Master's thesis
Master's Programme in Data Science

Multi-task deep learning models in visual fashion understanding

Mateusz Gierlach

November 2, 2020

Supervisor(s): Associate Professor Laura Ruotsalainen

Examiner(s): Associate Professor Laura Ruotsalainen
Assistant Professor Arto Klami

University of Helsinki
Faculty of Science

P. O. Box 68 (Pietari Kalmin katu 5)
00014 University of Helsinki

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty Faculty of Science	Koulutusohjelma — Utbildningsprogram — Degree programme Master’s Programme in Data Science

Tekijä — Författare — Author

Mateusz Gierlach

Työn nimi — Arbetets titel — Title

Multi-task deep learning models in visual fashion understanding

Työn laji — Arbetets art — Level Master’s thesis	Aika — Datum — Month and year November 2, 2020	Sivumäärä — Sidantal — Number of pages 63

Tiivistelmä — Referat — Abstract

Visual fashion understanding (VFU) is a discipline which aims to solve tasks related to clothing recognition, such as garment categorization, garment’s attributes prediction or clothes retrieval, with the use of computer vision algorithms trained on fashion-related data. Having surveyed VFU-related scientific literature, I conclude that, because of the fact that at the heart of all VFU tasks is the same issue of visually understanding garments, those VFU tasks are in fact related. I present a hypothesis that building larger multi-task learning models dedicated to predicting multiple VFU tasks at once might lead to better generalization properties of VFU models. I assess the validity of my hypothesis by implementing two deep learning solutions dedicated primarily to category and attribute prediction. First solution uses multi-task learning concept of sharing features from additional branch dedicated to localization task of landmarks’ position prediction. Second solution does not share knowledge from localization branch. Comparison of those two implementations confirmed my hypothesis, as sharing knowledge between tasks increased category prediction accuracy by 53% and attributes prediction recall by 149%. I conclude that multi-task learning improves generalization properties of deep learning-based visual fashion understanding models across tasks.

——

ACM Computing Classification System (CCS):
Computing methodologies → Machine learning → Learning paradigms → Multi-task learning
Computing methodologies → Machine learning → Machine learning approaches → Neural networks
General and reference → Document types → Surveys and overviews

Avainsanat — Nyckelord — Keywords

multi-task learning, visual fashion understanding, deep learning

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Contents

1. Introduction

After e-commerce boom, with massive sales increases in online shopping, the field of visual clothing analysis started getting increased attention, as fashion industry began to look for new ways to add value to customers around their base product. Research progress in the field of visual fashion understanding has partially been driven by the enormous commercial potential and can be attributed to a wide spectrum of possible industry applications. The field has drawn more attention from the research community in recent years.

1.1 Description of the discipline

Computer vision (CV) has made great progress in recent decade, with applications in many industries. In the context of understanding fashion, which is a visual medium, relevant descriptor features cannot be captured by text. Therefore, there is a need for application of vision-based methods for automatic feature extraction from fashion images. Fashion is a domain in which data is represented almost exclusively by images which makes it a natural application domain for computer vision techniques. However, applying CV methods in fashion context (pictures of clothing on people) has been challenging because of significant in-picture variance of person's pose, lighting, human's body proportions and camera angle.

CV-based visual fashion understanding field (VFU) consists of many different sub-problems. We can distinguish **five most important visual fashion understanding tasks**, all of which are a part of actual industry applications, such as shop assistant. Firstly, **category classification**, deals with predicting what type of clothing garment has been detected, e.g. dress/trousers etc. Secondly, **attributes prediction** focuses on finding detailed descriptions for each detected clothing category, e.g. fabric - tweed, shape - cropped. Thirdly, **clothes detection** focuses on finding areas of the picture with clothing garments. The fourth task is **landmark and pose estimation**, which deals with predicting essential points on a person's body to understand his/her pose. The fifth task is **clothes retrieval**, which deals with finding similar pieces of clothing in a database, based on a visual query.

Practical applications in VFU are possible by leveraging data in order to better understand individual clothing items in images, as well as improve understanding of relations between garments. Such applications are already in use by leading players in fashion technology industry. VFU tasks, however, are not easy and major challenges were encountered, while trying to improve performance of algorithms in visual fashion field [33]. The most significant challenges are: 1) Large variance of clothes attributes, such as garment's texture type or shape, 2) Deformation of clothes in actual, real-world pictures (e.g. not containing the whole garment in one picture), 3) Variation of scenarios under which a picture was taken, such as pose of the person wearing clothes in the picture, lightening or position of the camera.

Visual fashion understanding (VFU; also called: fashion recognition, fashion image analysis) discipline aims to solve tasks such as described above, as well as many others in the area, by leveraging fashion-related, domain-specific data and vision-based algorithms.

Fashion garment (also: clothing piece) is an individual piece of clothing that can be recognized in a fashion-related picture. Garments are fundamental elements in VFU models. Recognition and understanding of garments are the most crucial issues in the VFU discipline.

Multi-task learning is an approach in machine learning, where a model learns to optimize loss functions for multiple tasks at the same time. Features/representations are passed between task-specific branches, which leads to knowledge sharing between tasks.

In my work, I mainly focus on multi-task learning methods in the VFU discipline.

1.2 Historical and modern methods

Historically, the field of visual fashion understanding focused mainly on clothing recognition sub-problem and used methods such as SIFT [21] or HOG [5]. Performance of such predictive models were unsatisfactory, as models had access to limited representations, defined by human only [19].

After neural networks started obtaining exceptional results across many computer vision tasks, researchers began looking into its potential applications in the area of visual fashion understanding. These deep learning techniques, and in particular convolutional neural networks [9] (CNNs), are behind all the major improvements in the field. Nowadays, CNNs are considered state-of-the-art methods across all visual fashion tasks, such as attribute prediction or shop retrieval. CNN-based models outperformed prior methods by huge margins. Success of CNNs can be attributed to having much broader representation capabilities, as feature representations are being found automatically, without any human involvement in the process. Therefore, that led to more discriminative properties of built models. These successes were also made possible through availability of large datasets related to visual fashion (described in Section 2.1). It allowed for creation of universal benchmarks which are a foundation for comparison between different neural models.

