Ph.D. DISSERTATION

Cross-Modal Representation Learning : Joint and Distributed Embedding

공동 임베딩 및 분산 임베딩 방식을 통한 교차 모달 표현 학습 방법 연구

BY

Dae Ung Jo

FEBRUARY 2022

DEPARTMENT OF ELECTRICAL AND
COMPUTER ENGINEERING
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

Cross-Modal Representation Learning : Joint and Distributed Embedding

공동 임베딩 및 분산 임베딩 방식을 통한 교차 모달 표현 학습 방법 연구

지도교수 최 진 영

이 논문을 공학박사 학위논문으로 제출함

2021 년 10 월

서울대학교 대학원

전기 컴퓨터 공학부

조 대 웅

조 대 웅의 공학박사 학위논문을 인준함

2021 년 12 월

위 원 장	조 남 익
부위원장	최 진 영
위 원	곽 노 준
위 원	정 교 민
위 원	최 종 원

Abstract

In this dissertation, we propose two methods to overcome problems that may occur in cross-modal representation learning. First, in order to overcome the problem that the existing joint embedding based model is difficult to learn relation among data from heterogeneous modalities, we propose a cross-modal representation learning model adopting the distributed embedding method. The proposed model first learns intra-modal association by training a specialized embedding space for each modality with single-modal representation learning. Then the proposed model learns cross-modal association by introducing associator, which connects the embedding spaces of multiple modalities. To separate the learning process of intra-modal association and cross-modal association, the model parameters involved in intra-modal association are not updated during training of cross-modal association. Through the two-step learning process, the proposed model can well perform cross-modal representation learning among heterogeneous modalities. Furthermore, the proposed model has the advantage of utilizing unpaired data for learning. We validated the proposed method in the cross-modal data generation task between visual and auditory modalities, which is one of the heterogeneous modal relationships. The proposed method achieves improved performance compared to the existing joint-embedding based models.

Second, though cross-modal paired data is essential for cross-modal representation learning, securing a sufficient number of paired data is too difficult in practical applications. To mitigate data shortage problem, we propose an active learning method for cross-modal representation learning. In particular, we propose active learning for image-text retrieval, which is one of the most popular applications related to cross-modal representation learning. Since the existing active learning scenario for image-text retrieval can not be applied to the recent image-text retrieval benchmarks, we first propose an active learning scenario feasible for the recent benchmarks. In contrast to the existing scenario where a category label for a given image-text pair data is queried to the human experts, in the proposed scenario, unpaired image or text data are given and human experts are requested to pair the unpaired data. We also proposed an active learning algorithm for the proposed scenario. The proposed algorithm selects the data that is expected to have the most influence on the max-hinge triplet loss function, which is mainly adopted loss function in recent image-text retrieval method. To this end, we define the condition that data can influence the loss function, and estimate the influence score (referred to as HN-Score) of the data on the loss function based on the defined condition. The proposed algorithm selects the data of the highest score. We validate the effectiveness of the proposed active learning algorithm through the various experiments on recent image-text retrieval benchmarks.

Keywords: cross-modal representation learning, distributed embedding, active learning, deep neural network

Student Number: 2016-20977

Contents

List of Figures

List of Tables

Chapter 1

Introduction

As humans utilize the multi-sensory signals to think and judge something in their daily life, learning multi-modal data is also very important in machine learning area. The most basic step to understand multi-modal data is to learn the relationship among multi-modal data, referred to as multi-modal association learning. Through multi-modal association learning, useful machine learning tasks such as cross-modal generation or cross-modal retrieval can be performed. Multi-modal association learning can be realized through cross-modal representation learning. In cross-modal representation learning, multi-modal data are embedded into a latent space, while semantically correlated data are embedded closely to each other in the embedding space.

Many studies have been proposed for cross-modal representation learning. Those studies have achieved a great improvements in various machine learning tasks such as cross-modal data generation [1–9], cross-modal data retrieval [10–15, 15–20] and recognition [5, 21]. Nevertheless, learning cross-modal representations is still one of the difficult topics in machine learning area. In this dissertation, we focused on the two major problems of cross-modal representation learning and proposed methods to mitigate those problems.

The first problem is that learning cross-modal representations among heterogeneous modalities is very difficult. Note that we referred to modalities with very different characteristics as heterogeneous modality. Learning between vision and auditory modality can be an example. The word *know* and *no* have similar pronunciation, but they have different meanings. Like this, similarity in the auditory domain does not guarantee similarity in the visual domain, it may be difficult to learn correlations for these modals. We think that the existing joint embedding method, which projects multi-modal data into one embedding space, have a limitation to solve the aforementioned problem. In the joint embedding method, paired data from heterogeneous modality should be close to each other in the joint embedding space. However, as mentioned above, since similarity in one modality does not guarantee similarity in another modality, forcing close embedding of heterogeneous data can be problematic. To mitigate the limitation of the joint embedding method, we propose an approach that adopts distributed embedding spaces. In proposed approach, each modality is encoded in each embedding space separately by the variational auto-encoder (VAE) [22] and the distributed embedding spaces are associated with the other modalities via associators. Through the proposed structure, not only the relationship among heterogeneous modality can be learned, but also unpaired data can be utilized for learning.

The second problem is that securing a sufficient amount of paired data for cross-modal representation learning is difficult. Paired data is essential for cross-modal representation learning. However, collecting paired data requires a lot of money and time than single modal data, since paired data should be collected from multiple sources and the correlation between them should be checked. Therefore, in actual machine learning application, it is difficult to secure sufficient amount of paired data for cross-modal representation learning. To mitigate the data shortage problem, we employed active learning concept to multi-modal data to cost-efficiently collect the paired data, especially for image-text retrieval task. Since active learning studies for image-text retrieval is not such exploited, we defined the active learning scenario for the multi-modal dataset with unpaired data first. Then we propose active learning algorithm for the developed scenario. Inspired by the loss function usually adopted in image-text retrieval research [14], the proposed method select the samples which are the most likely to be a hard negative sample for the existing paired dataset. To this end, we defined the conditions for data to be a hard negative for a existing paired data, and estimated scores for data by reflecting that the given data satisfies defined conditions.

The contribution of this dissertation is to propose methods that can complement the two problems of cross-modal representation learning. First, for robust cross-modal association learning among heterogeneous modalities, we proposed a distributed embedding method that allocates an embedding space separately for each modality rather than the existing joint embedding method. Second, to mitigate the data shortage problem in multi-modal task, we proposed an active learning scenario and algorithm for cross-modal representation learning method to obtain paired data cost-effectively, especially for image-text retrieval.

Chapter 2

Preliminary

2.1 Associative Learning in Human Brain

There have been many studies that investigate associative learning from the perspective of neuroscience [23–26]. One of the most popular study about the associative learning is Pavlov's work [27, 28]. In their experiment, the unconditioned stimulus is to give a dog some food that make him salivate, while the neutral stimulus is to let the dog hear sounds of a metronome. If food and sounds are provided simultaneously and consistently, the dog learns association between the food and the sound. Then the dog salivates only by hearing the sound of a metronome even though the sounds of a metronome is a neutral stimulus and salivating is an unconditional stimulus.

In the recent study which tried to analyze associative learning at the cellular substrate level [29, 30], they introduce the associative memory cells to describe brain neurons which are mainly involved in integration and storage of associated signals. A brain learns associated information by enhancing the strength of the synapses between coactivated associative memory cells activated by associated signals. According to [29],

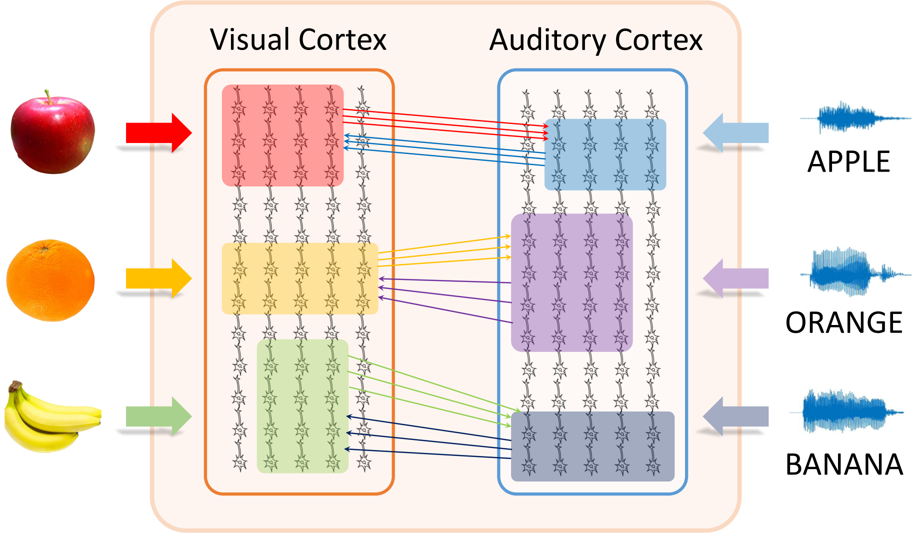


Figure 2.1: **Example of how the brain memorizes multi-modal associative data** Intra-modal associative memory cells (highlighted cells) in the visual cortex are trained to discriminate various sensory data (images of an apple, an orange and a banana) through their mutual innervations. Intra-modal associative memory in the auditory cortex is also trained in the same manner. When associative information between visual and auditory data is provided, co-activation of associative memory cells induces mutual synapse innervations between associative memory cells, and thus, cross-modal associative memory cells are trained.

detailed associative learning process in the brain includes intra-modal and cross-modal association processes. Figure 2.1 illustrates the two associative learning processes. The intra-modal association process is to make humans familiar with single-modal sensory information. An example of single modal associative learning is the process of remembering the image of a fruit in the visual cortex while observing the fruit. Then, without any help from a teacher, the image is memorized by itself and becomes familiar to the person. This is an unsupervised learning process. The process of remembering the name of the fruit and memorizing it in the auditory cortex works in a similar manner. In the cross-modal association process, when the image and name of the fruit enter into the sensory organs at the same time, the process of learning the relationship between them proceeds.

The conventional artificial neural network is an engineering model inspired by the biological mechanism of the brain. Parameters of those networks are usually updated by Hebbian learning rule where weight connections between firing nodes for input data are strengthened [31]. The Hopfield network and Boltzmann machine are representative examples [32]. The Hopfield network models associative memory of human, thus network is trained to memorize specific patterns. Even if the input is incomplete, The Hopfield network can restore incomplete data through recurrent iteration. The Boltzmann machine is a stochastic version of the Hopfield network, which can learn a latent representation for input data through its hidden nodes.

2.2 Cross-modal Representation Learning

One of the major issues in machine learning is exploiting multi-modal data for various applications, such as data generation [3, 22, 33, 34], retrieval [10] and recognition [5, 21]. There are a lot of studies that extract modality independent cross-modal representation by finding the joint representation of multi-modal data [1]. The joint representation is utilized in diverse applications such as handling a missing modality [3, 4] or accomplishing better performance than models trained on single-modal data [5, 21].

2.2.1 Cross-modal Data Generation

The research related to cross-modal data generation can be categorized into two groups [1]. One is a method mapping data from diverse modalities to the joint latent space.

[2] proposes the extended version of a Variational auto-encoder [22] which combines distribution parameters from encoders and calculate integrated distribution parameters. [3] also a variant of Variational auto-encoder for hand pose estimation with multimodal data. The model proposed by [3] chooses the input modality and the output modality pair and train the corresponding encoder and decoder pair at every iteration. [4] trains an auto-encoder that takes RGB images, depth images and semantic images as its network input, then the trained model can a generate complete depth image and semantic image from an RGB image and partial depth and semantic image. [21] builds a deep-belief network structure that maps audio data and lip images into the common hidden node for audio-visual speech recognition. [5] extends the RBM structure to reflect the sequential characteristic of a speech dataset.

The other group comprises methods that encode the corresponding data to the latent space of each modality but enforce similarity constraints to corresponded latent vectors. [7] trains domain specific encoders and decoders, allowing encoders and decoders from different modality to be combined, then, the model is able to generate an unseen data pair by combining the encoders and decoders. [8] extracts low-level representation from original data first. Then they trains auto-encoders for each modality and enforces similarity constraints to embedding spaces of each auto-encoders for correlated data pair. In [9], a model is trained to maximize the similarity of an image feature and a vectorized label to infer a proper label for a given image.

2.2.2 Image-text Retrieval

Image-text retrieval (ITR) is an popular machine learning application where a model retrieves the most semantically relevant text (image) in the data base when a query image (text) is given. VSE++ [14] is the most popular algorithms for find-grained ITR. VSE++ extracts feature from image and text, then estimates the similarity between image and text through inner-product. Then a retrieval model is trained by the hinge-based triplet ranking loss [35–38]. However, when training a model, a mini-batch from an find-grained ITR benchmark includes one relevant (positive) sample and a number of irrelevant (negative) samples. Thus gradient can be biased to the negative term of the triplet loss. To mitigate this problem, VSE++ only reflects the hard negative sample in the mini-batch to the loss term. SCAN [15] improves the retrieval performance by estimating the similarity between image and text more precisely. Whereas VSE++ extracts the one global feature from image and text data, SCAN extracts lots of local features from sub-regions of image [39] and words in text. Then SCAN calculates similarity between each region (word) and full text (image) and aggregates them. Such algorithm is referred to as fine-grained ITR algorithm. IMRAM [16] refines local features by iteratively fusing local features with the proposed memory units. In addition to the aforementioned studies, numerous studies have been proposed [15, 17–20].

2.3 Active Learning

2.3.1 Pool-based Active Learning

Recently in the deep learning field, active learning has emerged as one way to efficiently collect supervised data. The key idea of active learning is to efficiently improve the performance of the target model by actively selecting the data to be labeled through an algorithm, rather than randomly selecting the data to be labeled. Figure 2.2 shows the general process of active learning. Through active learning, it is expected to achieve better performance even if we pay the same amount of annotation budget compared to random selection (in several papers, random selection is referred to as passive learning).

There are many scenarios for active learning [40], but recent studies have mainly considered a pool-base active learning scenario. In the pool-based active learning, lots of labeled data and unlabeled data are given, and the machine queries a large number

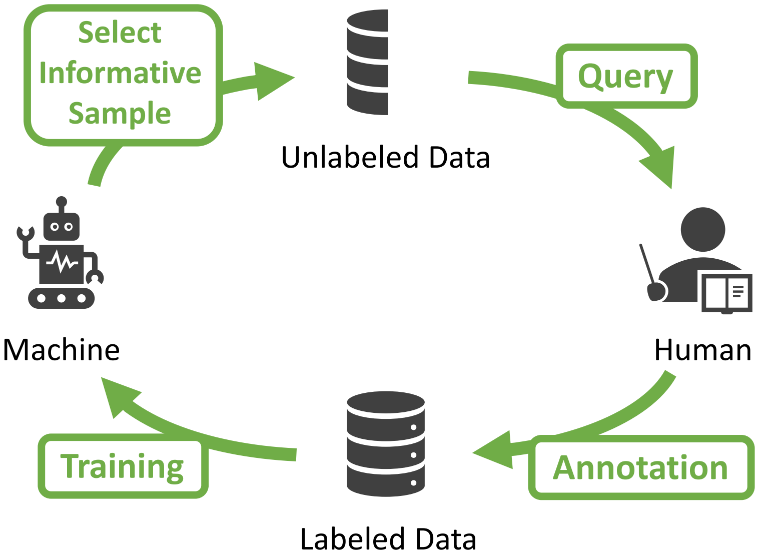


Figure 2.2: **General process of active learning.** Given labeled data and unlabeled data, the machine is trained with labeled data. Then machine selects the set of informative samples from the unlabeled data which are expected to improve the performance of the machine when human annotates them. The machine queries selected samples for human to annotate them. Finally, human annotates the queried samples and incorporate them into an existing labeled dataset. Those processes are repeated until the annotation budget is exhausted or target performance is achieved.

of informative samples to humans at once. Since deep learning algorithm requires a huge amount of training data, pool-based scenario is generally considered. The most important part of active learning is how to select the most informative samples from the unlabeled dataset. To this end, many studies are trying to design a function referred to as a selection function or an acquisition function, in consideration with correlation between unlabeled sample and labeled dataset, model prediction, additional module, and others.

In addition, most of the active learning studies consider the single-modal case. In the single-modal case, unlabeled data is simply composed of one data, and a human provides an annotation according to a task (e.g. class label, bounding box, segmentation ...). The most representative validation task for active learning is image classification. Most of the recent AL studies validated their algorithm on binary or multi-class classification problem.

2.3.2 Single-modal Active Learning

Recently, the active learning [40] has been applied to various deep learning tasks such as object detection [41, 42], person re-identification [43], multi-task learning [44], named entity recognition [45], human pose estimation [46], action localization [47], and biomedical image analysis [48, 49]. Recent active learning methods can be categorized into two types [50]. The one is the uncertainty-sampling methods [41–43, 45, 48, 51–53] which select the most uncertain (also referred to as informative) samples for the target model. The other is representative subset methods [54, 55], which select the representative subset from the unlabeled data pool.

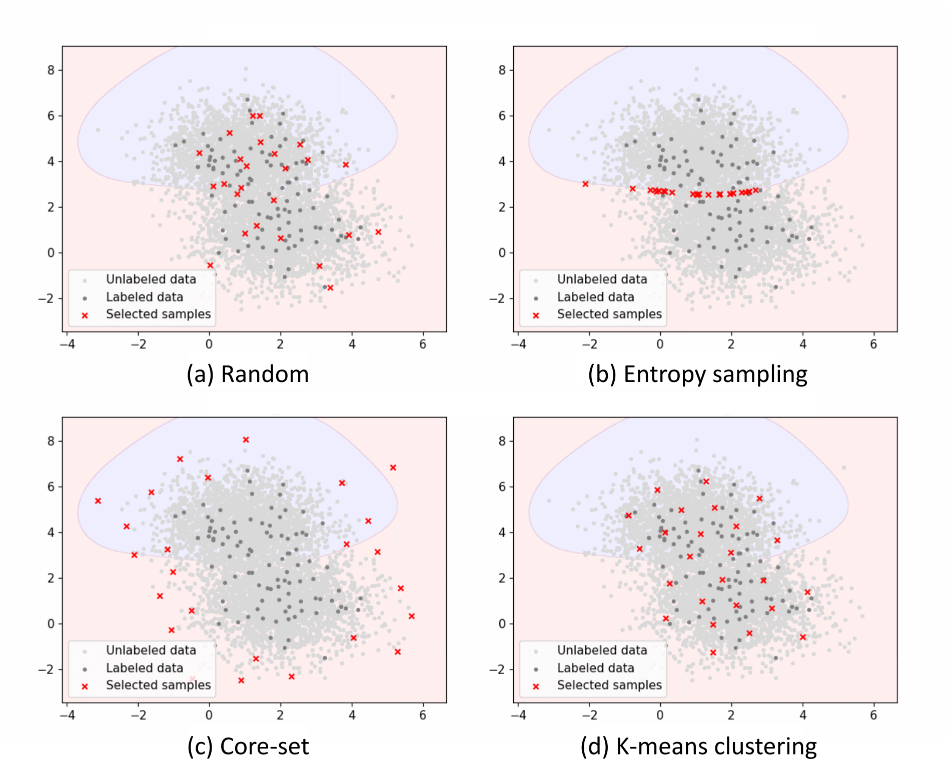
The core of the uncertainty-sampling methods is to estimate the uncertainty information for the unlabeled sample. In case the target model infers a probability distribution (e.g., image classification), classical methods such as entropy [56] or variational ratio [57] of probability distribution can be used as uncertainty estimators. Despite its simplicity, classical methods are still utilized in many deep learning applications and show prominent performance [41, 45–48, 53, 58].

However, most deep learning applications, such as object detection [59] or human-pose estimation [60], infer deterministic results, instead of probabilistic results. In addition, the classical uncertainty-based active learning methods have a problem with scalability to high dimensional data and a huge number of model parameters [51]. Therefore, recent studies attempt to efficiently estimate the sample-uncertainty in a deep model. Dropout-based method [51] performs multiple forward passes with dropout layers [61] to predict the sample uncertainty. Ensemble-based method [52] utilizes multiple deep neural networks, which have the same structure but are differently initialized. The method proposed in [42] estimates the sample-uncertainty by predicting the loss value of the sample. To this end, they add the additional network (loss prediction module) to the target model. Since the loss function should be defined in any deep learning tasks, this method can be applied to any deep learning tasks.

The representative subset methods select the representative subset from the unlabeled data pool. Core-set approach [54] formulated the active learning problem as k-Center problem [62]. The goal of the k-Center problem is to select the k points that maximize the minimum distance among the selected points and its nearest centers. Then the Core-set method solves the problem via integer programming. Variational adversarial active learning method [55] selects unlabeled data that are not similar to labeled data by training VAE [22] and discriminator adversarially. k-centered clustering algorithms such as k-medoid clustering [63] can be utilized for selecting representative samples by choosing cluster-center [64].

Some studies [65, 66] have attempted to combine the two strategies mentioned before. The method proposed in [65] selects samples which have high uncertainty while preserving the distribution of the dataset. The method proposed in [66] improves its original one [65] by considering the easiness of a sample.

In Figure 2.3 and 2.4, we present visualized results of applying several active learning algorithms to the 2-dimensional binary classification problem with a SVM classifier. Red cross, light-gray dots, and dark-gray dots represent selected samples, unlabeled samples, and labeled samples, respectively. The colored area denotes the predicted area of the classifier. For the Entropy sampling, most of the samples are selected near the classification boundary. In contrast, Core-set and K-means clustering method select evenly across the entire data set.



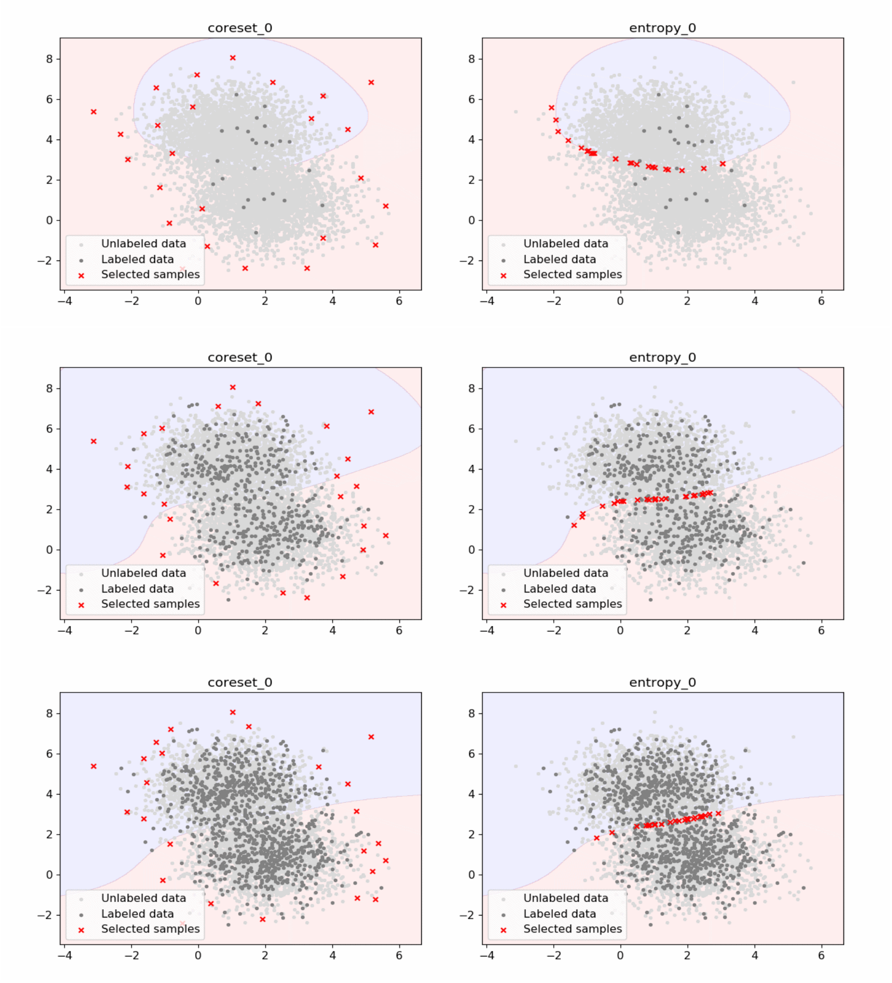
(a) Random

(b) Entropy sampling

(c) Core-set

(d) K-means clustering

Figure 2.3: An example of sample selection based on four query strategy (Random sampling, Entropy sampling, Core-set, and K-means clustering) in a 2-dimensional binary classification problem with a SVM classifier. Red cross, light-gray dots, and dark-gray dots represent selected samples, unlabeled samples, and labeled samples, respectively. The colored area denotes the predicted area of the classifier.



