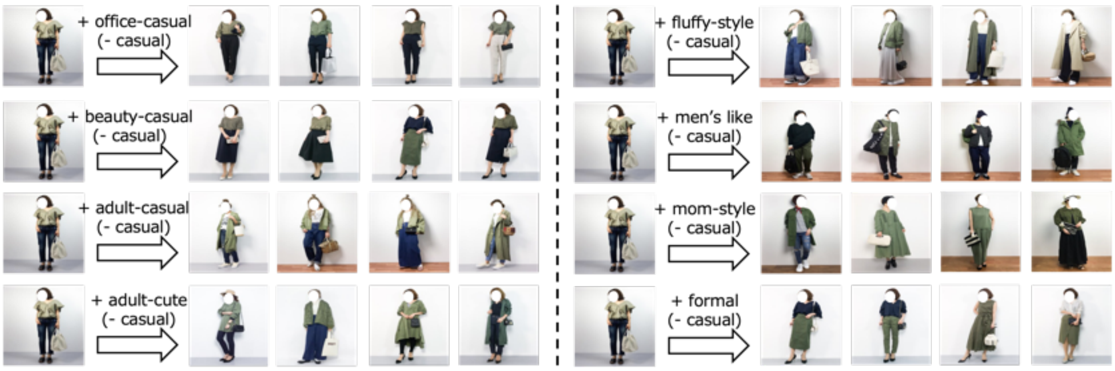


Figure 5.7: Example of “khaki” and “casual” outfit images retrieved by addition and subtraction

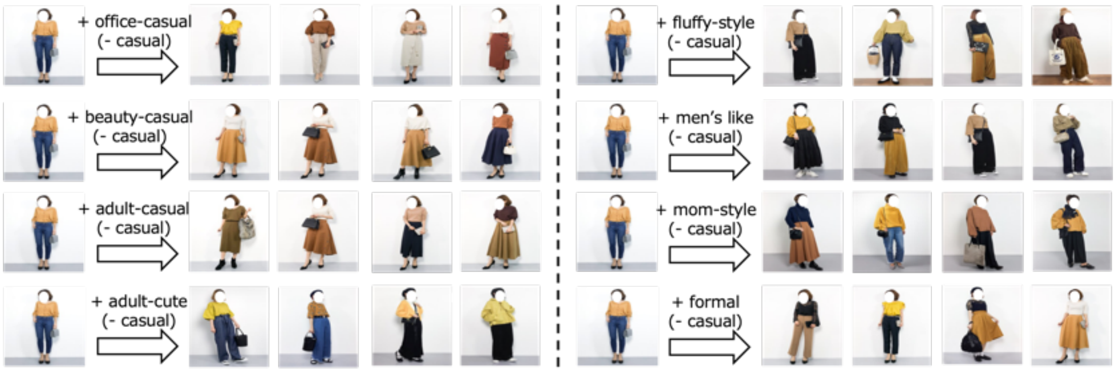


Figure 5.8: Example of “yellow” and “casual” outfit images retrieved by addition and subtraction

Thus, the type of atmosphere indicated by each ambiguous expression can be easily grasped using the image retrieval function. This function allows users to search for various variations of clothing, such as “casual,” “office-casual,” “beauty-casual,” and “adult-cute,” while maintaining the user’s preferred hue.

5.5.2 Attribute Activation Map

An AAM can be created by calculating the relevance scores of the embedded representation for each region of the image and of the representation of the target tag and representing these in a heatmap. Examples of AAMs are shown in Figure 5.9.

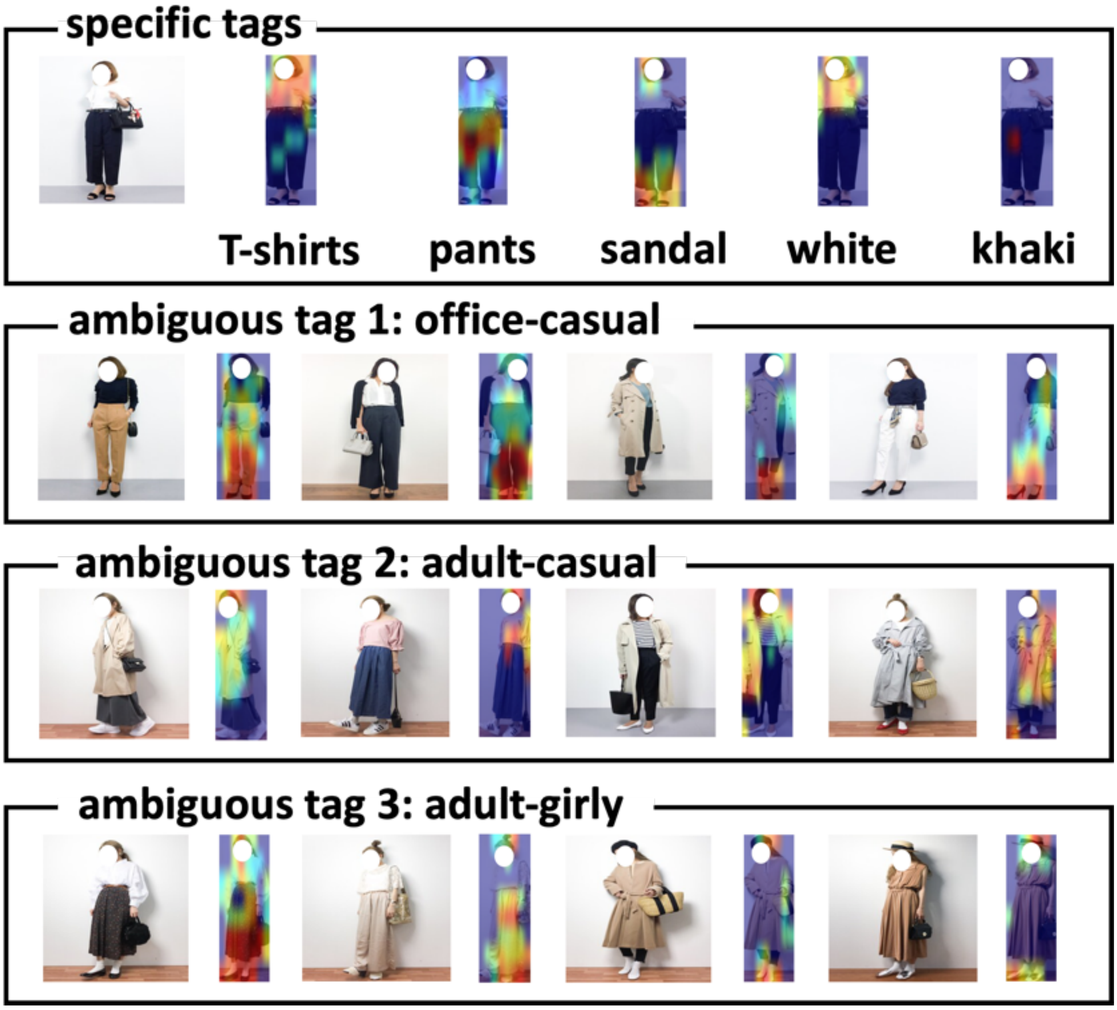

specific tags
T-shirts
pants
sandal
white
khaki
ambiguous tag 1: office-casual
ambiguous tag 2: adult-casual
ambiguous tag 3: adult-girly

Figure 5.9: Example of AAM

First, we verified the validity of the proposed model by observing the results for specific tags. The results show that "t-shirts," "trousers," "sandals," and "white" tags, which are attached to the target image, are colored in the appropriate places. In contrast, tags, such as "khaki", which are not relevant to the target fashion image and had lower relevance scores for all of the regions in the image. This indicates that the results are reasonable. Furthermore, for example, when we look at the ambiguous tags, the top items tend to be the key points for "adult-casual" coordinates. Additionally, "adult-girly" are associated with rounded items; however, in the case of coordinates that include items such as berets and straw hats, these items are the key points. Therefore, the region of interest in the full-body outfit image can be found by applying the results of the proposed model.

5.6 Discussion

5.6.1 Ablation Study

Tables 5.4–5.5 list the results of the ablation study. Here, "w/o all" is the same as simple VSE in [86], "w/o hw" represents PVSE-4 without the heuristic weighting for tags, "w/o gwap" represents PVSE without the grid weight map and GWAP processing, "w/o part" represents PVSE without the GWAP for parts, i.e., GWAP for the foreground, "w/o N-p ang" represents PVSE-4 without N-pair angular loss (triplet loss is adopted based on VSE in [8, 86]), and "ours" represents the complete PVSE-4.

Table 5.4: Summary of the ablation study results from evaluation experiment 1

	P@5	P@10	P@15	N@5	N@10	N@15
w/o all	0.412 ±.010	0.373 ±.002	0.355 ±.004	0.387 ±.010	0.378 ±.005	0.368 ±.005
w/o hw	0.790 ±.012	0.702 ±.005	0.658 ±.008	0.733 ±.011	0.707 ±.004	0.688 ±.009
w/o gwap	0.806 ±.032	0.729 ±.030	0.698 ±.008	0.722 ±.062	0.731 ±.028	0.703 ±.029
w/o part	0.857 ±.032	0.734 ±.030	0.713 ±.008	0.787 ±.062	0.773 ±.028	0.742 ±.029
w/o N-p ang	0.561 ±.013	0.512 ±.012	0.479 ±.008	0.520 ±.010	0.512 ±.010	0.496 ±.008
ours	**0.927** ±.001	**0.831** ±.002	**0.765** ±.002	**0.849** ±.002	**0.831** ±.001	**0.797** ±.001

First, the results in Table 5.4 reveal that the high accuracy of the proposed model is achieved because of the contributions of all mechanisms. Evaluation experiment 1 evaluates whether images and tags can be projected to appropriate positions in the target projection space. Thus, because this experiment's accuracy is directly linked to the accuracy of image retrieval and reordering, it is obvious that each mechanism is indispensable.

By contrast, the results in Table 5.5 indicate that in evaluation experiment 2, the accuracy is linked to the accuracy of the AAM, and the accuracy of "w/o part" was the highest for the head and upper-body. This result suggests that the operation that mixes the K-dimensional embedded representations (vectors) obtained for each part using the grid weight map in Eqs. (5.13)–(5.14) does not work perfectly. The proposed model performs contrastive learning on KL-dimensional

Table 5.5: Summary of the ablation study results from evaluation experiment 2

	Head		Upper-body		Lower-body		Shoes	
	P@5	N@5	P@5	N@5	P@5	N@5	P@5	N@5
w/o all	0.398 ±.165	0.363 ±.151	0.715 ±.054	0.646 ±.051	0.892 ±.002	0.808 ±.002	0.187 ±.024	0.173 ±.024
w/o hw	0.439 ±.037	0.398 ±.034	0.768 ±.065	0.683 ±.066	0.854 ±.105	0.683 ±.066	0.378 ±.091	0.366 ±.090
w/o gwap	0.740 ±.112	0.671 ±.106	0.911 ±.076	0.824 ±.067	0.919 ±.028	0.831 ±.027	0.212 ±.112	0.198 ±.104
w/o part	**0.837** ±.069	**0.766** ±.063	**0.962** ±.010	**0.869** ±.009	0.931 ±.018	0.842 ±.015	0.316 ±.111	0.297 ±.104
w/o *N*-p ang	0.373 ±.082	0.343 ±.075	0.556 ±.097	0.492 ±.092	0.764 ±.097	0.696 ±.085	0.250 ±.096	0.233 ±.096
ours	0.752 ±.062	0.689 ±.062	0.848 ±.015	0.768 ±.015	**0.977** ±.006	**0.884** ±.004	**0.485** ±.048	**0.465** ±.047

image and tag set vectors. However, when the relevance scores are computed and the AAM is created, these KL-dimensional vectors are weighted and aggregated into K-dimensional vectors. Thus, this phenomenon is caused by the operation of AAM creation, and the targeted task for learning is not entirely the same. However, from the results of Tables 5.2–5.3, it is clear that the accuracy is at least better than that of other comparison models. In addition, the fact that the results for "w/o part" are higher than those of "w/o gwap" emphasizes the validity of the direction of the proposed model, i.e., learning focused on specified parts. Furthermore, we note that because the "w/o gwap" and "w/o part" models cannot acquire embedded representations corresponding to parts, the target tasks of this study (image retrieval and re-ordering tasks focusing on specified parts) are impossible. For these reasons, this limited result does not imply the proposed model itself is not effective. Investigating a method for computing the relevance score that improves the accuracy of evaluation experiment 2 is an issue for future work.

5.6.2 On the Computational Complexity

Figure 5.10 shows the results of comparing the time complexity and space complexity of each model evaluated in the experimental evaluation section.

From the results illustrated in Figure 5.10, the space computational complexity does not increase compared to the conventional VSE model, regardless of how finely the parts are divided.

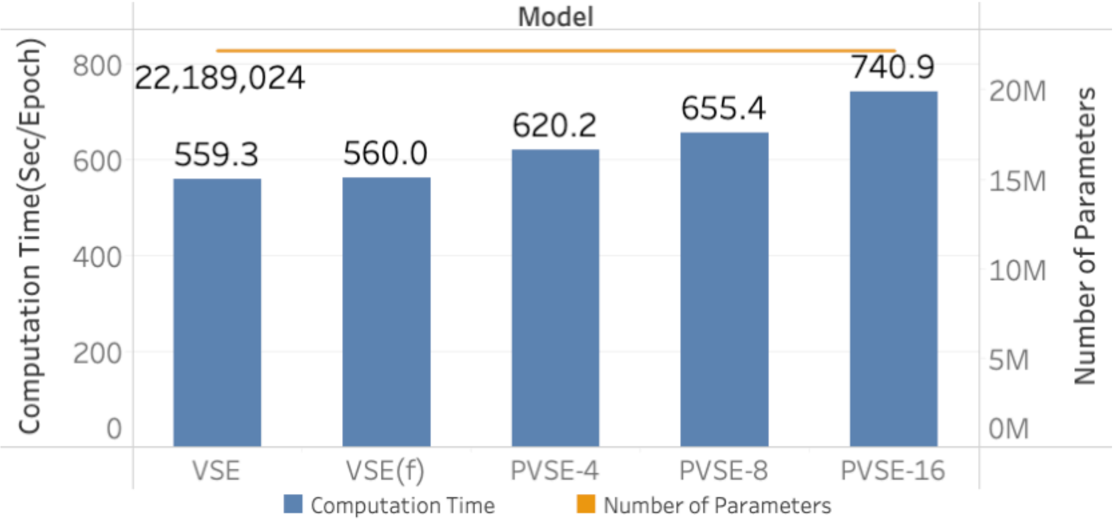


Figure 5.10: Summary of computational complexity (computation time and number of parameters)

Thus, the proposed model could be trained regardless of the memory specifications. Furthermore, the amount of training data required did not increase because the number of parameters did not increase. The time complexity increased by approximately 10-30%. These results are because the backbone model does not need to be separately applied to each item in the full-body image.

Under the experimental conditions of this study, the number of backbone model parameters accounted for more than 98% of the VSE model, and it accounts for most of the total training time of the entire model, even in the case where forward propagation of the backbone model is performed only once for each image. Thus, a structure in which backbone models are independently applied to each part requires significantly more computation time than an additional 10-30%. This suggests that our proposed method achieves per-part learning with a minimal increase in computation time. This leads to a decrease in the throughput, which is highly beneficial when considering real-world services.

5.6.3 On the Model Structure

A full-body outfit image has the unique characteristics mentioned in Section 5.2.1. These unique characteristics have hindered the realization of a partial fashion intelligence system because one could not be realized simply by using existing models from other domains. To handle these characteristics, the proposed model has a structure that contains GWAP processing with a grid weight map, in which one part corresponds to each dimension of the embedded representation of the image and the natural language expression. A partial fashion intelligence system was realized for the first time by the contribution of this direct and novel structure. To our knowledge, even if previous image-embedding models exist that that separate and extract the image features of each part or factor, such as color and shape, models that separate the dimensions of the natural language expression's embedded representation by corresponding parts do not. Moreover, studies that present such a wide range of practical applications for fashion interpretation that utilize the advantages of this partial learning method do not exist.

Indeed, individual techniques such as GWAP and the backbone model have already been proposed individually. However, we argued for the need for a partial fashion intelligence system, and a new PVSE model was constructed by combining these individual techniques with various modifications and improvements. Consequently, five applications can be provided using one relatively simple model, and this high versatility is one of the essential contributions of this study. As a result of the contribution of this study, it is expected that many studies will be conducted to realize similar systems using more complex models. Thus, this study has a vital role as a starting point for partial fashion intelligence systems.

In addition, the proposed feature extraction mechanism for each part based on the combined grid weight map and GWAP can be applied to many backbone models, mainly CNN and ViT-related models, regardless of the type. Therefore, this proposed mechanism can be applied to a backbone model suitable for the problems targeted by our paper's audience (analysts), and to more powerful backbone models that will be proposed in the future. This flexibility is also an important advantage. Moreover, the results of the multifaceted evaluation experiments confirm that the proposed model yielded higher accuracy than other methods in the embedding task of full-body outfit images. Consequently, it is suggested that the proposed model and the mechanism included in the proposed model may become the standard for embedding full-body outfit images.

5.6.4 On the Loss Function

This study adopted the N-pair angular loss when training the proposed model. However, many VSE models use triplet loss [8, 10, 73, 74, 81, 83, 86, 182] and N-pair loss [181, 183, 184], demonstrate the validity of N-pair angular loss, and explain the accuracy of other types of loss functions.

Tables 5.6-5.7 list the results of the evaluation experiments for each loss function.

Table 5.6: Summary of loss function type evaluation values from evaluation experiment 1

	P@5	P@10	P@15	N@5	N@10	N@15
triplet	0.560 ±.001	0.505 ±.007	0.469 ±.003	0.521 ±.002	0.510 ±.006	0.490 ±.004
N-pair	0.796 ±.020	0.714 ±.016	0.660 ±.013	0.737 ±.017	0.718 ±.015	0.690 ±.014
single angular	0.124 ±.009	0.118 ±.007	0.114 ±.010	0.113 ±.010	0.115 ±.006	0.114 ±.008
batch angular	0.918 ±.005	0.830 ±.009	0.764 ±.004	0.843 ±.004	0.829 ±.007	0.794 ±.004
N-pair angular	**0.927** ±.001	**0.831** ±.002	**0.765** ±.002	**0.849** ±.002	**0.831** ±.001	**0.797** ±.001

Table 5.7: Summary of loss function type evaluation values from evaluation experiment 2

	Head		Upper-body		Lower-body		Shoes	
	P@5	N@5	P@5	N@5	P@5	N@5	P@5	N@5
triplet	0.413 ±.121	0.382 ±.112	0.506 ±.052	0.449 ±.050	0.850 ±.067	0.772 ±.060	0.227 ±.125	0.210 ±.120
N-pair	0.465 ±.094	0.430 ±.087	0.834 ±.040	0.755 ±.036	0.887 ±.056	0.804 ±.051	0.257 ±.105	0.243 ±.103
single angular	0.192 ±.057	0.176 ±.053	0.393 ±.088	0.421 ±.054	0.393 ±.088	0.351 ±.082	0.101 ±.011	0.092 ±.006
batch angular	0.547 ±.115	0.502 ±.113	0.786 ±.047	0.708 ±.044	0.949 ±.019	0.860 ±.016	0.267 ±.086	0.256 ±.089
N-pair angular	**0.752** ±.062	**0.689** ±.062	**0.848** ±.015	**0.768** ±.015	**0.977** ±.006	**0.884** ±.004	**0.485** ±.048	**0.465** ±.047

The results in Tables 5.6-5.7 show that the N-pair angular loss adopted for training the proposed model exhibits the best accuracy compared with the other loss functions. Although omitted for space reasons, the experimental evaluation results, as in Section 5.4.3, show that the N-pair angular loss was the most effective. N-pair angular loss is a loss function that combines the N-pair loss and batch angular loss by hyperparameter λ. Based on the comparison of the results of single and batch angular losses, the batch angular loss is much higher, suggesting that a large number of negative samples must be used for a single anchor sample to render the angular loss more powerful. Additionally, inspired by [166], the batch angular loss was exceptionally high when combined with N-pair loss. This suggests that the proposed model can be more robust when combined with learning from both the N-pair and angular perspectives because it requires simultaneous mapping of a complex tag set that includes rich ambiguous tags and a complex image consisting of the combination of many parts.