Deep learning methods are highly suitable to be applied in VFU, because of their specific characteristics. Firstly, neural architectures allow for efficient learning, as they use hierarchical layer decomposition of weights with non-linear activations. These activations are able to learn hierarchical relations in discriminative manner, because weights in subsequent layers are influenced by each other. That leads to models having exponential discriminative power in relation to the number of parameters in the network. Secondly, universal composition structure of neural architectures might use variety of structural decompositions in order to understand context around phenomena they are modeling. That allows for creation of tailored algorithms for different problems [3]. Thirdly, automatic representation in deep neural networks gives them ability to find patterns in visual data through finding automatic representations, without the need for any formal problem description.

1.3 Solutions

In Chapter 2, I focus on describing current state-of-the-art methods for tasks across the VFU field. Vast majority of those solutions use convolutional neural networks (CNNs) for extraction of visual features. It is beyond the scope of this work to provide detailed explanation of CNNs [9] and it is assumed reader has knowledge in that area. Two of the most common architectures of CNNs are VGG-16 [24] and ResNet [11].

I describe my survey of published VFU methods in Chapter 2. Some of them aim to make fashion models generalize better across tasks in the VFU field. Other focus just on obtaining satisfactory performance in a narrowly defined task, as can be measured by accuracy. Regardless of that, it is valuable to learn successes and challenges of those methods, before reading about my novel contribution, as those solutions serve as a logical foundation to further analysis in Chapter 3.

Before diving into VFU solutions, in Section 2.1 I describe most important datasets in the field, which are used by algorithms described further. Methods dedicated to categorization tasks are described in Section 2.2. Methods dedicated to localization tasks are described in Section 2.3. Issues related to retrieval tasks are described in Section 2.4. State-of-the-art methods which focus on sharing knowledge between tasks, in order to get closer to more generalizable fashion understanding models, are described in Section 2.5. Those methods introduce concepts such as landmark pooling [19], feature map upsampling [16] or localization-aided attribute awareness [12, 2]. Performance analysis of those state-of-the-art algorithms is a good starting point for further discussion.

1.4 Goal of this work

When researchers propose novel VFU solutions with an aim of improving understanding of fashion, through introduction of new elements in neural architectures, they are mostly using traditional metrics such as accuracy or recall to measure these improvements. While these approaches work well for comparing different algorithms between each other, as they use constant benchmarks, they do not provide any notion of measuring algorithm's generalization abilities for understanding fashion across tasks.

At the heart of all VFU tasks, described in Section 1.1 is understanding fashion garments. No matter what is the end goal of an algorithm, all of the VFU-related algorithms try to learn an accurate representation of what a garment is by looking at this problem from different perspectives. I strongly believe that a path towards more general fashion model leads through more task-agnostic understanding of clothing garments. I believe that algorithms designed for different tasks could benefit from sharing information between each other and try to optimize loss functions jointly. Optimal representations could be used for many applications and shared between tasks so that one task could benefit from others. Losses could be averaged in each training epoch, which would lead to performance improving across many VFU tasks.

In this work, I try to understand how to leverage multi-task learning approach in order to build more generalizable fashion models. I focus on surveying modern deep learning-based state-of-the-art approaches designed for solving VFU tasks. I believe that VFU tasks are related to each other and building multi-task learning models can be beneficial for improving generalization across visual fashion understanding field.

In order to validate this hypothesis, I built two solutions. The first one uses the multi-task learning approach by sharing features between tasks, while second one does not. After making necessary implementation experiments, I assess whether knowledge sharing between tasks contributes to improved metrics for the underlying task and therefore to generalization ability across wider visual fashion understanding domain.

2. Methodologies

In order to understand how different deep learning methods could be used in order to improve generalization properties of models in VFU, we need to first understand how different, essential elements contribute to the algorithm's performance. In this chapter I describe my survey of those concepts, as discussed in the scientific literature. In Section 2.1, I describe most commonly used VFU datasets of today, as well as comparison benchmarks which are based on those datasets. In Section 2.2, I describe most important solutions from the research literature, which focus on categorization-based VFU tasks, such as garment type or attribute prediction. In Section 2.3 I focus on describing localization-related issues in VFU, such as importance of landmarks or differences between segmentation and detection. In Section 2.4, I describe VFU tasks of retrieval and search. Finally, in Section 2.5, I describe novel state-of-the-art solutions in VFU, which share knowledge between tasks through multi-task learning.

2.1 VFU datasets and benchmarks

Datasets proved to be one of the driving forces in development of artificial intelligence (AI). ImageNet [6] extensively contributed to progress in computer vision field by providing a universal and massively labeled dataset for image recognition and classification. However, ImageNet provided only course-grained category annotations with single label per image. When dealing with real-world scenarios, applications need to be trained on domain-specific, multi-label datasets with fine-grained descriptions of recognized items' attributes [10].

Recent advancements in clothing understanding have been heavily influenced by availability of field-specific clothing image datasets [13, 28]. However, before the introduction of DeepFashion [19], datasets lacked comprehensive annotations required to perform any advanced supervised learning process on them. This disadvantage made traditional datasets not particularly useful in any practical VFU application.

2.1.1 DeepFashion dataset

DeepFashion [19] is a clothes dataset of over 800K images, comprehensively annotated with categories, attributes, clothing landmarks, and cross-pose/cross-domain pairs. Those pairs are corresponding-to-each-other images taken across varied domains, shot angles or person's poses. Introduction of massive DeepFashion allowed for further research into visual fashion field and opened the possibility of any real-world application in the domain.

DeepFashion (DF) solved the problem, which existed in earlier visual fashion datasets, which consisted of only limited amount of attributes, bounding boxes or cross-domain relationships, and were not sufficient for effective learning for attribute or landmark prediction. What makes DeepFashion a much better dataset for these types of tasks, are its distinctive properties. Firstly, DF contains rich, fine-grained annotations of categories and attributes, as there are 50 categories and over 1000 descriptive attributes available [19]. Such massive attributes are essential in order to represent large clothing properties. Scale and massiveness of attribute annotations can be seen in Table 2.1. As we can see in the Table, DeepFashion is a much larger and more densely annotated dataset than its competitors. Secondly, landmark annotations replaced previously used bounding boxes, which allows for better localization properties. Each image is labeled with 4-8 landmarks. Landmark annotations help with dealing with deformations or variation of poses. Landmarks are described in more detail in Section 2.3.1. Technicality of landmark labeling is shown in Figure 2.1. Thirdly, DF has high domain variance of picture types, as the dataset includes variety of fashion pictures ranging from professional online store pictures to consumer photos on social media. Fourthly, cross-domain information availability is DF's another strong property, as it contains over 300000 cross-pose or cross-domain pairs between pictures from different domains containing the same clothing garment. These consumer-to-shop pairs (consumer picture, professional shop photograph) help with bridging the information gap between domains. Fifthly, another DF's advantage is the size of the dataset (over 800000 data points), which was by far the largest domain dataset introduced when compared to available resources before the arrival of DeepFashion. Sixthly, public availability of the dataset allows for the research community to be able to use it in order to introduce new methods in the field.