(a) Core-set

(b) Entropy sampling

Figure 2.4: An example of sample selection based on two query strategy (Entropy sampling, Core-set) in a 2-dimensional binary classification problem with a SVM classifier depending on the number of initial labeled data. Red cross, light-gray dots, and dark-gray dots represent selected samples, unlabeled samples, and labeled samples, respectively. The colored area denotes the predicted area of the classifier.

2.3.3 Multi-modal Active Learning

Collecting supervised data for multi-modal tasks such as image-text retrieval is more difficult and costly than for single-modal tasks. Therefore, AL is a more attractive research topic for multi-modal tasks than for single-modal tasks, but AL for deep multi-modal tasks have not been exploited much. [67] proposed COSLAQ, an AL algorithm for coarse-grained ITR based on [11]. Given two relevant image-text pair, COSLAQ calculates intra-modal (image-image and text-text) and cross-modal (image-text) similarities between them. Then algorithm measures variance of similarities and selects two relevant image-text pairs that have the highest variance of similarities. Note that COSLAQ queries human experts whether two pairs belong to the same class or not. However, COSLAQ can not be applied to find-grained ITR since each relevant pairs in find-grained ITR dataset belongs to different classes. [68] proposed MMQL, an AL algorithm for multi-modal classification based on reinforcement learning. In a training phase, MMQL trains individual network for each modality. The results from each modality network are concatenated and treated as state for reinforcement learner. Then reinforcement learner learns a binary action whether to query the input multi-moal data to a human or not.

Chapter 3

Distributed Embedding Model

3.1 Contribution

The brain combines multisensory information to understand the surrounding situation. Through various sensory experiences, humans learn the relationships between multisensory data and understand the experienced situation. This mechanism to learn the relationship among multiple stimuli is called associative learning [23–25]. Because of the associative learning mechanism, humans can robustly understand and perceive their surrounding situations even when only some of the modalities are available.

In the field of machine learning, utilizing multi-modality is also important issues because of its usefulness in a wide range of applications [1, 69]. As a representative example, object recognition and scene understanding methods based on multi-modal data outperform the methods using only single-modal data [5, 21]. Moreover, one can generate the synthesized data for a missing or desired modality [3, 4, 7, 34, 70, 71]. The multi-modal data association is one of the fundamental steps to understand the relationships among multi-modal data. Recently, along with the advances of deep learn-

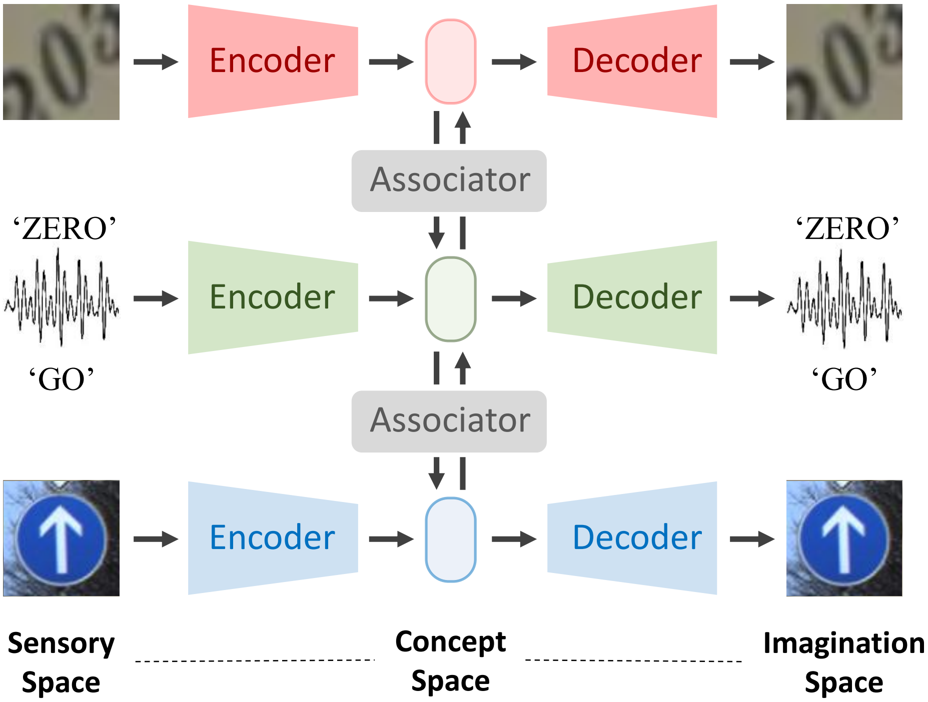


Figure 3.1: Conceptual illustration of the proposed AVAE. AVAE has modality-specific encoders and decoders for each modality (image, voice). Each modality has its own embedding space, which is painted with a different color (red, green, blue). The embedding spaces are connected via the proposed *associator* which associates two different modalities.

ing, many studies have attempted to solve the multi-modal data association problem by deep learning algorithms [1]. The studies have adopted an approach that encodes multi-modal data into a joint embedding space to memorize common features among multiple modalities [2–5, 21].

However, as pointed out by [8], most existing studies did not consider the case that the characteristic of each modality is very different from others. The encoding in the joint embedding space is hard to represent all characteristics of the heterogeneous modalities or could be biased to a dominant modality. Furthermore, the capacity of the joint embedding will be saturated as modalities increase and so encounters a scalability problem. To mitigate the limitation of the joint embedding space, we propose an approach that adopts distributed embedding spaces. In proposed approach, as shown in Figure 3.1, each modality is encoded in each embedding space separately by the variational auto-encoder (VAE) [22] and the distributed embedding spaces are associated with the other modalities via associators.

The proposed structure is implemented with a deep neural network with multiple variational auto-encoders and variational associators [6]. The loss function to train the network is derived by the variational inference framework. In experiments, the effectiveness and performance are evaluated through comparison with the existing methods and self-analysis using various datasets including voice and visual data. In addition, by self-experiments, the advantage of our structure is verified on generalization ability for semi-supervised learning, scalability of the network, and flexibility of distributed embedding space dimensions.

3.2 Motivation

For cross-modal representation learning, the most recent studies have adopted joint embedding schemes, where data of each modality is embedded into the joint embedding space. Joint embedding methods achieved good performances in many applications. However, Joint embedding methods often has trouble in learning cross-modal relationship among *heterogeneous* modalities. Heterogeneous means that characteristics of modalities are very different such as the case of vision and audio data. For heterogeneous modalities, similarity in the one modality does not guarantee similarity in the other modality. Thus, with joint embedding schemes, it is difficult to learn the relationship among heterogeneous modalities. But humans are proficient at learning relationships among even heterogeneous modalities. Therefore, I obtained a motif from the process of associative learning mechanism of the human brain.

According to recent studies [29, 30], associative learning process in the brain includes intra-modal and cross-modal association processes (detailed explanations are provided in Section 2.1). The intra-modal association process is to make humans familiar with single-modal sensory information. On the other hand, the cross-modal association process is accomplished to enhance the strength of the synapses connecting multi-modal information to be associated.

From the process of the associative learning in human brain, I obtained two key concepts: (1) Intra-modal information is memorized in each cortex, and (2) Intra-modal memorization and cross-modal memorization are separated. From the point of view of machine learning, those two key concepts can be interpreted as (1) Distributed latent space for each modality, and (2) Separate intra-modal association and cross-modal association in training phase. Note that the conventional methods adopting the joint embedding space are in conflict with the aforementioned key concepts. With the joint embedding space, every multi-modal data are embedded into the same spaces (conflict with key concept 1), and intra-modal association and cross-modal association are simultaneously trained during training phase (conflict with key concept 2). This discrepancy might be the reason why the joint embedding schemes are hard to learn the relationship among heterogeneous modalities.

Therefore, the goal of this chapter is to develop a novel scheme that can robustly learn cross-modal representation among heterogeneous modalities, motivated by associative learning mechanism of human brain. To this end, I establish the Bayesian formulation of these two association processes and to realize them in a variational auto-encoder framework.

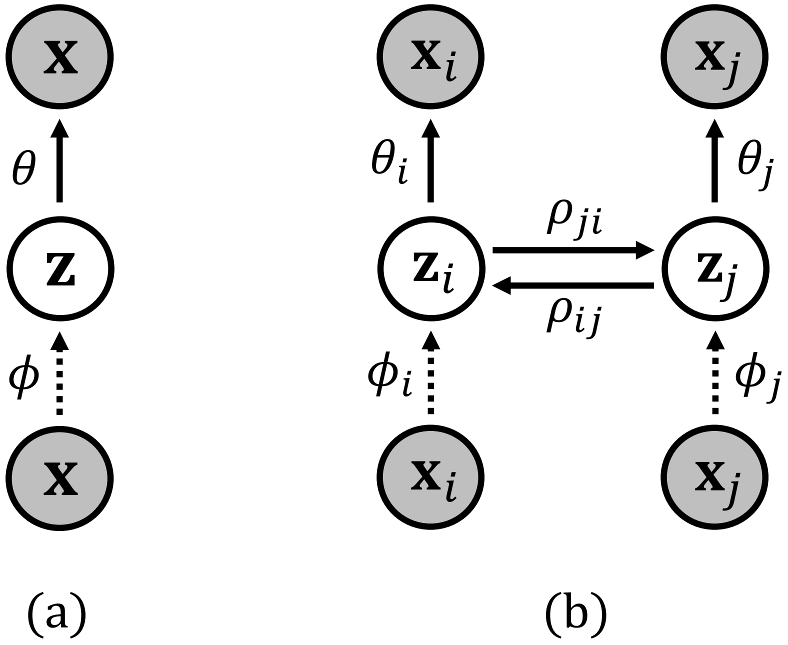


Figure 3.2: **Graphical models for intra-modal and cross-modal association.** Observable variables are illustrated as shadowed circles. θ, ϕ, ρ are distribution parameters: θ for true distribution, ϕ for variational distribution, and ρ for cross-modal association model. Subscripts denote modality. Dotted lines indicate variational approximation of true probability distribution. **(a) Intra-modal association** Latent variable $\mathbf{z}$ is obtained by $\mathbf{x}$ through $q_\phi(\mathbf{z}|\mathbf{x})$ and $\mathbf{x}$ is inferred from $\mathbf{z}$ through $p_\theta(\mathbf{x}|\mathbf{z})$ **(b) Cross-modal association between two modalities** The cross-modal association model has mutual connections between latent variables $\mathbf{z}_i$ and $\mathbf{z}_j$.

3.3 Graphical Modeling

3.3.1 Graphical Model of Intra-modal Association

Intra-modal association is the process of memorizing single-domain information. To efficiently memorize a vast amount of information, the model needs to extract the expressive features of the data. One way to make the encoding model remember the features of the data in an unsupervised manner is to formulate a mathematical model reconstructing the original sensory data from the encoded information. Figure 3.2 (a) shows Bayesian graphical model to formulate the intra-modal association to memorize a distribution of the latent variable $\mathbf{z}_i$ associated with the input variable $\mathbf{x}_i$ for an observation in modality i. In the Bayesian framework, the objective is to infer model parameter θ_i of posterior distribution $p_{\theta_i}(\mathbf{z}_i|\mathbf{x}_i)$.

One of the most popular approaches to approximate an intractable posterior is the variational inference method. In this method, the variational distribution $q_{\phi_i}(\mathbf{z}_i|\mathbf{x}_i)$ approximates the true posterior $p_{\theta_i}(\mathbf{z}_i|\mathbf{x}_i)$ by minimizing the Kullback-Leibler divergence, $D_{KL}\left(q_{\phi_i}(\mathbf{z}_i|\mathbf{x}_i)\|p_{\theta_i}(\mathbf{z}_i|\mathbf{x}_i)\right)$. According to [22], the minimization of Kullback-Leibler divergence $D_{KL}\left(q_{\phi_i}(\mathbf{z}_i|\mathbf{x}_i)\|p_{\theta_i}(\mathbf{z}_i|\mathbf{x}_i)\right)$ can be replaced with the maximization of the evidence lower bound, given by

$$\begin{aligned}\mathcal{L}(q_{\phi_i}(\mathbf{z}_i|\mathbf{x}_i)) = &- D_{KL}(q_{\phi_i}(\mathbf{z}_i|\mathbf{x}_i)\|p_{\theta_i}(\mathbf{z}_i)) \\ &+ \mathbb{E}_{q_{\phi_i}(\mathbf{z}_i|\mathbf{x}_i)}[\log p_{\theta_i}(\mathbf{x}_i|\mathbf{z}_i)],\end{aligned} \tag{3.1}$$

where $\mathbb{E}_{q_{\phi_i}(\mathbf{z}_i|\mathbf{x}_i)}$ indicates expectation over distribution $q_{\phi_i}(\mathbf{z}_i|\mathbf{x}_i)$.

3.3.2 Graphical Model of Cross-modal Association

In this section, we design a graphical model to represent the cross-modal association mechanism as in Figure 3.2 (b). Without loss of generality, we consider a path from modality i to j. From observations of an associated variable pair $(\mathbf{x}_i, \mathbf{x}_j)$, the distribution parameter ρ_{ji} is inferred to model the association between $\mathbf{z}_i$ and $\mathbf{z}_j$.

For a given observation pair $(\mathbf{x}_i, \mathbf{x}_j)$, the cross-posterior distribution $p_{\theta_i,\rho_{ji}}(\mathbf{z}_j|\mathbf{x}_i)$ is defined by marginalization for $\mathbf{z}_i$ as

$$p_{\theta_i,\rho_{ji}}(\mathbf{z}_j|\mathbf{x}_i) = \int p_{\rho_{ji}}(\mathbf{z}_j|\mathbf{z}_i)p_{\theta_i}(\mathbf{z}_i|\mathbf{x}_i)\,d\mathbf{z}_i. \tag{3.2}$$

To establish the cross-modal association model, we define a variational distribution for cross-posterior distribution $q_{\phi_i,\rho_{ji}}(\mathbf{z}_j|\mathbf{x}_i)$. Then, to infer the distribution parameters (ϕ_i, ρ_{ji}), we minimize Kullback-Leibler divergence between $p_{\theta_j}(\mathbf{z}_j|\mathbf{x}_j)$ and $q_{\phi_i,\rho_{ji}}(\mathbf{z}_j|\mathbf{x}_i)$. To avoid clutter, subscripts for the distribution parameters are omitted in the remainders of this section. Kullback-Leibler divergence between $p(\mathbf{z}_j|\mathbf{x}_j)$ and $q(\mathbf{z}_j|\mathbf{x}_i)$ is given by

$$D_{KL}(q(\mathbf{z}_j|\mathbf{x}_i)\|p(\mathbf{z}_j|\mathbf{x}_j)) = \log p(\mathbf{x}_j) - \mathcal{L}(q(\mathbf{z}_j|\mathbf{x}_i)), \tag{3.3}$$

where

$$\mathcal{L}(q(\mathbf{z}_j|\mathbf{x}_i)) = \int q(\mathbf{z}_j|\mathbf{x}_i) \log \frac{p(\mathbf{x}_j)p(\mathbf{z}_j|\mathbf{x}_j)}{q(\mathbf{z}_j|\mathbf{x}_i)}\, d\mathbf{z}_j. \tag{3.4}$$

Since log-evidence $\log p(\mathbf{x}_j)$ is independent to the model parameter, the target problem is identical to maximizing the evidence lower bound $\mathcal{L}(q(\mathbf{z}_j|\mathbf{x}_i))$. With probabilistic tricks, $\mathcal{L}(q(\mathbf{z}_j|\mathbf{x}_i))$ can be decomposed as following.

$$\begin{aligned}
\mathcal{L}(q(\mathbf{z}_j|\mathbf{x}_i)) &= \int q(\mathbf{z}_j|\mathbf{x}_i) \log \frac{p(\mathbf{x}_j)p(\mathbf{z}_j|\mathbf{x}_j)}{q(\mathbf{z}_j|\mathbf{x}_i)}\, d\mathbf{z}_j \\
&= \int q(\mathbf{z}_j|\mathbf{x}_i) \log \frac{p(\mathbf{z}_j)}{q(\mathbf{z}_j|\mathbf{x}_i)}\, d\mathbf{z}_j + \int q(\mathbf{z}_j|\mathbf{x}_i) \log \frac{p(\mathbf{x}_j)p(\mathbf{z}_j|\mathbf{x}_j)}{p(\mathbf{z}_j)}\, d\mathbf{z}_j \\
&= -D_{KL}(q(\mathbf{z}_j|\mathbf{x}_i)\|p(\mathbf{z}_j)) + \int q(\mathbf{z}_j|\mathbf{x}_i) \log p(\mathbf{x}_j|\mathbf{z}_j)\, d\mathbf{z}_j \\
&= -D_{KL}(q(\mathbf{z}_j|\mathbf{x}_i)\|p(\mathbf{z}_j)) + \mathbb{E}_{q(\mathbf{z}_j|\mathbf{x}_i)}[\log p(\mathbf{x}_j|\mathbf{z}_j)].
\end{aligned} \tag{3.5}$$

In the last line in Eq. (3.5), the first term is a negative KL divergence term that leads $\mathbf{z}_j$ given by $\mathbf{x}_j$ to have similar distribution with a prior distribution of target modality. The expectation term in Eq. (3.5) minimizes the reconstruction error of decoded output from $\mathbf{z}_j$ fired from $\mathbf{x}_i$, which also promotes the inference for ρ_{ji}. By the similar steps, we can easily derive the opposite association from modality j to modality i.

3.4 Realization

3.4.1 Cross-modal Association Network

We accomplish a realization of the aforementioned intra-modal and cross-modal association models by extending the Variational Auto-Encoder framework (VAE) [22]. Figure 3.3 illustrates the proposed cross-modal association network for modality i and j. Although only two modalities are considered in this paper, the proposed model can be applied to the association among three or more modalities also. In the proposed structure, the *encoder* produces the parameter of $q_{\phi_i}(\mathbf{z}_i|\mathbf{x}_i = x_i)$, and the *decoder* produces the parameter of $p_{\theta_i}(\mathbf{x}_i|\mathbf{z}_i = z_i)$. The encoder and decoder are realized by deep neural networks. Likewise, the latent space associating models $p_{\rho_{ji}}(\mathbf{z}_j|\mathbf{z}_i = z_i)$ and $p_{\rho_{ij}}(\mathbf{z}_i|\mathbf{z}_j = z_j)$ are also realized by deep neural networks, which are called by *associator*. Thus, the intra-modal association network contains several auto-encoders, each of which considers one of the multiple modalities only. The latent spaces of the auto-encoders are connected by *associators* in a pairwise manner, which configure the cross-modal association network.

The proposed network is trained in the two phases: intra-modal training phase and cross-modal training phase. In the intra-modal training phase, the auto-encoder in each modality is trained separately by minimizing the approximated version of the negative evidence lower bound in Eq. (3.1). As derived in [22], variational distributions are assumed by the centered isotropic multivariate Gaussian distribution. For a given observation sample x_i, the encoder E_i produces the mean μ_{ϕ_i} and the variance σ_{ϕ_i} for a Gaussian distribution of $q_{\phi_i}(\mathbf{z}_i|\mathbf{x}_i = x_i)$. Then, the latent vector z_i is sampled as $z_i = \mu_{\phi_i} + \sigma_{\phi_i} * \epsilon$ and $\epsilon \backsim N(0, I)$. Similarly, the decoder D_i also produces the mean μ_{θ_i} and the variance σ_{θ_i} for a Gaussian distribution of $p_{\theta_i}(\mathbf{x}_i|\mathbf{z}_i = z_i)$. Then, the reconstruction vector $\hat{x}_i$ is sampled as $\hat{x}_i = \mu_{\theta_i} + \sigma_{\theta_i} * \epsilon$ and $\epsilon \backsim N(0, I)$.

Using the samples, the empirical loss for auto-encoder can be derived as

$$\begin{aligned}
&\mathcal{L}_{int}(\theta_i, \phi_i; x_i) \\
&= -\mathbb{E}_{q_{\phi_i}(\mathbf{z}_i|x_i)}[\log p_{\theta_i}(x_i|\mathbf{z}_i)] + \lambda_{int}^{'} D_{KL}(q_{\phi_i}(\mathbf{z}_i|x_i)||p_{\theta_i}(\mathbf{z}_i)), \\
&= ||x_i - \hat{x}_i||_2^2 - \lambda_{int} \sum_{k}^{H}(1 + \log \sigma_{\phi_i(k)}^2 - \mu_{\phi_i(k)}^2 - \sigma_{\phi_i(k)}^2).
\end{aligned} \tag{3.6}$$

where λ_{int} is a user-defined parameter and H is the dimension of the latent variable $\mathbf{z}_i$. $\mu_{\phi_i(k)}$ and $\sigma_{\phi_i(k)}^2$ denote the k-th element of μ_{ϕ_i} and $\sigma_{\phi_i}^2$.

Detailed derivation step for each loss term is given as following. Given a latent vector z_i encoded by an input sample x_i, the probability distribution $p_{\theta_i}(\mathbf{x}_i|\mathbf{z}_i = z_i)$ is assumed to be a Gaussian with mean x_i and variance cI for a constant scalar c. Then the negative log-likelihood of sampled output $\hat{\mathbf{x}}_i$ of decoder D_i for input

$$\begin{aligned}
-\log p_{\theta_i}(\hat{x}_i|\mathbf{z}_i = z_i) &= -\log C^{'} e^{-\frac{1}{2c}(\hat{x}_i - x_i)^T(\hat{x}_i - x_i)} \\
&= C + \frac{1}{2c}\|\hat{x}_i - x_i\|_2^2,
\end{aligned} \tag{3.7}$$

where $C, C^{'}$ are appropriate constants. From the results of [22], Kullback-Leibler term in Eq. (3.6) can be written as:

$$D_{KL}(q_{\phi_i}(\mathbf{z}_i|x_i)\|p_{\theta_i}(\mathbf{z}_i)) = -\frac{1}{2}\sum_{k=1}^{H}(1 + \log \sigma_{\phi_i(k)}^2 - \mu_{\phi_i(k)}^2 - \sigma_{\phi_i(k)}^2). \tag{3.8}$$

Since the constant C can be ignored in Eq. (3.7), Eq. (3.6) can be obtained by combining the two terms in Eq. (3.7) and Eq. (3.8) with a weighting parameter λ_{int}. Figure 3.4 shows the example of training flow for intra-modal association.

After the convergence of the intra-modal training phase, the following cross-modal training phase proceeds to train the associators while freezing the weights of the auto-encoders. In the same way as in the intra-modal training phase, for a given observation pair x_i and x_j, the encoders E_i and E_j produce the latent vectors z_i and z_j, respectively. In addition, associators A_{ji} and A_{ij} produce the latent vectors z_{ji} and z_{ij} using inputs z_i and z_j, respectively. Thereafter, the decoders D_i and D_j produce the reconstruction vectors $\hat{x}_{ij}$ and $\hat{x}_{ji}$ from z_{ij} and z_{ji}, respectively.

Using the samples, the empirical loss for A_{ji} is designed according to Eq. (3.5) as follows:

$$\begin{aligned}
&\mathcal{L}_{crs}(\rho_{ji}; x_i, x_j) \\
&= -\mathbb{E}_{q_{\rho_{ji}}(\mathbf{z}_{ji}|x_i)}[\log p_{\theta_j}(x_j|\mathbf{z}_{ji})] + \lambda'_{crs} D_{KL}(q_{\phi_i,\rho_{ji}}(\mathbf{z}_{ji}|x_i)||p_{\theta_j}(\mathbf{z}_{ji})) \\
&= ||x_j - \hat{x}_{ji}||_2^2 - \lambda_{crs} \sum_k^H (1 + \log \sigma^2_{\rho_{ji}(k)} - \mu^2_{\rho_{ji}(k)} - \sigma^2_{\rho_{ji}(k)}).
\end{aligned} \tag{3.9}$$

where λ_{crs} is a user-defined parameter and H is the dimension of the latent variable $\mathbf{z}_{ji}$. ($\mu_{\rho_{ji}}$, $\sigma^2_{\rho_{ji}}$) are parameters for Gaussian distribution $q_{\phi_i,\rho_{ji}}(\mathbf{z}_{ji}|x_i)$ produced by A_{ji}.