5.6.5 On Future Actual Service Application

The proposed system can automatically obtain answers to ambiguous and difficult-to-answer questions from users, such as the aforementioned questions 1 to 5. We believe that users' online fashion lives will be enriched and more enjoyable after its introduction into the image search system or by offering it in the chat tool, as illustrated in Figure 5.1.

Furthermore, this system not only can be used on e-commerce or social media sites but also as a reference when store clerks advise users in actual stores. Thus, an application using the proposed model can be a powerful tool both online and offline. In addition, this study is expected to contribute to the solution of social problems, for instance by reducing waste, because users who need these items are better able to choose the correct ones.

5.7 Conclusion of this Chapter

In this study, we devised a PVSE model that can obtain an embedded representation for each of the multiple parts included in the full-body outfit image and attached rich tags. Each dimension of the resulting embedded representation of the image and tags corresponded to a single part of the full-body outfit. This feature of the proposed model maintains the three functions of the previous VSE model while adding two new functionalities. The proposed model outperforms the conventional model in terms of accuracy in multiple types of evaluation experiments. We confirm that these advantages can be obtained with a considerably slight increase in computational complexity. The proposed model is expected to be used in real-world applications, such as supporting users' purchasing activities on e-commerce sites and their browsing activities and learning about fashion on social media.

The aspects that have not been clarified regarding the necessity and effectiveness of this study include the use of heuristic weighting when transforming the tag set to the embedded representation. The results of the quantitative and qualitative evaluations on this dataset are reasonable;

however, it would be ideal if the heuristic part could be eliminated from the model training algorithm. In addition, the development of a more precise method of calculating the relevance scores for the AAM creation is a possible future task. Moreover, for future research, we aim to build models that can robustly learn fashion knowledge from the datasets, including various poses and backgrounds. Furthermore, while the basic model of the current fashion intelligence system is still a simple structure, the contribution of this study opens a novel research field, and it is expected that various complex models will be proposed to interpret fashion-specific ambiguous expressions. For example, in the method proposed in this study, semantic segmentation is performed as preprocessing. However, future tasks include the development of a model that treats the segmentation task in the same way as the core tasks of the partial fashion intelligence system.

Chapter 6

Detailed Fashion Image Analysis by Dual Gaussian Visual-Semantic Embedding

This chapter proposes a new visual-semantic embedding model that can estimate each embedding representation as a probability distribution. By utilizing the proposed model, we show that analyzing fashion-specific ambiguous expressions in more detail is possible.

6.1 Purpose of this Chapter

Nowadays, users refer to other people's fashion outfit images through e-commerce sites and social media and incorporate them into their own fashion and purchasing activities. Therefore, the fashion industry must support users' online search for fashion images to increase their interest in fashion and willingness to purchase. In the fashion field, outfits and items are described in subjective and abstract expressions such as "casual," "adult casual," "beautiful casual," and "formal." Many users perceive these expressions with difficulty, which discourages them from trying new fashion.

In response to this problem, as mentioned in Chapter 4, we proposed a system to support the interpretation of these terms through various application systems that apply VSE [8]. In this system, fashion images and attributes (tags) are mapped in the same embedding space to obtain an embedded representation. However, images and tags are processed to be mapped to only one point in the embedding space. Thus, abstract tags, such as casual, which have a wide range of meanings that can be interpreted differently by each person, and concrete tags, such as jeans, are treated in the same way and mapped to only one point. This is desirable because it does not represent the breadth of the meaning (diversity) each tag or image has in the real world in the destination space.

In contrast, Gaussian embedding, as represented by Word2Gauss [185], embeds each element as a probability distribution, assuming a (multidimensional) Gaussian distribution behind each embedded representation of the mapped object. This method is expected to contribute to quantifying and interpreting terms that are subjective, abstract, and not easy to interpret, which are unique to fashion. In this study, we propose dual Gaussian VSE (DGVSE), which embeds images and tags in the same space as a multidimensional Gaussian distribution. This method represents the embedding representation of each image and tag as a multidimensional Gaussian distribution, and the parameters of the embedded representation (mean vector and covariance matrix) estimated from the end-to-end model are used to enable multifaceted applications. We experimentally demonstrate the effectiveness of the proposed model using data accumulated in real services and presenting various functions that support the reduction of fashion-specific interpretation difficulties using the results obtained. The effectiveness of the proposed model is also demonstrated through a theoretical and empirical consideration of the several types of distances that can be included in the loss function.

The main contributions of this chapter are summarized as follows: 1) We propose a DGVSE model that can embed images and tags as probability distributions in the same space with a new end-to-end architecture. 2) Through analysis experiments using real data, we present various applications of DGVSE to support user fashion interpretation. 3) Theoretical and empirical considerations of several types of distances included in the loss function show the effectiveness of the proposed model.

6.2 Related Research: Gaussian Embedding

Gaussian embedding is a method that solves the problem of acquiring embedded representations by estimating the embedded representations as a (multidimensional) Gaussian distribution. Luke et al. proposed Word2Gauss [185], a model to solve the problem that the embedded representations obtained with the existing Word2Vec model [186, 187] failed to consider the spread of word meanings. For example, the words “Bach,” “composer,” and “man” should be ordered in order of breadth of meaning, “man ¿ composer ¿ Bach,” and this relationship is automatically captured. The strength of this method is that it can estimate the variance, or breadth of meaning of words, and by observing the estimated variance it is possible to obtain knowledge that cannot be obtained with Word2Vec. Research has developed Word2Vec into a method for embedding concepts (such as “animal”) and words (such as “dog” or “cat”) as probability distributions [188].

Many studies that extend network embedding to Gaussian embedding have also been published [189, 190, 191]. Other studies related to Gaussian embedding have developed to image recognition [192]. Face recognition is one of the major tasks in which Gaussian embedding shows great contributions [193, 194, 195]. Several studies aim to use this strength of Gaussian embedding to obtain the variance of an item in marketing areas, such as recommendation systems, and to gain important knowledge [196, 197, 198].

Some studies have extended the VSE model to Gaussian embedding. Ren et al. [182] proposed Gaussian visual-semantic embedding (GVSE). In this model, the embedded representation of a word is first learned by pre-training with GloVe [199] (focusing only on the text data attached to the image). Then, the embedding representation obtained from the pre-training is fixed as the mean vector of word embeddings, and all other parameters in the GVSE (including the covariance matrix) are estimated. In the loss function, the Mahalanobis distance is used to measure the distance between words (probability distributions) and images (points). However, the problems with this model are that it is not an end-to-end model and does not consider any image information when learning embedded representations for words. It uses the Mahalanobis distance as the distance measure in the loss function (see below). To use methods, such as GloVe, for pre-training, it is necessary to have a situation in which a large number of words are assigned to each data item, as is the case with text data. Although the data in this study are tagged, the number of tags per image is not as large as the number of words in text data, making it difficult to learn with these methods. Mukherjee and Hospedales [200] proposed a method similar to GVSE at very close to the time GVSE was proposed. That method also pre-trained the mean vector and covariance matrix for the word’s embedded representation by Word2Gauss with the dataset specific to text data such as Wikipedia corpus [201]. Therefore, same as GVSE, this method also encounters problems because it is not end-to-end learning and those related to the difference between tags and sentence (words) data.

In contrast, to solve all these problems, we propose DGVSE. This end-to-end model considers image information when learning embedded representations of words and uses a measure other than the Mahalanobis distance for the distance included in the loss function. The contributions of the study are its detailed analysis of variance, which has not been done in previous studies, and showing various applications of DGVSE.

6.3 Distance Functions Inducing Embedded Spaces

In this section, we consider the distance type to be adopted when mapping images and attributes to the embedding space in the proposed model.

Let $(\mathcal{X}, \mathcal{F}, \nu)$ be a measure space, where $\mathcal{X} \subseteq \mathbb{R}^d$ denotes the sample space, $d \in \mathbb{N}$ is the dimension of the sample space, $\mathcal{F}$ is the σ-algebra of measurable events, and ν is a positive probability measure. The set of the positive probability measure $\mathcal{P}$ is defined as

$$\mathcal{P} = \left\{ f(\boldsymbol{x}) \middle| f(\boldsymbol{x}) \geq 0 \ (\forall \boldsymbol{x} \in \mathcal{X}) \text{ and } \int_{\mathcal{X}} f(\boldsymbol{x}) d\nu(\boldsymbol{x}) = 1 \right\}. \tag{6.1}$$

In the following, we assume that $d\nu(\boldsymbol{x}) = d\nu = d\boldsymbol{x}$.

Definition 1 (Mahalanobis Distance). *Let $P \in \mathcal{P}$ be the probability distribution with mean vertical vector $\boldsymbol{\mu} \in \mathbb{R}^d$ and positive-definite covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$. The Mahalanobis distance $d_M : \mathcal{P} \times \mathcal{X} \rightarrow [0, \infty)$ between P and some point (vertical vector) $\boldsymbol{x} \in \mathcal{X}$ is defined as*

$$d_M(P, \boldsymbol{x}) := \sqrt{(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})}. \tag{6.2}$$

Proposition 1 (Closed form of Mahalanobis distance between Gaussian distributions). *Let $P, Q \in \mathcal{P}$ be two Gaussian distributions with mean vertical vectors $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1 \in \mathbb{R}^d$ and the same positive-definite covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$. The closed form of the Mahalanobis distance between P and Q is given as*

$$d_M(P, Q) = \sqrt{(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)}. \tag{6.3}$$

Many methods, including the conventional GVSE method, employ the Mahalanobis distance to gauge the distance between a point and a distribution. However, assuming that the covariance matrices of two points are identical is overly restrictive, especially when aiming for a diverse Gaussian distribution with varying variances across all points in the embedding space. Consequently, we explore the embedding space generated by alternative distance functions.

Definition 2 (Kullback − Leibler divergence). *The Kullback − Leibler divergence or KL divergence $D_{KL} : \mathcal{P} \times \mathcal{P} \rightarrow [0, \infty)$ is defined between two Radon − Nikodym densities p and q of*

ν-absolutely continuous probability measures by

$$D_{KL}[p \| q] := \int_{\mathcal{X}} p \ln \frac{p}{q} d\nu = \int_{\mathcal{X}} p(\boldsymbol{x}) \ln \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x}. \tag{6.4}$$

Because the KL divergence is asymmetric (i.e., $D_{KL}[p\|q] \neq D_{KL}[q\|p]$), the following symmetrization is often used to treat it as a distance function.

Definition 3 (Jeffreys divergence)**.** *The Jeffreys divergence $D_J : \mathcal{P} \times \mathcal{P} \to [0, \infty)$ is defined between two Radon-Nikodym densities p and q of ν-absolutely continuous probability measures by*

$$\begin{aligned} D_J[p\|q] &:= D_{KL}[p\|q] + D_{KL}[q\|p] \\ &= \int_{\mathcal{X}} p(\boldsymbol{x}) \ln \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x} + \int_{\mathcal{X}} q(\boldsymbol{x}) \ln \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} d\boldsymbol{x}. \end{aligned} \tag{6.5}$$

Proposition 2 (Closed form of Jeffreys divergence between Gaussian distributions)**.** *Let $P, Q \in \mathcal{P}$ be two Gaussian distributions with mean vertical vectors $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1 \in \mathbb{R}^d$ and positive-definite covariance matrix $\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1 \in \mathbb{R}^{d\times d}$. The closed form of the Jeffreys divergence between P and Q is given as*

$$\begin{aligned} D_J[P\|Q] &= \frac{1}{2}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top (\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) \\ &\quad + \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}_1 - 2\boldsymbol{I}_d\right), \end{aligned} \tag{6.6}$$

where $\boldsymbol{I}_d \in \mathbb{R}^{d\times d}$ is the d-dimensional identity matrix.

Proof. Because both P and Q are Gaussian distributions, the probability density functions are given by

$$\begin{aligned} p(\boldsymbol{x}) &= \frac{1}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}_0|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_0)\right\}, \\ q(\boldsymbol{x}) &= \frac{1}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}_1|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_1)\right\}. \end{aligned}$$

The closed form of $D_{KL}[P||Q]$ is given as

$$D_{KL}[P||Q] = \mathbb{E}_P[\ln p(\boldsymbol{x}) - \ln q(\boldsymbol{x})] = \frac{1}{2}\ln\frac{|\Sigma_1|}{|\Sigma_0|} - \frac{1}{2}\mathrm{tr}\,(\boldsymbol{I}_d)$$
$$+ \frac{1}{2}\left\{(\mu_0 - \mu_1)^\top \Sigma_1^{-1}(\mu_0 - \mu_1) + \mathrm{tr}\left(\Sigma_1^{-1}\Sigma_0\right)\right\}, \tag{6.7}$$

where $\mathbb{E}_P[\cdot]$ expresses the expected value for P. Similarly,

$$D_{KL}[Q||P] = \frac{1}{2}\ln\frac{|\Sigma_0|}{|\Sigma_1|} - \frac{1}{2}\mathrm{tr}\,(\boldsymbol{I}_d)$$
$$+ \frac{1}{2}\left\{(\mu_1 - \mu_0)^\top \Sigma_0^{-1}(\mu_1 - \mu_0) + \mathrm{tr}\left(\Sigma_0^{-1}\Sigma_1\right)\right\}. \tag{6.8}$$

From Eq. (6.7) and (6.8), we have

$$\begin{aligned} &D_{KL}[P||Q] + D_{KL}[Q||P] \\ &\quad = \frac{1}{2}(\mu_0 - \mu_1)^\top(\Sigma_0^{-1} + \Sigma_1^{-1})(\mu_0 - \mu_1) \\ &\qquad + \frac{1}{2}\mathrm{tr}(\Sigma_1^{-1}\Sigma_0 + \Sigma_0^{-1}\Sigma_1 - 2\boldsymbol{I}_d). \end{aligned}$$

□

Remark 1. *From Proposition 1 and Proposition 2, we can see that $d_M^2(P, Q) = 2D_J[P||Q]$ if P and Q are the Gaussian distributions with identical covariance matrix.*

Remark 2. *It is obvious that VSE with Mahalanobis distance induces variance-agnostic embedded space.*

Therefore, the symmetric KL divergence content where the variance-covariance matrices of the probability distributions being compared are perfectly matched coincides with the Mahalanobis distance. This means that estimating each parameter using the Mahalanobis distance risks not estimating the parameters well. Specifically, when measuring the distance between distribution P and distribution Q, the calculation is performed under the assumption that the variance of distribution P coincides with that of distribution Q. When measuring the distance between distribution P and distribution R, the calculation is performed under the assumption that the variance of distribution P coincides with that of distribution R. Thus, the parameters included in the model are estimated in a situation where the variance of distribution P changes in various ways (variance ignorance), depending on the pairs. In this situation, there is a risk that stable parameter estimation cannot be performed.

In machine learning methods, theoretical analysis suggests that care should be taken when adopting the Mahalanobis distance as a loss function (to measure the distance between a point and a probability distribution).

Definition 4 (Wasserstein Distance). *Given two probability distributions P and Q on two Polish spaces $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ and a positive lower semi-continuous cost function $c, \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$, optimal transport focuses on solving the following optimization problem:*

$$\inf_{\pi \in \Pi(P,Q)} \inf_{\mathcal{X} \times \mathcal{Y}} c(\boldsymbol{x}, \boldsymbol{y}) d\pi(\boldsymbol{x}, \boldsymbol{y}), \tag{6.9}$$

where $\Pi(P, Q)$ is the set of measures on $\mathcal{X} \times \mathcal{Y}$ with marginals P and Q. When $\mathcal{X}$ and $\mathcal{Y}$ are subspaces in $\mathbb{R}^d$ and $c(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x} - \boldsymbol{y}||^l$, where $||\boldsymbol{x}||^l$ is l-norm for vector $\boldsymbol{x}$ with $l \geq 1$, Eq. (6.9) *induces a distance over the set of measures with finite moment of order l, known as the l-Wasserstein distance W_l:*

$$W_l(P, Q) := \left(\inf_{\pi \in \Pi(P,Q)} \int_{\mathcal{X} \times \mathcal{Y}} ||\boldsymbol{x} - \boldsymbol{y}||^l d\pi(\boldsymbol{x}, \boldsymbol{y}) \right)^{\frac{1}{l}}, \tag{6.10}$$

or equivalently

$$W_l^l(P, Q) := \inf_{\boldsymbol{X} \sim P, \boldsymbol{Y} \sim Q} \mathbb{E}\left[||\boldsymbol{X} - \boldsymbol{Y}||^l \right]. \tag{6.11}$$

Proposition 3 (Closed form of the Wasserstein Distance Between Gaussian Distributions [202, 203]). *Let $P, Q \in \mathcal{P}$ be two Gaussian distributions with mean vertical vectors $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1 \in \mathbb{R}^d$ and positive-definite covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$, where $d \in \mathbb{N}$. The closed form of the* 2*-Wasserstein distance between P and Q is given as*

$$\begin{aligned} W_2^2(P, Q) := &(\mu_0 - \mu_1)^\top (\mu_0 - \mu_1) \\ &+ \operatorname{tr}\left(\Sigma_0 + \Sigma_1 - 2\left(\Sigma_0^{\frac{1}{2}} \Sigma_1 \Sigma_0^{\frac{1}{2}} \right) \right). \end{aligned} \tag{6.12}$$

Based on the above theoretical basis as well, the proposed method assumes a multidimensional Gaussian distribution for both images and tags, allowing the use of various measures such as KL divergence content and 2-Wasserstein distance.

6.4 Geometry of Embedded Space

In general, it is known that the space of probability distributions is not an Euclidean space, but rather constitutes a Riemannian manifold [204, 205]. This means that the space of probability distributions is non-Euclidean, which does not guarantee the validity of general vector operations. Therefore, even if the embedding vectors $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are obtained by DGVSE, they cannot be used directly for applications that combine multiple individual tags or embedded representations of images such as tags embedding and image retrieval. Thus, it is necessary to apply appropriate transformations to the embedded vectors to obtain new parameter vectors to allow operations such as those in Euclidean space. Geometrically, these transformations are called coordinate transformations, and parameter coordinates that allow linear operations are called affine coordinate systems.