Table 2.1: Scale and richness of annotations in DeepFashion [19]

	Where To Buy It [13]	DARN [12]	**DeepFashion** [19]
Number of images	79K	183K	**over 800K**
Number of categories and attributes	11	179	**1050**
Numbers of pairs	39K	91K	**over 300K**
Localization annotation	No	No	**Yes: 4-8 landmarks**
Public availability	No	No	**Yes**

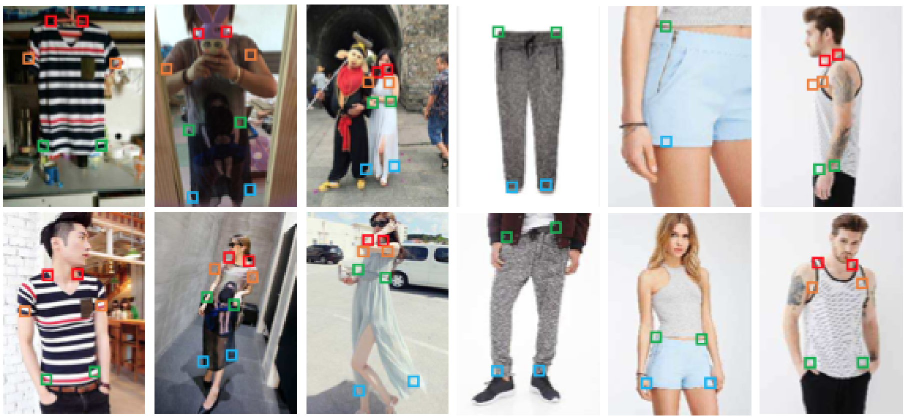


Figure 2.1: Landmark annotations in DeepFashion [19]

Besides DeepFashion's many advantages, the dataset has its flaws as well. Firstly, there is only single clothing garment per picture, which might result in models being skewed towards recognizing only the main garment in the picture [19]. Secondly, there are 4-8 landmarks per picture, which is an improvement compared to bounding boxes but it is still not precise enough to accurately describe localization of the garment in the picture. Thirdly, every clothing category shares the same landmark structure. Fourthly, there are no per-pixel masks on picture, which might make model have difficulty recognizing multiple clothes close to each other.

2.1.2 DeepFashion benchmark

Besides all of the usability of DeepFashion in training algorithms for better performance, the dataset also contributes benchmarks for three visual fashion sub-problems, namely: clothing attribute recognition, in-shop clothes retrieval and cross-domain clothes retrieval [19]. These benchmarks, which are comparison frameworks for particular tasks, allow for different algorithms and methods to be compared between each other against the constant baseline. Later in my work, I use benchmarks of DeepFashion extensively in order to test how different elements of an algorithm contribute to its prediction performance. Testing against the high-quality benchmark can be used in order to learn positive and negative properties of different neural network modules, which may lead to more powerful visual fashion understanding systems in image recognition and retrieval sub-problems.

Category and attribute prediction benchmark [19] contains over 63000 images and provides a testing framework for classifying 50 fine-grained categories and 1000 attributes. Category prediction performance evaluation uses a standard top-k classification accuracy metric. Attribute prediction is measured with top-k recall rates, which measure how many predicted attribute scores match with the ground truth labels. Attributes are segmented into five groups which describe whether the attribute relates to texture, fabric, shape, body part or style. Figure 2.2 graphically depicts properties of the benchmark. In the upper part of that Figure, example pictures for category and attribute groups are given. Categories and attributes are stored separately as categorical variables. In the lower part of that Figure, we can see that benchmark labels are not balanced, as some attributes are represented more densely.

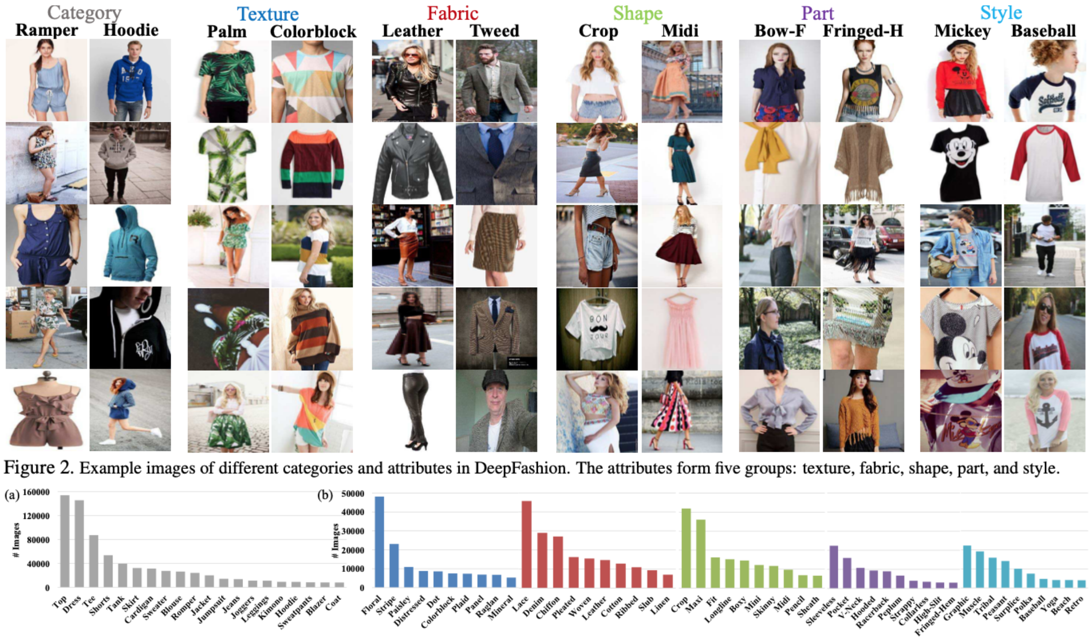


Figure 2.2: Category and attribute examples [19]

In-shop clothes retrieval benchmark [19] contains over 54000 images of 11735 clothing items from the professional fashion store and provides a testing framework for checking whether two in-shop images contain the same clothing item. It is particularly important when the photo is shared outside of the shop and visual query is the only method for the customer to obtain more information about the garment. Performance is measured by top-k retrieval accuracy, which calculates the rate of successful retrieval processes. In-shop clothes retrieval task is visually summarized in Figure 2.3. On the left side of that Figure, we can see input query images. System can be queried by such pictures and it returns the most similar retrieved images from the same in-shop domain, what can be seen on the right side of the Figure.