Detailed derivation step for Eq. (3.9) is similar to the auto-encoder case. The probability distribution $p_{\theta_j}(\mathbf{x}_j|\mathbf{z}_{ji} = z_{ji})$ is assumed to be a Gaussian distribution with mean x_j and variance cI when z_{ji} is given, and then the negative log-likelihood for sampled output $\hat{x}_{ji}$ of decoder D_j for input z_{ji} can be written as:

$$\begin{aligned}
-\log p_{\theta_j}(\hat{x}_{ji}|\mathbf{z}_{ji} = z_{ji}) &= -\log C' e^{-\frac{1}{2c}(\hat{x}_{ji}-x_j)^T(\hat{x}_{ji}-x_j)} \\
&= C + \frac{1}{2c}\|\hat{x}_{ji} - x_j\|_2^2.
\end{aligned} \tag{3.10}$$

From the results of [22], Kullback-Leibler term in Eq. (3.9) can be written as:

$$D_{KL}(q_{\phi_i,\rho_{ji}}(\mathbf{z}_{ji}|x_i)\|p_{\theta_j}(\mathbf{z}_{ji})) = -\frac{1}{2}\sum_{k=1}^{H}(1 + \log \sigma^2_{\rho_{ji}(k)} - \mu^2_{\rho_{ji}(k)} - \sigma^2_{\rho_{ji}(k)}). \tag{3.11}$$

Since the constant C can be ignored in Eq. (3.10), the loss in Eq. (3.9) in the paper can be obtained by combining the two terms in Eq. (3.10) and Eq. (3.11) with a weighting parameter λ_{crs}. Figure 3.5 shows the example of training flow for cross-modal association.

The loss $\mathcal{L}_{crs}(\rho_{ij}; x_i, x_j)$ for A_{ij} is given in the same form of A_{ji} except the index. Note that all μ's and σ's in Eq. (3.6) and Eq. (3.9) are the functions of weights ($\mathbf{w}$)

in encoders, decoders, or associators. Hence, the weights of the proposed network are trained by the negative direction of the gradient of the losses with respect to the weights ($\nabla_{\mathbf{w}}\mathcal{L}(\cdot)$).

3.4.2 Advantages of the Proposed Model

Owing to the newly introduced *associator*, the proposed model can associate heterogeneous modalities effectively. Reckless coalescence of heterogeneous data may have a fatal impact on associative learning such as the problem that shared latent vectors can be biased to the dominant modality. However, in our model, the associator acts as a translator between heterogeneous modalities and thus the characteristics of each latent space are preserved. Furthermore, in contrast to the existing models which adopt a shared latent space for the different modalities [2–4, 21], our structure can provide a flexible dimensional encoding in each latent space depending on the complexity of each modality. This provides better cross-modal data association results. Figure 3.6 (a) illustrates the advantage of flexible dimension.

The proposed model easily incorporates additional modalities while maintaining the existing modalities. That is, a new modality can be added via training of only a new associator between an existing auto-encoder and a new auto-encoder. Though the associator only associates the new modality with one of the existing modalities, the model can associate the new modality with the rest of the modality by passing through multiple associators. Figure 3.6 (b) illustrates the advantage of incorporating additional modalities.

Finally, in contrast to existing models which always require paired data for cross-modal association, our structure can train the associator with the only small amount of paired data in a semi-supervised manner after learning each auto-encoder using unpaired data independently. Since obtaining paired data for cross-modal association is more expensive than obtaining unpaired data, our model is cost-effective. Furthermore,

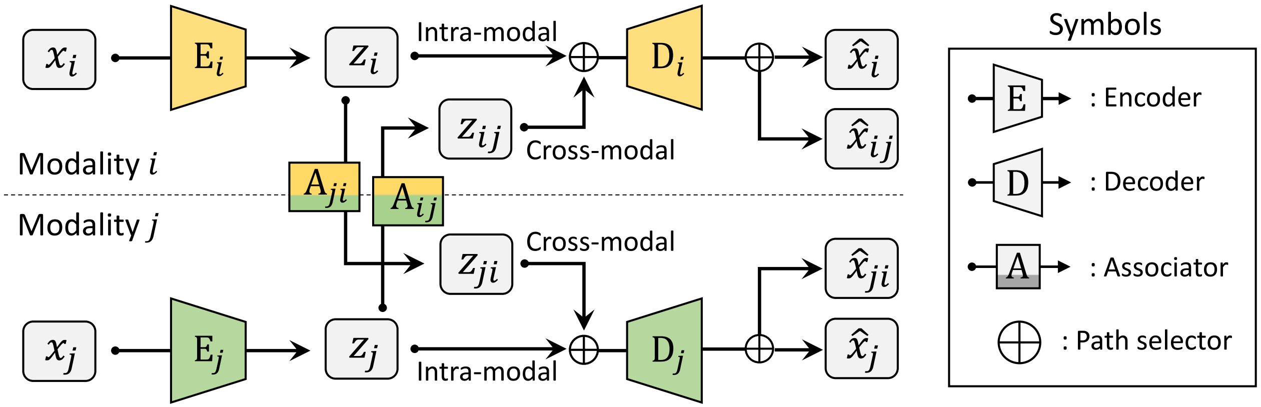


Figure 3.3: Overall structure of the proposed method for the modalities i and j. For an observation sample x_i for the variable $\mathbf{x}_i$, the intra-modal network of modality i encodes x_i into a latent vector z_i for the latent variable $\mathbf{z}_i$ through the encoder E_i and decodes z_i to $\hat{x}_i$ through decoder D_i. In the case of association from modality i to j, a sample x_i is encoded to z_{ji} through the encoder E_i and the associator A_{ji}. Then, z_{ji} is decoded to $\hat{x}_{ji}$ through D_j. The procedure for the opposite direction is performed in the same way.

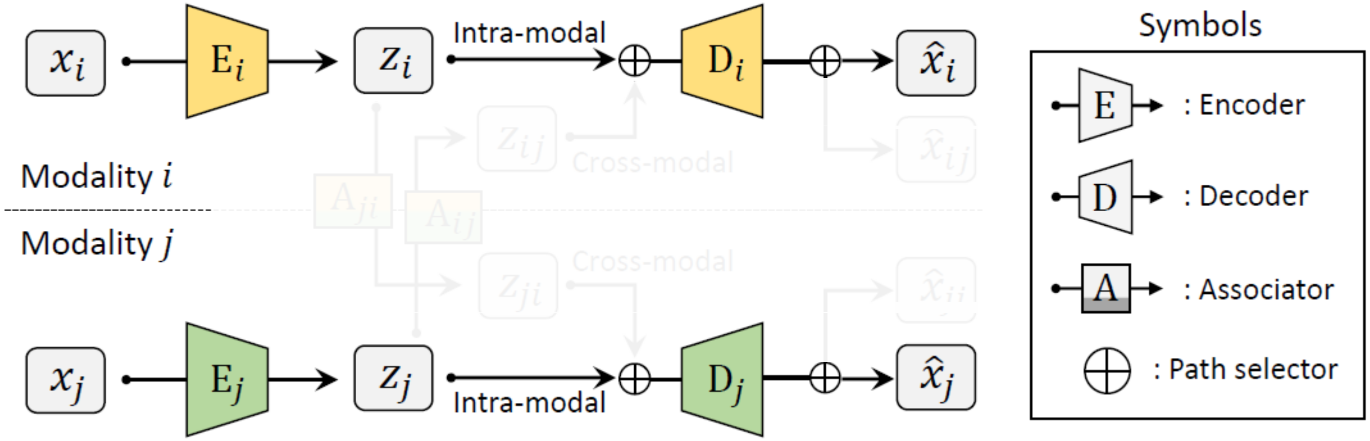


Figure 3.4: Example of training flow for intra-modal association.

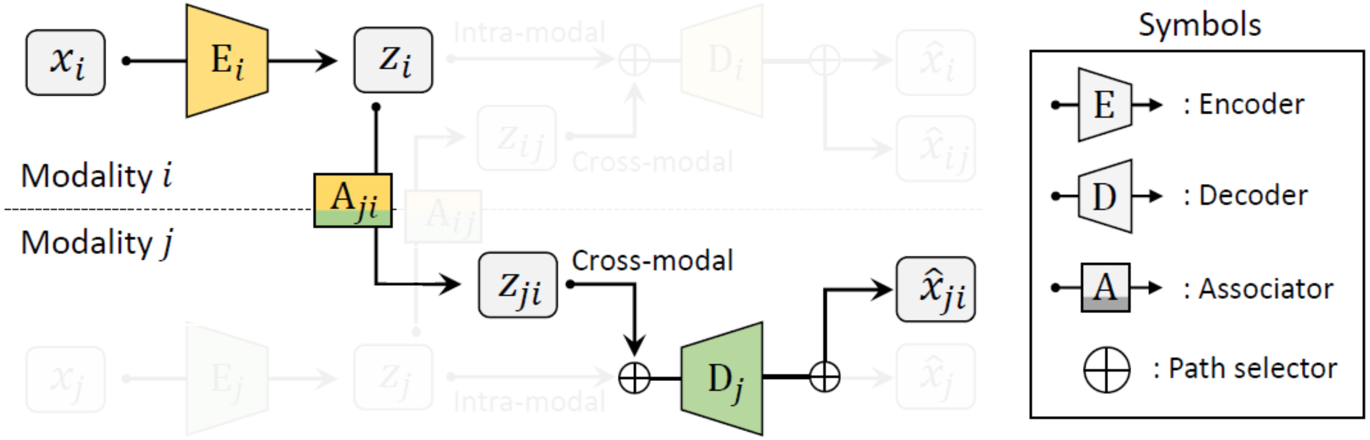
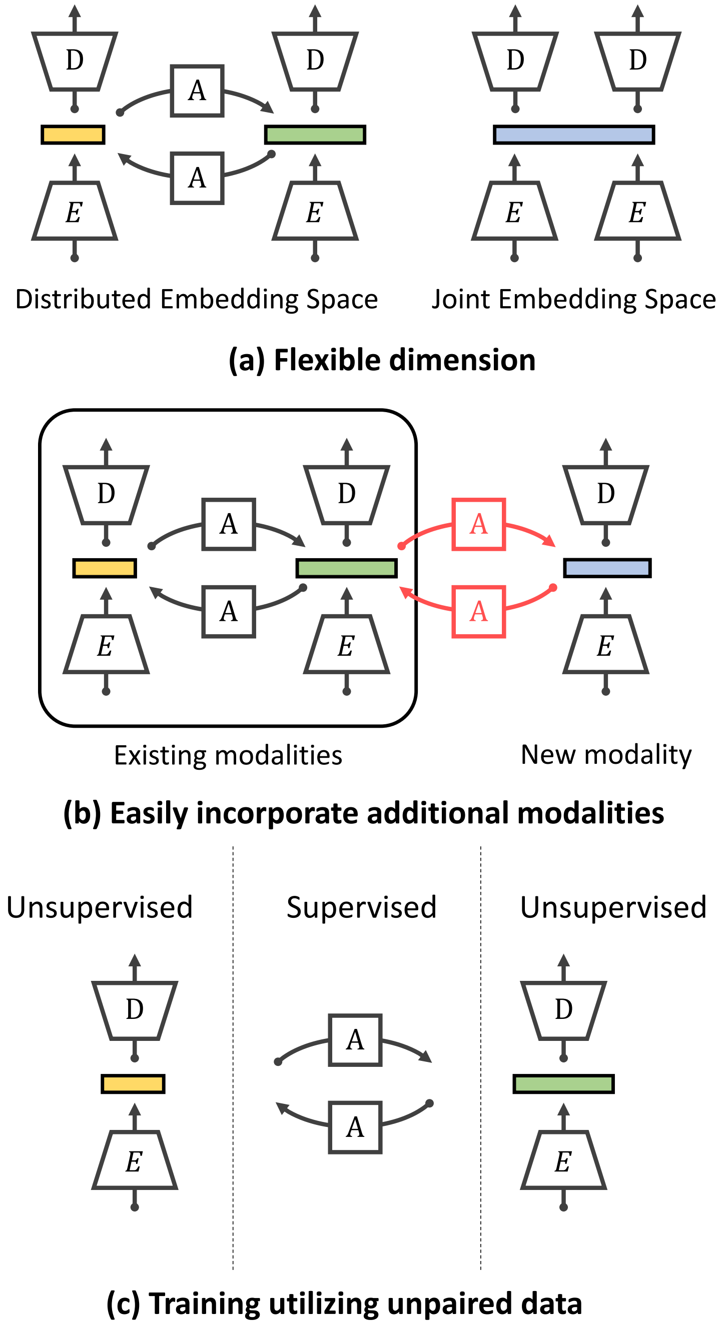


Figure 3.5: Example of training flow for cross-modal association.



(a) Flexible dimension

(b) Easily incorporate additional modalities

(c) Training utilizing unpaired data

Figure 3.6: Illustration of the advantages of the proposed network.

our model is plausible in that, when a person learns a cross-modal association, the paired examples are rarely given by a teacher after the person has become familiar with each modality via self-experience without a teacher. Figure 3.6 (c) illustrates the advantage of semi-supervised learning with unpaired data.

3.5 Experiment

3.5.1 Implementation Details

3.5.1.1 Datasets

Google Speech Commands (GSC) [72]: As the data for the auditory modality, we used the GSC dataset, which consists of 105,829 audio samples containing utterances of 35 short words. Each audio sample is one-second-long and encoded with a sampling rate of 16KHz. Among 35 words, we chose 14 words, including words for each digit ('ZERO' to 'NINE') and four traffic commands ('GO,' 'STOP,' 'LEFT,' 'RIGHT'). The chosen set has 54,239 samples. We extracted the Mel-Frequency Cepstral Coefficient (MFCC) from each audio clip to generate an audio feature. MFCC has been widely used in the processing of voice data because it reflects the human auditory perception mechanism well [73–75]. The resulting features are 40×101 matrices. We randomly divided the original dataset into training, validation and test sets at the ratio of 8:1:1.

German Traffic Sign Recognition Benchmark (GTSRB) [76]: For the visual data that correspond to the traffic commands in GSC, we used the GTSRB dataset, which consists of 51,839 RGB color images illustrating 42 kinds of traffic signals. In particular, to evaluate the performance on pairs of traffic sign images and voice commands in GSC, we chose four pair sets, where each pair set has similar semantic meaning, i.e., ('Ahead only,' 'GO'), ('No entry for vehicle,' 'STOP'), ('Turn left and ahead,' 'LEFT'), and ('Turn right and ahead,' 'RIGHT'). The first and the second element are taken from GTSRB and GSC dataset, respectively. Then, to prevent the four signs from occupying the entire latent space, we chose additional sign images in GTSRB such as ’No overtaking,’ ’Entry to 30kph zone,’ ’Prohibit overweighted vehicle,’ ’No-waiting zone,’ and ’Roundabout’. The chosen set includes 10,709 samples. All of the chosen signs have a circular backboard. The size of each image varies from 15×15 to 250×250 pixels for each RGB channel in the original dataset. In our experiments, we resized all images into 52×52.

MNIST [77]: We used the MNIST dataset as the corresponding visual data to the GSC for each digit. The MNIST consists of center-aligned 28×28 gray-scale images for handwritten digits from 0 to 9. The dataset contains 60k and 10k samples for the training set and testing set, respectively.

SVHN [78]: We used the SVHN dataset as another visual modality. Even though SVHN and MNIST are equally categorized by digits, their capturing environments are very different from each other. The SVHN consists of 32×32 RGB images for digits from 0 to 9. The dataset contains 73257 and 26032 samples for the training set and testing set, respectively.

Fashion-MNIST (F-MNIST) [79]: To validate that the proposed model can associate even not semantically related datasets, we used the F-MNIST dataset. We associated F-MNIST with the MNIST and the GSC dataset. After this association learning, we can imagine the clothing items (F-MNIST) from their numberings (MNIST). The F-MNIST consists of center-aligned 28×28 gray-scale images assigned with a label from 10 kinds of clothing such as T-shirt, Trouser and Sneaker. The dataset contains 60k and 10k samples for the training set and testing set, respectively.

3.5.1.2 Network Architecture

Table 3.1, Table 3.2, and Table 3.3 describe the network architectures of classifiers and auto-encoders for each dataset. The input data are forwarded from the top layer to the bottom layer in Tables. The Tuples in Shape column denote (*the number of input channels*, *the number of output channels*, *kernel height*, *kernel width*) for Conv layers, (*kernel height*, *kernel width*) for Linear layers, while the tuples in Stride and Padding columns denote (*vertical steps*, *horizontal steps*). The last linear layer in each encoder returns μ, σ^2. Dropout ratio for all dropout layers is set to 0.5.

All the associators used in experiments have the identical structure. Each associator consists of multiple linear and ReLU layers: $(I \times 2(I+O))$, $(2(I+O) \times 2(I+O)) * 4$ and $(2(I+O) \times 2O)$ where I and O denote dimension of latent space of input modality and output modality, respectively.

3.5.1.3 Evaluation Metric

Since aforementioned datasets have no direct matching relationships, we cannot measure cross-likelihood $p(\mathbf{x}_1|\mathbf{x}_2)$ for paired sample $(\mathbf{x}_1, \mathbf{x}_2)$ used in recent works [2,80]. In our work, we used the classification accuracy for the reconstructed results as the evaluation metric of the association models. The quality of results reconstructed by an association model can be a valid measure to evaluate the association model since the quality of the reconstructed results is acceptable to both the human and the classifier. Table 3.4 shows the performance of the classifiers trained with the each dataset, which shows sufficient performance for evaluating the reconstructed results of the compared encoders. For the GSC dataset, we get performance comparable to the 88.2% [72].

(a) MNIST and F-MNIST

Layer	**Shape**	**Stride**	**Padding**	**Activation**
Linear	(784, 256)	-	-	ReLU
Linear	(256, 10)	-	-	Softmax

(b) GTSRB

Layer	**Shape**	**Stride**	**Padding**	**Activation**
Conv.	(3, 16, 3, 3)	(1, 1)	(1, 1)	-
Pool.	(16, 16, 2, 2)	(2, 2)	(0, 0)	ReLU
Conv.	(16, 32, 3, 3)	(1, 1)	(1, 1)	-
Dropout	-	-	-	-
Pool.	(32, 32, 2, 2)	(2, 2)	(0, 0)	ReLU
Linear	(5408, 256)	-	-	ReLU
Dropout	-	-	-	-
Linear	(256, 9)	-	-	Softmax

Table 3.1: Network architecture of classifier for MNIST, F-MNIST and GTSRB dataset.

(a) GSC

Layer	Shape	Stride	Padding	Activation
Conv.	(1, 16, 9, 21)	(1, 1)	(4, 10)	ReLU
Dropout	-	-	-	-
Pool.	(16, 16, 2, 2)	(2, 2)	(0, 0)	-
Conv.	(16, 16, 5, 11)	(1, 1)	(2, 5)	ReLU
Dropout	-	-	-	-
Pool.	(16, 16, 2, 2)	(2, 2)	(0, 0)	-
Linear	(4000, 256)	-	-	ReLU
Linear	(256, 14)	-	-	Softmax

(b) SVHN

Layer	Shape	Stride	Padding	Activation
Conv.	(3, 32, 5, 5)	(1, 1)	(1, 1)	ReLU
Pool.	(32, 32, 2, 2)	(2, 2)	(0, 0)	-
Conv.	(32, 64, 5, 5)	(1, 1)	(1, 1)	ReLU
Pool.	(64, 64, 2, 2)	(2, 2)	(0, 0)	-
Conv.	(64, 128, 5, 5)	(1, 1)	(1, 1)	ReLU
Pool.	(128, 128, 2, 2)	(2, 2)	(0, 0)	-
Conv.	(128, 128, 5, 5)	(1, 1)	(1, 1)	ReLU
Pool.	(128, 128, 2, 2)	(2, 2)	(0, 0)	-
Linear	(1152, 2048)	-	-	ReLU
Linear	(2048, 2048)	-	-	ReLU
Linear	(2048, 10)	-	-	Softmax

Table 3.2: Network architecture of classifier for GSC and SVHN dataset. For GSC and SVHN dataset, batch normalization layers are added after every Conv. layers.

(a) MNIST and F-MNIST

Layer	Shape	Activation
Linear	(784, 128)	ReLU
Linear	(128, 128)	-
Reparam.	-	-
Linear	(64, 128)	ReLU
Linear	(128, 784)	Sigmoid

(b) GSC

Layer	Shape	Activation
Linear	(4040, 512)	ReLU
Linear	(512, 128)	-
Reparam.	-	-
Linear	(64, 512)	ReLU
Linear	(512, 4040)	HardTanh

(c) GTSRB

Layer	Shape	Activation
Linear	(8112, 512)	ReLU
Linear	(512, 128)	-
Reparam.	-	-
Linear	(64, 512)	ReLU
Linear	(512, 8112)	Sigmoid

(d) SVHN

Layer	Shape	Activation
Linear	(3072, 512)	ReLU
Linear	(512, 128)	-
Reparam.	-	-
Linear	(64, 512)	ReLU
Linear	(512, 8112)	Sigmoid

Table 3.3: Network architecture of auto-encoders for each dataset. For GSC, GTSRB, SVHN datasets, batch normalization layers are added after every linear layers except for the last layer.

Table 3.4: Performance of the classifier trained with the each dataset and reconstruction performance of VAE. dim($\mathbf{z}$) denotes the dimension of latent space of VAE.

Dataset	Acc (%)	VAE (%)	dim($\mathbf{z}$)
MNIST	97.97	96.12	64
F-MNIST	89.22	80.54	64
SVHN	93.73	78.22	64
GSC	88.65	81.93	64
GTSRB	98.53	95.70	64
GTSRB	-	95.50	256

3.5.2 Intra-Modal Association

As mentioned in the problem statements section, it is also essential for the proposed model to learn the intra-modal association that encodes single modal input data into the latent space. For a fair comparison to existing works, we trained encoders and decoders for each dataset with the fixed dimension of latent space ($\dim(\mathbf{z}) = 64$). In addition, to show the advantages of the proposed model where the dimension of the latent space can be flexibly designed according to the complexity of target modality, we trained additional auto-encoder whose latent space dimension is 256 for GTSRB dataset.

Table 3.4 shows the performance of the classifiers and the intra-modal association network implemented by VAE. Performance of VAE is also measured by the classifier on the results reconstructed by VAE. As shown in the table 3.4, the voice data in the GSC dataset shows much degraded accuracy, which means that the voice data are hard to be reconstructed than other modalities. Since F-MNIST has confused classes such as pullover, coat, and shirt, performance on F-MNIST dataset is also degraded.

3.5.3 Cross-Modal Association

Cross-modal problem is defined to develop a model that can generate the sample of target modality from a given sample of source modality, where samples are semantically associated. We evaluated the proposed model on five scenarios: (1) Association between MNIST and GSC, (2) GTSRB and GSC, (3) F-MNIST and GSC, (4) F-MNIST and MNIST, (5) SVHN and MNIST. Scenario (1), (2) and (3) are for association between heterogeneous datasets, i.e. voice and image datasets. Scenario (3) and (4) is for association between datasets which have no semantic relations between classes. Scenario (5) is for association between semantically related datasets, through two datasets have different dataset characteristics like image size. Figure 3.7 illustrates all scenarios. In order to train cross-modal association, we used randomly paired training samples from each dataset belonging to the correlated class. For example, we paired a randomly chosen sample in ’0’ class of MNIST dataset with a randomly chosen sample in ’ZERO’ class of GSC dataset.

To evaluate the proposed associator, the following methods were compared: **VAE** and **VAE-CG** are variants of the standard VAE. When training VAE, we constructed the training data with vectors concatenated with data from two associated modalities, whereas one modality in the concatenated vector for 50% of training data was set to zero-vector to learn the case of the missing modality. **VAE-CG** is trained to generate the target modality sample from a given input sample of other modality. VAE-CG has to be trained only by supervised data with input and output pairs. Joint Multimodal Variational Auto-Encoder (**JMVAE**) [80] has two kinds of latent spaces: one is for each modality and the other is for jointly encoding of two modalities. The joint latent space is shared for association between two modalities. The training for encoding in the joint latent space is done to minimize Kullback-Leibler divergence between the latent vector of each encoder and the joint latent vector of the joint encoder. In

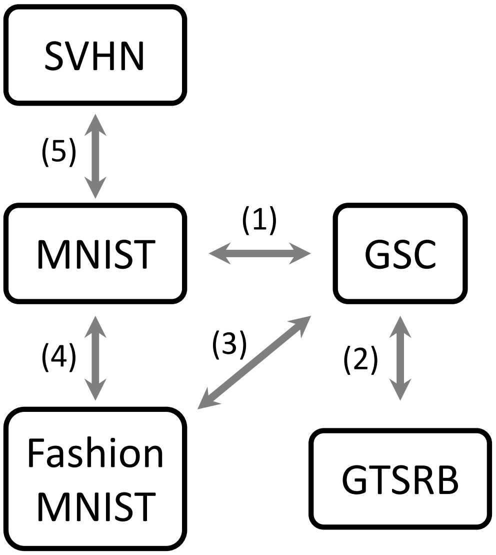


Figure 3.7: Illustration of experiment scenarios for cross-modal generation.

comparison, the hyper-parameter α was set to 0.01 for whole scenarios. Cross-modal Variational Auto-Encoder (**CVA**) [3] is an extension of VAE for cross-modal data. In CVA, the latent space is shared between two modalities. In the training process, the selected sample pair are trained alternately throughout iteration. Multimodal Variational Auto-Encoder (**MVAE**) [2] is also a variant of VAE for cross-modal data. MVAE uses the standard VAE for each modality, but each latent space is associated via a shared latent space expressing the unified distribution of the association modalities. We trained MVAE by using the sub-sampled training paradigm presented in their paper.