Consider an exponential family, the generalization of the Gaussian distributions, expressed in the following form:

$$p(\boldsymbol{x};\boldsymbol{\theta}) \coloneqq \exp\left\{\boldsymbol{\theta}^\top \boldsymbol{z} + k(\boldsymbol{x}) - \psi(\boldsymbol{\theta})\right\}, \tag{6.13}$$

where $\boldsymbol{x}$ is a random variable, $\boldsymbol{\theta} = (\theta^{(1)}, \ldots, \theta^{(n)})^\top$ is an n-dimensional vector parameter, $h_i(\boldsymbol{x})$ are n functions of $\boldsymbol{x}$, which are linearly independent, $k(\boldsymbol{x})$ is a function of $\boldsymbol{x}$, and ψ corresponds to the normalization factor. Here, let $\boldsymbol{z} = (z_1, \ldots, z_i, \ldots, z_n)^\top = (h_1(\boldsymbol{x}), \ldots, h_i(\boldsymbol{x}), \ldots, h_n(\boldsymbol{x}))^\top$ be a new vector random variable and $d\nu(\boldsymbol{z})$ be a measure in the sample space $\mathcal{Z} \subseteq \mathbb{R}^{\backslash}$ defined as

$$d\nu(\boldsymbol{z}) \coloneqq \exp\{k(\boldsymbol{x})\}d\boldsymbol{x}. \tag{6.14}$$

Then, Eq. (6.13) is rewritten as

$$p(\boldsymbol{x};\boldsymbol{\theta})d\boldsymbol{x} = \exp\left\{\boldsymbol{\theta}^\top \boldsymbol{z} - \psi(\boldsymbol{\theta})\right\} d\nu(\boldsymbol{z}), \tag{6.15}$$

$$p(\boldsymbol{z};\boldsymbol{\theta}) = \exp\left\{\boldsymbol{\theta}^\top \boldsymbol{z} - \psi(\boldsymbol{\theta})\right\}. \tag{6.16}$$

The family of distributions $\mathcal{M} = \{p(\boldsymbol{z};\boldsymbol{\theta})\}$ forms a J-dimensional manifold, where $\boldsymbol{\theta}$ is a coordinate system. Because $\psi(\boldsymbol{\theta})$ is a normalization factor, we have

$$\int_{\mathcal{Z}} p(\boldsymbol{z};\boldsymbol{\theta})d\nu(\boldsymbol{z}) = 1, \tag{6.17}$$

$$\psi(\boldsymbol{\theta}) = \log \int_{\mathcal{Z}} \exp(\boldsymbol{\theta}^\top z) d\nu(z). \tag{6.18}$$

Here, a dually flat Riemannian structure is introduced in $\mathcal{M}$ using $\psi(\boldsymbol{\theta})$. The affine coordinate system is $\boldsymbol{\theta}$, which is called the natural parameter, and the dual affine parameter is given by the Legendre transformation $\boldsymbol{\theta}^* = \nabla\psi(\boldsymbol{\theta})$, which is the expectation of z denoted by $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^\top$ as

$$\boldsymbol{\eta} := \boldsymbol{\theta}^* = \mathbb{E}[z] = \int_{Z} z p(z; \boldsymbol{\theta}) d\nu(z). \tag{6.19}$$

Here, $\boldsymbol{\eta}$ is called the expectation parameter. Hence, $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are two affine coordinate systems connected by the Legendre transformation.

Example 1 (Univariate Gaussian distribution)**.** *The probability density function of the Gaussian distribution with mean μ and variance σ^2 is given as*

$$p(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}. \tag{6.20}$$

Let $\boldsymbol{\xi} = (\xi^{(1)}, \xi^{(2)})^\top$ where

$$\xi^{(1)} = h_1(x) = x,$$
$$\xi^{(2)} = h_2(x) = x^2.$$

Here, we can see that $\xi^{(1)}$ and $\xi^{(2)}$ are dependent, but are linearly independent. We further introduce new parameters $\boldsymbol{\theta} = (\theta^{(1)}, \theta^{(2)})^\top$ as

$$\theta^{(1)} = \frac{\mu}{\sigma^2},$$
$$\theta^{(2)} = -\frac{1}{2\sigma^2}.$$

Then, Eq (6.20) *is written in the standard form as*

$$p(\boldsymbol{\xi}; \boldsymbol{\theta}) = \exp\{\boldsymbol{\theta}^\top \boldsymbol{\xi} - \psi(\boldsymbol{\theta})\}. \tag{6.21}$$

The convex function $\psi(\boldsymbol{\theta})$ is given by

$$\begin{aligned}\psi(\boldsymbol{\theta}) &= \frac{\mu^2}{2\sigma^2} + \log\left(\sqrt{2\pi}\sigma\right)\\ &= -\frac{(\theta^{(1)})^2}{4\theta^{(2)}} - \frac{1}{2}\log(-\theta^{(2)}) + \frac{1}{2}\log\pi.\end{aligned}$$

Finally, the dual affine coordinates $\boldsymbol{\eta}$ are given as

$$\eta_1 = \mu, \quad \eta_2 = \mu^2 + \sigma^2. \tag{6.22}$$

6.5 Methodology

The DGVSE model assumes a multidimensional Gaussian distribution behind embedded representations of full-body outfit images and attributes and allows embedding each of them (as a probability distribution) in the same space while considering the spread of meaning. The points to identify the model from the conventional GVSE are: 1) it is an end-to-end model, and 2) it considers image information when learning embedded representations of words. In addition, based on the theoretical basis in section 6.3, 3) it can estimate not only words but also embedded representations of images as probability distributions, and 4) the distance included in the loss function is not Mahalanobis distance. Moreover, as mentioned in section 6.4, 5) natural parameters are introduced when combining multiple individual tag distributions.

6.5.1 Model Architecture

The model structure is shown in Figure 6.1.

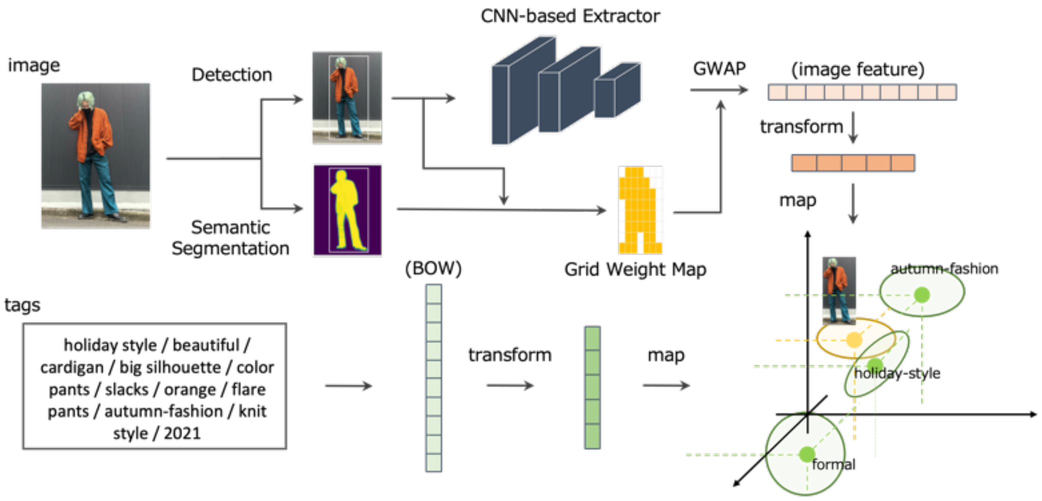


Figure 6.1: Structure of a prototype of our dual Gaussian visual-semantic embedding model proposal

The basic structure is based on the VSE in the fashion intelligence system [8]. We then assume a multidimensional Gaussian distribution behind the embedded representations of both images and tags. This allows us to estimate the mean and variance (semantic spread) for the embedded representation of each image and tag. This model solves the following problems of the conventional GVSE model: 1) it is not an end-to-end learning method, 2) it only looks at the co-occurrence of words and ignores image information when learning the embedded representation (mean) of words, and 3) it assumes a multidimensional Gaussian distribution only for words (hence the problem of measuring the distance between a point and the probability distribution using the Mahalanobis distance). In particular, the third problem is theoretically presented in the aforementioned section 6.3.

6.5.2 Parameter Optimization

The dataset for this study consists of a single full-body outfit image to which multiple tags are assigned. First, embedding (foreground-centered learning) based on CNN and grid weight map is performed on an arbitrary image I to obtain an image embedded representation of foreground $\mathbf{x} \sim \mathcal{N}(\mu_I, \mathbf{\Sigma}_I)$. Here, $\mathbf{x}, \mu_I \in \mathbb{R}^d$, $\mathbf{\Sigma}_I \in \mathbb{R}^{d \times d}$, where d is the dimension of the embedded space, μ_I is mean vertical vector, and $\mathbf{\Sigma}_I$ is the assumed spherical covariance matrix for the image I. Furthermore, the tags set T assigned to an arbitrary image I is embedded to obtain the tags embedded representation $\mathbf{v} \sim \mathcal{N}(\mu_T, \mathbf{\Sigma}_T)$. Here, $\mathbf{v}, \mu_T \in \mathbb{R}^d$, $\mathbf{\Sigma}_T \in \mathbb{R}^{d \times d}$, where μ_T is mean vertical vector and $\mathbf{\Sigma}_T$ is the assumed spherical covariance matrix for the tags T.

In computing this tag embedding representation $\mathbf{v}$, the probability distributions of embedding representations for the individual tags in arbitrary tags T are combined using the aforementioned method that introduces natural parameters. We now define a tag set $T = \{t_1, \cdots, t_n, \cdots, t_{N_T}\}$ consisting of all N_T individual tags. Then, the natural parameters $\Theta_n := (\theta_n^{(1)}, \theta_n^{(2)})^\top \in \mathbb{R}^d + \mathbb{R}^{d \times d}$ for the individual tag embedding representation $\mathbf{a}_{t_n} \sim \mathcal{N}(\mu_{t_n}, \mathbf{\Sigma}_{t_n})$ are calculated by the following Eq. (6.23)-(6.24):

$$\theta_n^{(1)} = \mathbf{\Sigma}_{t_n}^{-1} \mu_{t_n}, \tag{6.23}$$

$$\theta_n^{(2)} = -\frac{1}{2} \mathbf{\Sigma}_{t_n}^{-1}. \tag{6.24}$$

Here, $\mathbf{a}_{t_n}, \mu_{t_n} \in \mathbb{R}^d$, $\mathbf{\Sigma}_{t_n} \in \mathbb{R}^{d \times d}$, where μ_{t_n} is a mean vertical vector and $\mathbf{\Sigma}_{t_n}$ is assumed a spherical covariance matrix for the individual tag t_n. Furthermore, natural parameters are calculated for all individual tags in T, and their centroid Θ_T is calculated. Using the calculated centroid $\Theta_T := (\theta_T^{(1)}, \theta_T^{(2)})^\top \in \mathbb{R}^d + \mathbb{R}^{d \times d}$, the following Eq. (6.25)-(6.26) gives the parameters included in Gaussian distribution $\mathcal{N}(\mu_T, \mathbf{\Sigma}_T)$ for the embedded representation of the tag set T:

$$\mu_T = -\frac{1}{2} \theta_T^{(2)^{-1}} \theta_T^{(1)}, \tag{6.25}$$

$$\mathbf{\Sigma}_T = -\frac{1}{2} \theta_T^{(2)^{-1}}. \tag{6.26}$$

This method of computation, which introduces natural parameters, allows the distribution to be synthesized while accurately accounting for the variance of the elements (individual tags) that make up the set (tags).

The basic policy of learning DGVSE is to learn so that the probability distribution of embedded representations in the image and the probability distribution of embedded representations in terms of tags attached to the image are close. Therefore, the parameter estimation is achieved by optimizing the following contrastive loss Eq. (6.27) [8, 74, 86]:

$$\begin{aligned} \mathcal{L}(\mathrm{O}) = &\sum \max\left(0, m + d(\mathbf{x}^+, \mathbf{v}^+) - d(\mathbf{x}^+, \mathbf{v}^-)\right) \\ &+ \sum \max\left(0, m + d(\mathbf{v}^+, \mathbf{x}^+) - d(\mathbf{v}^-, \mathbf{x}^+)\right), \end{aligned} \tag{6.27}$$

where $\mathrm{O} = \{V, \mathbf{W}_I, \mathbf{W}_T\}$ is a set of target parameters to be optimized, V is a parameter set contained in CNN, $\mathbf{W}_I \in \mathbb{R}^{d \times r}$ is a transform matrix from an image feature vector obtained from a CNN-based extractor to the image embedded representation (r is the number of dimensions of the final convolutional layer of the CNN), $\mathbf{W}_T \in \mathbb{R}^{d \times s}$ is a transform matrix from a bag-of-words representation for tags T to a tags embedded representation (s is the number of tags in the entire dataset), m is a margin, $d(\mathbf{x}, \mathbf{y})$ indicates the distance between vectors $\mathbf{x}$ and $\mathbf{y}$, and $\beta > 0$ is a positive hyperparameter to adjust the importance of the background regularization term. The superscript sign + of A^+ indicates that A is a variable related to the positive sample, and $-$ of A^- indicates that A is a variable related to the negative sample.

Because we assume a multidimensional Gaussian distribution behind both the image and the embedded representation of the tag, the distance measure in the embedding space must be able to consider the mean and covariance matrix. Therefore, the following three distance measures are adopted in this study. While the Mahalanobis distance $d_{\mathrm{M}}(\mathbf{x}, \mathbf{y})$ is also adopted for comparison, the KL divergence $d_{\mathrm{KL}}(\mathbf{x}, \mathbf{y})$ and 2-Wasserstein distance $d_{\mathrm{W}_2^2}(\mathbf{x}, \mathbf{y})$ are adopted based on theoretical grounds to measure the distance between vectors $\mathbf{x}$ and $\mathbf{y}$:

$$\begin{aligned}
d_{\mathrm{M}}(\mathbf{x}, \mathbf{v}) &= \sqrt{(\mu_I - \mu_T)^\top \Sigma^{-1} (\mu_I - \mu_T)}, && (6.28)\\
d_{\mathrm{KL}}(\mathbf{x}, \mathbf{v}) &:= D_{\mathrm{KL}}(\mathbf{x}, \mathbf{v}) = \frac{1}{2} \ln \frac{|\Sigma_I|}{|\Sigma_T|} - \frac{1}{2} \mathrm{tr}\,(\boldsymbol{I}_d) \\
&+ \frac{1}{2} \left\{ (\mu_T - \mu_I)^\top \Sigma_I^{-1} (\mu_T - \mu_I) + \mathrm{tr}\left(\Sigma_I^{-1} \Sigma_T \right) \right\}, && (6.29)\\
d_{\mathrm{W}_2^2}(\mathbf{x}, \mathbf{v}) &:= W_2^2(\mathbf{x}, \mathbf{v}) = (\mu_I - \mu_T)^\top (\mu_I - \mu_T) \\
&+ \mathrm{tr}\left(\Sigma_I + \Sigma_T - 2 \left(\Sigma_I^{\frac{1}{2}} \Sigma_T \Sigma_I^{\frac{1}{2}} \right)^{\frac{1}{2}} \right), && (6.30)
\end{aligned}$$

where the joint covariance matrix Σ in Eq. (6.28) is defined as $\frac{1}{2}(\Sigma_I + \Sigma_T)$ for convenience in this study.

The assumption of probability distributions on both sides is expected to solve the problem of dispersion ignorance that occurs when the Mahalanobis distance is used to measure the distance between a point and a distribution, which is adopted in GVSE. In subsequent sections of the experiment, we will observe and discuss how the results change with each distance measure. One of the main contributions of this study is the detailed theoretical analysis of the differences between the three typical distance scales adopted, based not only on the observed experimental results but also on the results.

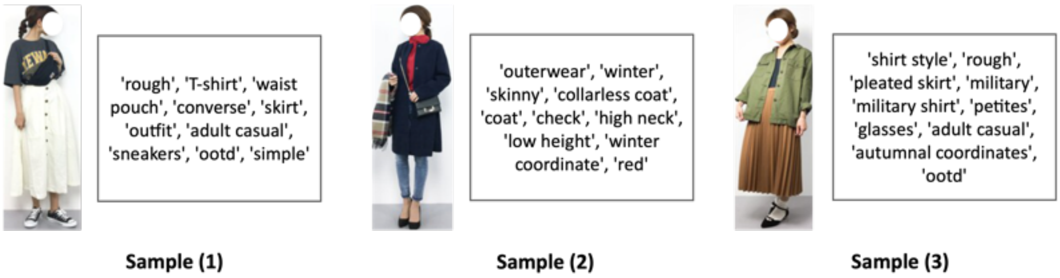
DGVSE is trained through the above process of parameter optimization. By considering the probability distribution of embedded representations of the resulting images and tags, DGVSE not only retains the useful functions of the traditional fashion intelligence system, such as search and sorting, but also enables the interpretation of abstract fashion terms related to dispersion, which is not possible with the previous methods.

6.6 Experimental Analysis

To evaluate the effectiveness of the proposed DGVSE model, we applied it to actual posted full-body outfit image data and the tags information attached to each image accumulated in WEAR [5], a fashion coordination application including social media features.

6.6.1 Experimental Settings

The number of full-body outfit images in the experimental data was 15,740, and the number of unique tags attached to all images was 1,104. All models in the target images were female, and the backgrounds of all images contained relatively little noise. An example of a full-body outfit image and its tags are shown in Figure 6.2 below.