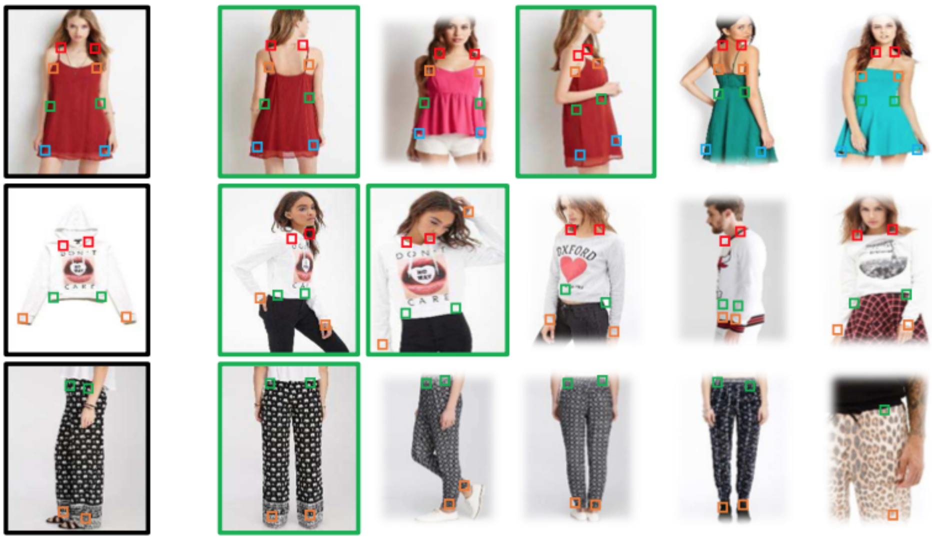


Figure 2.3: In-shop retrieval task visualization [19]

Consumer-to-shop retrieval benchmark [19] contains over 250000 image pairs between domain of consumer pictures and shop photographs and provides a testing framework for checking whether pictures from different domains contain the same garment. Performance is measured by top-k retrieval accuracy. Consumer-to-shop clothes retrieval task is visually summarized in Figure 2.4. As we can see on the left side of the Figure, system can be queried with a picture and it returns the most similar retrieved images across available domains (what can be seen on the right side of the Figure), with the green color marking the correct between-domains garment matching.

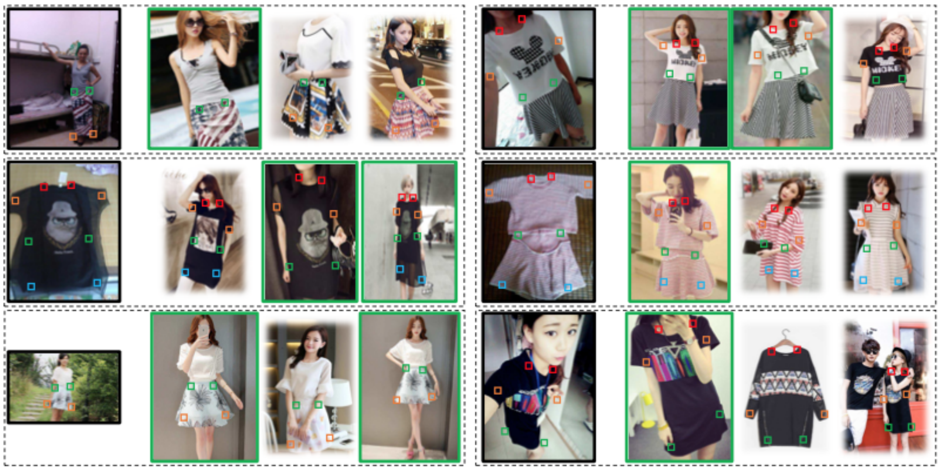


Figure 2.4: Consumer-to-shop retrieval task visualization [19]

2.1.3 DeepFashion2 dataset

A good VFU system should be able to recognize a clothing garment correctly across different domains (consumer pictures, professional shop studio photographs etc.). One of the main challenges of applying VFU methods in real-world applications is the issue of clothes ambiguity across domains, as occlusion or deformations of pictures in popular datasets [19, 20] make it extremely hard to learn cross-domain relationships for clothes. DeepFashion2 [8] is a novel dataset that aims to solve this issue.

DeepFashion2 [8] (DF2) builds on top of DeepFashion [19] and addresses its problems, described in Section 2.1.1. DeepFashion2 dataset contains 491K images and 801K clothing items, with 13 clothing categories and rich, fine-grained annotations. There are 43.8K clothes identities, where each identity is a set of almost-identical clothing garments. Clothes from the same identity across different domains (consumer, commercial) form a cross-domain pair. There are 873K such pairs. DeepFashion2 remains the largest and most densely annotated VFU dataset to date, as of writing of this paper, and it facilitates further research in VFU.

DeepFashion2 has seven unique properties, that make it stand out from other VFU datasets [19, 10, 20, 33, 13]. Firstly, DF2 is of large size and contains 801K clothing items, 491K images, 43.8K identities (with on average 12.7 items related to it) and 873K consumer-commercial pairs [8]. Secondly, DF2 has rich annotations of style, scale, viewpoint, occlusion, bounding box, dense landmarks, zooming, per-pixel masks. Thirdly, for each of the 4 main properties (scale, occlusion, zoom-in, viewpoint), a difficulty level is assigned, e.g. a garment shown from the back will have a high-level of viewpoint difficulty. Fourthly, DF2 allows for support of multiple tasks of visual fashion understanding: clothes detection, landmark and pose estimation, instance segmentation, and cross-domain clothes retrieval. Fifthly, DF2 might possibly contain multiple clothing items in one picture. Sixthly, DF2 has different landmark structure for different clothing categories, e.g. skirt needs different amount of landmarks in a different structure than a shirt does (an average category contains 23 landmarks, compared to 6 in DeepFashion [19]). Seventhly, high variance of scale, occlusion, zooming and viewpoint in the dataset allows for diversity of examples in the training set.