To evaluate the flexibility of encoding dimension in our model, we have conducted an experiment where each modality is encoded in a different dimensional space from the other. **ours-flex** has large dimension of latent space for GTSRB dataset ($dim(\mathbf{z}) = 256$). Except for ours-flex, all compared models use the same VAE of which the latent space dimension is 64.

Table 3.5: Evaluation of cross-modal association models. Accuracy is measured by the classifier for the reconstructed data of the target modality from the input data of the other modality. Bold font denotes the best performance for each case.

	Classification Accuracy (%)								
Model	SVHN → MNIST	F-MNIST → MNIST	MNIST → F-MNIST	MNIST → GSC	GSC → MNIST	F-MNIST → GSC	GSC → F-MNIST	GTSRB → GSC	GSC → GTSRB
VAE	38.73	47.36	41.86	10.34	28.61	12.83	22.18	35.93	19.44
VAE-CG	66.05	82.41	83.84	32.46	66.62	29.87	63.79	28.43	55.00
JMVAE	64.93	**83.49**	88.14	28.15	62.31	**47.58**	51.23	41.02	65.18
CVA	57.01	76.51	85.88	24.61	65.04	18.70	59.73	31.02	77.78
MVAE	31.18	62.65	77.62	23.04	46.70	13.52	33.24	28.06	69.17
ours	**74.20**	82.02	**94.26**	**59.47**	**88.66**	43.95	**77.84**	**58.89**	**77.87**
ours-flex	-	-	-	-	-	-	-	**61.39**	**80.56**

Table 3.5 shows the evaluation result of the proposed model and the compared models for the cross-modal association. The proposed model accomplishes significant enhancement from the compared algorithms for most of the scenarios. Interestingly, in the challenging scenarios such as the association between heterogeneous modalities, for instance, between voice (GSC) and image data (MNIST, GTSRB), the proposed model achieves a remarkable improvement compared to the existing methods.

3.5.4 Qualitative Results

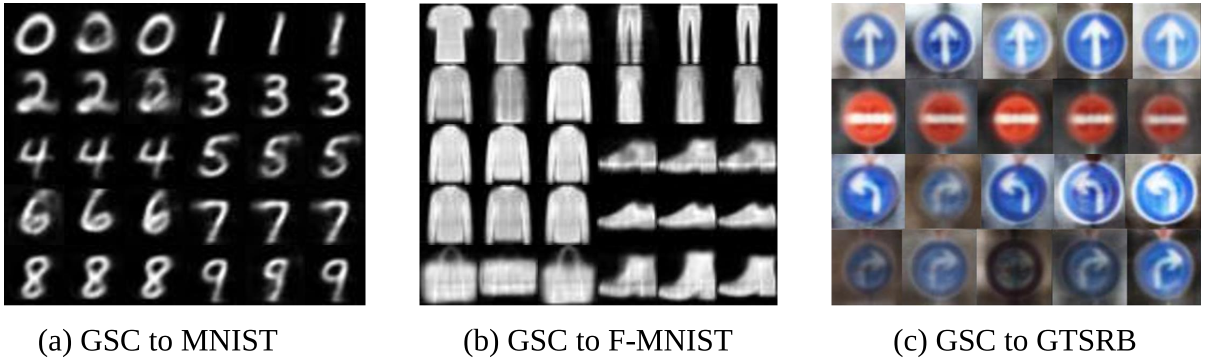
Figure 3.8 presents the qualitative results of our model for data generation from GSC dataset. Figure 3.8 (a) and (b) show 3 generated images for each 'number' command of GSC. Figure 3.8 (c) shows 5 generated images for each 'traffic command' of GSC.

3.5.5 Application: 3D Hand pose estimation

We have conducted additional experiments for 3D hand pose estimation on Rendered Hand pose Dataset (RHD) [81]. RHD dataset provides 320×320 RGB image, depth map, segmentation map and 21 keypoints for each hand. The dataset contains 41258 training and 2728 testing samples. The association target is to generate 3D keypoints from the RGB image. The evaluation metric is the average End-Point-Error (EPE), which measures Euclidean distance between ground truth keypoints and estimated keypoints. We used the same encoders and decoders structure to CVA and added our associator. The proposed model achieves **13.15**, which outperforms recent 3D hand pose estimation algorithms such as CVA (19.73) and HPS (30.42) [81]. Figure 3.9 shows qualitative results for 3D hand pose estimation.

3.5.6 Utilizing Unpaired Data

We conducted an additional experiment to verify the effectiveness of the proposed associator in semi-supervised learning. Figure 3.10 illustrates a trend of performance



(a) GSC to MNIST

(b) GSC to F-MNIST

(c) GSC to GTSRB

Figure 3.8: Qualitative results on cross-modal association from auditory dataset (GSC) to visual dataset (MNIST, F-MNIST, GTSRB). **(a)** Image generation from GSC to MNIST dataset. **(b)** Image generation from GSC to F-MNIST dataset. Results in (a) and (b) shows 3 generated images for each ‘number’ command of GSC. **(c)** Image generation from GSC to GTSRB dataset. Each row in (c) shows 5 generated images for each ‘traffic’ command of GSC.

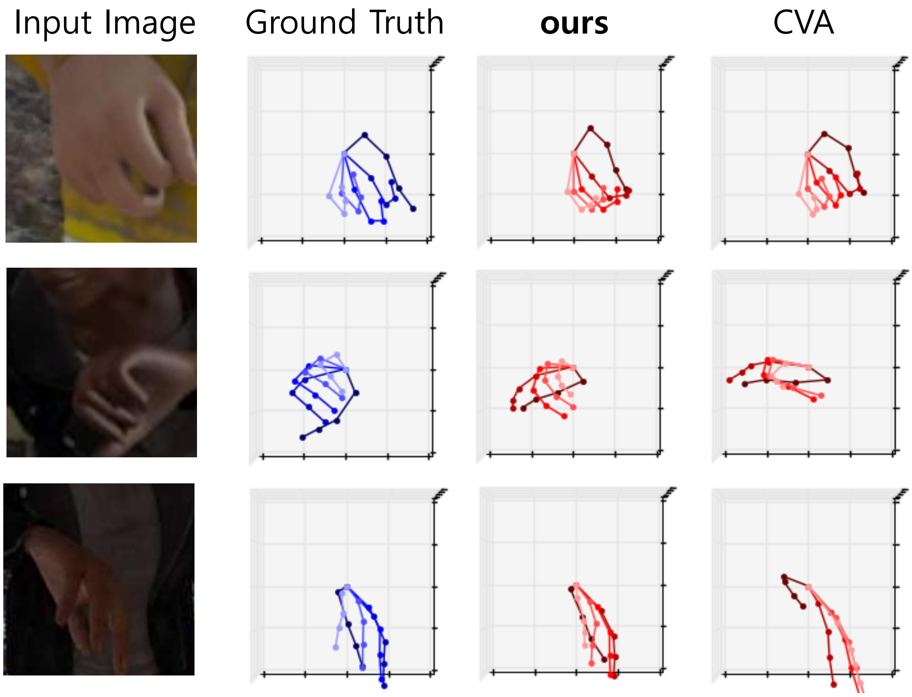


Figure 3.9: Qualitative results for 3D hand pose estimation on RHD dataset. Each column corresponds to input images, ground truth 3D keypoints, estimated 3D keypoints in order from left to right.

variation depending on the proportion of paired data, from 100% to 1% in the GSC → MNIST scenario. The result shows that the proposed associator can achieve eminent performance with only a small proportion of paired data (5%) in a semi-supervised manner.

3.5.7 Scalability

The proposed structure can easily expand a new modality while maintaining the existing modalities. That is, a new modality can be added via training of only a new

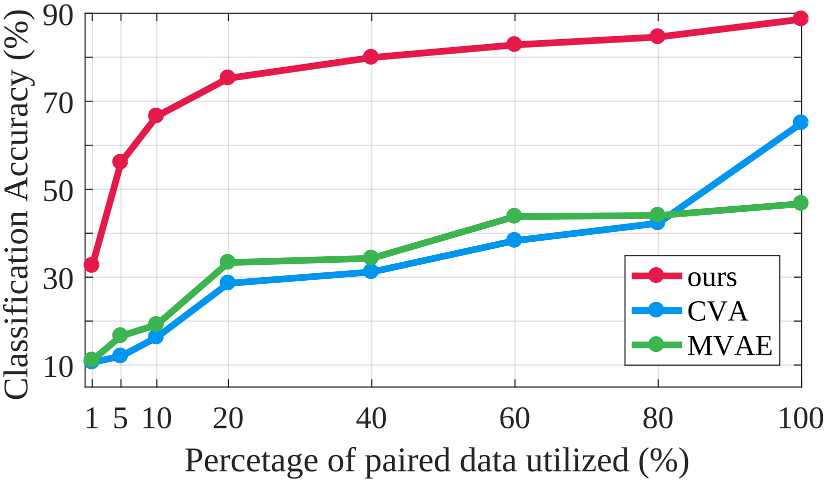


Figure 3.10: **Utilizing Unpaired Data**. Performance variation while reducing the proportion of paired data from 100% to 1% in the GSC → MNIST. Our method can achieve much better performance with only 5% paired data than the existing methods with 80% paired data.

Table 3.6: Performance of the proposed model in the case of cascading association and direct association.

GSC → MNIST	GSC → F-MNIST → MNIST
88.66	76.99

associator between an existing auto-encoder and the new auto-encoder. Since the associator connect only two latent spaces, if the existing network associates N modality, N associators need to be trained newly. In our model, this inefficiency can be mitigated by cascading association through multiple associators. Table 3.6 compares the results of cascading association and direct association for the example of MNIST, F-MNIST and GSC dataset. The F-MNIST is utilized as a medium between MNIST and

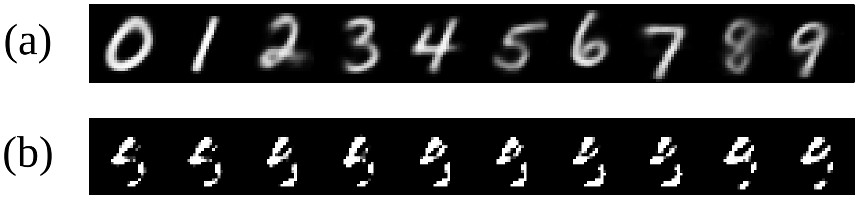


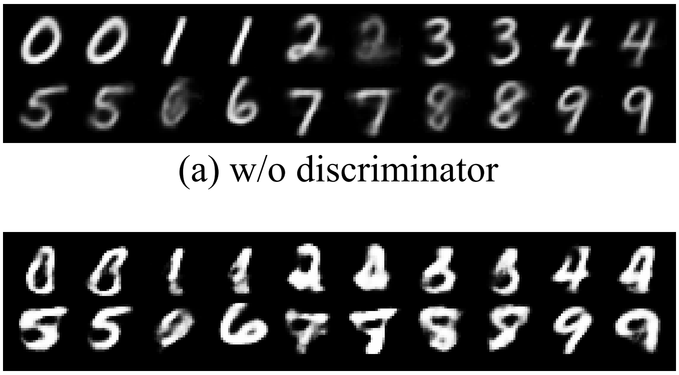
Figure 3.11: Quantitative results when the proposed model is trained **(a) with / (b) without** weight-freezing at training phase.

GSC. Although the cascading association has some performance degradation, it still has good performance compared to other algorithms presented in Table 3.5.

3.5.8 Effect of Weight-freezing

After the convergence of the intra-modal training phase, the following cross-modal training phase proceeds to train the associators while freezing the weights of the auto-encoders. By weight-freezing scheme, the proposed model can learn relationship among heterogeneous modalities and easily incorporates additional modalities while maintaining the existing modalities.

Figure 3.11 shows the quantitative results of MNIST images generated from GSC dataset, when the proposed model is trained with (a) weight-freezing and (b) without weight-freezing. When the proposed model is trained without weight-freezing (Figure 3.11 (b)), generated results converges to the similar results whatever the input audio data are given. This phenomenon is referred to as mode collapse problem. Since simultaneous training of intra-modal and cross-modal association between image and audio dataset, the model first learns mean images of MNIST datasets but fails to generate diverse outputs. On the other hand, when weight-freezing is applied, the model first learns decoding ability through intra-modal association learning (relative easy task) and freezes parameters of decoder. Therefore, cross-generated results are well



(a) w/o discriminator

(b) w/ discriminator

Figure 3.12: Quantitative results when the proposed model is trained **(a) without / (b) with** discriminator.

generated even the model learns relationships between heterogeneous modalities.

3.5.9 Adversarial Learning for Generator

Since the proposed model is based on VAE structure, generated results seem blurry. To obtain clear results, we introduced discriminator to the proposed model. Figure 3.13 shows the intra-modal association network of the proposed network with discriminator. While intra-modal association learning, discriminator is trained to discriminate input data (real data) and self-reconstructed data from decoder (fake data). Training generator (encoder and decoder) and discriminator was performed alternately. Figure 3.12 shows quantitative results when the proposed model is trained without / with discriminator. Discriminator version generates more clear images but slightly distorted. This is well known problem of adversarial learning. Therefore, it would be proper to modify the proposed model according to the purpose.

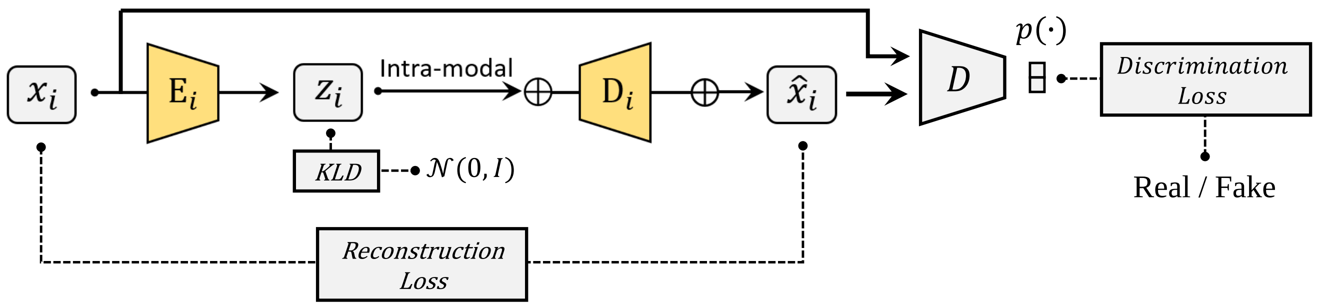


Figure 3.13: Intra-modal association network with discriminator.

3.6 Summary

We proposed a novel multi-modal association network structure that consists of multiple modal-specific auto-encoders and associators for cross-modal association. By adopting the associators, the proposed multi-modal network can incorporate new modalities easily and efficiently while preserving the encoded information in the latent space of each modality. In addition, the proposed network can effectively associate even heterogeneous modalities by designing each latent space independently and can be trained by a small amount of paired data in a semi-supervised manner. Based on the validation of our structure in experiments, future work can attempt to implement a large-scale multi-modal association network for practical use.

Chapter 4

Cross-modal Active Learning

4.1 Contribution

To increase the applicability of deep learning networks in various machine learning tasks, it is essential to collect sufficient high-quality data on target applications. However, the collection of a sufficient amount of data is a cost-consuming task. Active learning (AL) is one of the methods to collect data cost-efficiently. AL assumes a situation where only a small number of data can be annotated by the annotator with a limited budget although a large amount of unlabelled data are given. In this situation, it is important to actively select samples that should be annotated for cost-efficiently training of the target model. With the limited number of annotated samples, the target model can achieve better performance with the actively selected samples via AL than the randomly selected samples.

In multi-modal applications, collecting data is more cost-consuming than the single-modal applications. In terms of reducing annotator's labor, AL for multi-modal tasks can be much more beneficial than that for single-modal tasks. However, many previous AL studies have considered single-modal tasks such as image classification, image segmentation [55] and AL for the multi-modal applications has not been exploited much yet. In this paper, we focus on AL for multi-modal applications, especially on image-text retrieval (ITR), which is one of the most popular multi-modal tasks.

In ITR, given a query image, an ITR model should retrieve relevant text from a database, and vice versa for a query text. To train a model for ITR, most methods usually employ the contrastive learning scheme that leads a model to yield high similarity for a relevant image-text pair and low similarity for an irrelevant image-text pair. Thus, the training data for ITR contains lots of *relevant image-text pairs* and their *category labels* to predict relevance for other image-text pairs in the dataset.

In previous AL methods for ITR, annotators provide category labels for queried pairs [67]. However, according to recent ITR studies [14, 15, 15–20], category label becomes less important in a training phase. The recent ITR studies have targeted challenging benchmarks, where relevant pairs are finely-categorized into numerous categories and each category contains very few pairs [82, 83]. Thus, during a training phase, in most cases, most relevant pairs in a training mini-batch have different categories from others. For this reason, many ITR studies assume that each relevant pair in the training dataset has its own category different from the categories of the other pairs, which does not need to compare its category label to the others. Therefore, the category label is no longer utilized in the training phase and so asking a category label to annotators is meaningless for finely-categorized benchmarks.

In this paper, first, we set up a new AL scenario that is feasible to finely-categorized ITR benchmarks. In our scenario for the finely-categorized benchmarks, relevant pairs without category labels are used instead of category-annotated data. Thus, an unpaired image is regarded as an unlabelled sample that is used for a query sample to request its paired text from the annotator. For the new scenario, we develop an AL algorithm of which key idea is to select unpaired images that are expected to produce a large training loss at a training phase. Samples causing a high loss can be regarded as hard samples to the current model. Thus, the model trained with the hard samples can achieve better performance than the model trained with randomly selected samples. To this end, we utilize the triplet ranking loss function adopted in recent ITR studies that emphasize the hard negative samples [14]. Then we design an AL algorithm that selects images that can be a hard negative for as many texts as possible from the paired dataset. To determine a hard negative image for a text in the paired dataset, we suggest a scoring function to measure the 'hard negativeness' of each unpaired image sample for the given texts. Our AL algorithm selects image samples in the order of the highest score. Through extensive experiments on the Flickr30K [83] and MS-COCO [82], we validate the effectiveness of the proposed AL algorithm.

The contribution of the paper can be summarized as follow.

- We set up a novel AL scenario that is feasible to finely-categorized ITR benchmarks. In the scenario, a set of unpaired images is given and annotators provide paired texts for the limited number of images selected by an AL algorithm.
- We propose an AL algorithm for our AL scenario, which can cost-effectively construct paired data beneficial for training the model to perform ITR tasks.
- We validate the proposed AL algorithm through extensive evaluation and self-ablation studies on the Flickr30K and MS-COCO.

4.2 Proposed Active Learning for ITR

4.2.1 New AL Scenario for ITR

In this section, we set up a new AL scenario considering the characteristics of the finely-categorized dataset. In e-th epoch of the scenario, AL algorithm actively selects b valuable images $Q^{(e)}$ from the unpaired image data set $X^{(e)}$. Then an annotator provides a proper text for each image in $Q^{(e)}$, which yields a paired set $P^{(e)}$. Then $P^{(e)}$ is added to the set of accumulated relevant pairs $P_a^{(e)}$, whereas $Q^{(e)}$ is subtracted from $X^{(e)}$, resulting in $X^{(e+1)}$. Then a retrieval model $\mathcal{M}$ is trained by the accumulated paired dataset, i.e., $P_a^{(e+1)} = P_a^{(e)} \cup P^{(e)}$.

To represent initial states, we set the initial epoch index to zero, i.e., $e = 0$. Algorithm 1 and Figure 4.1 describe a detailed procedure of the proposed AL scenario for ITR. In the scenario above, an unpaired image data set is given and the annotator provides paired texts for the queried images. The reverse scenario of Algorithm 1, where an unpaired text data set is given and the annotator provides images for the queried texts, can be defined in a similar manner.

4.2.2 Key Concept of Proposed AL Algorithm

An AL algorithm selects unpaired images that are expected to largely improve the performance of a model. Our key idea for the AL algorithm is to select unpaired images that will have a large training loss at the training phase. Samples causing a large loss can be regarded as hard samples to the current model. When using the same number of training samples, the model trained with the hard samples can achieve better performance than the model trained with randomly selected samples. In our AL algorithm, we employ the triplet ranking loss modified to emphasize the hard negative samples [14]. Based on the characteristics of the triplet ranking loss, we define conditions that an image is determined as a hard negative image for a text in the paired

Algorithm 1 Proposed AL scenario for ITR

Input:

$\mathcal{M}^{(0)}$: Initial retrieval model

$P_a^{(0)}$: Initial accumulated paired dataset

$X^{(0)}$: Initial unpaired image dataset

b: The number of images to be selected at each epoch

E: Maximum epoch

Functions:

$train(\mathcal{M}, P_a)$: Train model $\mathcal{M}$ with dataset P_a

$AL(X, b, \cdot)$: Actively selects b images from X

$Annotator(Q)$: Annotate images in Q

Procedure:

```
$\mathcal{M} \leftarrow train(\mathcal{M}^{(0)}, P_a^{(0)})$
for $e = 0$ to $E - 1$ do
    # ACTIVE SAMPLE SELECTION
        $Q^{(e)} = \{x_i\}_{i=1}^{b} = AL(X^{(e)}, b, \cdot)$; Algorithm 2
        $P^{(e)} = \{(x_i, t_i)\}_{i=1}^{b} = Annotator(Q^{(e)})$
        $P_a^{(e+1)} = P_a^{(e)} \cup P^{(e)}$
        $X^{(e+1)} = X^{(e)} \setminus Q^{(e)}$
    # EVALUATION
        $\mathcal{M} \leftarrow train(\mathcal{M}^{(0)}, P_a^{(e+1)})$
end for
return $\mathcal{M}, P_a^{(E)}$
```

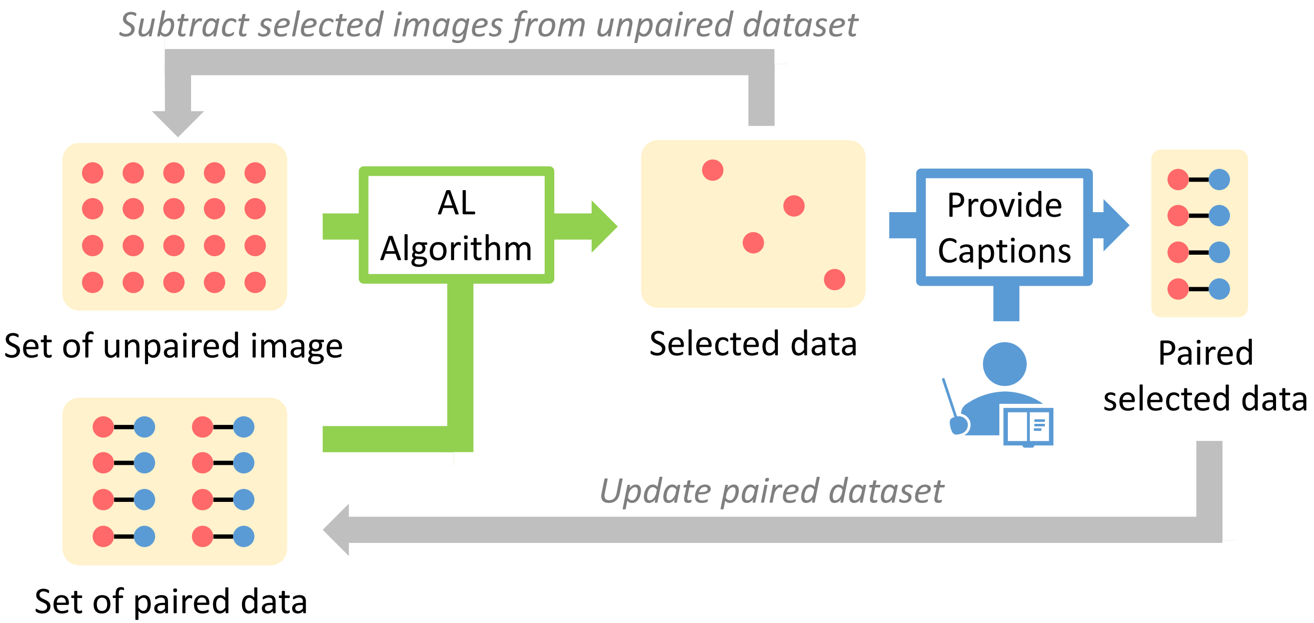


Figure 4.1: Illustration for the proposed AL scenario where paired data are collected from unpaired images.

dataset. Then we propose a scoring function that measures the 'hard negativeness' of each unpaired image sample for the texts in the paired dataset accumulated during AL. Finally, we propose the AL algorithm that selects the samples in the order of the highest score.