'rough', 'T-shirt', 'waist pouch', 'converse', 'skirt', 'outfit', 'adult casual', 'sneakers', 'ootd', 'simple'
'outerwear', 'winter', 'skinny', 'collarless coat', 'coat', 'check', 'high neck', 'low height', 'winter coordinate', 'red'
'shirt style', 'rough', 'pleated skirt', 'military', 'military shirt', 'petites', 'glasses', 'adult casual', 'autumnal coordinates', 'ootd'

Sample (1) Sample (2) Sample (3)

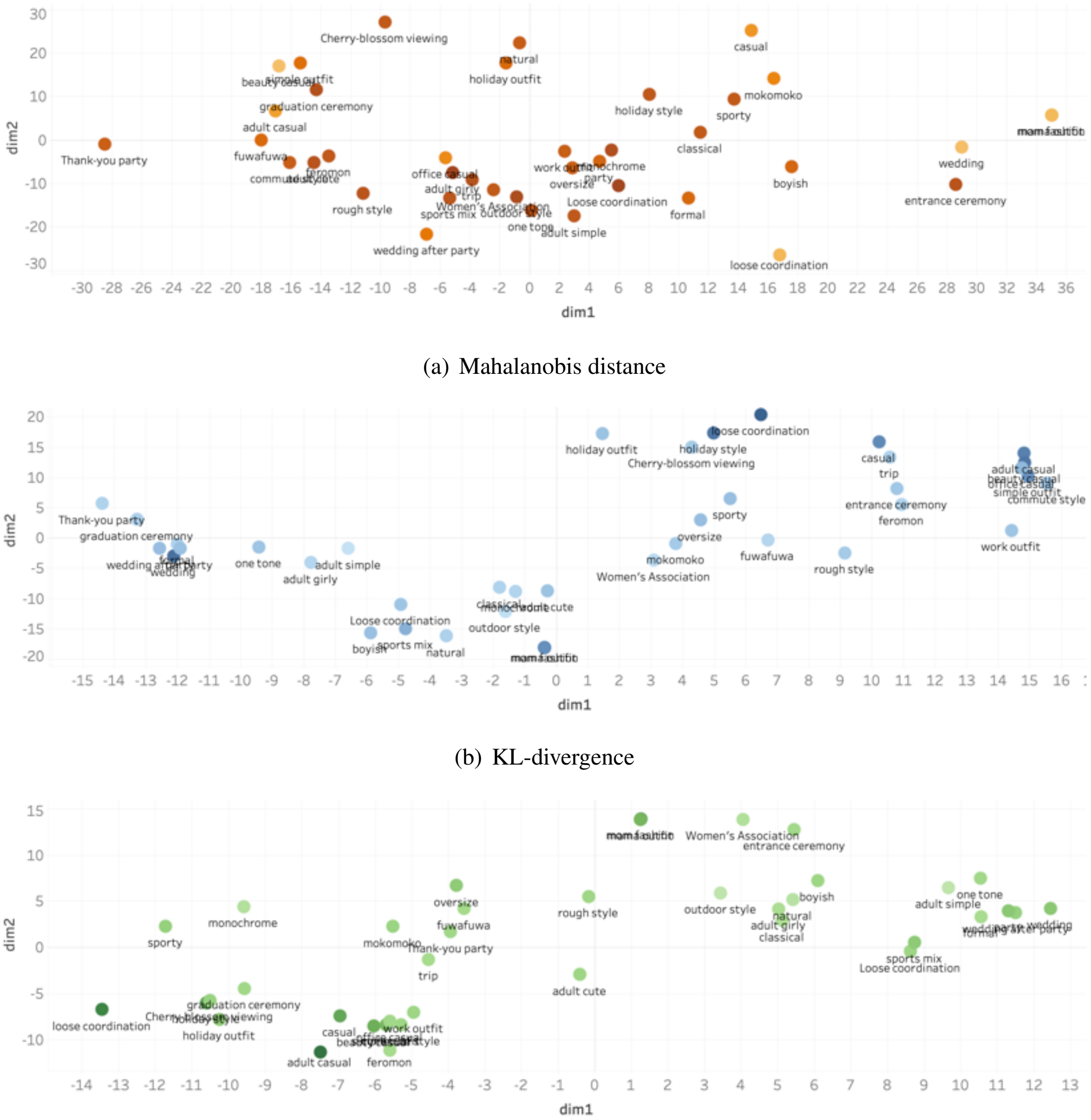
Figure 6.2: Example of samples in the target dataset

The embedded representation dimension included in the VSE model d was set to 64. The learning rate was 0.001, and the number of epochs was 50. The batch size was 32, and the margin m was set to 0.2. We used GoogleNet and Inception V3 pre-trained on ImageNet [139] as the extractor to assess the impact of CNN. For preprocessing, SSD (extractor: MobileNet

V2 [140]) trained on Open Images [206]) was used for object detection, and FCN (extractor: ResNet [168]) trained on MS-COCO [141] was used for semantic segmentation.

6.6.2 Attribute Mapping

The average values of embedded tags obtained from the proposed model were compressed by t-SNE [108] and shown in a two-dimensional map in Figure 6.3 below. However, because it is impossible to see the results of mapping all tags because of the scope of the figure, some representative abstract attributes are extracted and mapped.



(a) Mahalanobis distance

(b) KL-divergence

(c) 2-Wasserstein distance

Figure 6.3: Mapping the result of compressing the average of tag embedded representations

Observing these figures makes it possible to grasp each expression's semantic relationship, taking the image information into consideration. Because the estimation results by the three distance measures are shown, it is possible to consider the validity and interrelationships of each distance. First, in the results obtained from KL divergence and 2-Wasserstein distance models, for example, the pairs of "wedding" and "wedding after-party," "mom outfit" and "mom fashion," "commute style" and "work outfit," "holiday outfit" and "holiday style," and "wedding" and "wedding after party" that have similar meaning to each other are in close proximity. Conversely, for the Mahalanobis distance model, these pairs are in many cases not near each other, so the validity of the results is questionable. These results suggest that adopting the Mahalanobis distance may be risky for the model and the problem under study.

For a more detailed interpretation, Figure 6.4 below shows an enlarged map (KL divergence) of the area around many of the tags associated with “casual.”

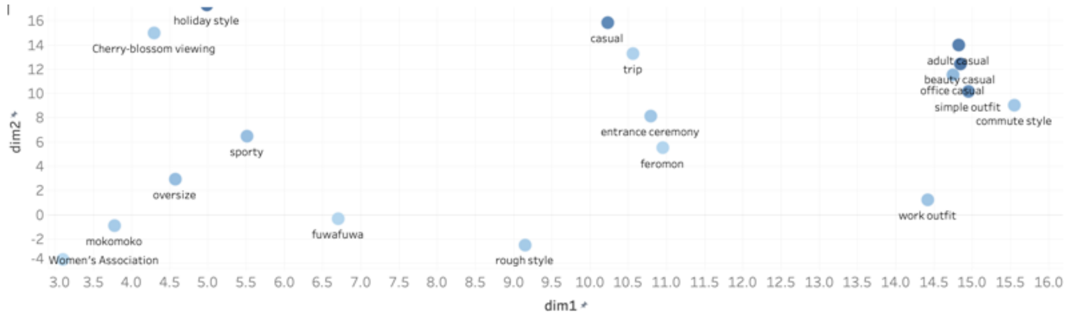


Figure 6.4: Enlarged map of the area around many of the tags associated with “casual” (KL-divergence)

Using this figure, it is possible to accurately understand the relationship between these abstract expressions, which have been used subjectively in the past when conversing about fashion. For example, the fact that “office casual” is more similar to “beauty casual” than to “adult casual” may have been an ambiguous fact for experts and non-specialists. It is also a new finding by quantitatively expressing each tag that “office casual” is closer to “simple outfit,” “commute style,” and “work outfit” than to “adult casual.” Furthermore, “office casual,” “adult casual,” and “beautiful casual” are quite similar in expression, and we can see that making them more “casual” brings them closer to “trip” and “enrollment ceremony” (suitable attire). Clearly, “fuwafuwa” and “mokomoko” are similar and onomatopoeic words used to describe near meanings such as “oversize” clothing and “rough style.”

The use of this map-based interpretation support method will reduce the difficulties in understanding fashion-specific ambiguous expressions. All users are expected to be able to talk and explain fashion using words specific to the fashion field and make decisions based on a common understanding. Furthermore, it is important to note that this map is not based only on word (tag) co-occurrence relations, as in Word2Gauss and GVSE, but is obtained by considering image information.

6.6.3 Interpretation of Variance

The proposed model can acquire both the mean and the variance for embedded representations of images and tags. Although the conventional GVSE model can also acquire only variance for words, the embedded representation for the acquired words is not observed and considered. In contrast, the variance obtained is also thoroughly observed and considered in this study.

6.6.3.1 Variance for Attributes

First, Table 6.1 below shows the top eight and bottom eight tags with the largest variance, their variance values, and the number of images (count) to which they are attached in the whole images.

Table 6.1: Summary of tags with large and small variance of embedded representation

(a) Mahalanobis distance

rank	tag	variance	count	rank	tag	variance	count
1	Ease-up Outfits	1.0369	147	1095	Petites	0.9144	2618
2	Hair Bands	1.0331	235	1096	Mom's Outfits	0.9137	811
3	Fall Colors	1.0280	86	1097	Wedding	0.9088	697
4	Gaucho Pants	1.0258	74	1098	Beautiful Casual	0.9085	1717
5	Big Silhouette	1.0254	194	1099	Spring Outfits	0.9074	3111
6	Bun Hairstyles	1.0222	37	1100	Denim	0.9069	3116
7	Floral Blouse	1.0221	14	1101	Knitwear	0.9062	2792
8	Drawstring Bags	1.0215	84	1102	Black	0.9036	2344

(b) KL-divergence

rank	tag	variance	count	rank	tag	variance	count
1	Petites	1.4426	2618	1095	Normcore	0.8587	8
2	Knitwear	1.3853	2792	1096	Black Lace	0.8587	23
3	Adult Women	1.3736	1624	1097	tweedmill	0.8578	5
4	Fall Outfits	1.3735	2291	1098	Beige Pants	0.8569	81
5	Otona Casual	1.3686	2759	1099	Dandy Glasses	0.8569	95
6	Little Recommend	1.3565	673	1100	Character T-shirt	0.8554	26
7	Black	1.3562	2344	1101	Red Converse	0.8549	39
8	Ballet Shoes	1.3495	1829	1102	Red Socks	0.8533	41

(c) 2-Wasserstein distance

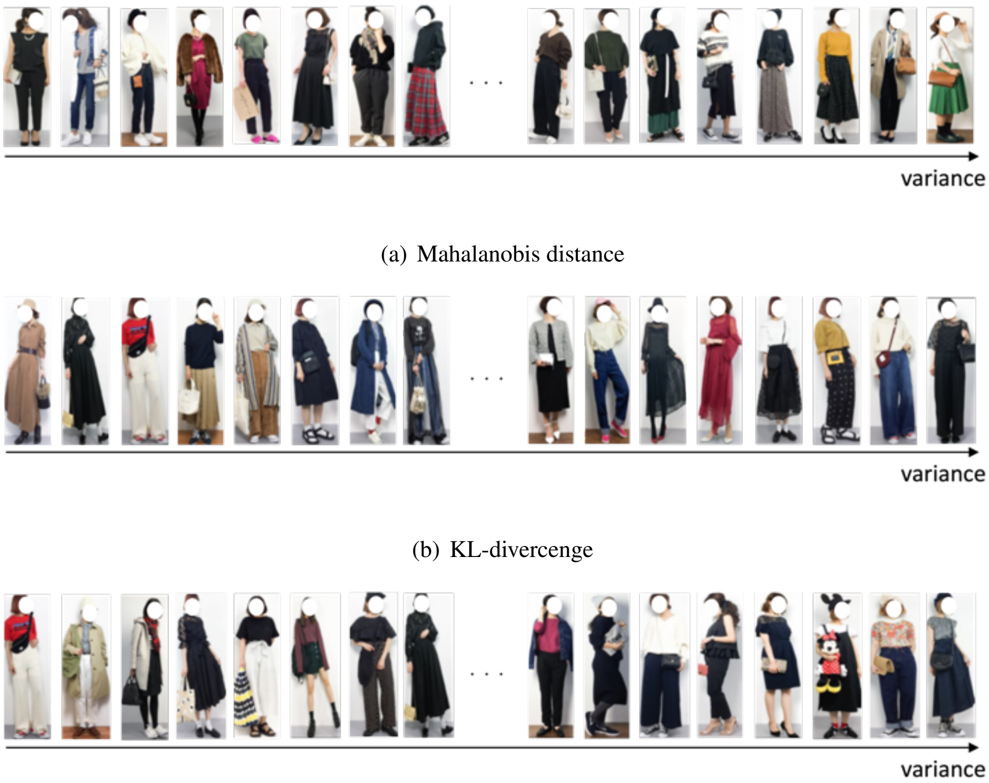
rank	tag	variance	count	rank	tag	variance	count
1	Petites	1.4597	2618	1095	Danton	0.8490	20
2	Black	1.3552	2344	1096	Black Skirt	0.8483	85
3	Knitwear	1.3466	2792	1097	Cable-Knit	0.8428	6
4	Adult Women	1.3442	1624	1098	Gaucho Pants	0.8426	74
5	Fall Outfits	1.3400	2291	1099	Beige Pants	0.8396	81
6	Adult Casual	1.3375	5680	1100	Voluminous Skirts	0.8393	22
7	Ballet Shoes	1.3157	1829	1101	ZOZO Summer Sale	0.8370	36
8	Otona Casual	1.3140	2759	1102	Normcore	0.8347	8

First, a comparison of the tables obtained by the three measures shows that many of the tags judged to have particularly high variance are overlapped in DGVSE models based on KL divergence and 2-Wasserstein distance. Conversely, the Mahalanobis distance model results differ significantly from those of the other two models. Combined with the results for the mean mentioned in the previous section, this suggests that KL divergence and 2-Wasserstein distance are similar measures and that Mahalanobis distance may be very different compared to the other two measures.

The results of KL divergence and 2-Wasserstein distance show that the variance tends to be larger for tags with a larger frequency of occurrence (count) and smaller for tags with a smaller frequency of occurrence (count). The top tags mostly contain abstract attributes, colors, and highly versatile items, such as "denim," while the bottom tags mostly contain specific items not included in many outfits. Naturally, the variance increases for highly versatile items and colors (specific tags), because they are included in various outfits (i.e., assigned to many images). Therefore, a certain degree of correlation with the number of times a tag is assigned may be a basis for increasing the validity of the results. For example, the tag "petites" is not assigned to fashion items or atmosphere but to images of models with a small height. Therefore, the variance of tags, such as "knitwear," "denim," "adult casual," and "otona casual," that are directly related to fashion are smaller than that of tags, despite these tags appearing more frequently than "petites." This suggests the validity of the obtained variance.

6.6.3.2 Variance for Images

We discuss the variance of the images obtained from DGVSE. It was difficult to understand the relationship between images and dispersion only by observing them. Therefore, we compared the variance estimated by DGVSE with each statistic related to the tags assigned to the images. The results are shown in the following Figure 6.6.



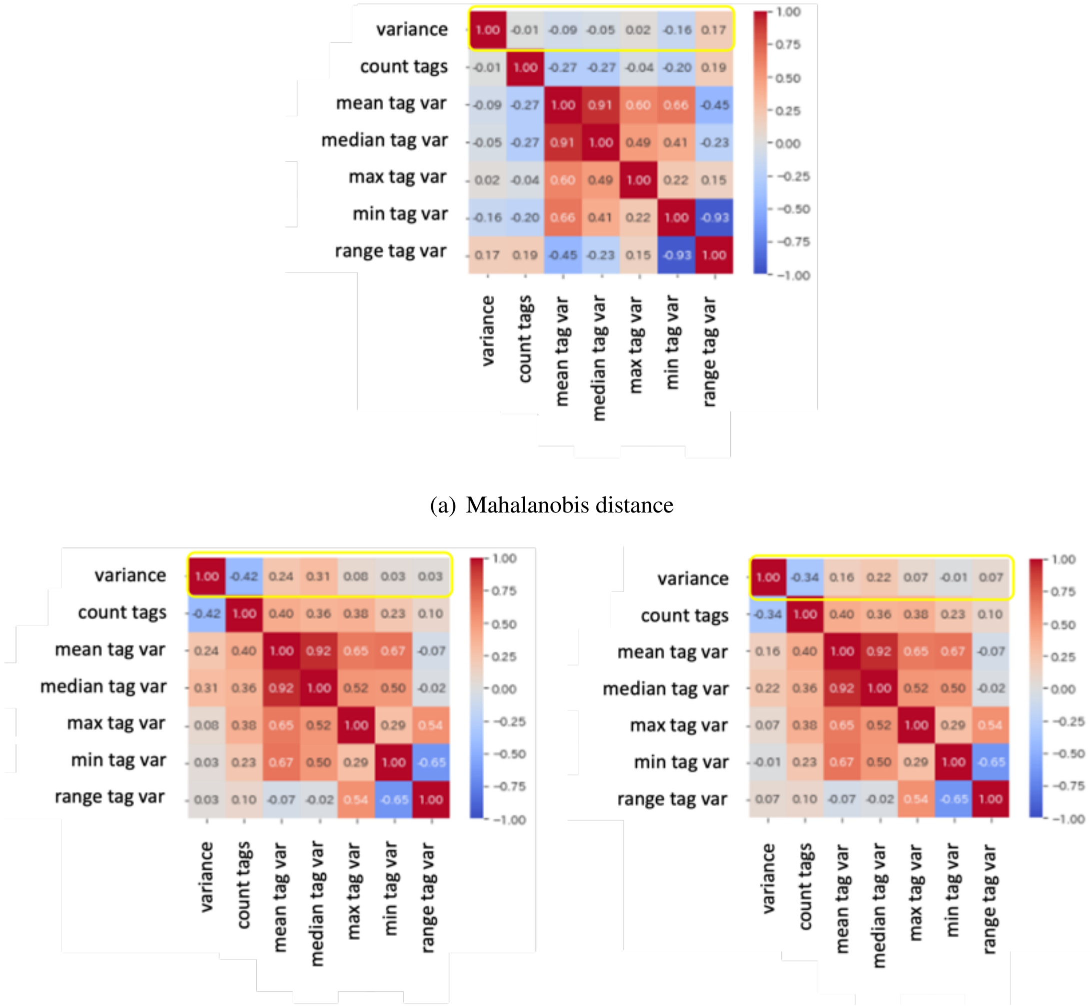
(a) Mahalanobis distance

(b) KL-divercenge

(c) 2-Wasserstein distance

Figure 6.5: Images sorted by the variance of image embedded representation

Observing this figure makes it possible to see which images are more semantically broad and narrower. However, it was difficult to understand the relationship between images and dispersion just by observing them. Therefore, we compared the variance estimated by DGVSE with each statistic related to the tags assigned to the images. The results are shown in the following Figur 6.6.



(a) Mahalanobis distance

(b) KL-divergence

(c) 2-Wasserstein distance

Figure 6.6: Correlation coefficient matrix with each statistic associated with the tags assigned to the image

Results show that the variance of images obtained from the model adopting KL divergence and 2-Wasserstein distance is negatively correlated with the number of tags attached to the image, and no significant correlation was observed with other statistics. Conversely, the model adopting Mahalanobis distance showed little correlation with any of the statistics. This result can be attributed to the fact that when training the model, the placement of images with many tags in the embedding space is defined in detail by the tags. Conversely, images with few tags are ambiguous in terms of the placement specified by the tags, suggesting that they may require to be positioned well in relation to other images and tags. Some tags that should potentially be attached are often not assigned because the general user assigns tags. The results on image variance obtained from DGVSE are expected to be useful in understanding whether the tags assigned are insufficient to describe the target fashion image.

These results suggest that in these situations, where the types and number of tags assigned to each image are diverse, the variance estimated for an image tends to be determined by the amount of information that can be obtained from the tags information for the model (ambiguity of the tags as a whole). In this study, we used the accumulated data as is, but it may be possible to learn a better embedding space by implementing tag completion and other measures. In such a case, the proposed method is expected to focus on images with large variance and tag completion to reduce the variance and ultimately contribute to acquiring a delicate and accurate embedding space.

6.6.4 Image Retrieval

DGVSE allows for estimating the variance of the embedded representation of each image or tag and maintains the original application methods of VSE in [8].

Figure 6.7 below shows an example of image retrieval results based on tag and image operations. We present results using a model adopting the 2-Wasserstein distance (out of the KL divergence model and 2-Wasserstein distance model), which can learn a reasonable embedding space based on the multifaceted analysis described above.