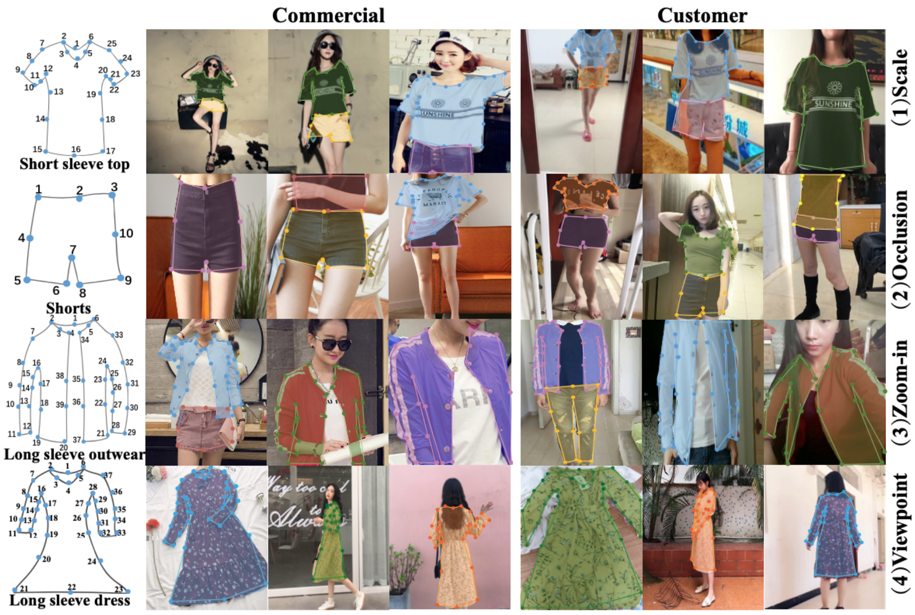


Figure 2.5: Variety of DeepFashion2 dataset [8]

Variety of DeepFashion2 can be seen in Figure 2.5. As we can see, there is a high variety of viewpoints, zoom-in properties, occlusion and scale. Also, different landmark structures are assigned to different clothes categories. Identities across rows can come either from commercial or consumer domain [8]. Cross-domain pairs are also shown. Each clothing item has its landmarks and pixel mask.

2.1.4 DeepFashion2 benchmarks

DeepFashion2 [8] also provides novel benchmarks for VFU tasks of clothes detection, pose estimation, segmentation and clothes retrieval. Various researchers were working on providing universal benchmarks for different VFU tasks [20, 27], but DeepFashion2 is the first work to introduce a unified approach for all of these tasks.

High variance of values in regards to scale, occlusion, zoom-in and viewpoint properties, as well as richness of annotations, with each clothing item containing information about its bounding box, dense landmarks, per-pixel masks and a consumer-commercial pair mapping, allowed for creators of DeepFashion2 to create 4 benchmarks based on the dataset [8]. **Clothes detection benchmark** predicts a category and a bounding box for each detected clothing in a picture. **Landmark estimation benchmark** predicts landmark keypoints for each detected clothing. **Segmentation benchmark** predicts pixel-wise locations of detected clothes, with every pixel being assigned to one of the categories or to background. **Commercial-consumer clothes retrieval benchmark** aims to find a commercial-domain image paired to each detected clothing item in a consumer-domain photograph. Lack of clothes attribute annotations and any benchmark related to it remains to be DeepFashion2's greatest weakness.

2.1.5 iFashion-Attribute dataset

iMaterialist Fashion Attribute Dataset (or iFashion-Attribute) [10] is fashion images dataset of more than 1M images with dense annotations of 228 attributes grouped into 8 types of attributes. It is the largest available attribute-centered fashion dataset, which makes it suitable to be used in real-world problems. Release of the dataset encouraged more research into attribute prediction in the VFU field.

Table 2.2: Multi-label attribute annotations in iFashion-Attribute; reproduced [10]

Attribute	Amount of labels	Value examples
Category	914K	Heels, Cargo Pants, Jeans
Color	895K	Black, Gold, Green
Gender	1013K	Male, Female, Neutral
Material	701K	Nylon, Patent, Cotton
Neckline	722K	Turtlenecks, U-Necks, Square Necked
Pattern	325K	Camouflage, Checkered, Floral
Sleeve	734K	Sleeveless, Short Sleeves, Long Sleeved
Style	610K	Asymmetric, Vintage Retro, Summer

iFashion-Attribute is purposed solely for clothing attributes prediction, providing multi-label annotations for each of the 8 groups of attributes: category, color, gender, material, neckline, pattern, sleeve, style. Table 2.2 depicts scale of annotations for each attribute group, as well as example values for those groups. Another property of iFashion-Attribute is the fine-grained labeling. High granularity of attributes is beneficial for building more precise models, that are able to represent complex phenomena from the visual fashion world. However, fine-grained labels also make it harder for algorithms to discriminate between classes correctly, when visual differences between classes are minor and examples within one class have high variance [10]. As an example, as we can see in Figure 2.6, "plaid" and "checkered" examples for red shirts are more similar to each other, then each of them are to other examples within their own class.

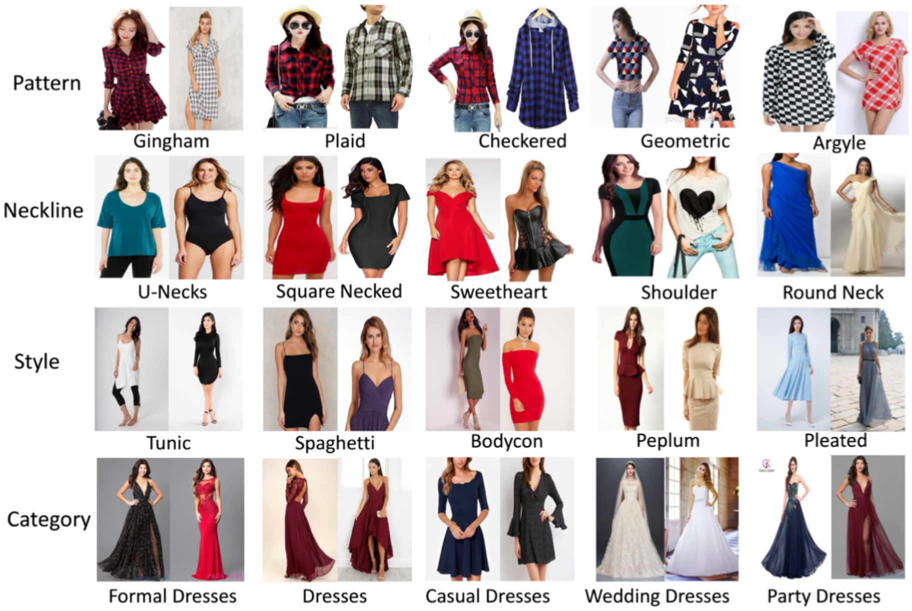


Figure 2.6: Fine-grained attributes in iFashion-Attributes [10]