4.2.3 Loss Function for Training ITR Model

To perform ITR, a retrieval model is trained to yield high similarity for a relevant pair, and low similarity for an irrelevant pair. To this end, the triplet ranking loss is usually adopted for the model training [84, 85]. Especially, we consider the max of hinges loss function [14] that emphasizes the hard negative samples, i.e., only the most irrelevant pair which gives the highest similarity score is penalized. For a relevant pair (x, t) for image x and text t, the max of hinges loss is defined by

$$\begin{aligned} l(x,t) =& \max_{t'}[\alpha + s(x,t') - s(x,t)]_+ + \max_{x'}[\alpha + s(x',t) - s(x,t)]_+ \\ =& [\alpha + s(x,t^{(-)}) - s(x,t)]_+ + [\alpha + s(x^{(-)},t) - s(x,t)]_+, \end{aligned} \tag{4.1}$$

where x' (or t') represents any one image (or text) except for x (or t) in the training mini-batch. The hard negative image (or text) is denoted as $x^{(-)} = \arg\max_{x'} s(x', t)$ $(t^{(-)} = \arg\max_{t'} s(x, t'))$. $s(x, t)$ denotes the cosine similarity function between x and t, and $[x]_+$ denotes the hinge function as: $[x]_+ = \max(x, 0)$. α is a margin for the ranking loss.

4.2.4 Proposed Hard Negative Conditions

According to Eq. 4.1, the hard negative image $x^{(-)}$ makes the loss large and so can be chosen as a valuable image for AL. For a given relevant pair sample (x, t), we can obtain the hard negative sample $x^{(-)}$ from Eq. 4.1. However, in the AL, because only unpaired samples are given, we can not obtain the hard negative sample $x^{(-)}$ from Eq. 4.1. To circumvent this, we propose an approximate condition to choose the hard negative unpaired image for a certain text in a given paired data. For convenience, the condition is referred to as '*hard negative condition*'.

To establish the hard negative condition, let Z (or T) be a set of images (or texts) in the set of accumulated relevant pairs P_a. When describing the procedure in each epoch, we omit the superscript '(e)' for simplicity. x_i denotes the i-th image in X. Additionally, z_j (or t_j) as the j-th sample in Z (or T). Note that the same subscript for z and t means that they are relevant. Then, we define the hard negative condition of x_i regarding t_j as below.

Hard negative condition: $x_i \in X$ is determined as the hard negative image of t_j if $s(x_i, t_j) > \xi_j(t_j)$, where $\xi_j(t_j)$ is a threshold to be designed depending on t_j.

ξ_j can be designed in several ways. The first way is to design a threshold ξ_j, where we aim to choose x_i such that it should have higher similarity with t_j than all other images z_l in $Z \setminus \{z_j\}$. To this end, we design ξ_j as

$$\xi_j = \max_{z_l} \{s(z_l, t_j) \mid z_l \in Z \setminus \{z_j\}\}. \tag{4.2}$$

Z (full-batch) is the image set of P_a and so its size is large, which requires heavy computation when using Eq. 4.2. To reduce the computation, we design a relaxed version of ξ_j using a small subset (mini-batch) of Z as

$$\xi_j = \max_{z_l} \{s(z_l, t_j) \mid z_l \in Z_s\}, \tag{4.3}$$

where Z_s is a randomly chosen subset of $Z \setminus \{z_j\}$. We also define $T_s \subset T$ as a text subset corresponded to Z_s.

Another way for relaxing Eq. 4.2 is to replace the max function with the top-k max function which returns k-th largest value, that is, ξ_j is designed as

$$\xi_j = k^{\text{th}} \max_{z_l} \{s(z_l, t_j) \mid z_l \in Z \setminus \{z_j\}\}. \tag{4.4}$$

Eq. 4.2, 4.3, and 4.4 are referred to as a combination of Full-batch and Top-1 condition, Mini-batch and Top-1 condition, and Full-batch and Top-k condition, respectively. In a similar manner, Mini-batch + Top-k condition can be defined by replacing a max function of Eq. 4.3 with top-k function. In the ablation study, we validate the effectiveness of the threshold designs and choose one considering both computation and performance.

4.2.5 Proposed AL Algorithm

Utilizing the hard negative condition, we propose a scoring function that measures the 'hard negativeness' of each unpaired image sample for the texts in the paired dataset, accumulated during AL. The proposed scoring function of x_i counts the number of $t_j \in T$ for which x_i satisfies the hard negative condition. To this end, the hard negativeness score function $h_i(x_i)$ is defined by

$$h_i(x_i) = \sum_{t_j \in T} w_{ij} \cdot \mathbf{1}(s(x_i, t_j) > \xi_j), \tag{4.5}$$

where $\mathbf{1}(\cdot)$ is indicator function that returns 1 if the input condition is satisfied, otherwise returns 0. w_{ij} is a aggregation weight for x_i and t_j. For the mini-batch condition, T in Eq. 4.5 is replaced with T_s.

Note that when $w_{ij} = 1$ for all t_j, then h_i merely counts the number of text for which x_i satisfies the hard negative condition. This is referred to as the **Counting** weight. On the other hand, we can suppose to give more weight to the harder negatives as follows: $w_{ij} = [s(x_i, t_j) - \xi_j]_+$. In this case, w_{ij} aims to give weights for (x_i, t_j) such that $s(x_i, t_j) > \xi_j$. This is referred to as the **Surplus** weight.

Finally, the proposed AL algorithm selects b images from X, in the order of the highest score. Algorithm 2 describes the detailed procedure of the proposed AL algorithm with the combination of Full-batch and Top-1 conditions and the Surplus weight. Figure 4.2 shows an example of the score calculation.

Algorithm 2 Proposed AL Algorithm for ITR

Input:

$\mathcal{M}$: Retrieval model

P_a: Accumulated paired dataset

X: Unpaired image dataset

b: The number of images to be selected at each epoch

Functions:

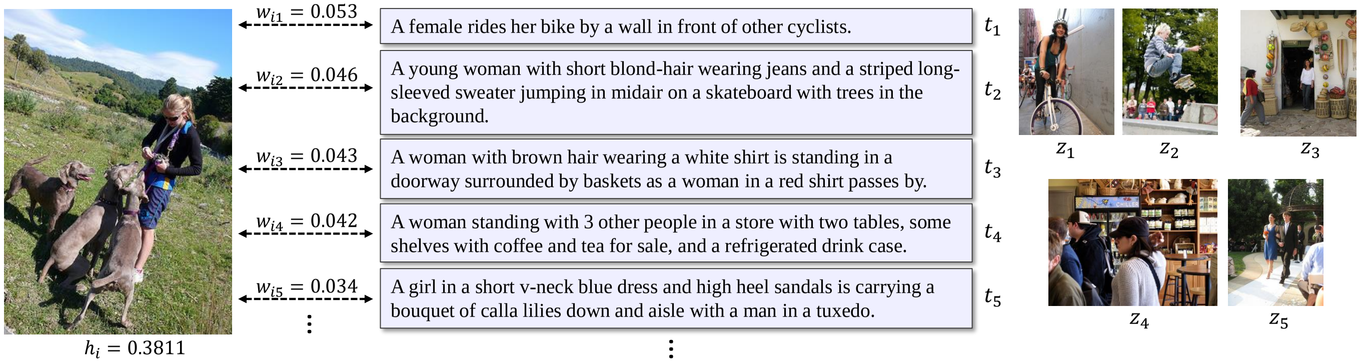
$s(x, t)$: Calculate similarity between image x and text t

$[x]_+$: Return maximum value between x and 0

$\mathbf{1}(c)$: Return 1 / 0 if condition c is true / false

Procedure:

```
$n_l = |P_a|, n_u = |X|$
Split $P_a = \{(z_j, t_j)\}_{j=1}^{n_l}$ into
$Z = \{z_j\}_{j=1}^{n_l}, T = \{t_j\}_{j=1}^{n_l}$
# CALCULATE THRESHOLD
for $j = 1$ to $n_l$ do
    $t_j \leftarrow j$-th sample of $T$
    $z_j \leftarrow j$-th sample of $Z$
    $\xi_j = \max\{s(z_l, t_j) \mid z_l \in Z \setminus \{z_j\}\}$
end for
# CALCULATE SCORE
for $i = 1$ to $n_u$ do
    for $j = 1$ to $n_l$ do
        $t_j \leftarrow j$-th sample of $T$
        $w_{ij} = [s(x_i, t_j) - \xi_j]_+$
    end for
    $h_i = \sum_{j=1}^{n_L} w_{ij} \cdot \mathbf{1}(s(x_i, t_j) > \xi_j)$
end for
$Q \leftarrow$ select $b$ images with the highest $h_i$ from $X$
Return $Q$
```


$w_{i1} = 0.053$
$w_{i2} = 0.046$
$w_{i3} = 0.043$
$w_{i4} = 0.042$
$w_{i5} = 0.034$
$h_i = 0.3811$
A female rides her bike by a wall in front of other cyclists.
A young woman with short blond-hair wearing jeans and a striped long-sleeved sweater jumping in midair on a skateboard with trees in the background.
A woman with brown hair wearing a white shirt is standing in a doorway surrounded by baskets as a woman in a red shirt passes by.
A woman standing with 3 other people in a store with two tables, some shelves with coffee and tea for sale, and a refrigerated drink case.
A girl in a short v-neck blue dress and high heel sandals is carrying a bouquet of calla lilies down and aisle with a man in a tuxedo.
t_1
t_2
t_3
t_4
t_5
z_1
z_2
z_3
z_4
z_5

(a) Unpaired image $x_i \in X$

(b) Texts $t_j \in T$ sorted with the highest weight w_{ij}

(c) Corresponding images $z_j \in Z$ to t_j

Figure 4.2: Example of the images selected by the proposed AL algorithm. **(a)** For an unpaired image x_i, **(b)** Calculated weight w_{ij} for text data $t_j \in T$ for x_i. w_{ij} can be calculated according to the designed threshold for the hard negative condition (Section 4.2.4) and the aggregation weight (Section 4.2.5). Then a score h_i for x_i can be obtained by sum w_{ij} along j. **(c)** Corresponding images for texts in (b). t_j and z_j are a relevant pair.

4.3 Experiments

4.3.1 Settings

4.3.1.1 Dataset

We have evaluated the proposed algorithm on MS-COCO [82] and Flickr30K [83] datasets, which are popular fine-grained categorized benchmarks for ITR. **MS-COCO** contains $82,783$ training images and $40,504$ validation images. Each image has five captions. Following [86], we utilized only $5,000$ for validation and $5,000$ images for testing from the original validation set. **Flickr30K** contains $31,014$ images and five captions are provided for each image. Following [86], we split the dataset into $29,000$ training images, $1,014$ validation images, and $1,000$ testing images. Since each image has five captions, we generated five positively relevant pairs. But many recent studies [15, 16] assumed that each pair among the five pairs has a different category from the others although five pairs share the same image. This data processing might not be a fatal problem for training the ITR model. However, for AL task, the image sharing might be problematic. Thus, we used only one caption for each image for training. On the other hand, for test and validation, we used all five captions for each image.

4.3.1.2 Retrieval Model

For validation of the proposed method, we employed Iterative Matching with Recurrent Attention Memory network (IMRAM) [16], one of the state-of-the-art ITR models, as our retrieval model. For computational efficiency, we used Text-IMRAM. The hyper-parameters were set following [16].

4.3.1.3 Feature Extraction

For each image, following [39], we extracted 36 local features using Faster R-CNN [59] with ResNet-101 backbone [87] pretrained on Visual Genome dataset [88]. Each local feature vector for image has 2048-dimension. For text data, following [15], each word in sentence was embedded into 300-dimensional vector first. Then we utilized bi-directional GRU [89] to extract final feature for each word. Final feature vector for word has 1024-dimension.

4.3.1.4 Training Scheme

At each epoch of the proposed scenario, we trained a retrieval model from scratch, with Adam optimizer [90] during 40 epochs. Learning rate was initially set to 0.0002 and decayed to 0.00002 at 20 epoch. We performed validation during last 10 epoch, and chose the model with the best validation performance.

4.3.1.5 AL Settings

Randomly selected 30% of the entire paired data were assigned to $P_a^{(0)}$. Then the remaining 70% images were assigned to $X^{(0)}$. We set a maximum epoch of the AL scenario to $E = 3$ and b to 5% of the cardinality of the entire data set. Therefore, after completing the scenario, $|P_a^{(E)}|$ becomes 45% of the cardinality of the entire dataset.

4.3.1.6 Hyper-parameters

For Flickr30K, a threshold ξ was determined with combination of Full-batch and Top-1 condition. For MS-COCO, we determined ξ with combination of Mini-batch and Top-1 condition to reduce the computational complexity. For both datasets, aggregation weight w was determined with the Surplus weight.

4.3.1.7 Evaluation

We evaluated the retrieval model on two tasks. In a Image Retrieval task, a model retrieves images given a text query. In a Text Retrieval task, a model retrieves texts given an image query. Performance was measured by Recall at K (R@K) metric. We evaluated the model with $K = 1$, 5, and 10.

For each epoch of the scenario, we trained the model with the accumulated paired data set ($P_a^{(e+1)}$ for e-th epoch) and reported the test performance of the trained model. For MS-COCO, the model was validated by both testing on full 5K test images (COCO-5K) and averaging the results over five subsets of 1K test images (COCO-1K).

4.3.2 Ablation and Self Study

4.3.2.1 Validation on Hyper-parameters

We evaluated the proposed algorithm with various hard negative conditions (Section 4.2.4) and aggregation weights (Section 4.2.5). Table 4.1 shows the R@1 performance at each epoch of AL scenario for Flickr30K. To evaluate the overall performance over epoch, we also reported R@1-sum metric that sums all R@1 performances over epoch for both text retrieval and image retrieval task.

According to the results in Table 4.1, the dominant combination that achieves the best performance at every epoch does not exist. But the combination of Top-1 condition and Surplus weight achieves the best R@1-sum performance by 298.5 and 298.6, about 2 $\sim$ 6 better than the other combinations. Therefore, we mainly considered Top-1 condition and Surplus weight as our default setting. Unless otherwise specified for the top-k condition and the aggregation weight, the Full-batch / Mini-batch version of the proposed algorithm indicates the algorithm to which a combination of Full-batch / Mini-batch and Top-1 condition and Surplus weight are applied.

Each of Full-batch and Mini-batch versions has its own advantages and disadvantages. Full-batch version has advantages in performance. For R@1-sum, R@5-sum, and R@10-sum values, Full batch version achieves 298.5/520.0/602.9, whereas Mini-batch version achieves 298.6/520.2/600.7, respectively. Though both show similar R@1-sum and R@5-sum performance, the Full-batch version is slightly better in the R@10-sum, Mini-batch version is computationally efficient. Mini-batch version requires similarity calculations of $|Z_s|\cdot|T_s|+|X|\cdot|T_s|$, whereas Full-batch version requires $|Z|\cdot|T|+|X|\cdot|T|$. In our experimental setting on Flickr30K, $|Z_s|=|T_s|=2560$ is considerably smaller than $|Z|=|T|\in[8700,13050]$. (For MS-COCO, $|Z|\in[24834,37252]$) Therefore, a trade-off in performance and computational cost can be negotiated between full and mini-batch versions.

4.3.2.2 Validation on Relaxed Condition

In the Section 4.2.4, we proposed relaxed conditions including Mini-batch condition (Eq 4.3) and Top-k condition (Eq 4.4). To validate the relaxation effect of proposed conditions, we chose hard negative unpaired images satisfying the relaxed condition for at least one text in the paired dataset. Then we compared the ratio of the hard negative images in the unpaired image dataset with those of the non-relaxed conditions. High ratio means that the relaxation effect is large.

Table 4.2 shows the ratio of hard negative images at each epoch of the AL scenario, depending on hard negative conditions. According to Table 4.2, the ratio of hard negative images increases when Mini-batch condition is applied and k of Top-k condition increases. This implies that the combination of Mini-batch and Top-k conditions gives the largest relaxation effect.

It is interesting to note that the ratio of the hard negative images decreases as the AL scenario progresses. The more training paired data accumulated as the AL scenario progresses, any unpaired image is more likely to be similar to the images in the accumulated paired dataset. In addition, the retrieval model is trained to predict

Condition	Weight	Text Retrieval (R@1)				Image Retrieval (R@1)				
		$e = 0$	$e = 1$	$e = 2$	$e = 3$	$e = 0$	$e = 1$	$e = 2$	$e = 3$	R@1-sum
Full-batch + Top-1	Surplus	40.2	42.9	44.8	46.4	29.1	30.2	32.3	32.8	**298.5**
	Count	40.2	42.1	44.7	45.3	29.1	29.9	32.0	33.5	296.6
Full-batch + Top-5	Surplus	40.2	42.4	42.8	44.9	29.1	30.3	31.1	32.7	293.5
	Count	40.2	41.4	44.0	45.4	29.1	30.4	31.6	32.8	294.8
Full-batch + Top-10	Surplus	40.2	42.2	43.0	45.8	29.1	29.8	31.3	32.9	294.2
	Count	40.2	41.7	45.0	46.7	29.1	29.9	31.4	32.9	296.8
Mini-batch + Top-1	Surplus	40.2	41.9	45.1	47.1	29.1	30.2	32.2	33.0	**298.6**
	Count	40.2	42.6	44.7	45.0	29.1	30.1	32.1	32.5	296.3
Mini-batch + Top-5	Surplus	40.2	41.1	45.0	45.7	29.1	30.0	31.9	33.3	296.1
	Count	40.2	43.2	44.0	45.7	29.1	30.5	31.2	32.6	296.4
Mini-batch + Top-10	Surplus	40.2	40.8	44.1	45.4	29.1	29.6	31.5	32.3	292.9
	Count	40.2	41.2	44.8	44.6	29.1	29.9	31.6	32.6	293.8

Table 4.1: R@1 performance of the proposed AL algorithm at each epoch of AL scenario, according to the hard negative condition and aggregation weight for the Flickr30K.

Condition	% of Hard neg. Images		
	$e = 0$	$e = 1$	$e = 2$
Full-batch + Top-1	55.04	44.85	39.72
Full-batch + Top-5	92.64	89.85	88.17
Full-batch + Top-10	98.27	97.51	96.98
Mini-batch + Top-1	56.23	49.71	47.11
Mini-batch + Top-5	95.26	93.94	92.76
Mini-batch + Top-10	99.24	99.14	99.19

Table 4.2: Percentage of the hard negative unpaired images in the unpaired image dataset depending on the hard negative condition.

low similarity between texts and negative images in the accumulated paired dataset. Therefore, the retrieval model is likely to predict low similarity between any unpaired image and texts in the accumulated dataset. In consequence, the number of unpaired images that satisfy the condition will decrease.

4.3.2.3 Validation on Score Function

In the Section 4.2.5, we proposed the scoring function. If scores have similar values for all unpaired images, the images selected in the order of the highest score are not different from the randomly selected images. To observe the detailed distribution of score values, we present histograms of score values calculated for unpaired images. Figure 4.3 shows the histograms depending on the aggregation weights and the hard negative conditions, at $e = 0$ of AL scenario. The vertical axis indicates the number of images and the horizontal axis denotes the score values h_i. According to the results, the scores are diversely distributed. Therefore, the samples selected in the order of the

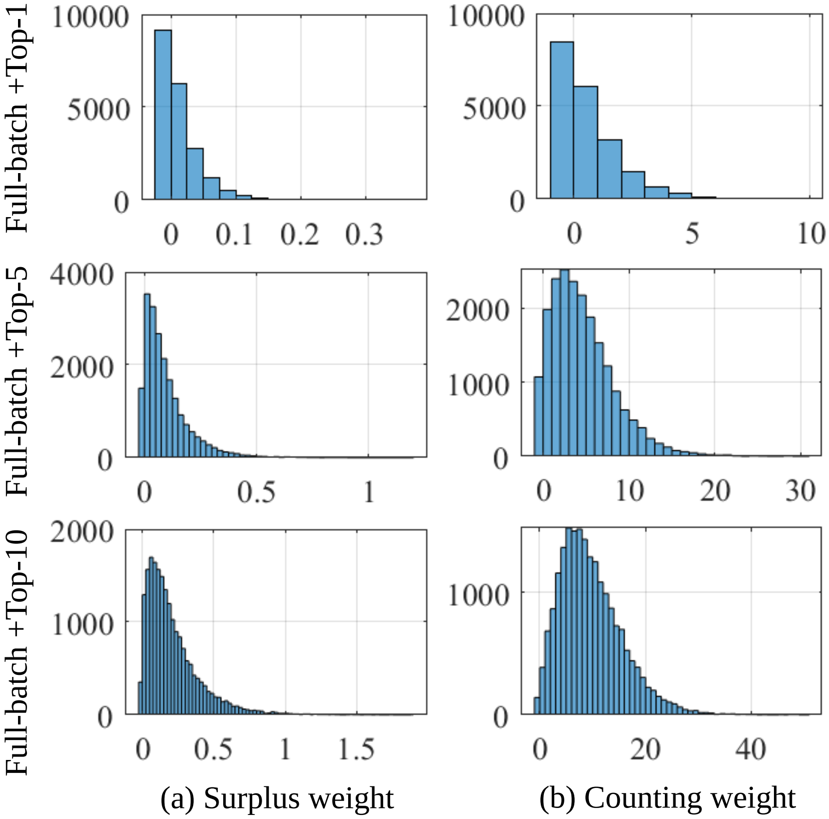


Figure 4.3: Histogram of score values of unpaired images depending on the aggregation weights (columns) and the hard negative conditions (rows).

highest score are distinct from the randomly selected samples. Full histogram results are provided in the Appendix 4.5.4.

4.3.2.4 $|Z_s|$ for Mini-batch Condition

For the mini-batch version of the proposed algorithm, the cardinality of subset Z_s, $|Z_s|$, needs to be determined. To validate the effect of $|Z_s|$, we evaluated the Mini-batch version algorithm by increasing $|Z_s|$ from 640 to 5120.

Figure 4.4 shows R@1-sum (top), R@5-sum (middle), R@10-sum (bottom) per-

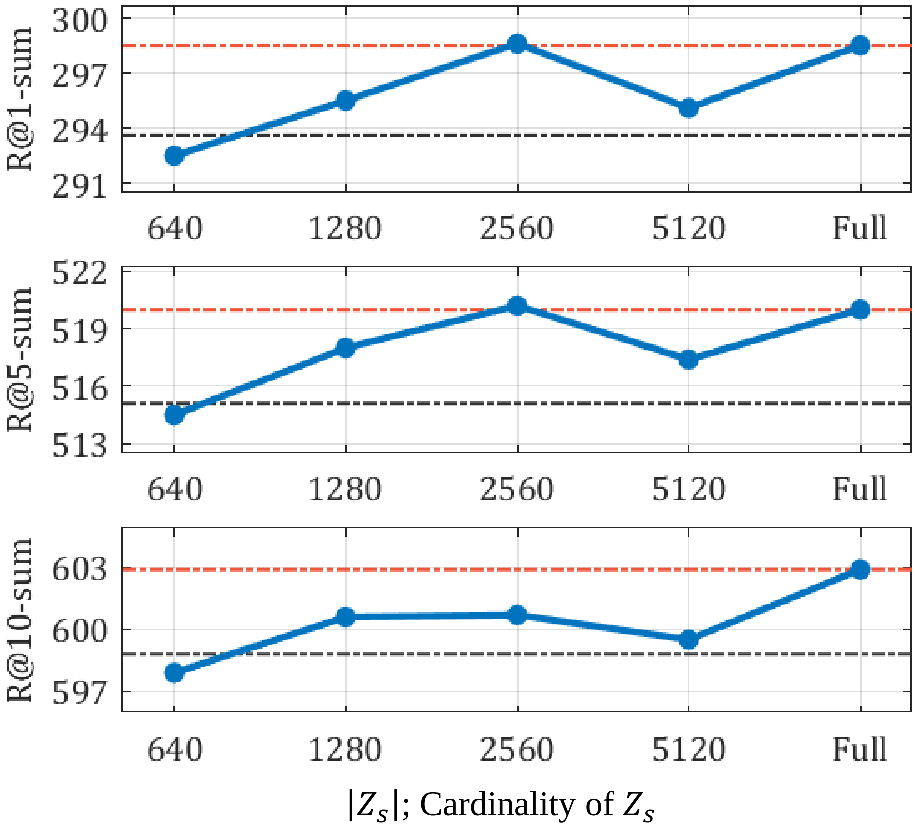


Figure 4.4: Performance of the proposed method with Mini-batch version, according to the cardinality of Z_s.

formances. Red and black dotted lines denote the performance of the Full-batch version and random selection, respectively. When $|Z_s|\geq$ 1280, the mini-batch version always shows better performance than random selection. The best performance is achieved when $|Z_s|=$ 2560, which is equivalent to 40 times of the training mini-batch size and 20% of the full-batch version. We fixed $|Z_s|=$ 2560 for Mini-batch version during the other experiments.