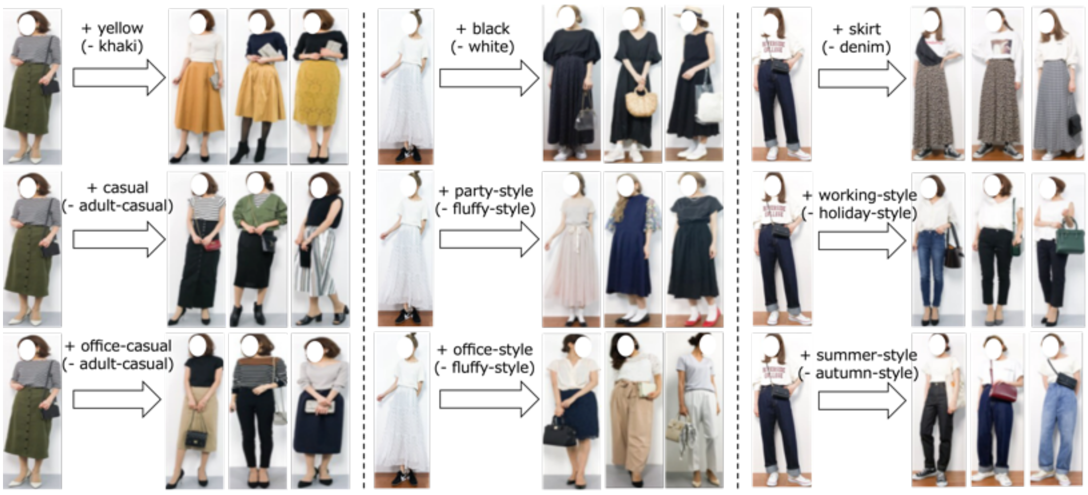


Figure 6.7: Example of image retrieval (2-Wasserstein distance)

First, the above three examples of using the image retrieval function with specific tags are shown. For example, if we remove the “khaki” tag from a query image (wearing a striped shirt and a khaki-colored skirt) that has been assigned the “khaki” tag and the “adult casual” tag and instead assign the “yellow” tag to the image with a yellow skirt, the image with a yellow skirt is retrieved, and this full-body outfit image stays in the adult casual category. In addition, the example in the two lines below shows the search results using abstract tags. For example, if the “working-style” tag is added instead of “holiday-style” to the image (on the right) with the tags “denim,” “holiday-style,” and “autumn-style,” it can search for images that are appropriate for the office while retaining the overall color scheme. Similarly, if the “summer-style” tag is added instead of “autumn-style,” the whole color tone is kept, and outfits with short sleeves or thin-looking fabrics are retrieved.

Thus, DGVSE enables image retrieval utilizing tag and image operations, including fashion-specific abstract tags.

6.6.5 Image Re-ordering

Another function maintained from the original VSE is image re-ordering, which retrieves images that are more (less) relevant to the target tag by reordering images in the order of the strength of the relevance score calculated between the target tag and each image to which the tag is attached. An example of image re-ordering results using DGVSE is shown in Figure 6.8 below.

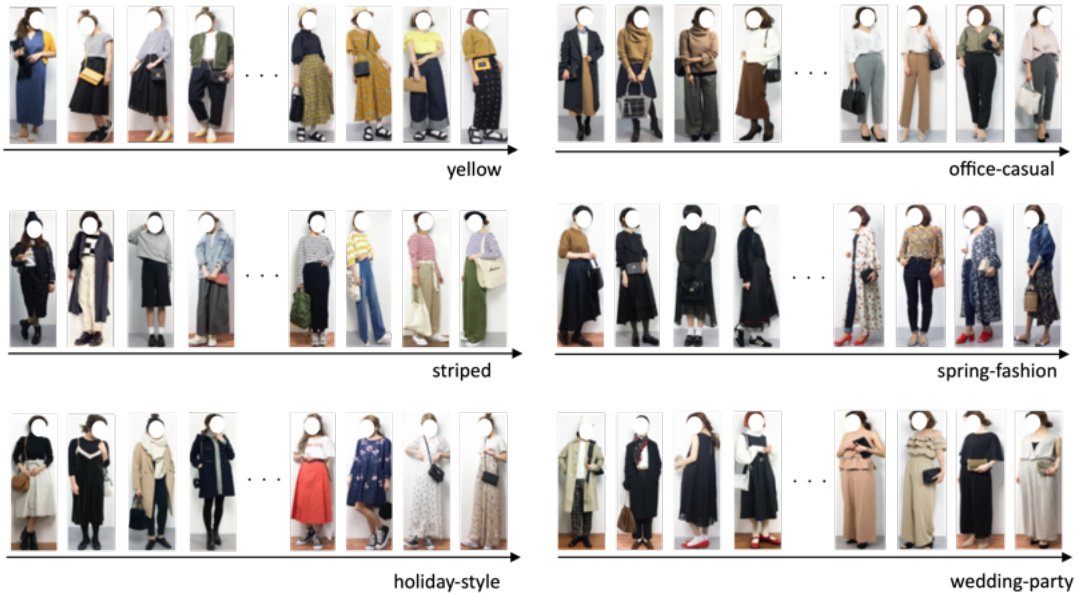


Figure 6.8: Example of image re-ordering (2-Wasserstein distance)

For example, outfits with a high relevance score to the specific “yellow” tag tend to have a high percentage of yellow areas in the entire image. Conversely, an outfit with a low relevance score will have fewer yellow areas. Other outfits with high relevance scores to the specific “striped” tag tend to have a high percentage of the border portion of the item in the whole image. Previously, it was only possible to display images with the “yellow” or “striped” tag in a batch; however, this application allows users to search for clothes that meet their detailed objectives, such as “I want to incorporate yellow in only one item” or “I want to find clothes that contain a border pattern throughout the entire outfit.” It is now possible to search for clothes in detail according to the detailed objectives of the user.

The results of image reordering using abstract expressions, such as “holiday style,” “office casual,” “spring fashion,” and “wedding,” and the relevance scores of tags related to usage scenes and seasons also have the potential to be used very effectively in actual services. For example, the results of “holiday style, images with strong relevance to “holiday style,” are generally warm in color and often contain items with patterns, soft silhouette clothes, and sneakers. Conversely, images with low relevance scores include items in specific colors, such as black and beige, and slightly more formal (office casual) items, such as tights and heels. The images with high relevance to “spring style” contained brightly colored and patterned items, while those with low relevance contained a lot of blacks. Using these results, users can ask themselves questions such as “what is a good outfit for the holidays?” and “what is appropriate for spring?” as well as “what is the best outfit for a holiday?”

Thus, DGVSE can reorder images utilizing tag and image operations, including fashion-specific abstract tags. These functions are expected to be effective in quantifying fashion-specific abstract terms and images and support users’ interpretation of fashion.

While the functions of image retrieval and image sorting are maintained from the conventional VSE the proposed method has a wider range of applications than the conventional methods because it also allows interpretation using variance as mentioned above. This is the strength of DGVSE in this application, and it is expected to be more widely effective in supporting user understanding of fashion in real applications.

6.7 Discussion

6.7.1 On the Loss Function

In section 6.3, we theoretically showed the risk that the Mahalanobis distance between points (images) and probability distributions (attributes) adopted in GVSE cannot implement stable learning. Moreover, many measures can be introduced by assuming variance even for images and by making it possible to measure the distance between variance against variance. In this study, we adopted the Mahalanobis distance, KL divergence, and 2-Wasserstein distance and confirmed the validity and appropriateness of each measure through multifaceted experiments. Based on these results, it is necessary to further investigate which distance should be adopted.

6.7.1.1 On the Risks of the Mahalanobis Distance

The results of the mapping experiment using the mean values of the embedded representations obtained showed that 1) KL divergence and the 2-Wasserstein distance model produced reasonable maps in which similar expressions (e.g., “mom-cordinate” and “mom-fashion”) were gathered in close proximity. In contrast, the Mahalanobis distance model produced a questionable mapping. 2) Similarly, the results for the variance of the tags obtained show that KL divergence and the 2-Wasserstein distance model have a larger variance for abstract expressions and for items that are included in many outfits and slight for items that are not included in many outfits, indicating that the results were somewhat valid. Conversely, the Mahalanobis distance model gave quite different results. 3) The results for the variance of the acquired images from KL divergence and the 2-Wasserstein distance model differed significantly from those of the Mahalanobis distance model.

The Mahalanobis distance between the two probability distributions is expressed by Eq. (6.28). If we look at the part of the joint covariance matrix Σ in this equation, we can see that it plays only a scaling role when taking the inner product of the differences of mean vectors between the two distributions. Specifically, when this part (the sum of the two distributional covariance matrices) is large, the overall Mahalanobis distance increases, and when it is small, the overall Mahalanobis distance decreases. Moreover, it plays no other role. This makes it questionable whether adopting this type of distance is appropriate to understand the spread of meanings of images and tags. Particularly, in the case of data similar to that used in this study, where the frequency of occurrence of each tag is highly skewed, the loss of the entire training batch containing the tag can be reduced by reducing the variance of the tag with the highest frequency of occurrence. Specifically, learning is expected to proceed so that the variance is adjusted by looking at the frequency of occurrence of tags attached to images in the entire data set. Moreover, the experimental results showed a complete tendency to do this. Adopting the Mahalanobis distance may be risky for this targeted problem and the proposed DGVSE model.

6.7.1.2 On the Similarities between KL divergence and 2-Wasserstein Distance

The respective results obtained from the models adopting KL divergence and 2-Wasserstein distance were very similar.

It is known that the KL divergence measures the distance from the midpoints for multimodal distributions. More precisely, given $p \in \mathcal{P}$, we search for the distribution $\hat{p}$ that minimizes the divergence from p to a smooth submanifold $\mathcal{S} \subset \mathcal{P}$,

$$\hat{p} = \arg\min_{q \in \mathcal{S}} D_{KL}[p \| q]. \tag{6.31}$$

Then, the best approximation $\hat{p}$ in the closure of $\mathcal{S}$ satisfies $\hat{p}(\boldsymbol{x}) = 0$ for $\boldsymbol{x}$ at which $p(\boldsymbol{x}) = 0$, and this property is called zero-forcing. Note that for the reverse KL divergence $D_{KL}[q \| p]$, the best approximation $\hat{p}$ in the closure of $\mathcal{S}$ satisfies $\hat{p}(\boldsymbol{x}) \neq 0$ for x at which $p(x) \neq 0$. However, l-Wasserstein distance is robust for multimodal distributions [207, 208].

However, the results obtained from both models in this study were similar. Therefore, the distributions may estimate as unimodal embedded representations of each tag or image, at least for the data of this study. It is also suggested that the distance measures selected in this study (KL divergence and 2-Wasserstein distance) do not have any major problems. Thus, the distance to be used should be appropriately selected according to the target problem and the characteristics of the data while observing the theoretical and empirical results.

6.7.2 On the Covariance Matrix

In this study, we assumed a multidimensional Gaussian distribution with only the diagonal components' values behind the embedded representation. That is, the parameters were estimated, assuming that each component follows an independent Gaussian distribution. This method is used in many Gaussian embedding methods, such as Word2Gauss and GVSE. If the non-diagonal components also have values, the number of parameters increases to "({number of images} + {number of tags}) × {number of dimensions of the embedded representation}^{2}." However, if only the diagonal components have values, the number of parameters can be reduced to "({number of images} + {number of tags}) × {number of dimensions of the embedded representation}."

This study assumes a "spherical" situation where all diagonal components have the same value. This method reduces the number of parameters to "{number of images} + {number of tags}." This method is used in many Gaussian embedding methods, such as Word2Gauss and GVSE. However, a "diagonal" situation in which all diagonal components have different values can also be considered, but the number of parameters increases to "({number of images} + {number of tags}) × {number of dimensions of the embedded representation}." Even if the analyst adopts the diagonal covariance matrix and different variances are obtained for each dimension, the marketing insight that can be gained from the results is small. Thus, the analyst wants to know one value per image or tag (the overall trend) and is unlikely to gain many benefits if different values are obtained for each dimension individually.

Therefore, the method of estimating a "spherical" covariance matrix adopted in this study is considered the best in achieving the objective of this study, which is to propose a method that is useful for marketing.

6.7.3 On the Strengths of the Proposed Model

Because VSE, GVSE, and DGVSE have been proposed in previous studies, it is necessary to clarify their differences and understand which is stronger in which situations. The following table summarizes the characteristics of each method.

Table 6.2: Summary comparing the characteristics of each method

	VSE	GVSE	DGVSE
embedding of image (mean)	○	△ (without visual information)	○
embedding of attribute (mean)	○	○	○
embedding of image (variance)	×	○	○
embedding of attribute (variance)	×	×	○

As shown in the above table, DGVSE is the one that can capture more from a single model. However, VSE has the advantage of requiring fewer parameters to be estimated, although it is not able to capture variance. If the purpose is only to use the original functions of VSE, such as search and sorting, the choice should be based on the accuracy of the target task. However, because this study aims to create term maps and visualize the spread of meanings, DGVSE is suitable for this purpose.

6.8 Conclusion of this Chapter

In this chapter, we proposed a new model, DGVSE, a kind of new fashion intelligence system that enables the interpretation of abstract fashion attribute information. The proposed model can be added to the VSE function of mapping images and tags in the same space and can estimate the variance of both embedded representations. The conventional method of GVSE faces the following problems: 1) it is not an end-to-end learning method, 2) it only observes the co-occurrence of words and ignores image information when learning the embedded representation (mean) of words, and 3) it assumes a multidimensional Gaussian distribution only for words (hence the problem of measuring the distance between a point and the probability distribution using the Mahalanobis distance). Notably, the proposed model solves these problems. The risk that loss functions that include the Mahalanobis distance cannot be learned stably has been theoretically and empirically demonstrated. We showed how multifaceted analysis contributes to reducing fashion-specific ambiguity and complexity by applying the proposed method to a real service dataset. We expect this method to be used in the real world to develop systems that support user purchasing activities in online fashion searches.

Chapter 7

Discussion

This chapter clarifies the roles and differences of the four proposed models in this study. Particularly, we discuss the fashion intelligence system newly defined in this study and consider its social impact.

7.1 On the Differences between the Proposed Models

We have provided a detailed discussion of each model in each chapter of this study. In this chapter, we present the overall discussion of this study.

In this study, a total of four models were proposed. The proposed model 1 is an explainable recommendation model with a fundamentally different application from the other three models. However, it is necessary to clarify the usage of the other three models. The following is a summary of the features of each model.

Proposed Model 2 (Chapter 4): VSE [8]

- A simple model to learn the full-body outfit image and attached expressions at once.
- By adding various innovations such as negative sampling, full-body images and tags with ambiguous expressions can be mapped to the same space.

Proposed Model 3 (Chapter 5): PVSE [9]

- A model capable of acquiring embedded representations that correspond to any one part in each dimension.
- Added ability to search and sort images focusing only on specified parts.

Proposed Model 4 (Chapter 6): DGVSE [10]

- A model that assumes a multidimensional normal distribution behind the embedded representation of images and tags.
- End-to-end learning allows us also to consider image information when estimating the mean and variance-covariance matrix of the embedded representation of the tag.

The following table summarizes the characteristics of each model.

Table 7.1: Summary of the features of each proposed VSE model

Feature / Models	VSE	PVSE	DGVSE
Accuracy of Embedded Space	○	◎	△
Accuracy of AAM	○	◎	○
Analysis with variance	×	×	○

Table 7.2: Summary of the features of each VSE model

Feature / Models	VSE	PVSE	DGVSE
Q1. What happens if I make this outfit “formal”?	○	○	○
Q2. How “casual” is this outfit?	○	○	○
Q3. What is the “street” point in this outfit?	○	○	○
Q4. What happens if I make the upper body of this outfit “formal”?	×	○	×
Q5. How “casual” is this jacket?	×	○	×

As this table provides a more accurate and broader range of answers, we believe that the proposed model 2 (PVSE model) is a good choice for the fashion intelligence system. However, the proposed model 3 (DGSE model) has the unique feature of obtaining a covariance matrix for each mapped element. Therefore, selecting the proposed model 3 is effective if you want to analyze the usage and the breadth of meaning of items and expressions in detail.

In addition, the preprocessing required to train each model is as follows.

Proposed Model 2 (Chapter 4): VSE [8]

- It is possible to perform with only preprocessing by semantic segmentation, which separates the foreground (the area in which the person is in the picture) and background.

Proposed Model 3 (Chapter 5): PVSE [9]

- Need preprocessing with semantic segmentation separable into regions per item.

Proposed Model 4 (Chapter 6): DGVSE [10]

- It is possible to perform with only preprocessing by semantic segmentation, which separates the foreground (the area in which the person is in the picture) and background.

Thus, the proposed model 1 (VSE model) only needs to separate the background from the human in the pre-processing (semantic segmentation) step and does not need to perform segmentation for each item as in the pre-processing step for PVSE. Notably, open-source software, which is readily available worldwide, can be used for this purpose. Therefore, selecting the proposed Model 1 for the initial implementation stage would not be a wrong choice for a company.

7.2 On the Social Impact of the Fashion Intelligence System

The proposed system can be used on e-commerce sites and social media and as a valuable reference and support tool for store clerks assisting customers in physical stores. Furthermore, this system can be applied to fields such as architecture, art, furniture, and cuisine using data that consists of images and information such as sentences and words associated with the images. Particularly, it is expected to be effective for applications in fields wherein ambiguous expressions are used to describe objects. Additionally, it offers a means to address societal challenges and make meaningful contributions, such as waste reduction, by empowering consumers to make informed choices among the products offered by companies.

One of the key advantages of the system proposed in this research is its versatility, as it caters to both professionals and non-professionals in offline and online settings, extending its utility to virtually any field. When deployed as a real-world service, users will have the opportunity to discover items and outfits that genuinely resonate with their preferences, thereby holding the potential to alleviate social issues and foster positive societal impact.

Chapter 8

Conclusion

This chapter describes the conclusion of this study and future avenues for research.

8.1 Conclusion of this Study

This study aimed to propose a system to support users' understanding and interpretation of items in the fashion industry and to enhance online usability and user satisfaction. Based on this objective, we proposed the following two approaches.

- Approach 1: Focusing on the interaction between the company and users; a recommendation system with explainability (proposed model 1) was proposed to provide persuasive and convincing recommendations [7].
- Approach 2: Targeting the evaluation and specification stages of users' purchasing processes, a fashion intelligence system capable of automatically interpreting ambiguous fashion expressions and retrieving answers to user queries was developed [8, 9, 10].

Both proposals were applied to real-world service data, and their effectiveness and usefulness were demonstrated through multifaceted evaluation and analysis experiments.

Specifically, the results confirmed that proposed model 1 (Chapter 3) provides highly accurate recommendations and offers various ways to use the results, including the ability to provide reasons for recommendations. By leveraging diverse side information and mitigating computational complexity, this contribution enables users to understand the positive aspects of recommended items and fosters a compelling recommendation system. Furthermore, the fashion intelligence system (Chapters 4-6) uses VSE modeling techniques as its foundation, and we proposed three types of VSE models in this study.

Furthermore, the fashion intelligence system (Chapters 4-6) utilizes VSE modeling techniques as its foundation, and we proposed three types of VSE models in this study.

Proposed model 2 VSE can map a massive amount of full-body outfit images with abundant tags containing various ambiguous expressions into the same space using foreground-centered learning, background regularization, and other schemes [8].

Proposed model 3 Partial VSE that enables sensitive learning of each part [9].

Proposed model 4 Dual Gaussian VSE that enables the analysis of the meaning and diversity of mapped elements, such as outfits, items, and ambiguous expressions [10].

In Chapter 4, we proposed the fashion intelligence system, which is the key to the rest of the study. To realize the fashion intelligence system, we proposed a VSE (proposed model 2) that maps a full-body image composed of various parts and backgrounds, and a tag set, including both concrete and ambiguous expressions, into the same space at once, and various applications based on the VSE technology.