2.1.6 FLD dataset

Fashion Landmark Dataset (FLD) [20] is a subset of DeepFashion dataset [19], and consists of over 123000 images with diverse poses, as well as substantial scale variations and zoom-in properties. Each image contains only one clothing garment. Each garment is annotated with 8 fashion landmarks and their visibility rating. Pose and zoom-in deviations from the norm are additionally labeled as normal/medium/large, in order to logically distinguish images with different scale of these deviations. FLD provides no other type of fashion annotations, besides landmarks. In order to tackle challenges of garment localization described in Section 2.3, images of FLD dataset contain substantial pose and scale variations, as we can see in Figure 2.7.

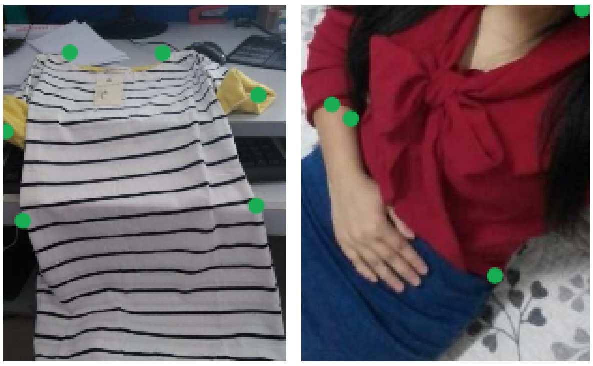


Figure 2.7: Landmark locations in FLD [20]

2.1.7 DARN dataset

The main goal of retrieval in visual fashion is to find garment in the shop domain image that corresponds to a query image in the consumer domain . A major challenge in performing successful image retrieval in the visual fashion understanding discipline is bridging the gap between domains, through modeling the discrepancy between domains. Problems related to that issue are described in Section 2.4.1. Lack of training sets which would have direct annotations between same clothing items in different domains, made this task historically extremely challenging.

Learning cross-domain mappings started to be possible with the introduction of DARN dataset [12] by Huang. DARN dataset consists of a large database of pairs of clothing items, being depicted in a consumer photograph and in an online shop picture. Availability of these direct cross-domain mappings is what makes any retrieval task possible. DARN dataset includes around 450K online shopping store images with clothing items and around 90K counterpart street images with exact same garments. Some examples of these online-offline pairs can be seen in Figure 2.8. All images come from either online clothing stores or consumer portals. Images show how clothes look in the real world and their depictions reflect diversity of clothes appearance, which allows for their modeling. Images in DARN show garments in different contexts and scenarios, as they include variations in pose, background or lighting. This appearance diversity is what helps retrieval systems to cross the domain gap.

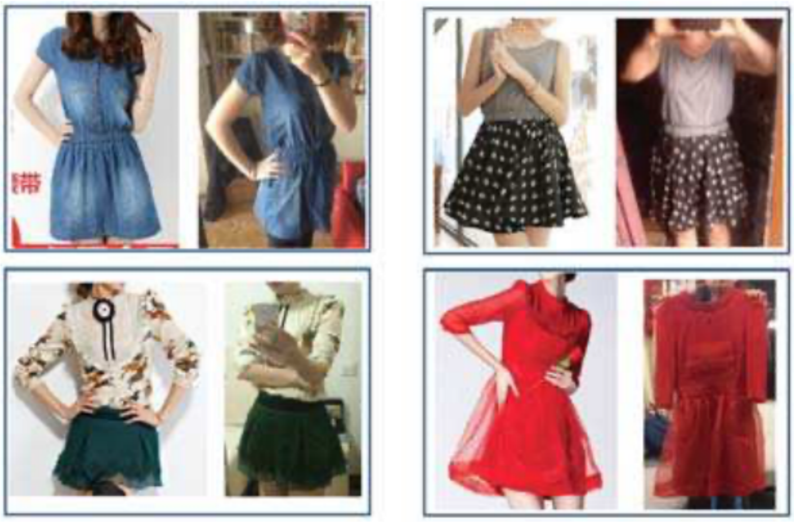


Figure 2.8: Online-offline pairs [12]

Beyond street-shop pairs, the dataset includes also fine-grained attribute annotations for each garment. Each image has 5-10 attribute type groups (also: attribute categories, not to be mistaken with garment categories) [12]. Some of these attribute categories are garment color, collar pattern, shape and length of sleeve etc. Possible values for these attribute types can be seen in Table 2.3.

Table 2.3: Attribute examples in DARN; reproduced [12]

Attribute groups	Value examples	Amount of unique values
Clothes Button	Double Breasted, Pullover	12
Clothes Category	T-Shirt, Skirt, Leather Coat	20
Clothes Color	Black, White, Red, Blue	56
Clothes Length	Regular, Long, Short	6
Clothes Pattern	Pure, Stripe, Lattice, Dot	27
Clothes Shape	Slim, Straight, Cloak, Loose	10
Collar Shape	Round, Lapel, V-Neck	25
Sleeve Length	Long, Three-Quarter, Sleeveless	7
Sleeve Shape	Puff, Raglan, Petal, Pile	16

In summary, there are three distinctive features of DARN dataset: large scale of the dataset, availability of online-offline pairs and fine-grained attributes. Massive scale allows for training efficient neural networks for the task, fine-grained attributes allow for learning semantic representations for clothing, while online-offline pairs provide a possibility to learn the dissimilarity metric between domains [12]. Combining all of these features together makes visual retrieval tasks in VFU possible.

2.1.8 ModaNet dataset

A breakthrough in clothing segmentation happened with publishing of the paper by Zheng, where a new dataset called ModaNet [33] was introduced. ModaNet is the largest dataset designed for fashion garment detection and segmentation. Its high granularity of mask annotations, which can be seen on Figure 2.9, combined with relatively large scale as for the segmentation-focused dataset makes researchers consider ModaNet as the state-of-the-art dataset for fashion item segmentation and detection.