4.3.3 Comparisons

Since the proposed AL scenario is the new one, there are no existing works. Hence we compared the proposed method with the baseline with random selection and our AL scenario version using Core-set algorithm for the single-modal AL [54]. Detailed implementation of our modified Core-set version is provided in the Appendix 4.5.2. For the MS-COCO, Core-set version was not compared due to the GPU memory limitation.

Figure 4.5 shows the evaluation results. We can see that the proposed method achieves the best performance in most epochs, compared to both random baseline and Core-set. When performing image-text retrieval, only the similarity between image and text is considered. Thus, the proposed method that selects images by considering the similarity between image-text seems to select more valuable samples rather than the Core-set method that considers only the relationship between images. In fact, as shown in Figure 4.2, the images corresponding to texts having the highest similarity to x_i are not very similar to x_i.

An interesting point is that, at the first epoch (when the percentage of paired data is 35%), the proposed algorithm achieves much better $\mathrm{R}@1$ performance than the other algorithms by $[0.4 \sim 1.6]$. Since the proposed algorithm selects hard negative images for the retrieval model, the hard negative images decrease as epochs progress. Thus the proposed algorithm needs to select harder negative images than the previously selected hard negative images. However, it is much more difficult to select hard negative images in later epochs of AL scenario. Therefore, the proposed method shows especially high performance in the first epoch of the AL scenario.

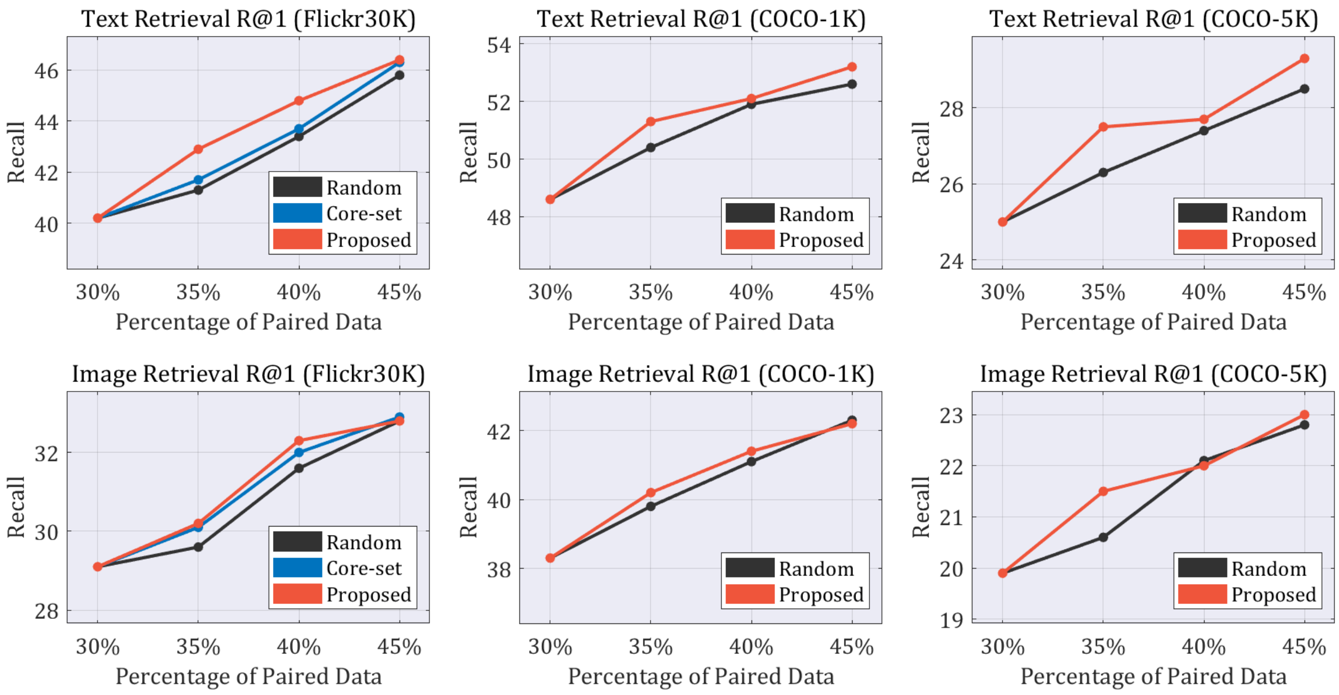


Figure 4.5: Evaluation results on Flickr30K and MS-COCO. Each graph shows R@1 retrieval performance at each epoch of the AL scenario. x-axis represents the ratio of paired data to the entire dataset, i.e. $|P_a|/(|P_a|+|X|)$, and y-axis denotes the R@1 performance.

4.4 Summary

In this paper, first, we have suggested a new AL scenario for ITR where unpaired images are given and the annotator provides their corresponding paired texts. Then we have developed our own AL algorithm that chooses unpaired samples that are expected to have a large training loss, especially triplet ranking loss [14] in our work. The key components of our AL algorithm are 1) hard negative conditions to mine the hard negative images for constructing new paired data; 2) the scoring formula that weighs the number of texts satisfying the hard negative conditions, which is used as the criterion to determine the hard negative images. We demonstrated the effectiveness of the proposed algorithm through extensive experiments for self&ablation studies, and comparisons on Flickr30K and MS-COCO.

Dataset	#Image	#Text	#Class
NUS-WIDE-1.5K	1,521	1,521	30
LabelMe	2,688	2,688	8
Wikipedia	2,866	2,866	20
Pascal-VOC	6,146	6,146	20
Flickr8K	8,091	40,455	8,091
Flickr30K	31,014	155,070	31,014
MS-COCO	123,287	616,435	123,287

Table 4.3: Configuration of the popular ITR benchmarks.

4.5 Appendix

4.5.1 Configuration of ITR Datasets

ITR dataset contains relevant image-text pairs categorized into several classes. In datasets such as Wikipedia [91], LabelMe [92], Pascal VOC2007 [93] and NUS-WIDE [94] utilized in [67], data are categorized according to high-level (coarse) semantics. Therefore, each category contains many relevant pairs. we referred to those datasets as coarsely-categorized datasets. But recent ITR studies have validated their algorithm at more challenging datasets such as Flickr [83] and MS-COCO [82]. These datasets, on the other hand, contain data that are categorized according to low-level (fine) semantics. Thus, data samples are discriminated more thoroughly and precisely than the coarsely-categorized datasets. In other words, the number of category has increased, but the number of data samples has relatively decreased compared to coarsely-categorized datasets. Table 4.3 shows the configuration of the popular coarsely-categorized datasets used in [67] (top side) and finely-categorized datasets (bottom side).

Algorithm	Text Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Core-set-mean	45.2	74.8	**84.3**	**33.2**	59.2	69.8
Core-set-BoW	**46.3**	**75.0**	83.7	32.9	**59.3**	**70.3**
Random	45.8	74.2	84.3	32.8	59.8	69.7
Proposed	46.4	75.1	84.4	32.8	59.9	70.9

Table 4.4: Performance of Core-set depending on the feature extraction method.

4.5.2 Implementation Details for Core-set

Core-set algorithm [54] extracts one global feature vecror from an image sample, and then utilizes distances between any two feature vectors. However, IMRAM model extracts local feature vectors from 36 local regions in an image. Therefore, in order to apply Core-set method to our setting, it is necessary to extract one global feature vector for an image.

We obtain the global feature vector via two methods. One is to extract a 2048 dimensional global feature vector by average 36 local feature vectors extracted from an image, referred to as 'Core-set-mean' in Table 4.4. The other one is to extract a 300 dimensional global feature vector of which element is the number of local feature vectors belonging to a local cluster region formed by K-means clustering, referred to as 'Core-set-BoW' in Table 4.4.

Table 4.4 shows the evaluation results at the last epoch of AL scenario, when above two methods are applied to Core-set. We also present the performance of proposed method and random selection. Core-set-mean and Core-set-BoW show comparable performances, but Core-set-BoW shows slightly better performances than Core-set-mean. Therefore, we compared Core-set-BoW with ours in the experiments.

4.5.3 Distribution of Selected Samples

To observe the tendency of samples selected by the proposed algorithm, we visualized the selected samples in tSNE embedding space. Figure 4.6 shows the visualization results of the samples selected by the proposed method (Top) and random selection (bottom) at each epoch of AL scenario. Blue and gray circles represent the selected and non-selected unpaired images, respectively. The selected images extracted by both our and random methods are evenly distributed over the entire region, which shows that our results also are not biased to a region, along with achieving meaningful improvement of performance.

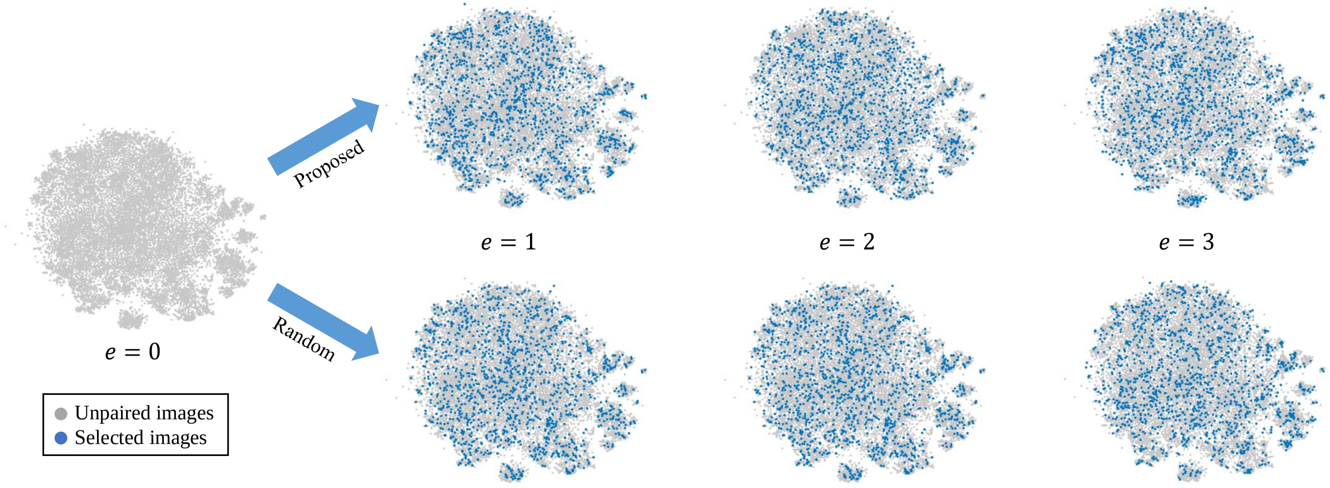
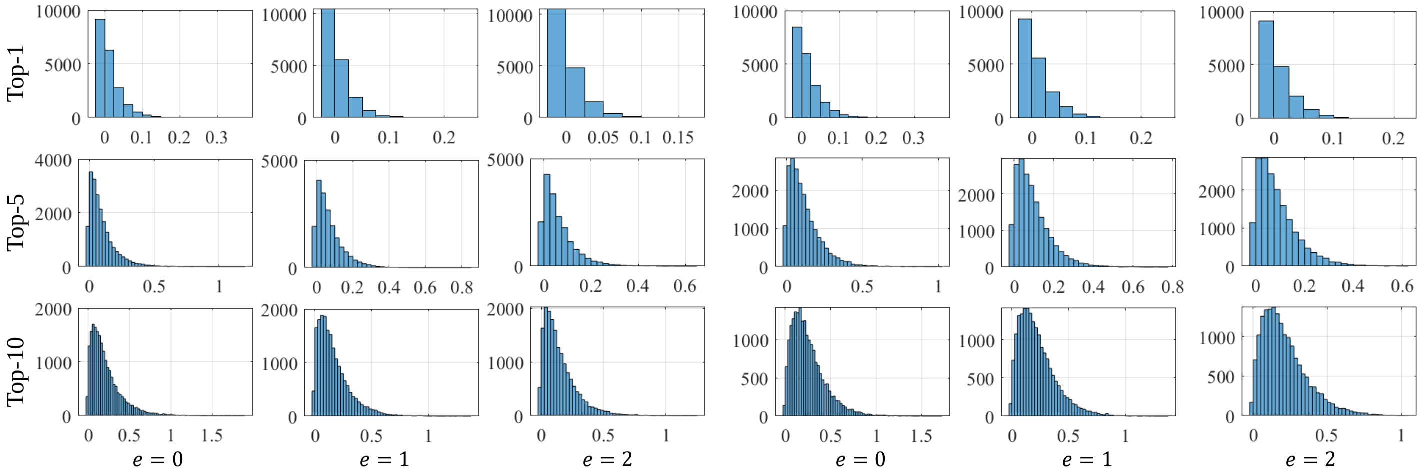


Figure 4.6: Visualization of images selected by the proposed method (Top) and the random selection (bottom) in the tSNE embedding space.

4.5.4 Full Experimental Results for Section 4.3.2.3

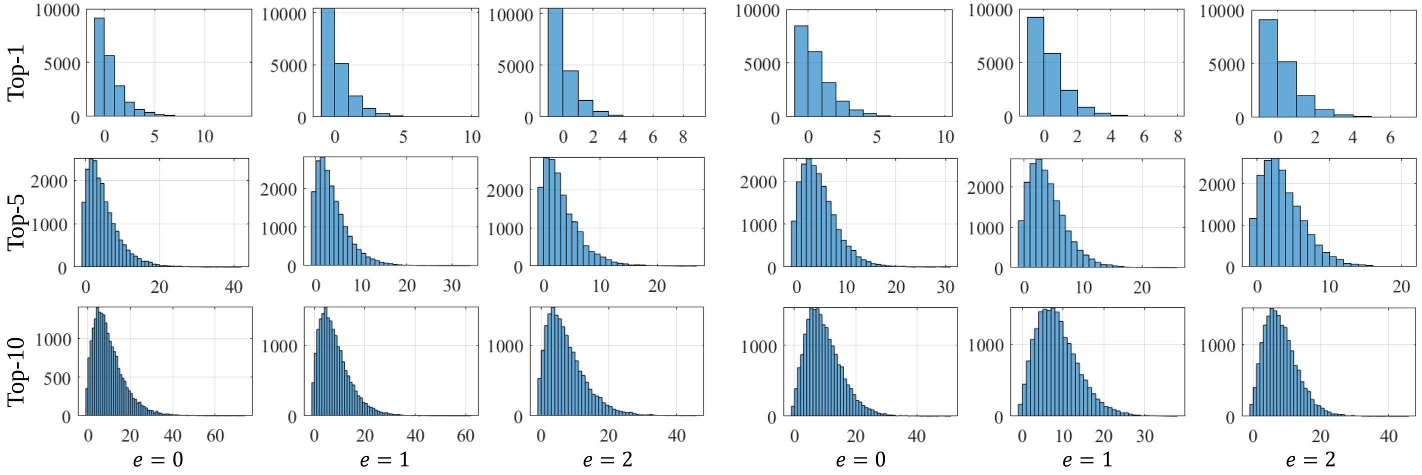
In the Figure 4.3 in the main document, we provide the histograms of scores at $e = 0$. Figure 4.7 shows full results at each iteration $e = 0, 1, 2$: (a) Full-batch + Top-k condition with Surplus weight and (b) Mini-batch + Top-k condition with Surplus weight. Figure 4.8 shows full results at each iteration $e = 0, 1, 2$: (a) Full-batch + Top-k condition with Counting weight and (b) Mini-batch + Top-k condition with Counting weight.



(a) Full-batch + Top-k

(b) Mini-batch + Top-k

Figure 4.7: Histogram of scores of unpaired images depending on the hard negative conditions with the Surplus weights. The leftmost bar indicates the number of images with $h_i = 0$.



(a) Full-batch + Top-k

(b) Mini-batch + Top-k

Figure 4.8: Histogram of scores of unpaired images depending on the hard negative conditions with the Surplus weights. The leftmost bar indicates the number of images with $h_i = 0$.

4.5.5 Full Experimental Results for Section 4.3.3

Figure 4.9 and 4.10 show R@5 and R@10 results of the proposed algorithm and comparisons.

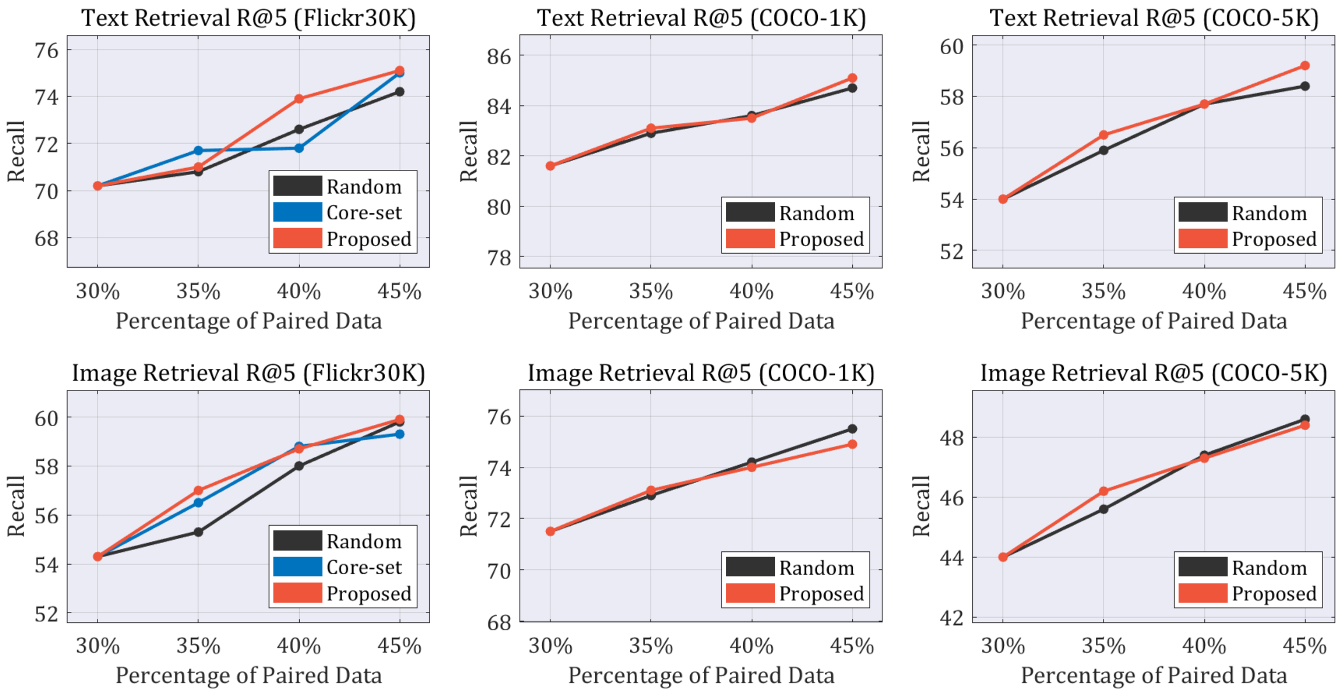


Figure 4.9: Evaluation results (R@5) of the proposed AL algorithm on Flickr30K and MS-COCO.

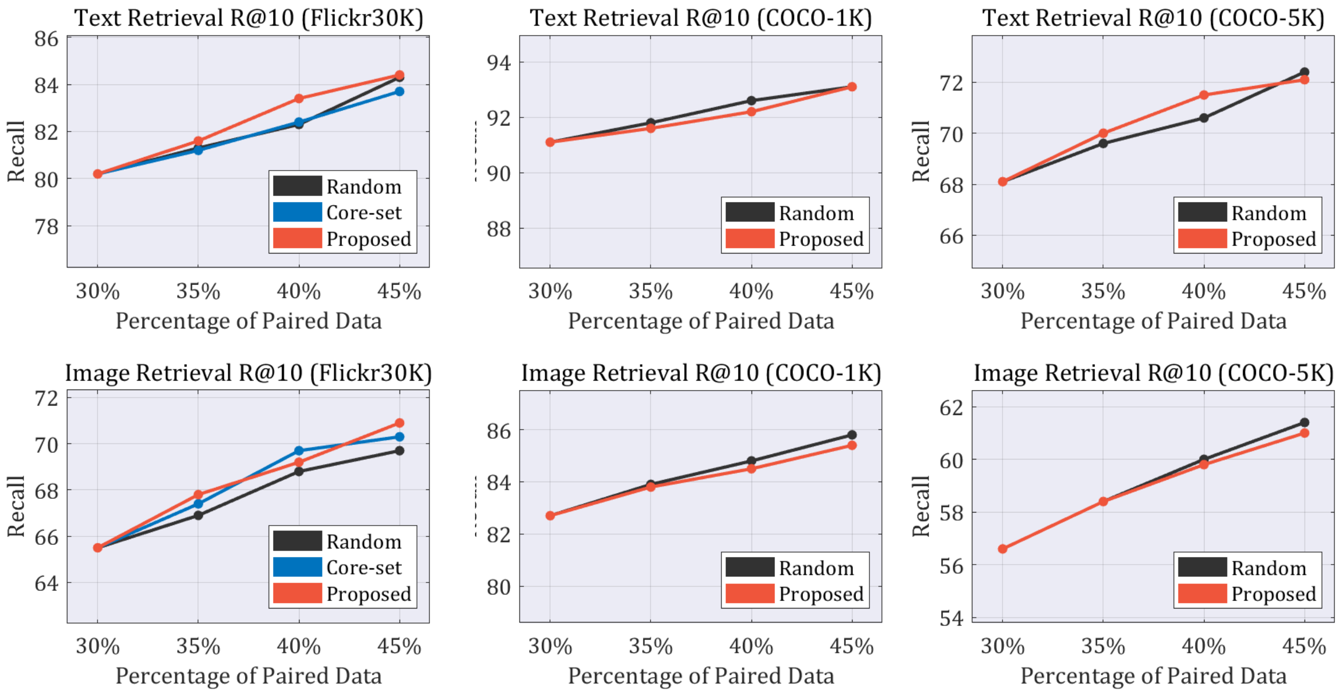


Figure 4.10: Evaluation results (R@10) of the proposed AL algorithm on Flickr30K and MS-COCO.

4.5.6 Full Experimental Results for Section 4.3.2.1

Table 4.5 and 4.6 provide R@5 and R@10 results for all combinations of the hard negative conditions presented in Section 4.3.2.1. Similar to the R@1 results in Table 4.1 in the main document, the combination of Top-1 conditions and Surplus weight still shows the best performances at R@5-sum and R@10-sum metric.

Figure 4.11, 4.12, 4.13, and 4.14 present full evaluation results with aggregation weight function is fixed. Figure 4.15, 4.16, 4.17, 4.18, 4.19, and 4.20 present full evaluation results with the hard negative condition is fixed.

Condition ξ_j	Weight w_{ji}	Text Retrieval (R@5)				Image Retrieval (R@5)				
		$e=0$	$e=1$	$e=2$	$e=3$	$e=0$	$e=1$	$e=2$	$e=3$	R@5-sum
Full-batch + Top-1	Surplus	70.2	71.0	73.9	75.1	54.3	57.0	58.7	59.9	**520.0**
	Count	70.2	70.9	73.3	75.0	54.3	56.2	58.2	59.8	517.8
Full-batch + Top-5	Surplus	70.2	70.5	74.1	73.5	54.3	56.1	57.7	59.5	515.7
	Count	70.2	70.7	73.3	74.9	54.3	56.3	58.5	59.1	517.2
Full-batch + Top-10	Surplus	70.2	70.9	72.6	74.7	54.3	55.7	57.6	59.2	515.1
	Count	70.2	54.3	74.3	74.4	54.3	56.1	58.0	59.2	516.7
Mini-batch + Top-1	Surplus	70.2	72.2	73.6	75.6	54.3	56.1	58.6	59.7	**520.2**
	Count	70.2	71.3	73.3	74.7	54.3	56.6	58.6	59.8	518.6
Mini-batch + Top-5	Surplus	70.2	70.8	73.2	74.2	54.3	56.3	57.6	59.1	515.6
	Count	70.2	70.8	73.2	74.2	54.3	56.3	57.6	59.1	515.9
Mini-batch + Top-10	Surplus	70.2	71.9	72.9	74.4	54.3	56.0	57.7	58.9	516.2
	Count	70.2	70.5	74.1	74.2	54.3	56.1	57.8	58.8	515.9

Table 4.5: R@5 performance of the proposed AL algorithm at each epoch of AL scenario, according to the hard negative condition and aggregation weight for the Flickr30K.