In Chapter 5, we proposed PVSE (proposed model 3), which enabled us to obtain features for each part of a full-body outfit, such as hairstyle, face, jacket, t-shirt, pants, and shoes, in contrast to proposed model 2, which learns a full-body image at once. We showed that proposed model 3 could realize an application that focuses on specified parts, which was not possible with proposed model 2, and can respond to the more detailed needs of the user.

In Chapter 6, we proposed the DGVSE model (proposed model 4), which maps each element not as a point but as a distribution in the projective space. Proposed model 4 shows a method to analyze the meaning of the elements to be mapped in detail and the breadth of their uses, which was impossible with proposed models 2 and 3. This is expected to contribute to a more detailed understanding of users' ambiguous images of fashion.

The fashion intelligence system, empowered by these VSE models, precisely maps images and tags in the same space. By leveraging the obtained embedded representations in various ways, the system expands users' knowledge of fashion, catering to both experts and non-experts, and supports a wide range of choices and actions. This technology aids users in their decision-making and activities through social media and other platforms, addressing the recent trend of using others' outfits as references. Given the historical context, the contributions of this research hold significant value in real-world services.

8.2 Future Works and Prospects

In future research, the work on explainable recommendations can be extended to incorporate side information with many-to-many relationships beyond favorite information. It is necessary to propose an algorithm that effectively models and learns these relationships with uncertainty and explore techniques to mitigate overlearning issues, which have been identified as challenges for graph neural networks.

Based on the new technology developed in this study, the proposed system currently operates only with datasets containing relatively clean images and tags assigned by users with a certain degree of expertise. Therefore, efforts should be made to enhance the system's applicability to data contributed by any user.

Furthermore, the proposed system has the potential to be applied in various fields, such as architecture, art, furniture, and cuisine, as long as the dataset comprises images and associated textual information (e.g., sentences, words). Its effectiveness is particularly expected in domains characterized by ambiguous expressions.

Acknowledgments

This study summarizes the research results conducted by the author at the Graduate School of Creative Science and Engineering, Waseda University.

First, I would like to express my deepest gratitude to Professor Masayuki Goto of the Department of Industrial & Management Systems Engineering, School of Creative Science and Engineering, Waseda University, who has guided me throughout my whole research career as a supervisor. For the nine years since I joined the Nepal-Japan Project in my second year as an undergraduate student, he has provided ample guidance and advice in both my research life and personal life. I continued my master's course and returned to the doctoral course largely due to my admiration for him. I also thank his enthusiastic and attentive guidance because of which I have been able to work positively on my research and compile my doctoral dissertation. I express my profound appreciation and gratitude to him.

I would like to thank Professor Yasushi Nagata and Professor Tomoji Kishi of the Department of Industrial & Management Systems Engineering, School of Creative Science and Engineering, Waseda University, and Professor Masao Ueda of the Department of Data Science, Faculty of Data Science, Yokohama City University, for their valuable comments and abundant advice from the viewpoint of their respective specialties. I express my deepest gratitude to them.

I would like to express my deepest gratitude to the professors of the Department of Industrial & Management Systems Engineering, School of Creative Science and Engineering, Waseda University, who provided me with a great deal of guidance in many areas, including lectures and research, since I was a student at the School of Creative Science and Engineering, Waseda University.

I would like to express my deepest gratitude to Associate Professor Haruka Yamashita of the Department of Information and Communication Sciences, Faculty of Science and Engineering, Sophia University, who has been a great help to me both publicly and privately since my undergraduate days. In particular, winning prizes in both the data analysis competition that I participated in with Mr. Leona Suzuki and Mr. Teppei Sakamoto during my undergraduate days and in the subsequent conference of Industrial Management and Engineering was a formative experience in my career. I express my deepest gratitude.

I would like to thank Dr. Megumi Matsutani, Dr. Yuki Saito, Mr. Masanari Kimura, Mr. Takuma Nakamura, and other members of the ZOZO Research for their support in balancing my studies and work and for their advice regarding the content of my research. Thanks to the favorable environment, I could complete my doctoral thesis while managing my daily work and parenting responsibilities. I express my deepest gratitude.

I am grateful to ZOZO, Inc. and ZOZO NEXT, Inc. for their support of the doctoral program for working adults, which has made it possible for me to pursue my long-cherished dream of entering the doctoral program. Moreover, thanks to their program, I could spend my student life without inconvenience until I completed my doctoral thesis. I express my deepest gratitude.

In addition, I would like to thank the individuals at DeNA, Inc. for graciously supporting my decision to pursue a doctoral degree.

My life in Goto Laboratory started when I lost to Professor Goto in a table tennis game, but it turned out to be a wonderful experience as I found great colleagues and discovered my passion for research. I express my sincere gratitude to all the people who have been involved.

I would like to express my heartfelt gratitude to my parents, my four younger siblings, Happy and Coco (our pet dogs), for raising and supporting me, and to my supportive friends. Finally, I sincerely thank my wife Haruna, my pet dog Nats, and my son Sosuke who always graciously supported and made me happy.

July, 2023

Ryotaro Shimizu

References

[1] Meta Platforms, Inc., Instagram. Retrieved from `https://www.instagram.com/`, Accessed Oct. 30, 2022.

[2] Meta Platforms, Inc., Facebook. Retrieved from `https://www.facebook.com/`, Accessed Oct. 30, 2022.

[3] Twiter, Inc., Twiter. Retrieved from `https://twitter.com/`, Accessed Oct. 30, 2022.

[4] ZOZO, Inc., ZOZOTOWN. Retrieved from `https://zozo.jp/`, Accessed Oct. 30, 2022.

[5] ZOZO, Inc., WEAR. Retrieved from `https://wear.jp/`, Accessed Oct. 30, 2022.

[6] ZOZO, Inc., WEAR (pirosh's fashion list). Retrieved from `https://wear.jp/pirosh`, Accessed Oct. 30, 2022.

[7] Ryotaro Shimizu, Megumi Matsutani, and Masayuki Goto. An explainable recommendation framework based on an improved knowledge graph attention network with massive volumes of side information. *Knowledge-Based Systems*, 239:107970, 2022. ISSN 0950-7051.

[8] Ryotaro Shimizu, Yuki Saito, Megumi Matsutani, and Masayuki Goto. Fashion intelligence system: An outfit interpretation utilizing images and rich abstract tags. *Expert Systems with Applications*, page 119167, 2022. ISSN 0957-4174.

[9] Ryotaro Shimizu, Takuma Nakamura, and Masayuki Goto. Fashion-specific ambiguous expression interpretation with partial visual-semantic embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3496–3501, Jun. 2023.

[10] Ryotaro Shimizu, Masanari Kimura, and Masayuki Goto. Fashion-specific attributes interpretation via dual gaussian visual-semantic embedding. *CoRR*, abs/2210.17417, 2022.

[11] Chandadevi Giri, Sheenam Jain, Xianyi Zeng, and Pascal Bruniaux. A detailed review of artificial intelligence applied in the fashion and apparel industry. *IEEE Access*, 7: 95376–95396, 2019.

[12] Youssra Riahi, Tarik Saikouk, Angappa Gunasekaran, and Ismail Badraoui. Artificial intelligence applications in supply chain: A descriptive bibliometric analysis and future research directions. *Expert Systems with Applications*, 173:114702, 2021. ISSN 0957-4174.

[13] Sanchi Arora and Abhijit Majumdar. Machine learning and soft computing applications in textile and clothing supply chain: Bibliometric and network analyses to delineate future research agenda. *Expert Systems with Applications*, 200:117000, 2022. ISSN 0957-4174.

[14] Hyungjung Kim, Woo-Kyun Jung, Young-Chul Park, Jae-Won Lee, and Sung-Hoon Ahn. Broken stitch detection method for sewing operation using cnn feature map and image-processing techniques. *Expert Systems with Applications*, 188:116014, 2022. ISSN 0957-4174.

[15] Jun-Ming Lu, Mao-Jiun J. Wang, Chien-Wen Chen, and Jyi-Hua Wu. The development of an intelligent system for customized clothing making. *Expert Systems with Applications*, 37(1):799–803, 2010. ISSN 0957-4174.

[16] C.K.H. Lee, K.L. Choy, G.T.S. Ho, and C.H.Y. Lam. A slippery genetic algorithm-based process mining system for achieving better quality assurance in the garment industry. *Expert Systems with Applications*, 46:236–248, 2016. ISSN 0957-4174.

[17] Ming-Kuen Chen, Ying-Han Wang, and Tsu-Yi Hung. Establishing an order allocation decision support system via learning curve model for apparel logistics. *Journal of Industrial and Production Engineering*, 31(5):274–285, 2014.

[18] Congying Guan, Shengfeng Qin, Wessie Ling, and Guofu Ding. Apparel recommendation system evolution: an empirical review. *International Journal of Clothing Science and Technology*, 28(6):854–879, 2016.

[19] Ting-Peng Liang and Yu-Hsi Liu. Research landscape of business intelligence and big data analytics: A bibliometrics study. *Expert Systems with Applications*, 111:2–10, 2018. ISSN 0957-4174.

[20] Samit Chakraborty, Md. Saiful Hoque, Naimur Rahman Jeem, Manik Chandra Biswas, Deepayan Bardhan, and Edgar Lobaton. Fashion recommendation systems, models and methods: A review. *Informatics*, 8(3), 2021. ISSN 2227-9709.

[21] Ruining He and Julian McAuley. Vbpr: Visual bayesian personalized ranking from implicit feedback. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 144–150, Feb. 2016.

[22] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence, Montreal*, volume 24, pages 452–461, Jun. 2009.

[23] Ruining He, Chunbin Lin, Jianguo Wang, and Julian McAuley. Sherlock: Sparse hierarchical embeddings for visually-aware one-class collaborative filtering. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pages 3740–3746, Jul. 2016.

[24] Yuki Saito, Takuma Nakamura, Hirotaka Hachiya, and Kenji Fukumizu. Exchangeable deep neural networks for set-to-set matching and learning. In *Proceedings of the Computer Vision – ECCV 2020*, pages 626–646, Aug. 2020. ISBN 978-3-030-58519-8.

[25] Wen Chen, Pipei Huang, Jiaming Xu, Xin Guo, Cheng Guo, Fei Sun, Chao Li, Andreas Pfadler, Huan Zhao, and Binqiang Zhao. Pog: Personalized outfit generation for fashion recommendation at alibaba ifashion. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2662–2670, Aug. 2019.

[26] Wei-Lin Hsiao and Kristen Grauman. Creating capsule wardrobes from fashion images. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7161–7170, Jun. 2018.

[27] Xishan Zhang, Jia Jia, Ke Gao, Yongdong Zhang, Dongming Zhang, Jintao Li, and Qi Tian. Trip outfits advisor: Location-oriented clothing recommendation. *IEEE Transactions on Multimedia*, 19(11):2533–2544, 2017.

[28] Sanghyuk Park, Minchul Shin, Sungho Ham, Seungkwon Choe, and Yoohoon Kang. Study on fashion image retrieval methods for efficient fashion visual search. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 316–319, Jun. 2019.

[29] Ayush Chopra, Abhishek Sinha, Hiresh Gupta, Mausoom Sarkar, Kumar Ayush, and Balaji Krishnamurthy. Powering robust fashion retrieval with information rich feature embeddings. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 326–334, Jun. 2019.

[30] Saddam Hussain, Muhammad Ahmad Zia, and Waqas Arshad. Additive deep feature optimization for semantic image retrieval. *Expert Systems with Applications*, 170:114545, 2021. ISSN 0957-4174.

[31] Yiming Gao, Zhanghui Kuang, Guanbin Li, Ping Luo, Yimin Chen, Liang Lin, and Wayne Zhang. Fashion retrieval via graph reasoning networks on a similarity pyramid. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–15, 2020.

[32] Sarah Ibrahimi, Nanne van Noord, Zeno Geradts, and Marcel Worring. Deep metric learning for cross-domain fashion instance retrieval. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3165–3168, Oct. 2019.

[33] Bojana Gajic and Ramon Baldrich. Cross-domain fashion image retrieval. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1950–1952, Jun. 2018.

[34] Sirion Vittayakorn, Kota Yamaguchi, Alexander C. Berg, and Tamara L. Berg. Runway to realway: Visual analysis of fashion. In *Proceedings of the 2015 IEEE Winter Conference on Applications of Computer Vision*, pages 951–958, Jan. 2015.

[35] Charles Corbi è re, Hedi Ben-Younes, Alexandre Ram é , and Charles Ollion. Leveraging weakly annotated data for fashion image retrieval and label prediction. In *Proceedings*

of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), pages 2268–2274, Oct. 2017.

[36] Jianfeng Dong, Zhe Ma, Xiaofeng Mao, Xun Yang, Yuan He, Richang Hong, and Shouling Ji. Fine-grained fashion similarity prediction by attribute-specific embedding learning. *IEEE Transactions on Image Processing*, 30:8410–8425, 2021.

[37] Xianjing Han, Xuemeng Song, Yiyang Yao, Xin-Shun Xu, and Liqiang Nie. Neural compatibility modeling with probabilistic knowledge distillation. *IEEE Transactions on Image Processing*, 29:871–882, 2020.

[38] Kenan E. Ak, Joo Hwee Lim, Ying Sun, Jo Yew Tham, and Ashraf A. Kassim. Fashionsearchnet-v2: Learning attribute representations with localization for image retrieval with attribute manipulation. *CoRR*, abs/2111.14145, 2021.

[39] Minchul Shin, Sanghyuk Park, and Taeksoo Kim. Semi-supervised feature-level attribute manipulation for fashion image retrieval. *CoRR*, abs/1907.05007, 2019.

[40] Yuxin Hou, Eleonora Vig, Michael Donoser, and Loris Bazzani. Learning attribute-driven disentangled representations for interactive fashion retrieval. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12127–12137, Oct. 2021.

[41] Xiting Wang, Yiru Chen, Jie Yang, Le Wu, Zhengtao Wu, and Xin Xie. A Reinforcement Learning Framework for Explainable Recommendation. In *Proceedings of the IEEE International Conference on Data Mining*, pages 587–596, Nov. 2018.

[42] Yongfeng Zhang and Xu Chen. *Explainable Recommendation: A Survey and New Perspectives*. Now Foundations and Trends, 2020.

[43] Yongfeng Zhang. Explainable Recommendation: Theory and Applications. *CoRR*, abs/1708.06409, 2017.

[44] Yongfeng Zhang, Yi Zhang, and Min Zhang. Sigir 2018 workshop on explainable recommendation and search (ears 2018). In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1411–1413, Jun. 2018. ISBN 9781450356572.

[45] Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 765–774, Jul. 2019. ISBN 9781450361729.

[46] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. Explaining Collaborative Filtering Recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, pages 241–250. Association for Computing Machinery, Dec. 2000.

[47] Krisztian Balog, Filip Radlinski, and Shushan Arakelyan. Transparent, Scrutable and Explainable User Models for Personalized Recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 265–274, 2019.

[48] James McInerney, Benjamin Lacker, Samantha Hansen, Karl Higley, Hugues Bouchard, Alois Gruson, and Rishabh Mehrotra. Explore, Exploit, Explain: Personalizing Explainable Recommendations with Bandits. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 31–39, Oct. 2018.

[49] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ”Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1135–1144, Aug. 2016.

[50] Fan Zhu, Min Jiang, Yiming Qiu, Chenglong Sun, and Min Wang. RSLIME: An Efficient Feature Importance Analysis Approach for Industrial Recommendation Systems. In *Proceedings of 2019 International Joint Conference on Neural Networks*, pages 1–6, Jul. 2019.

[51] Caio Nóbrega and Leandro Marinho. Towards explaining recommendations through local surrogate models. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pages 1671–1678, Apr. 2019.

[52] Jaspreet Singh and Avishek Anand. EXS: Explainable Search Using Local Model Ag-

nostic Interpretability. In *Proceedings of the 12th ACM International Conference on Web Search & Data Mining*, pages 770–773, Feb. 2019.

[53] Georgina Peake and Jun Wang. Explanation Mining: Post Hoc Interpretability of Latent Factor Models for Recommendation Systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2060–2069, Aug. 2018.

[54] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix Factorization Techniques for Recommender Systems. *Computer*, 42(8):30–37, 2009.

[55] Gregory Piatetsky-Shapiro. Discovery, Analysis, and Presentation of Strong Rules. *Knowledge Discovery in Databases*, 1991.

[56] Alexandre Heuillet, Fabien Couthouis, and Natalia D í az-Rodr í guez. Explainability in deep reinforcement learning. *Knowledge-Based Systems*, 214:106685, 2021. ISSN 0950-7051.

[57] Behnoush Abdollahi and Olfa Nasraoui. Using Explainability for Constrained Matrix Factorization. In *Proceedings of the 11th ACM Conference on Recommender Systems*, pages 79–83, Aug. 2017.

[58] Behnoush Abdollahi and Olfa Nasraoui. Explainable Restricted Boltzmann Machines for Collaborative Filtering. *CoRR*, abs/1606.07129, 2016.

[59] Sungyong Seo, Jing Huang, Hao Yang, and Yan Liu. Interpretable Convolutional Neural Networks with Dual Local and Global Attention for Review Rating Prediction. In *Proceedings of the 11th ACM Conference on Recommender Systems*, pages 297–305, Aug. 2017.

[60] Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. Neural Attentional Rating Regression with Review-Level Explanations. In *Proceedings of the 2018 World Wide Web Conference*, pages 1583–1592, Apr. 2018.

[61] Cai Xu, Ziyu Guan, Wei Zhao, Quanzhou Wu, Meng Yan, Long Chen, and Qiguang Miao. Recommendation by Users’ Multimodal Preferences for Smart City Applications. *IEEE Transactions on Industrial Informatics*, 17(6):4197–4205, 2021.

[62] Shaohua Tao, Runhe Qiu, Yuan Ping, and Hui Ma. Multi-modal Knowledge-aware Reinforcement Learning Network for Explainable Recommendation. *Knowledge-Based Systems*, 227:107217, 2021. ISSN 0950-7051.

[63] Zhaochun Ren, Shangsong Liang, Piji Li, Shuaiqiang Wang, and Maarten de Rijke. Social Collaborative Viewpoint Regression with Explainable Recommendations. In *Proceedings of the 10th ACM International Conference on Web Search & Data Mining*, pages 485–494, Feb. 2017.

[64] Amirreza Salamat, Xiao Luo, and Ali Jafari. HeteroGraphRec: A heterogeneous graph-based neural networks for social recommendations. *Knowledge-Based Systems*, 217: 106817, 2021. ISSN 0950-7051.

[65] Xiangnan He, Tao Chen, Min-Yen Kan, and Xiao Chen. TriRank: Review-Aware Explainable Recommendation by Modeling Aspects. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1661–1670, Oct. 2015.

[66] Konstantin Bauman, Bing Liu, and Alexander Tuzhilin. Aspect Based Recommendations: Recommending Items with the Most Valuable Aspects Based on User Reviews. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 717–725, Aug. 2017.

[67] Chao Yang, Weixin Zhou, Zhiyu Wang, Bin Jiang, Dongsheng Li, and Huawei Shen. Accurate and Explainable Recommendation via Hierarchical Attention Network Oriented Towards Crowd Intelligence. *Knowledge-Based Systems*, 213:106687, 2021. ISSN 0950-7051.