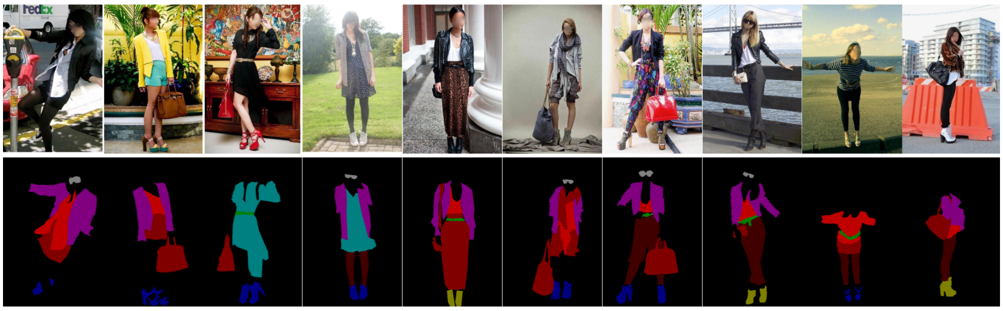


Figure 2.9: Mask annotation examples in ModaNet [33]

ModaNet is a fashion dataset containing over 55K fully-annotated, street fashion pictures, and has following properties. Firstly, dataset is dedicated primarily to garment segmentation task. Secondly, it provides fine-grained, pixel-level segmentation masks. Thirdly, dataset provides coordination information on polygons enclosing separate clothing garments. Fourthly, multi-item mask annotations on single image allow for algorithms to learn essential visual features in "border areas" between garments [33]. Fifthly, mask annotations can be transformed into bounding boxes in order to redefine the problem as fashion garment detection. Sixthly, one of 13 clothing category is assigned to each segmented garment. Seventhly, introduction of pose and shot angle variance, allows for models fitted on the dataset to generalize well across whole training data distribution. Eightly, variety of clothing types, appearances, styles and composition makes ModaNet domain-agnostic.

2.1.9 Exact Street2Shop dataset

Exact Street2Shop (E2S) dataset has been built by Kiapour et al. [13] and can be trained for finding street-to-shop domain mapping using pairs of exactly-matching garments in these two domains. The dataset consists of 40’000 pairs and is the largest available dataset of that type.

First type of images are street photographs of people wearing clothes in real-world uncontrolled setting. Those photographs were taken by unprofessional photographers. There are extreme variations of appearances of garments in photos from this domain, related to the quality of the picture, lighting, issue of indoor/outdoor environment, shapes and sizes of people’s bodies, human pose, camera angle, occlusion characteristics and whether a full head-to-toe picture is available [13]. Variations of garment depictions in street domain can be seen in Figure 2.10.

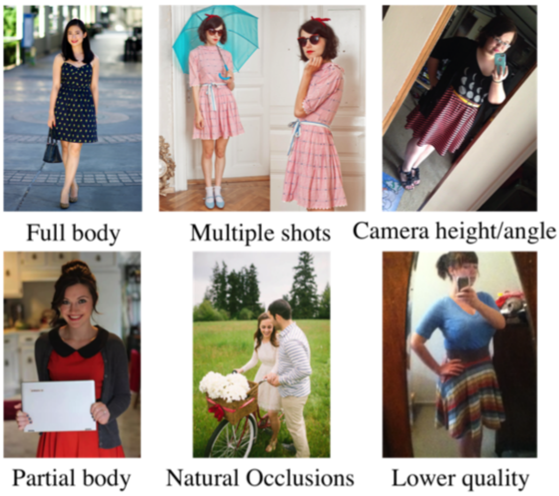


Figure 2.10: Street domain photograph variations [13]

Second type of images are shop photographs from online retail stores, depicting fashion garments either on a person, on a mannequin or in isolation. Those photographs were taken by professional photographers. Each individual garment might be linked to many pictures depicting it from different views. Variations of garment depictions in street domain can be seen in Figure 2.11. As we can see in the Figure, photographs in the street domain differ greatly from the shop domain. Learning cross-domain mapping is crucial to find a universal garment representation, which is the key issue in multi-task learning [13].

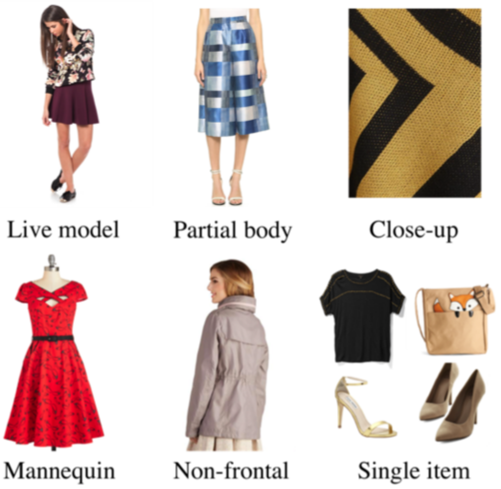


Figure 2.11: Shop domain photograph variations [13]

2.2 Categorization tasks

Categorization tasks are the most fundamental ones in visual fashion understanding. Two main categorization tasks in VFU are garment's category prediction and garment's attributes prediction. While categorization tasks might also be used as side tasks in some of multi-task learning methods (described in Section 2.5), in this chapter I describe solutions that focus solely on categorization.

2.2.1 Hierarchical modeling for multi-level categorization

Fashion industry is rapidly changing from season to season and over years. Therefore, it is highly unpractical to manually label huge fashion datasets. To cope with that problem, datasets usually use hierarchies of concepts to describe a fashion item (garment). A typical semantic structure of such hierarchy follows a following template: category - e.g. trousers, sub-category - e.g. jeans trousers (optional, as not all datasets use this additional categorization level), properties/attributes - e.g. high-waisted.

Such hierarchical nature of garment characterization cannot be easily modeled by standard convolutional neural networks, such as VGG [24] or ResNet [11], provided we would like to include in the model information on how class at one certain level affects classification chances of class at another level. One approach to go around this problem would be to train separate models for each categorization level, using prediction outputs from the higher-level model as an input of the lower-level model. However, it only considers all variables as independent of each other and does not explore the known structure of semantic concepts and information on how these concepts relate to each other between their respective levels in the hierarchy. Schematic visualization of the multiple-model take on the problem can be seen in Figure 2.12. Figure depicts scheme of three models, where dresses, shoes and coats are modeled separately. Researchers proposed a solution to the above-described problem, based on a custom neural network structure, which leverages embeddings of hierarchical labels [7]. This approach aims to model fashion items as a structured multi-level categorization task, which uses one global model for the whole concept structure and is consistent with intuitive way of analyzing visual fashion data.