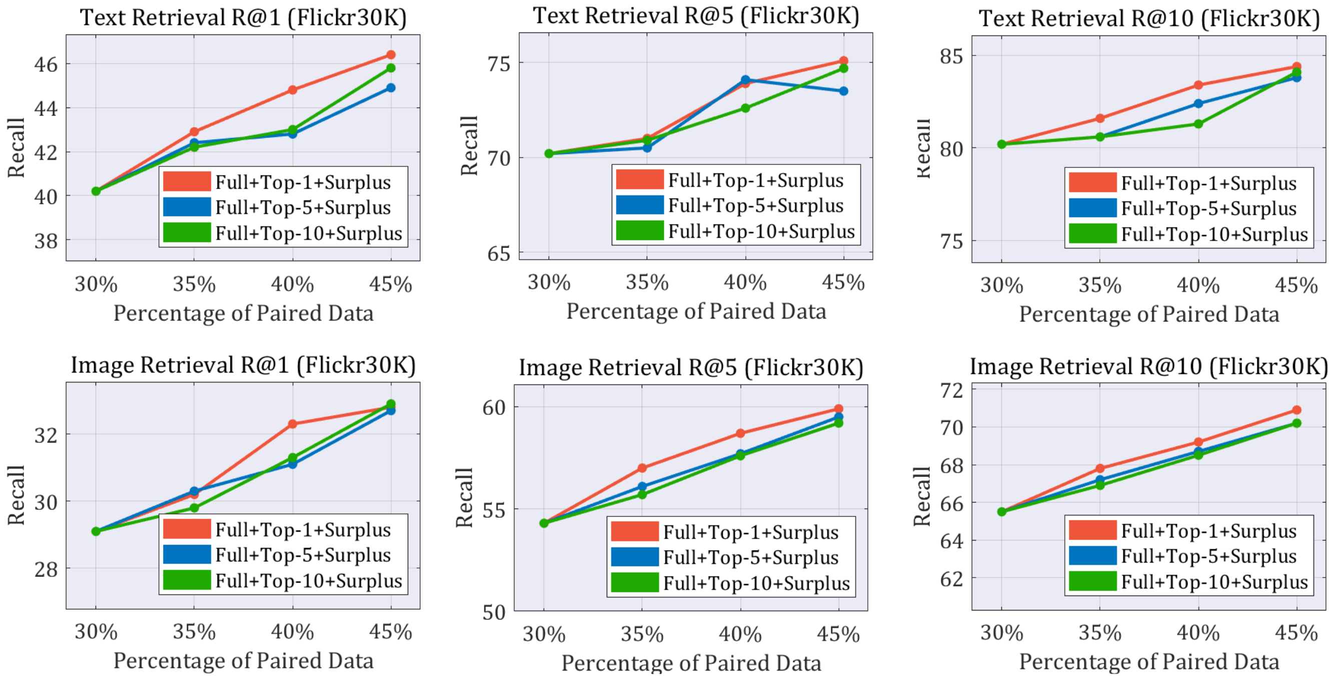


Figure 4.11: Ablation study for ξ_j.

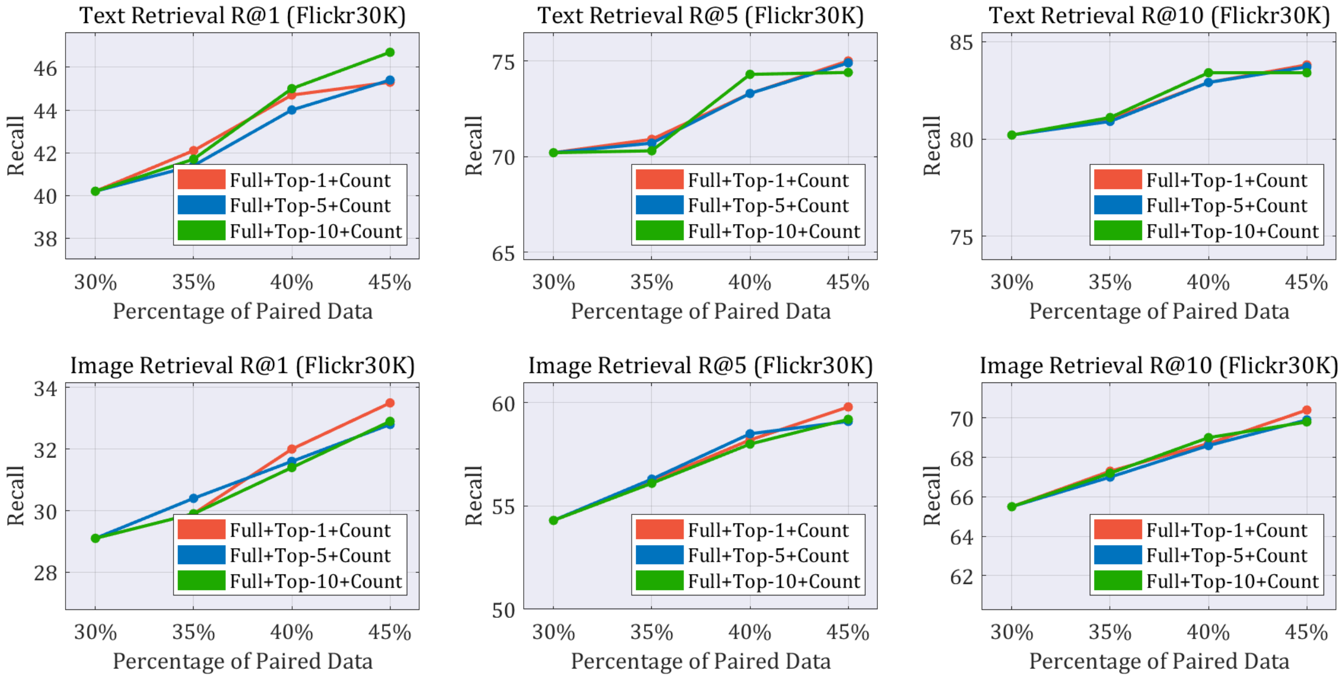


Figure 4.12: Ablation study for ξ_j.

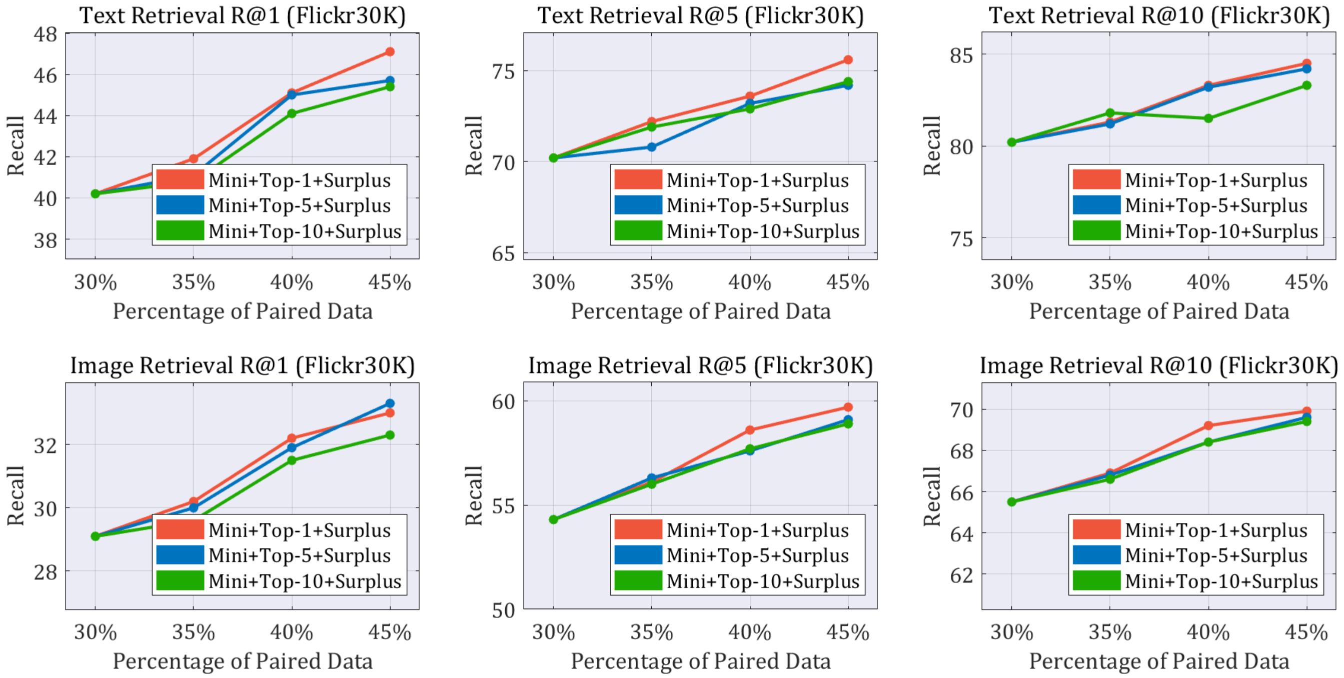


Figure 4.13: Ablation study for ξ_j.

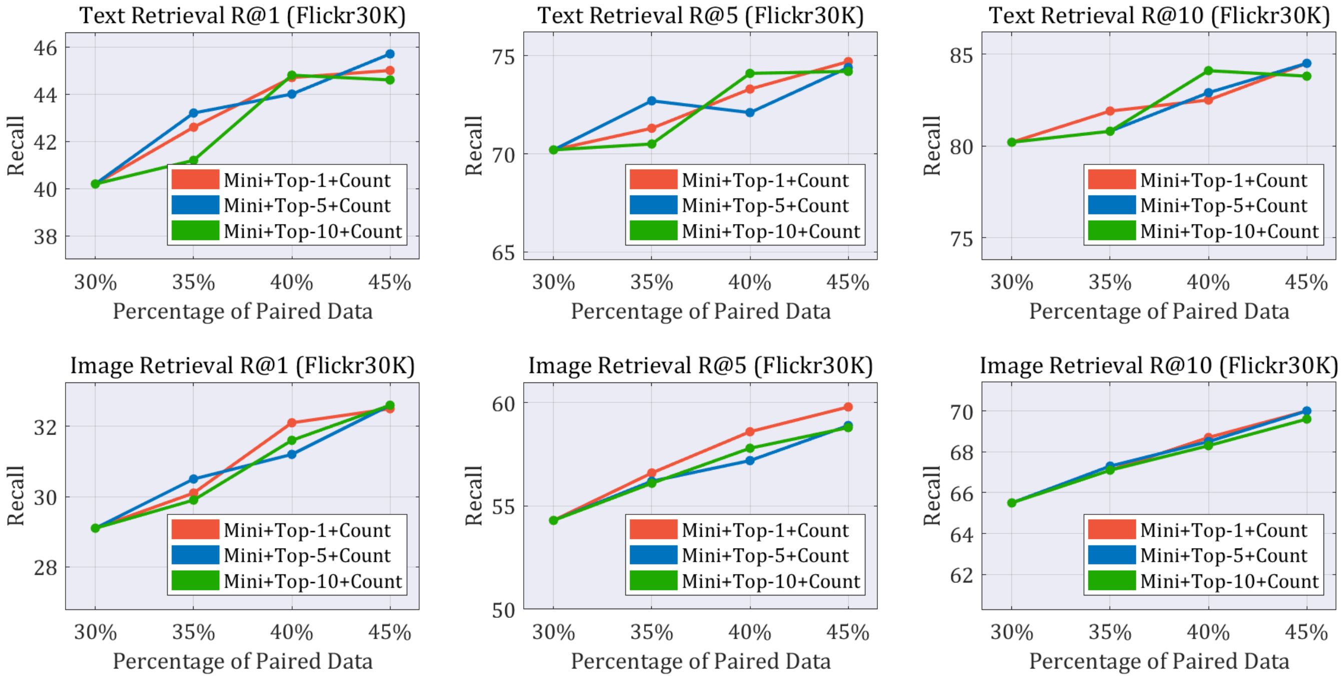


Figure 4.14: Ablation study for ξ_j.

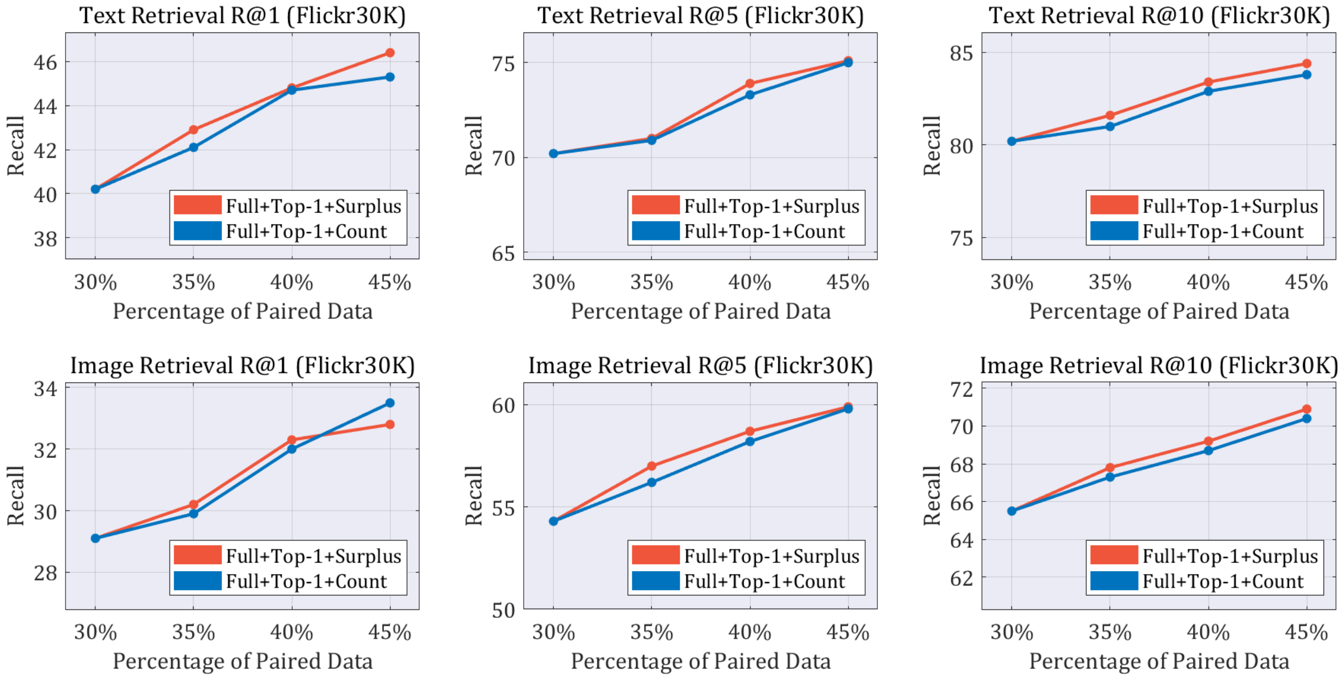


Figure 4.15: Ablation study for w_{ij}.

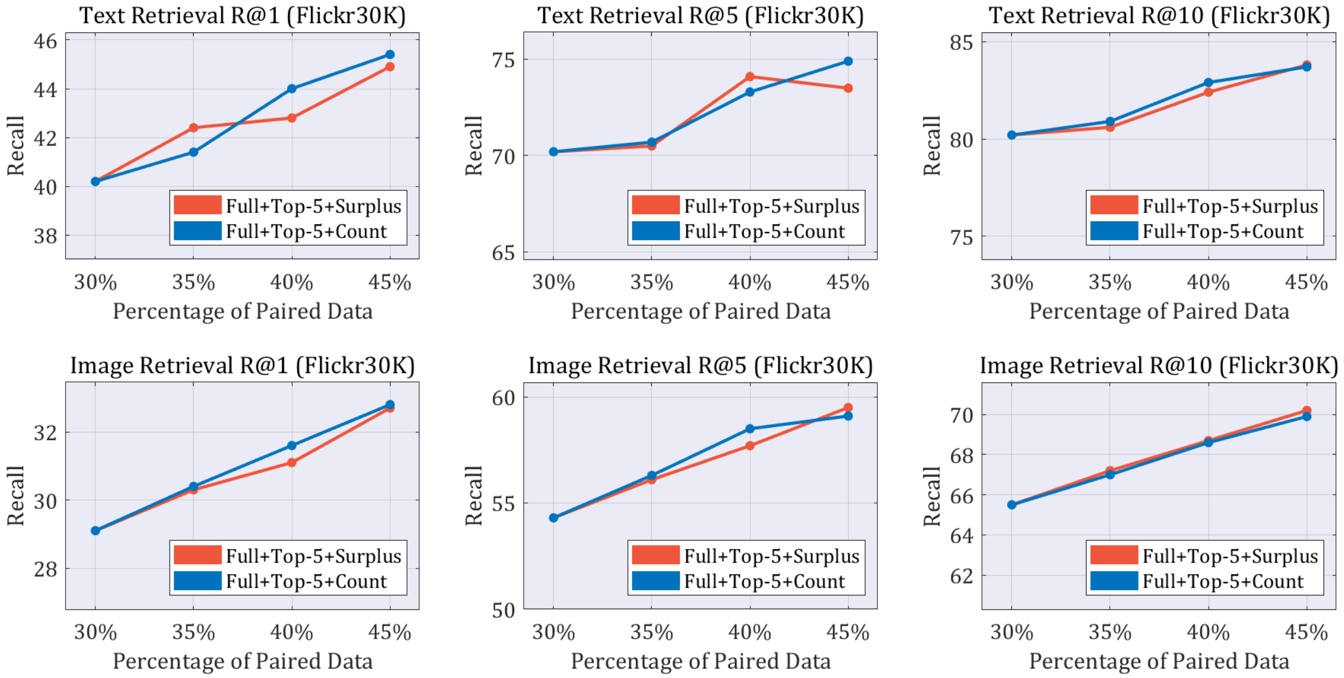


Figure 4.16: Ablation study for w_{ij}.

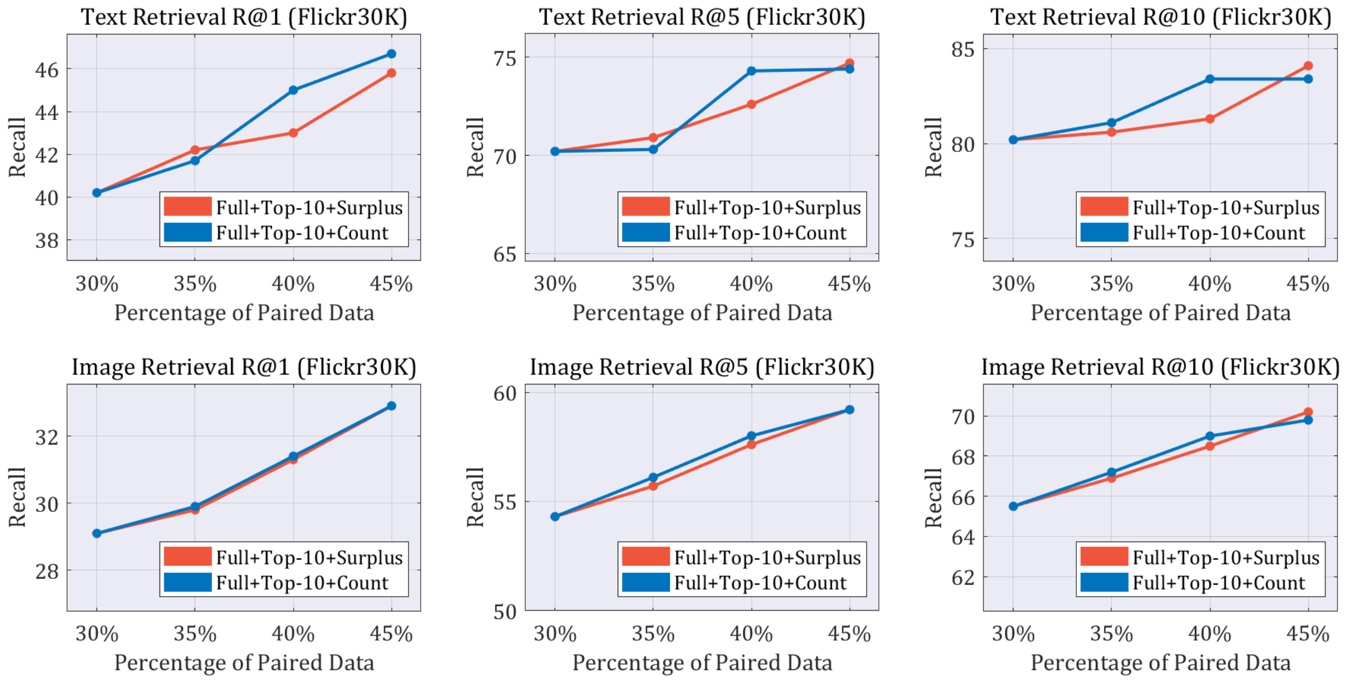


Figure 4.17: Ablation study for w_{ij}.

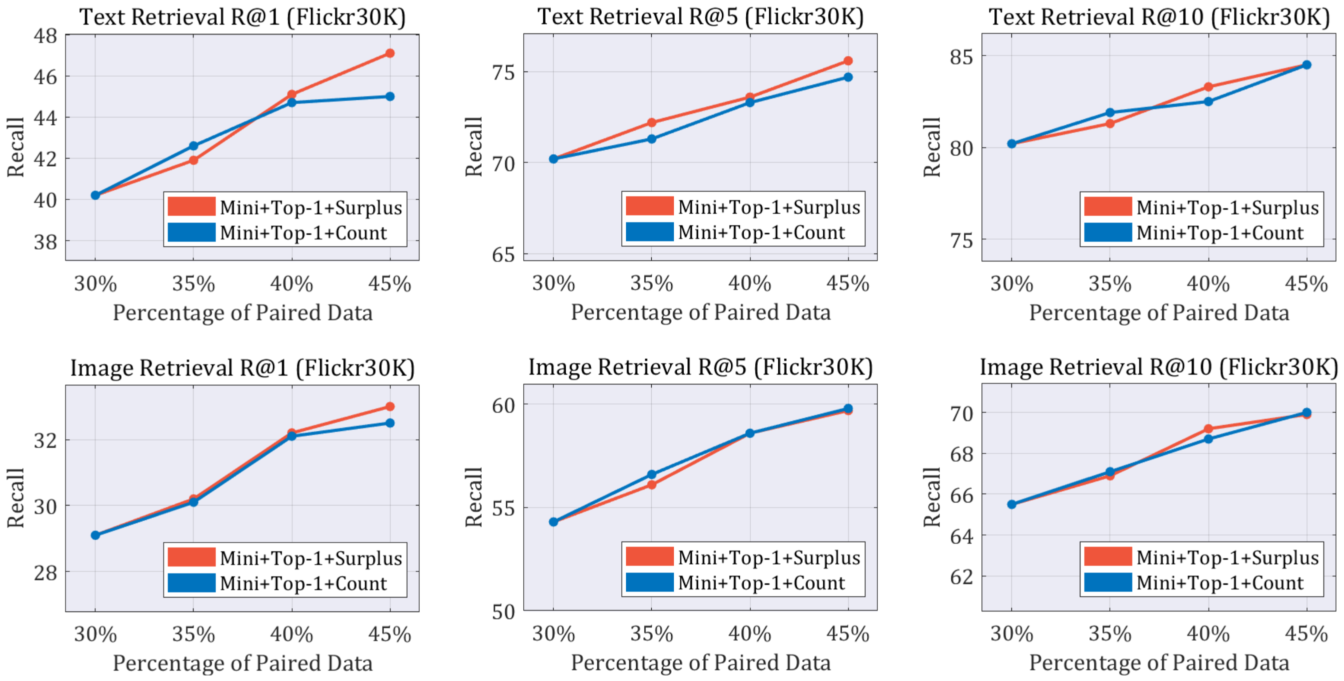


Figure 4.18: Ablation study for w_{ij}.