[68] Yikun Xian, Zuohui Fu, S. Muthukrishnan, Gerard de Melo, and Yongfeng Zhang. Reinforcement Knowledge Graph Reasoning for Explainable Recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 285–294, Jul. 2019.

[69] Yuan Zhang, Xiaoran Xu, Hanning Zhou, and Yan Zhang. Distilling Structured Knowledge into Embeddings for Explainable and Accurate Recommendation. *CoRR*, abs/1912.08422, 2019.

[70] Huijing Zhan, Jie Lin, Kenan Emir Ak, Boxin Shi, Ling-Yu Duan, and Alex C. Kot. A^3-fkg: Attentive attribute-aware fashion knowledge graph for outfit preference prediction. *IEEE Transactions on Multimedia*, 24:819–831, 2022.

[71] Min Hou, Le Wu, Enhong Chen, Zhi Li, Vincent W. Zheng, and Qi Liu. Explainable fashion recommendation: A semantic attribute region guided approach. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 4681–4688, Aug. 2019. ISBN 9780999241141.

[72] Yujie Lin, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Jun Ma, and Maarten de Rijke. Explainable outfit recommendation with joint outfit matching and comment generation. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1502–1516, 2020.

[73] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6602–6611, Jun. 2019.

[74] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference 2018 (BMVC)*, pages 1–13, Sep. 2018.

[75] Taotao Jing, Haifeng Xia, Jihun Hamm, and Zhengming Ding. Augmented multimodality fusion for generalized zero-shot sketch-based visual retrieval. *IEEE Transactions on Image Processing*, 31:3657–3668, 2022.

[76] Mengye Ren, Ryan Kiros, and Richard S. Zemel. Exploring models and data for image question answering. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, volume 2, pages 2953–2961, Dec. 2015.

[77] Andrea Frome, Greg S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, volume 2, pages 2121–2129, Dec. 2013.

[78] Zheng Zhang, Zhihui Lai, Zi Huang, Wai Keung Wong, Guo-Sen Xie, Li Liu, and Ling

Shao. Scalable supervised asymmetric hashing with semantic and latent factor embedding. *IEEE Transactions on Image Processing*, 28(10):4803–4818, 2019.

[79] Fumin Shen, Xiang Zhou, Jun Yu, Yang Yang, Li Liu, and Heng Tao Shen. Scalable zero-shot learning via binary visual-semantic embeddings. *IEEE Transactions on Image Processing*, 28(7):3662–3674, 2019.

[80] Li Niu, Jianfei Cai, Ashok Veeraraghavan, and Liqing Zhang. Zero-shot learning via category-specific visual-semantic mapping and label refinement. *IEEE Transactions on Image Processing*, 28(2):965–979, 2019.

[81] Zan Gao, Hongwei Wei, Weili Guan, Weizhi Nie, Meng Liu, and Meng Wang. Multi-granular visual-semantic embedding for cloth-changing person re-identification. *CoRR*, abs/2108.04527, 2021.

[82] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137, Jun. 2015.

[83] Yanbei Chen and Loris Bazzani. Learning joint visual semantic matching embeddings for language-guided retrieval. In *Proceedings of the Computer Vision – ECCV 2020*, pages 136–152, Aug. 2020. ISBN 978-3-030-58541-9.

[84] Ivona Tautkute, Tomasz Trzcinski, Aleksander P. Skorupa, Lukasz Brocki, and Krzysztof Marasek. Deepstyle: Multimodal search engine for fashion and interior design. *IEEE Access*, 7:84613–84628, 2019.

[85] Jianfeng Wang, Xiaochun Cheng, Ruomei Wang, and Shaohui Liu. Learning outfit compatibility with graph attention network and visual-semantic embedding. In *Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, Jul. 2021.

[86] Xintong Han, Zuxuan Wu, Phoenix X. Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S. Davis. Automatic spatially-aware fashion concept discovery. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1472–1480, Oct. 2017.

[87] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.

[88] Matt Turek. DARPA: Defense Advanced Research Projects Agency, Explainable Artificial Intelligence (XAI), Accessed 22 Jun. 2021. `https://www.darpa.mil/program/explainable-artificial-intelligence`, 2020.

[89] Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. KGAT: Knowledge Graph Attention Network for Recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 950–958, Aug. 2019.

[90] Zuoxi Yang and Shoubin Dong. HAGERec: Hierarchical Attention Graph Convolutional Network Incorporating Knowledge Graph for Explainable Recommendation. *Knowledge-Based Systems*, 204:106194, 2020. ISSN 0950-7051.

[91] Qingbin Liu, Guirong Bai, Shizhu He, Cao Liu, Kang Liu, and Jun Zhao. Heterogeneous Relational Graph Neural Networks with Adaptive Objective for End-to-End Task-Oriented Dialogue. *Knowledge-Based Systems*, 227:107186, 2021. ISSN 0950-7051.

[92] Xiangyu Song, Jianxin Li, Yifu Tang, Taige Zhao, Yunliang Chen, and Ziyu Guan. JKT: A joint graph convolutional network based Deep Knowledge Tracing. *Information Sciences*, 580:510–523, 2021. ISSN 0020-0255.

[93] Guotong Xue, Ming Zhong, Jianxin Li, Jia Chen, Chengshuai Zhai, and Ruochen Kong. Dynamic Network Embedding Survey. *CoRR*, abs/2103.15447, 2021.

[94] Petar Veli☒kovi☒, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. *CoRR*, abs/1710.10903, 2018.

[95] Cataldo Musto, Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. Context-aware graph-based recommendations exploiting Personalized PageRank. *Knowledge-Based Systems*, 216:106806, 2021. ISSN 0950-7051.

[96] Ke Ji and Hong Shen. Addressing cold-start: Scalable recommendation with tags and keywords. *Knowledge-Based Systems*, 83:42–50, 2015. ISSN 0950-7051.

[97] Tribikram Pradhan and Sukomal Pal. CNAVER: A Content and Network-based Academic VEnue Recommender system. *Knowledge-Based Systems*, 189:105092, 2020. ISSN 0950-7051.

[98] Abdulgabbar Saif, Mohd Juzaiddin Ab Aziz, and Nazlia Omar. Reducing explicit semantic representation vectors using Latent Dirichlet Allocation. *Knowledge-Based Systems*, 100:145–159, 2016. ISSN 0950-7051.

[99] Chien-Liang Liu and Xuan-Wei Wu. Fast recommendation on latent collaborative relations. *Knowledge-Based Systems*, 109:25–34, 2016. ISSN 0950-7051.

[100] Akaike Hirotugu. Information Theory and an Extension of the Maximum Likelihood Principle. *Selected Papers of Hirotugu Akaike* , 1998.

[101] Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2): 461–464, 1978.

[102] Masayuki Goto. Latent class models on business analytics. In *Proceedings of the 2019 IEEE International Conference on Big Data, Cloud Computing, Data Science & Engineering (BCD)*, pages 142–147, May 2019.

[103] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 2181–2187, Jan. 2015.

[104] Qigang Liu, Lifeng Mu, Vijayan Sugumaran, Chongren Wang, and Dongmei Han. Pair-wise ranking based preference learning for points-of-interest recommendation. *Knowledge-Based Systems*, 225:107069, 2021. ISSN 0950-7051.

[105] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 289–296, Jul.-Aug. 1999. ISBN 1558606149.

[106] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003. ISSN 1532-4435.

[107] Qingyao Ai, Vahid Azizi, Xu Chen, and Yongfeng Zhang. Learning Heterogeneous Knowledge Base Embeddings for Explainable Recommendation. *Algorithms*, 11(9):137, 2018.

[108] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[109] Bee des Bee (in ZOZOTOWN). Retrieved from `https://zozo.jp/brand/bee-des-bee/`, Accessed Jul. 21, 2021.

[110] 392 plusm (in ZOZOTOWN). Retrieved from `https://zozo.jp/brand/392/`, Accessed Jul. 21, 2021.

[111] Gunze. Retrieved from `https://www.gunze.jp/store/kids`, Accessed Jul. 21, 2021.

[112] Kodomo BEAMS. Retrieved from `https://www.beams.co.jp/kodomobeams/`, Accessed Jul. 21, 2021.

[113] URBAN RESEARCH Sony Label. Retrieved from `https://www.urban-research.jp/shop/sonny-label/`, Accessed Jul. 21, 2021.

[114] URBAN RESEARCH. Retrieved from `https://www.urban-research.jp/`, Accessed Jul. 21, 2021.

[115] BEAMS. Retrieved from `https://www.beams.co.jp/`, Accessed Jul. 21, 2021.

[116] earth music&ecology. Retrieved from `https://stripe-club.com/earth1999/`, Accessed Jul. 21, 2021.

[117] Jia Chen, Ming Zhong, Jianxin Li, Dianhui Wang, Tieyun Qian, and Hang Tu. Effective Deep Attributed Network Representation Learning With Topology Adapted Smoothing. *IEEE Transactions on Cybernetics*, pages 1–12, 2021.

[118] Amazon.com, Inc., Amazon. Retrieved from `https://www.amazon.co.jp/`, Accessed Oct. 30, 2022.

[119] Snap Vision Ltd., Snap Vision. Retrieved from `https://www.snapfashion.com/`, Accessed Oct. 30, 2022.

[120] Xiaonan Zhao, Huan Qi, Rui Luo, and Larry Davis. A weakly supervised adaptive triplet loss for deep metric learning. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3177–3180, Oct. 2019.

[121] Debopriyo Banerjee, Lucky Dhakad, Harsh Maheshwari, Muthusamy Chelliah, Niloy Ganguly, and Arnab Bhattacharya. Recommendation of compatible outfits conditioned on style. In *Advances in Information Retrieval*, pages 35–50. Springer International Publishing, 2022.

[122] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1096–1104, Jun. 2016.

[123] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In *Proceedings of the Computer Vision – ECCV 2016*, volume 9905, pages 21–37. Springer International Publishing, Oct. 2016.

[124] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, volume 1, pages 91–99, Dec. 2015.

[125] Zhiyuan Liang, Kan Guo, Xiaobo Li, Xiaogang Jin, and Jianbing Shen. Person foreground segmentation by learning multi-domain networks. *IEEE Transactions on Image Processing*, 31:585–597, 2022.

[126] Yiheng Liu, Wengang Zhou, Jianzhuang Liu, Guo-Jun Qi, Qi Tian, and Houqiang Li. An end-to-end foreground-aware network for person re-identification. *IEEE Transactions on Image Processing*, 30:2060–2071, 2021.

[127] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, Jun. 2015.

[128] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, Jun. 2016.

[129] Suo Qiu. Global weighted average pooling bridges pixel-level localization and image-level classification. *CoRR*, abs/1809.08264, 2018.

[130] Xinguang Xiang, Yajie Zhang, Lu Jin, Zechao Li, and Jinhui Tang. Sub-region localized hashing for fine-grained image retrieval. *IEEE Transactions on Image Processing*, 31: 314–326, 2022.

[131] Shabnam Daghaghi, Tharun Medini, Nicholas Meisburger, Beidi Chen, Mengnan Zhao, and Anshumali Shrivastava. A tale of two efficient and informative negative sampling distributions. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 2319–2329, Jul. 2021.

[132] Gordon E. Moon, Denis Newman-Griffis, Jinsung Kim, Aravind Sukumaran-Rajam, Eric Fosler-Lussier, and P. Sadayappan. Parallel data-local training for optimizing word2vec embeddings for word and graph embeddings. In *Proceedings of the 2019 IEEE/ACM Workshop on Machine Learning in High Performance Computing Environments (MLHPC)*, pages 44–55, Nov. 2019.

[133] Stergios Stergiou, Zygimantas Straznickas, Rolina Wu, and Kostas Tsioutsiouliklis. Distributed negative sampling for word embeddings. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, volume 31, pages 2569–2575, Feb. 2017.

[134] Long Chen, Fajie Yuan, Joemon M. Jose, and Weinan Zhang. Improving negative sampling for word representation using self-embedded features. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 99–107, Feb. 2018. ISBN 9781450355810.

[135] Yongqi Zhang, Quanming Yao, Yingxia Shao, and Lei Chen. Nscaching: Simple and efficient negative sampling for knowledge graph embedding. In *Proceedings of the 2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 614–625, Apr. 2019.

[136] Zhen Yang, Ming Ding, Chang Zhou, Hongxia Yang, Jingren Zhou, and Jie Tang. Understanding negative sampling in graph representation learning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1666–1676, Jul. 2020. ISBN 9781450379984.

[137] Suhyeon Kim, Haecheong Park, and Junghye Lee. Word2vec-based latent semantic analysis (w2v-lsa) for topic modeling: A study on blockchain technology trend analysis. *Expert Systems with Applications*, 152:113401, 2020. ISSN 0957-4174.

[138] Shuying Liu and Weihong Deng. Very deep convolutional neural network based image classification using small training sample size. In *Proceedings of the 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 730–734, Nov. 2015.

[139] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE conference on computer vision and pattern recognition*, pages 248–255, Jun. 2009.

[140] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, Jun. 2018.

[141] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the Computer Vision – ECCV 2014*, pages 740–755, Sep. 2014.

[142] Ming Zhang, Vasile Palade, Yan Wang, and Zhicheng Ji. Attention-based word embeddings using artificial bee colony algorithm for aspect-level sentiment classification. *Information Sciences*, 545:713–738, 2021. ISSN 0020-0255.

[143] Sheng Jin, Shangchen Zhou, Yao Liu, Chao Chen, Xiaoshuai Sun, Hongxun Yao, and Xian-Sheng Hua. Ssah: Semi-supervised adversarial deep hashing with self-paced hard sample generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(7): 11157–11164, Feb. 2020.

[144] Jian Zhang and Yuxin Peng. Ssdh: Semi-supervised deep hashing for large scale image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(1):212–225, 2019.

[145] Luyao Liu, Xingzhong Du, Lei Zhu, Fumin Shen, and Zi Huang. Learning discrete hashing towards efficient fashion recommendation. *Data Science and Engineering*, 3: 307–322, 2018.

[146] Cairong Yan, Yizhou Chen, and Lingjie Zhou. Differentiated fashion recommendation using knowledge graph and data augmentation. *IEEE Access*, 7:102239–102248, 2019.

[147] Wei-Lin Hsiao, Isay Katsman, Chao-Yuan Wu, Devi Parikh, and Kristen Grauman. Fashion++: Minimal edits for outfit improvement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5046–5055, Oct.-Nov. 2019.

[148] Xuemeng Song, Fuli Feng, Xianjing Han, Xin Yang, Wei Liu, and Liqiang Nie. Neural compatibility modeling with attentive knowledge distillation. In *Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 5–14, Jul. 2018.

[149] Wang-Cheng Kang, Eric Kim, Jure Leskovec, Charles Rosenberg, and Julian McAuley. Complete the look: Scene-based complementary product recommendation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10524–10533, Jun. 2019.

[150] Mariya I. Vasileva, Bryan A. Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David Forsyth. Learning type-aware embeddings for fashion compatibility. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 405–421, Sep. 2018.

[151] Xingxing Zou, Kaicheng Pang, Wen Zhang, and Waikeung Wong. How good is aesthetic ability of a fashion model? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21200–21209, Jun. 2022.

[152] Wei-Lin Hsiao and Kristen Grauman. Creating capsule wardrobes from fashion images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7161–7170, Jun. 2018.

[153] Xue Dong, Xuemeng Song, Fuli Feng, Peiguang Jing, Xin-Shun Xu, and Liqiang Nie. Personalized capsule wardrobe creation with garment and user modeling. In *Proceedings of the ACM International Conference on Multimedia*, pages 302–310, Oct. 2019.

[154] Wen Chen, Pipei Huang, Jiaming Xu, Xin Guo, Cheng Guo, Fei Sun, Chao Li, Andreas Pfadler, Huan Zhao, and Binqiang Zhao. Pog: Personalized outfit generation for fashion recommendation at alibaba ifashion. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pages 2662–2670, Aug. 2019.

[155] Zunlei Feng, Zhenyun Yu, Yongcheng Jing, Sai Wu, Mingli Song, Yezhou Yang, and Junxiao Jiang. Interpretable partitioned embedding for intelligent multi-item fashion outfit composition. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 15(2s), Apr. 2019.

[156] Yuncheng Li, Liangliang Cao, Jiang Zhu, and Jiebo Luo. Mining fashion outfit composition using an end-to-end deep learning approach on set data. *IEEE Transactions on Multimedia*, 19(8):1946–1955, Aug. 2017. ISSN 1520-9210.

[157] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, May 2021.

[158] Yaoyu Li, Hantao Yao, Tianzhu Zhang, and Changsheng Xu. Part-based structured representation learning for person re-identification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 16(4), 2020.

[159] Jianyuan Guo, Yuhui Yuan, Lang Huang, Chao Zhang, Jin-Ge Yao, and Kai Han. Beyond human parts: Dual part-aligned representations for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3641–3650, Oct.-Nov. 2019.

[160] Zhichen Zhao, Huimin Ma, and Shaodi You. Single image action recognition using semantic body part actions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3411–3419, Oct. 2017.

[161] Xishan Zhang, Jia Jia, Ke Gao, Yongdong Zhang, Dongming Zhang, Jintao Li, and Qi Tian. Trip outfits advisor: Location-oriented clothing recommendation. *IEEE Transactions on Multimedia*, 19(11):2533–2544, 2017.

[162] Han Xintong, Wu Zuxuan, Wu Zhe, Yu Ruichi, and S. Davis Larry. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7543–7552, Jun. 2018.

[163] Haoye Dong, Xiaodan Liang, Yixuan Zhang, Xujie Zhang, Xiaohui Shen, Zhenyu Xie, Bowen Wu, and Jian Yin. Fashion editing with adversarial parsing learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8117–8125, Jun. 2020.

[164] Jianbin Jiang, Tan Wang, He Yan, and Junhui Liu. Clothformer: Taming video virtual try-on in all module. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10799–10808, Jun. 2022.

[165] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, pages 1857–1865, Dec. 2016.

[166] Jian Wang, Feng Zhou, Shilei Wen, Xiao Liu, and Yuanqing Lin. Deep metric learning with angular loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2612–2620, Oct. 2017.

[167] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1386–1393, Jun. 2014.

[168] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Jun. 2016.

[169] Hangbo Bao, Li Dong, and Furu Wei. BEiT: BERT pre-training of image transformers. In *Proceedings of the International Conference on Learning Representations (ICLR)*, May 2022.

[170] Ricard Lado-Roig é and Marco A. P é rez. STB-VMM: Swin Transformer based Video Motion Magnification. *Knowledge-Based Systems*, 269:110493, 2023. ISSN 0950-7051.

[171] Shuangquan Zuo, Yun Xiao, Xiaojun Chang, and Xuanhong Wang. Vision transformers for dense prediction: A survey. *Knowledge-Based Systems*, 253:109552, 2022. ISSN 0950-7051.

[172] Mingrui Ma, Yuanbo Xu, Lei Song, and Guixia Liu. Symmetric transformer-based network for unsupervised image registration. *Knowledge-Based Systems*, 257:109959, 2022. ISSN 0950-7051.

[173] Lisai Zhang, Hongfa Wu, Qingcai Chen, Yimeng Deng, Joanna Siebert, Zhonghua Li, Yunpeng Han, Dejiang Kong, and Zhao Cao. VLDeformer: Vision – Language Decomposed Transformer for fast cross-modal retrieval. *Knowledge-Based Systems*, 252: 109316, 2022. ISSN 0950-7051.

[174] Wenjun Zhu, Zuyi Wang, Li Xu, and Jun Meng. Exploiting temporal coherence for self-supervised visual tracking by using vision transformer. *Knowledge-Based Systems*, 251: 109318, 2022. ISSN 0950-7051.

[175] Yuhang Liu, Han Wang, Zugang Chen, Kehan Huangliang, and Haixian Zhang. TransUNet+: Redesigning the skip connection to enhance features in medical image segmentation. *Knowledge-Based Systems*, 256:109859, 2022. ISSN 0950-7051.

[176] Tao Ruan, Ting Liu, Zilong Huang, Yunchao Wei, Shikui Wei, and Yao Zhao. Devil in the details: Towards accurate single and multiple human parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4814–4821, Jan.-Feb. 2019.

[177] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3260–3271, 2022.

[178] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(4):871–885, 2019.

[179] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self Correction for Human Parsing. Retrieved from `https://github.com/GoGoDuck912/Self-Correction-Human-Parsing`. Accessed Oct. 30, 2022, 2019.

[180] Kalervo Järvelin and Jaana Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 41–48, (2000).

[181] Muhammet Bastan, Arnau Ramisa, and Mehmet Tek. T-vse: Transformer-based visual semantic embedding. In *CVPR 2020 Workshop on Computer Vision for Fashion, Art, and Design*, Jun. 2020.

[182] Zhou Ren, Hailin Jin, Zhe Lin, Chen Fang, and Alan Yuille. Joint image-text representation by gaussian visual-semantic embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 207–211, Oct. 2016.

[183] Yu Yang, Seungbae Kim, and Jungseock Joo. Explaining deep convolutional neural networks via latent visual-semantic filter attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8333–8343, Jun. 2022.

[184] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 139, pages 8748–8763, 2021.

[185] Luke Vilnis and Andrew McCallum. Word representations via gaussian embedding. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Jul. 2015.

[186] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[187] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 26, Dec. 2013.

[188] Chen Qian, Fuli Feng, Lijie Wen, and Tat-Seng Chua. Conceptualized and contextualized gaussian embedding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35 (15):13683–13691, 2021.

[189] Yujun Chen, Juhua Pu, Xingwu Liu, and Xiangliang Zhang. Gaussian mixture embedding of multiple node roles in networks. *World Wide Web*, 23, 03 2020.

[190] Shizhu He, Kang Liu, Guoliang Ji, and Jun Zhao. Learning to represent knowledge graphs with gaussian embedding. In *Proceedings of the ACM International on Conference on Information and Knowledge Management (CIKM)*, pages 623–632, Oct. 2015.

[191] Aleksandar Bojchevski and Stephan Günnemann. Deep gaussian embedding of graphs: Unsupervised inductive learning via ranking. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Apr. 2018.

[192] Qilong Wang, Peihua Li, and Lei Zhang. G2DeNet: Global gaussian distribution embedding network and its application to visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6507–6516, Jul. 2017.

[193] Yuying Zhao and Weihong Deng. Dual gaussian modeling for deep face embeddings. *Pattern Recognition Letters*, 2022.

[194] Yichun Shi and Anil Jain. Probabilistic face embeddings. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 6901–6910, Oct.-Nov. 2019.

[195] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5709–5718, Jun. 2020.

[196] Luyi Ma, Jianpeng Xu, Jason H.D. Cho, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. NEAT: A label noise-resistant complementary item recommender system with trustworthy evaluation. In *Proceedings of the IEEE International Conference on Big Data (Big Data)*, pages 469–479, Dec. 2021.

[197] Ludovic Dos Santos, Benjamin Piwowarski, and Patrick Gallinari. Gaussian embeddings for collaborative filtering. In *Proceedings of the International ACM SIGIR Conference*

on Research and Development in Information Retrieval (SIGIR), pages 1065–1068, Aug. 2017.

[198] Junyang Jiang, Deqing Yang, Yanghua Xiao, and Chenlu Shen. Convolutional gaussian embeddings for personalized recommendation with uncertainty. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2642–2648, Aug. 2019.

[199] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Oct. 2014.

[200] Tanmoy Mukherjee and Timothy Hospedales. Gaussian visual-linguistic embedding for zero-shot recognition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 912–918, Nov. 2016.

[201] Jeff Pasternack and Dan Roth. The wikipedia corpus. 2008.

[202] DC Dowson and BV666017 Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

[203] Asuka Takatsu. On wasserstein geometry of gaussian measures. *Probabilistic approach to geometry*, 57:463–472, 2010.

[204] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2000.

[205] Shun-ichi Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

[206] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128:1956–1981, 2020.

[207] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial

networks. In *Proceedings of the International Conference on Machine Learning*, pages 214–223, Aug. 2017.

[208] Soheil Kolouri, Phillip E Pope, Charles E Martin, and Gustavo K Rohde. Sliced wasserstein auto-encoders. In *Proceedings of the International Conference on Learning Representations*, Apr. 2018.

Research Achievements

Type	Title, Name of Journal, Date of Publication, Co-author
Article	（corresponding-author）
◎	Partial Visual-Semantic Embedding: Fashion Intelligence System with Sensitive Learning, Knowledge-Based Systems, （Under Review） April 2023, Ryotaro Shimizu, Takuma Nakamura, Masayuki Goto
◎	Fashion Intelligence System: An Outfit Interpretation Utilizing Images and Rich Abstract Tags, Expert Systems with Applications, Vol.213(C), 119167, March 2023, Ryotaro Shimizu, Yuki Saito, Megumi Matsutani, Masayuki Goto
◎	Fashion-Specific Attributes Interpretation via Dual Gaussian Visual-Semantic Embedding, IEEE Transactions on Image Processing, （Under Review） November 2022, Ryotaro Shimizu, Masanari Kimura, Masayuki Goto
◎	An Explainable Recommendation Framework Based on an Improved Knowledge Graph Attention Network with Massive Volumes of Side Information, Knowledge-Based Systems, Vol.239, 107970, March 2022, Ryotaro Shimizu, Megumi Matsutani, Masayuki Goto
	Latent Variable Models for Integrated Analysis of Credit and Point Usage History Data on Rewards Credit Card System, International Business Research, Vol.13(3), pp.106-117, February 2020, Ryotaro Shimizu, Haruka Yamashita, Masao Ueda, Ranna Tanaka, Tetsuya Tachibana, Masayuki Goto

Type	Title, Name of Journal, Date of Publication, Co-author
	Proposal of a Purchase Behavior Analysis Model on an Electronic Commerce Site Using Questionnaire Data, Total Quality Science, Vol.4(1), pp.1-12, December 2018, Ryotaro Shimizu, Teppei Sakamoto, Haruka Yamashita, Masayuki Goto
	How Did the 2015 Political Crisis Affect Nepal in Economic and Social Respects?, Horizon Research Publishing Corporation, Vol.6(6), pp.571-582, 2018 年 10 月, Ryotaro Shimizu, Brenda Bushell, Masayuki Goto
	アンケートデータを考慮した EC サイトの購買履歴分析モデルの提案, 経営システム, Vol.27(2), pp.70-75, 2017 年 7 月, 清水良太郎, 坂元哲平, 山下遥, 後藤正幸
Itn CF ◎	（corresponding-author） Fashion-Specific Ambiguous Expression Interpretation with Partial Visual-Semantic Embedding, The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023 - 6th Workshop on Computer Vision for Fashion, Art, and Design, June 2023, Ryotaro Shimizu, Takuma Nakamura, Masayuki Goto
	Consumer Purchasing behavior model for Purchase History Data Stored in Credit Card, The 19th Asia Pacific Industrial Engineering and Management System Conference, December 2018, Ryotaro Shimizu, Haruka Yamashita, Ranna Tanaka, Masayuki Goto
	Proposal of a Purchase Behavior Analysis Model on EC Site Considering Questionnaire Data, The 15th Asian Network for Quality Congress, pp.98, September 2017, Ryotaro Shimizu, Teppei Sakamoto, Haruka Yamashita, Masayuki Goto

Type	Title, Name of Journal, Date of Publication, Co-author
	Research on the Political Crisis and its Impact on the Economic and Social Sectors in Nepal, 22nd International Interdisciplinary Conference on the Environment, June 2016, Ryotaro Shimizu, Masayuki Goto, Brenda Bushell
Domestic CF	（corresponding-author）
◎	ファッション特有の曖昧な表現を解釈するFashion Intelligence Systemの応用と今後の展開, 第37回人工知能学会全国大会, 2023年6月, 清水良太郎, 斎藤侑輝, 後藤正幸
◎	機械学習に基づくファッション特有の曖昧な表現を自動的に解釈するためのシステム, 日本経営工学会2022年秋季大会, 2022年11月, 清水良太郎, 斎藤侑輝, 松谷恵, 後藤正幸
	外部ドメインデータを活用した潜在顧客発見に向けた取り組み, Conference, on Computer Science for Enterprise 2021, 2021年12月, 清水良太郎, 柳圭祐
◎	ファッション系ECサイトにおける多様な補助情報を有したグラフ構造の学習アルゴリズムに関する一考察, 日本経営工学会2021年春季大会, 2021年5月, 清水良太郎, 松谷恵, 後藤正幸
	クレジットとポイントを併用可能な多機能クレジットカードにおける利用履歴データの統合分析モデルの提案, 第41回 情報理論とその応用シンポジウム, pp.442-447, 2018年12月, 清水良太郎, 山下遥, 上田雅夫, 田中藍奈, 後藤正幸
	ECサイトにおける購買履歴データとアンケートデータを融合した顧客の購買行動分析モデルの提案, 日本経営工学会2017年度学生発表会, pp.82-83, 2018年3月, 清水良太郎, 坂元哲平, 山下遥, 後藤正幸

Type	Title, Name of Journal, Date of Publication, Co-author
	アンケートデータを考慮したECサイトの購買履歴行動分析モデルの提案, 日本計算機統計学会第31回シンポジウム, pp.7-12, 2017年11月, 清水良太郎, 坂元哲平, 山下遥, 後藤正幸 アンケートデータを考慮したECサイトの購買履歴行動分析モデルの提案, 日本経営工学会2017年春季大会, pp.76-77, 2017年5月, 清水良太郎, 坂元哲平, 山下遥, 後藤正幸 ECサイトにおけるアンケートデータを考慮した購買行動分析モデルの提案, 平成28年度データ解析コンペティションJIMA予選会, 2017年2月, 清水良太郎, 坂元哲平, 山下遥, 後藤正幸
Lecture	（corresponding-author） 企業の研究における研究計画の作り方, Conference, on Computer Science for Enterprise 2021, 2021年12月, 清水良太郎 クレジットとポイントを併用可能な多機能クレジットカードにおける利用履歴データの統合分析モデルの提案, 価値創造マネジメントシンポジウム, 2018年12月, 清水良太郎, 山下遥, 上田雅夫, 田中藍奈, 後藤正幸
Article	（co-author） 最大次数が未知の多項式回帰モデルに対するスパース推定に関する一考察, 情報処理学会論文誌, (Under Review) 2023年5月, 井上一磨, 清水良太郎, 須子統太, 後藤正幸 ユーザの多様性を考慮したクラスタワイズ型XGBoostモデルの提案とその解釈方法に関する研究, 情報処理学会論文誌, (Under Review) 2023年5月, 三橋可奈, 清水良太郎, 山下遥

Type	Title, Name of Journal, Date of Publication, Co-author
	知識グラフに基づく説明可能な推薦のための効率的な学習アルゴリズムに関する一考察, 日本経営工学会論文誌, (Under Review) 2023 年４月, 楊冠宇, 清水良太郎, 山極綾子, 後藤正幸 Estimation of the Effects of the Multiple Treatment for Business Analytics Using Observational Data, Neural Computing & Applications, (Under Review) March 2023, Yuki Tsuboi, Yuta Sakai, Ryotaro Shimizu, Masayuki Goto Generation of Attractive Fashion Outfit Images Using Conditional StyleGan2-ADA Considering User Attribute Information, Fashion & Textile, (Under Review) March 2023, Oike Tatsuki, Haruka Yamashita, Ryotaro Shimizu Recommendation Item Selection Algorithm Considering the Recommendation Region in Embedding Space and New Evaluation Metric, Industrial Engineering & Management Systems, (Under Review) January 2023, Tomoki Amano, Ryotaro Shimizu, Masayuki Goto Hidden Semi-Markov Model-Based Advanced Analysis Method for Item Browsing Data Considering Duration of User Interests, Industrial Engineering & Management Systems, (Under Review) January 2023, Kirin Tsuchiya, Yuki Tsuboi, Ryotaro Shimizu, Masayuki Goto A Latent Class Analysis for Item Demand Based on Temperature Difference and Store Characteristics, Industrial Engineering & Management Systems, Vol.20(1), pp.35-47, March 2021, Yuto Seko, Ryotaro Shimizu, Gendo Kumoi, Tomohiro Yoshikai, Masayuki Goto
Itn CF	（co-author）

Type	Title, Name of Journal, Date of Publication, Co-author
	Multiple Treatment Effect Estimation for E-commerce Marketing Using Observational Data, The 22nd Asia Pacific Industrial Engineering and Management System Conference, November 2022, Yuki Tsuboi, Yuta Sakai, Ryotaro Shimizu, Masayuki Goto Extraction of fashion themes focusing hashtags based on Guided LDA, The 22nd Asia Pacific Industrial Engineering and Management System Conference, November 2022, Hiroya Furuta, Haruka Yamashita, Ryotaro Shimizu Recommendation Item Selection Algorithm Considering the Recommendation Region in the Embedding Space, The 22nd Asia Pacific Industrial Engineering and Management System Conference, November 2022, Tomoki Amano, Ryotaro Shimizu, Masayuki Goto An Efficient Path Search Algorithm for Explainable Recommendation Based on Knowledge Graph and Reinforcement Learning, The 22nd Asia Pacific Industrial Engineering and Management System Conference, November 2022, Guanyu Yang, Ryotaro Shimizu, Ayako Yamagiwa, Masayuki Goto A study on generation of images with many likes by Conditional StyleGAN2-ada considering user attribute information, The 20th Asian Network for Quality Congress, October 2022, Oike Tatsuki, Haruka Yamashita, Ryotaro Shimizu A study on cluster-wise XGBoost model considering user diversity and its interpretation approach, The 20th Asian Network for Quality Congress, October 2022, Kana Mitsuhashi, Haruka Yamashita, Ryotaro Shimizu

Type	Title, Name of Journal, Date of Publication, Co-author
Domestic CF	A method for item analysis considering duration of user interests based on a hidden semi-markov model, The 20th Asian Network for Quality Congress, October 2022, Kirin Tsuchiya, Yuki Tsuboi, Ryotaro Shimizu, Masayuki Goto A discussion on improving fraud detection performance by Generative Adversarial Networks Transactions, The 19th Asian Network for Quality Congress, October 2021, Guanyu Yang, Yuki Tsuboi, Ryotaro Shimizu, Gendo Kumoi, Masayuki Goto A Prediction Model of Item Demands Based on Temperature Difference and Store Characteristics, The 16th Asian Network for Quality Congress, pp.200, September 2018, Yuto Seko, Ryotaro Shimizu, Gendo Kumoi, Masayuki Goto, Tomohiro Yoshikai （co-author） ユーザの潜在的な購買意欲を考慮した機械学習モデルに基づくクーポン配布施策の効果検証モデル, 第37回人工知能学会全国大会, 2023年6月, 米田安希子, 清水良太郎, 桜井詩音, 川田心, 山下遥, 後藤正幸 画像情報及び言語情報に基づくファッションコーディネート投稿の推薦, 第37回人工知能学会全国大会, 2023年6月, 岩井理紗, 山下遥, 清水良太郎 FT-Transformerの精度向上と効率化に関する一考察, 第37回人工知能学会全国大会, 2023年6月, 磯村時将, 天野智貴, 清水良太郎, 後藤正幸 レビュー文書データを対象としたBERTとSHAPによる評価値向上要因分析モデル, 第37回人工知能学会全国大会, 2023年6月, 渡邊真己子, 山田晃輝, 清水良太郎, 鈴木佐俊, 後藤正幸

Type	Title, Name of Journal, Date of Publication, Co-author
	複数のECマーケティング施策を対象とした観察データに基づく効果推定手法, 情報理論とその応用シンポジウム, 2022年11月, 坪井優樹, 阪井優太, 清水良太郎, 後藤正幸 知識グラフと強化学習に基づく説明可能な推薦のための効率的な経路探索アルゴリズム, 情報理論とその応用シンポジウム, 2022年11月, 楊冠宇, 清水良太郎, 山極綾子, 後藤正幸 埋め込み空間における推薦領域を考慮した推薦アイテム獲得手法の提案, 日本経営工学会2022年 秋季大会, 2022年11月, 天野智貴, 清水良太郎, 後藤正幸 Guided LDAによるファッションテーマの抽出及び推薦への応用, 日本経営工学会2022年 秋季大会, 2022年11月, 古田博也, 山下遥, 清水良太郎 CNNを用いた能動学習におけるラベル付与データの選択手法に関する一考察, 日本計算機統計学会 第36回シンポジウム, 2022年11月, 益田恵里花, 山田晃輝, 清水良太郎, 後藤正幸 隠れセミマルコフモデルに基づくユーザの嗜好持続性を考慮した商品分析手法に関する一考察, 2022年度 人工知能学会全国大会（第36回）, 2022年6月, 土屋希琳, 坪井優樹, 清水良太郎, 後藤正幸 複数の施策を対象とした処置効果推定手法に関する一考察, 情報処理学会 第84回全国大会, 2022年3月, 坪井優樹, 阪井優太, 清水良太郎, 後藤正幸 Conditional StyleGAN2-adaによるユーザの属性情報を考慮した高評価画像の生成に関する研究, 情報処理学会 第84回全国大会, 2022年3月, 大池樹, 清水良太郎, 山下遥

Type	Title, Name of Journal, Date of Publication, Co-author
	ユーザの多様性を考慮したクラスタワイズ型機械学習モデルの提案とその解釈方法に関する研究, 情報処理学会 第 84 回全国大会, 2022 年 3 月, 三池可奈, 清水良太郎, 山下遥 対照群のデータを考慮した Transformed outcome method によるコンプライアーの予測とその評価に関する研究, 日本経営システム学会第 67 回全国研究発表大会, 2021 年 11 月, 夏堀雄太, 清水良太郎, 山下遥 仮想通貨取引データに対する敵対的生成ネットワークを用いた分類性能向上手法の検討, 日本経営工学会 2021 年 春季大会, 2021 年 5 月, 楊冠宇, 清水良太郎, 雲居玄道, 後藤正幸 最大次数が未知の多項式回帰モデルに対するスパース推定に関する一考察, 情報論的学習理論と機械学習研究会, 2018 年 11 月, 井上一磨, 清水良太郎, 須子統太, 後藤正幸 気象条件と店舗特性を考慮した商品別需要モデル構築に関する一考察, 日本経営工学会 2018 年春季大会, pp.8-9, 2018 年 5 月, 世古裕都, 清水良太郎, 雲居玄道, 後藤正幸, 吉開朋弘 Tweet データに基づく料理画像の魅力度定量化モデル, 日本経営工学会 2018 年春季大会, pp.66-67, 2018 年 5 月, 藤波英輝, 清水良太郎, 雲居玄道, 後藤正幸