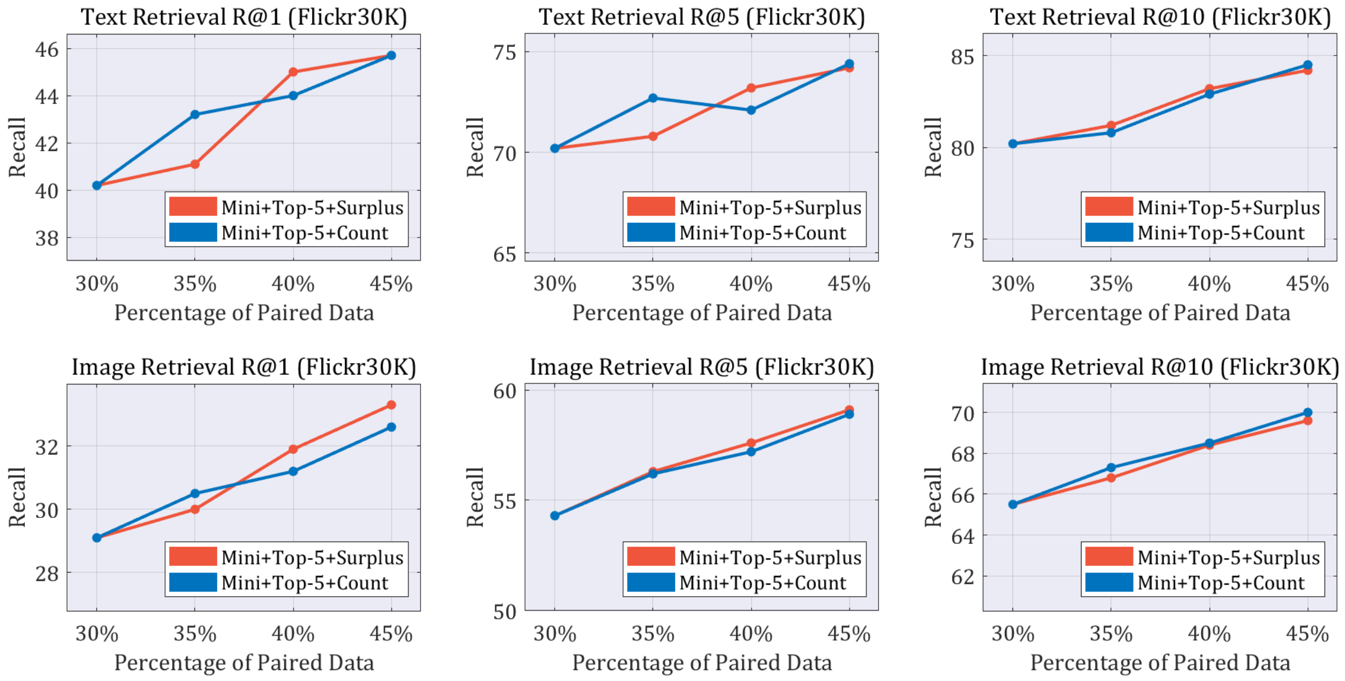


Figure 4.19: Ablation study for w_{ij}.

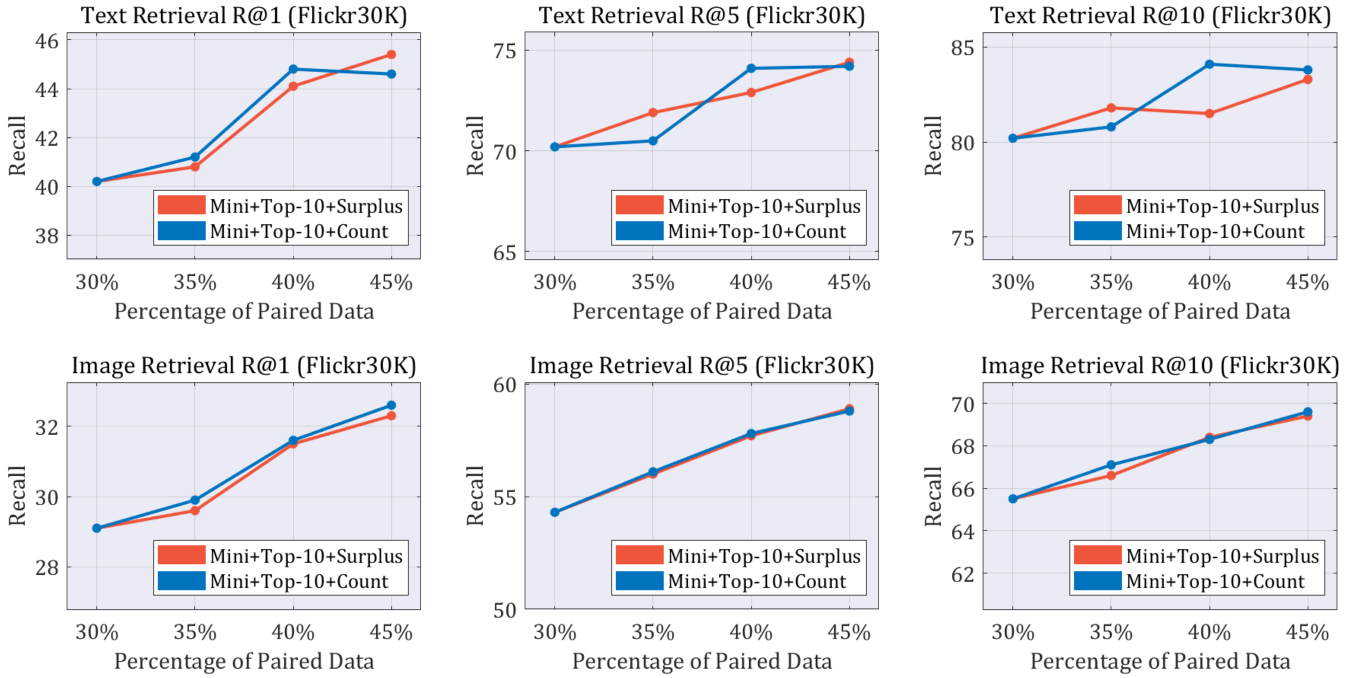


Figure 4.20: Ablation study for w_{ij}.

Condition ξ_j	Weight w_{ji}	Text Retrieval (R@10)				Image Retrieval (R@10)				
		$e=0$	$e=1$	$e=2$	$e=3$	$e=0$	$e=1$	$e=2$	$e=3$	R@10-sum
Full-batch + Top-1	Surplus	80.2	81.6	83.4	84.4	65.5	67.8	69.2	70.9	**602.9**
	Count	80.2	81.0	82.9	83.8	65.5	67.3	68.7	70.4	599.6
Full-batch + Top-5	Surplus	80.2	80.6	82.4	83.8	65.5	67.2	68.7	70.2	598.5
	Count	80.2	80.9	82.9	83.7	65.5	67.0	68.6	69.9	598.5
Full-batch + Top-10	Surplus	80.2	80.6	81.3	84.1	65.5	66.9	68.5	70.2	597.2
	Count	80.2	81.1	83.4	83.4	65.5	67.2	69.0	69.8	599.4
Mini-batch + Top-1	Surplus	80.2	81.3	83.3	84.5	65.5	66.9	69.2	69.9	**600.7**
	Count	80.2	81.9	82.5	84.5	65.5	67.1	68.7	70.0	600.3
Mini-batch + Top-5	Surplus	80.2	81.2	83.2	84.2	65.5	66.8	68.4	69.6	598.8
	Count	80.2	80.8	82.9	84.5	65.5	67.3	68.5	70.0	599.5
Mini-batch + Top-10	Surplus	80.2	81.8	81.5	83.3	65.5	66.6	68.4	69.4	596.5
	Count	80.2	80.8	84.1	83.8	65.5	67.1	68.3	69.6	599.2

Table 4.6: R@10 performance of the proposed AL algorithm at each epoch of AL scenario, according to the hard negative condition and aggregation weight for the Flickr30K.

Chapter 5

Conclusion

In this dissertation, we proposed methods to overcome the two major problems that may occur in cross-modal representation learning: (1) learning cross-modal association among heterogeneous modalities, and (2) lack of paired data.

First, to overcome the problem of learning cross-modal association among heterogeneous modalities, we proposed a cross-modal representation learning model adopting the distributed embedding method. Motivated by the association learning mechanism of human brain, the proposed method consists of two learning phase. The proposed model first learns intra-modal association by training a specialized embedding space for each modality with single-modal representation learning. Intra-model association model is realized by training conventional variational auto-encoder for each modalities. Then, the proposed model learns cross-modal association by introducing associator, which connects the embedding spaces of multiple modalities. Associator is realized by tiny networks with small number of fully connected layers, which connects latent spaces of each variational auto-encoders. To separate the learning process of intra-modal association and cross-modal association, the model-parameters involved to intra-modal association are not updated during training of cross-modal association. Through the two-step learning process, the proposed model has the advantage of learning relation among heterogeneous modalities, utilizing unpaired data for learning, and incorporating additional modality. We validated the proposed method on image-audio generation and 3d hand pose estimation tasks. The proposed method achieves improved performance compared to the existing joint-embedding based models.

Second, to mitigate the data shortage problem in cross-modal representation learning, we proposed an novel active learning scenario and algorithm for cross-modal representation learning. In particular, we targeted active learning for image-text retrieval, which is one of the most popular applications related to cross-modal representation learning. In the proposed scenario, unpaired image or text data are given and active learning algorithm selects the most informative unpaired data. Then selected data are queried to human experts to be paired. The proposed active learning algorithm selects the data that is expected to have the most influence on the max-hinge triplet loss function, which is mainly adopted loss function in recent image-text retrieval method. To this end, we define the condition that an image (text) can be the hard negative for the texts (images) in the existing paired data. Based on condition, we proposed HN-score for an unpaired image (text) which estimates an image (text) can be a hard negative for as many texts (images) in paired data as possible. Then the proposed algorithm selects the data of highest score. We validate the proposed active learning algorithm through the comparison to random selection and self-ablation studies on Flickr and MS-COCO dataset. As a future work, we will further validate proposed methods for the various benchmarks.

Bibliography

[1] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[2] M. Wu and N. Goodman, “Multimodal generative models for scalable weakly-supervised learning,” in *Advances in Neural Information Processing Systems*, 2018, pp. 5580–5590.

[3] A. Spurr, J. Song, S. Park, and O. Hilliges, “Cross-modal deep variational hand pose estimation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 89–98.

[4] C. Cadena, A. R. Dick, and I. D. Reid, “Multi-modal auto-encoders as joint estimators for robotics scene understanding.” in *Robotics: Science and Systems*, 2016.

[5] D. Hu, X. Li *et al.*, “Temporal multimodal learning in audiovisual speech recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3574–3582.

[6] D. U. Jo, B. Lee, J. Choi, H. Yoo, and J. Y. Choi, “Associative variational auto-encoder with distributed latent spaces and associators,” in *Proceedings of the*

AAAI Conference on Artificial Intelligence, vol. 34, no. 07, 2020, pp. 11 197–11 204.

[7] Y. Wang, J. van de Weijer, and L. Herranz, "Mix and match networks: encoder-decoder alignment for zero-pair image translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5467–5476.

[8] S. Chaudhury, S. Dasgupta, A. Munawar, M. A. S. Khan, and R. Tachibana, "Conditional generation of multi-modal data using constrained embedding space mapping," *arXiv preprint arXiv:1707.00860*, 2017.

[9] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov *et al.*, "Devise: A deep visual-semantic embedding model," in *Advances in neural information processing systems*, 2013, pp. 2121–2129.

[10] K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang, "A comprehensive survey on cross-modal retrieval," *arXiv preprint arXiv:1607.06215*, 2016.

[11] C. Kang, S. Liao, Y. He, J. Wang, W. Niu, S. Xiang, and C. Pan, "Cross-modal similarity learning: A low rank bilinear formulation," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 2015, pp. 1251–1260.

[12] V. E. Liong, J. Lu, Y.-P. Tan, and J. Zhou, "Deep coupled metric learning for cross-modal matching," *IEEE Transactions on Multimedia*, vol. 19, no. 6, pp. 1234–1244, 2016.

[13] F. Wu, X. Jiang, X. Li, S. Tang, W. Lu, Z. Zhang, and Y. Zhuang, "Cross-modal learning to rank via latent joint representation," *IEEE Transactions on Image Processing*, vol. 24, no. 5, pp. 1497–1509, 2015.

[14] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, “Vse++: Improving visual-semantic embeddings with hard negatives,” *arXiv preprint arXiv:1707.05612*, 2017.

[15] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, “Stacked cross attention for image-text matching,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 201–216.

[16] H. Chen, G. Ding, X. Liu, Z. Lin, J. Liu, and J. Han, “Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 655–12 663.

[17] Z. Wang, X. Liu, H. Li, L. Sheng, J. Yan, X. Wang, and J. Shao, “Camp: Cross-modal adaptive message passing for text-image retrieval,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5764–5773.

[18] Y. Song and M. Soleymani, “Polysemous visual-semantic embedding for cross-modal retrieval,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1979–1988.

[19] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, “Visual semantic reasoning for image-text matching,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4654–4662.

[20] Y. Huang, Q. Wu, C. Song, and L. Wang, “Learning semantic concepts and order for image and sentence matching,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6163–6171.

[21] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, “Multimodal deep learning,” in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.

[22] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.

[23] T. V. Bliss and G. L. Collingridge, “A synaptic model of memory: long-term potentiation in the hippocampus,” *Nature*, vol. 361, no. 6407, p. 31, 1993.

[24] D. V. Buonomano and M. M. Merzenich, “Cortical plasticity: from synapses to maps,” *Annual review of neuroscience*, vol. 21, no. 1, pp. 149–186, 1998.

[25] H. Van Praag, B. R. Christie, T. J. Sejnowski, and F. H. Gage, “Running enhances neurogenesis, learning, and long-term potentiation in mice,” *Proceedings of the National Academy of Sciences*, vol. 96, no. 23, pp. 13 427–13 431, 1999.

[26] N. M. Weinberger, “Specific long-term memory traces in primary auditory cortex,” *Nature Reviews Neuroscience*, vol. 5, no. 4, p. 279, 2004.

[27] I. P. Pavlov and W. Gantt, “Lectures on conditioned reflexes: Twenty-five years of objective study of the higher nervous activity (behaviour) of animals.” 1928.

[28] P. I. Pavlov, “Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex,” *Annals of neurosciences*, vol. 17, no. 3, p. 136, 2010.

[29] J.-H. Wang and S. Cui, “Associative memory cells and their working principle in the brain,” *F1000Research*, vol. 7, 2018.

[30] ——, “Associative memory cells: formation, function and perspective,” *F1000Research*, vol. 6, 2017.

[31] D. O. Hebb, *The organization of behavior.* na, 1961.

[32] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, “A learning algorithm for boltzmann machines,” *Cognitive science*, vol. 9, no. 1, pp. 147–169, 1985.

[33] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[34] Y. Yoo, S. Yun, H. J. Chang, Y. Demiris, and J. Y. Choi, “Variational autoencoded regression: high dimensional regression of visual data on complex manifold,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3674–3683.

[35] A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[36] R. Kiros, R. Salakhutdinov, and R. S. Zemel, “Unifying visual-semantic embeddings with multimodal neural language models,” *arXiv preprint arXiv:1411.2539*, 2014.

[37] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, “Grounded compositional semantics for finding and describing images with sentences,” *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 207–218, 2014.

[38] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, “Aligning books and movies: Towards story-like visual explanations by watching movies and reading books,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 19–27.

[39] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, “Bottom-up and top-down attention for image captioning and visual question an-

swering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

[40] B. Settles, "Active learning literature survey," University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep., 2009.

[41] H. H. Aghdam, A. Gonzalez-Garcia, J. v. d. Weijer, and A. M. López, "Active learning for deep detection neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3672–3680.

[42] D. Yoo and I. S. Kweon, "Learning loss for active learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 93–102.

[43] Z. Liu, J. Wang, S. Gong, H. Lu, and D. Tao, "Deep reinforcement active learning for human-in-the-loop person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6122–6131.

[44] K. Murugesan and J. Carbonell, "Active learning from peers," in *Advances in Neural Information Processing Systems*, 2017, pp. 7008–7017.

[45] Y. Shen, H. Yun, Z. C. Lipton, Y. Kronrod, and A. Anandkumar, "Deep active learning for named entity recognition," *arXiv preprint arXiv:1707.05928*, 2017.

[46] B. Liu and V. Ferrari, "Active learning for human pose estimation," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 4373–4382.

[47] F. Caba Heilbron, J.-Y. Lee, H. Jin, and B. Ghanem, "What do i annotate next? an empirical study of active learning for action localization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 199–216.

[48] Z. Zhou, J. Shin, L. Zhang, S. Gurudu, M. Gotway, and J. Liang, “Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally,” in *IEEE conference on computer vision and pattern recognition, Hawaii*, 2017, pp. 7340–7349.

[49] L. Yang, Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen, “Suggestive annotation: A deep active learning framework for biomedical image segmentation,” in *International conference on medical image computing and computer-assisted intervention.* Springer, 2017, pp. 399–407.

[50] S. Dasgupta, “Two faces of active learning,” *Theoretical computer science*, vol. 412, no. 19, pp. 1767–1781, 2011.

[51] Y. Gal, R. Islam, and Z. Ghahramani, “Deep bayesian active learning with image data,” in *Proceedings of the 34th International Conference on Machine Learning-Volume 70.* JMLR. org, 2017, pp. 1183–1192.

[52] W. H. Beluch, T. Genewein, A. Nürnberger, and J. M. Köhler, “The power of ensembles for active learning in image classification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9368–9377.

[53] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, “Cost-effective active learning for deep image classification,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 12, pp. 2591–2600, 2016.

[54] O. Sener and S. Savarese, “Active learning for convolutional neural networks: A core-set approach,” in *International Conference on Learning Representations*, 2018.

[55] S. Sinha, S. Ebrahimi, and T. Darrell, "Variational adversarial active learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5972–5981.

[56] C. E. Shannon, "A mathematical theory of communication," *Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[57] L. C. Freeman, *Elementary applied statistics: for students in behavioral science*. John Wiley & Sons, 1965.

[58] Y. Siddiqui, J. Valentin, and M. Nießner, "Viewal: Active learning with viewpoint entropy for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9433–9443.

[59] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[60] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7291–7299.

[61] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[62] R. Z. Farahani and M. Hekmatfar, *Facility location: concepts, models, algorithms and case studies*. Springer, 2009.

[63] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for k-medoids clustering," *Expert systems with applications*, vol. 36, no. 2, pp. 3336–3341, 2009.

[64] F. Zhdanov, "Diverse mini-batch active learning," *arXiv preprint arXiv:1901.05954*, 2019.

[65] Z. Wang and J. Ye, "Querying discriminative and representative samples for batch mode active learning," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 9, no. 3, p. 17, 2015.

[66] Y.-P. Tang and S.-J. Huang, "Self-paced active learning: Query the right thing at the right time," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 5117–5124.

[67] N. Gao, S.-J. Huang, Y. Yan, and S. Chen, "Cross modal similarity learning with active queries," *Pattern Recognition*, vol. 75, pp. 214–222, 2018.

[68] O. Rudovic, M. Zhang, B. Schuller, and R. Picard, "Multi-modal active learning from human data: A deep reinforcement learning approach," in *2019 International Conference on Multimodal Interaction*, 2019, pp. 6–15.

[69] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[70] J. Lim, Y. Yoo, B. Heo, and J. Y. Choi, "Pose transforming network: Learning to disentangle human posture in variational auto-encoded latent space," *Pattern Recognition Letters*, 2018.

[71] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. S. Kweon, "Learning to localize sound source in visual scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4358–4366.

[72] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[73] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques," *arXiv preprint arXiv:1003.4083*, 2010.

[74] B. Logan *et al.*, "Mel frequency cepstral coefficients for music modeling." in *ISMIR*, vol. 270, 2000, pp. 1–11.

[75] J. Rubin, R. Abreu, A. Ganguli, S. Nelaturi, I. Matei, and K. Sricharan, "Classifying heart sound recordings using deep convolutional neural networks and mel-frequency cepstral coefficients," in *Computing in Cardiology Conference (CinC), 2016*. IEEE, 2016, pp. 813–816.

[76] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "The German Traffic Sign Recognition Benchmark: A multi-class classification competition," in *IEEE International Joint Conference on Neural Networks*, 2011, pp. 1453–1460.

[77] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[78] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," 2011.

[79] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[80] M. Suzuki, K. Nakayama, and Y. Matsuo, "Joint multimodal learning with deep generative models," *arXiv preprint arXiv:1611.01891*, 2016.

[81] C. Zimmermann and T. Brox, “Learning to estimate 3d hand pose from single rgb images,” arXiv:1705.01389, Tech. Rep., 2017, https://arxiv.org/abs/1705.01389. [Online]. Available: https://lmb.informatik.uni-freiburg.de/projects/hand3d/

[82] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[83] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions,” *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[84] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, “Large scale online learning of image similarity through ranking,” 2010.

[85] A. Frome, Y. Singer, F. Sha, and J. Malik, “Learning globally-consistent local distance functions for shape-based image retrieval and classification,” in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[86] A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[87] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[88] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, “Visual genome: Connecting language and

vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, no. 1, pp. 32–73, 2017.

[89] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[90] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[91] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 251–260.

[92] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.

[93] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge 2007 (voc2007) results," 2007.

[94] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-wide: a real-world web image database from national university of singapore," in *Proceedings of the ACM international conference on image and video retrieval*, 2009, pp. 1–9.

Abstract

본 논문에서는 교차 모달 표현 학습에서 발생할 수 있는 문제점들을 개선하기 위한 두 가지 방법을 제안한다. 첫 째, 기존의 공동 임베딩 방식의 교차 모달 표현 학습 모델이 상이한 모달 데이터 사이의 표현을 학습하기 어려운 단점을 해결하기 위하여, 분산 임베딩 방식의 교차 모달 학습 모델을 제안한다. 분산 임베딩 방식의 학습 모델은 먼저 각 모달마다 독립적으로 단독 모달 표현 학습을 수행함으로써 각 모달마다 특화된 임베딩 공간을 학습한다. 그 후 교차 모달 표현을 학습하기 위해 여러 모달의 임베딩 공간사이를 연결하는 연상학습 모듈을 학습한다. 두 단계를 거치는 학습 과정을 통해 제안하는 모델은 상이한 모달들 간의 교차 모달 표현학습도 잘 수행할 수 있으며, 쌍이 주어지지 않은 교차 모달 데이터도 활용하여 학습할 수 있다는 장점을 가진다. 상이한 모달 관계 중 하나인 시각과 청각 모달 간의 데이터 생성 실험에서 제안하는 방법은 기존의 공동 임베딩 방식의 모델보다 향상된 성능을 검증하였다. 둘 째, 교차 모달 표현 학습을 위해서는 모달간 쌍을 이루는 데이터가 필수적이지만 실제 응용분야에서 충분한 수의 데이터 쌍을 확보하는 것은 어렵다. 이러한 문제점을 해결하기 위하여 교차 모달 표현 학습을 위한 능동적 학습 방법을 제안한다. 특히 교차 모달 표현 학습 관련 응용분야 중 하나인 이미지-텍스트 반환에 대한 능동적 학습을 제안한다. 기존의 이미지-텍스트 반환에 대한 능동적 학습 시나리오는 최신의 이미지-텍스트 반환 데이터셋에 적용하기 어렵기 때문에, 본 논문에서는 우선 최신의 데이터셋에 적합한 능동적 학습 시나리오를 먼저 제안한다. 주어진 이미지-텍스트 쌍 데이터에 대하여 사람에게 분류 라벨을 요청하는 기존의 시나리오와는 달리, 제안하는 시나리오는 쌍이 주어지지 않은 이미지 혹은 텍스트 데이터에 대하여 사람에게 나머지 모달리티의 데이터를 요청하여 쌍 데이터를 확보하는 것을 목표로 한다. 또한 제안하는 시나리오에 적합한 능동적 학습 알고리즘도 제안한다. 제안하는 알고리즘은 이미지-텍스트 반환에서 주로 사용되는 최대 힌지

트리플렛 손실함수에 가장 영향력을 많이 끼칠 것으로 생각되는 데이터를 선별한다. 이를 위해 특정 데이터가 손실함수에 영향력을 미칠 수 있는 조건을 정의하고, 정의된 조건에 기반하여 데이터가 손실함수에 미치는 영향력 점수를 추정한다. 제안하는 알고리즘은 영향력 점수가 가장 높은 순서대로 데이터를 선택하여 사람에게 나머지 쌍 데이터를 제공해줄 것을 요청한다. 최신의 이미지-텍스트 데이터셋에서의 제안하는 알고리즘이 무작위로 쌍 데이터를 확보하는 것보다 학습데이터 수 대비 향상된 성능을 달성하는 것을 보여주었다.

Keywords: 교차 모달 표현 학습, 분산형 임베딩, 능동적 학습, 심층 학습 네트워크

Student Number: 2016-20977