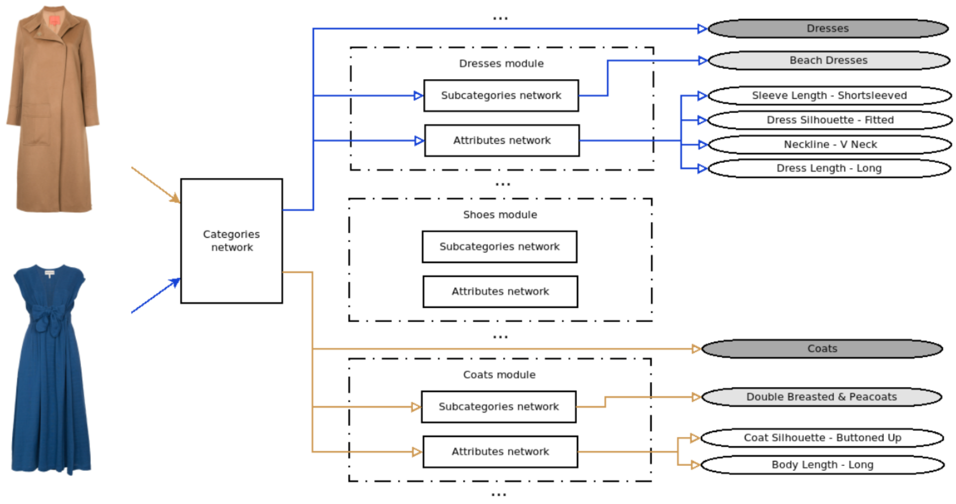


Figure 2.12: Schematic visualization of the approach based on separate models [7]

The goal of the multi-level categorization task is to perform classification across the whole structural tree and to predict classes for all levels. Visualization of working principle of the task can be seen in Figure 2.13. Usually all annotation levels, besides attributes, are exclusive to only one label. Therefore, main category and sub-category classifications can be seen as a multi-class classification, while on an attribute level it is a multi-label classification. Category can be inferred directly from sub-category (if we know "jeans trousers" is the sub-category, we can be 100% sure that the general category is "trousers"), while same attributes can be linked to many garments across many different high-level categories. Therefore, message passing connections, which denote variable influences, are not fully symmetrical between all category levels [7], and it follows the following rules: 1) Category's influence on sub-category, 2) Category's influence on attribute, 3) No influence of sub-category on attribute.

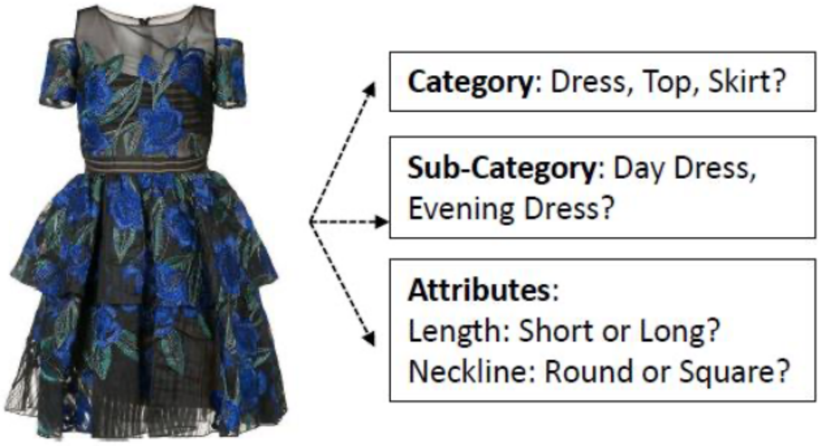

Category: Dress, Top, Skirt?
Sub-Category: Day Dress, Evening Dress?
Attributes:
Length: Short or Long?
Neckline: Round or Square?

Figure 2.13: Working principle for the multi-level categorization model in VFU [7]

Network description

The proposed unified model network structure for multi-level categorization problem, proposed by Ferreira et al. [7], can be seen in Figure 2.14. First part of the network is the ResNet-50 [11] CNN network, pretrained on the ImageNet [6], which extracts visual features from fashion pictures. ResNet's output then flows into three paralel fully-connected layers, one for each level in the hierarchy.

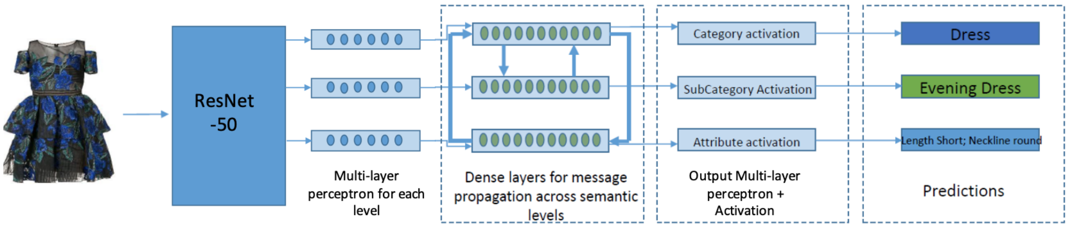

ResNet -50
Multi-layer perceptron for each level
Dense layers for message propagation across semantic levels
Category activation
SubCategory Activation
Attribute activation
Output Multi-layer perceptron + Activation
Dress
Evening Dress
Length Short; Neckline round
Predictions

Figure 2.14: Unified multi-level categorization network structure [7]

The next module of the network, which is a message passing block, is the most important module of the whole network responsible for hierarchical modeling of concept relations between levels. Message propagation connects only particular level layers. Information which benefits one level of the connection could also benefit the other end, and the same rule applies in reverse - therefore, propagation of messages between connected layers is bi-directional [7]. All layers in this module are dense, with L2 normalization being applied on them to decrease the over-fitting risk. Activation functions are then applied to outputs of layers of message propagation modules: softmax to multi-class classification for category and sub-category layers, sigmoid to multi-label classification in attribute layers. At the end, predictions are made for three variable levels, and based on them, cross-entropy loss functions are calculated for each level separately.

Performance evaluation

Evaluation of the built network has been performed by comparing its results to other solutions, such as separate baseline model or to a similar unified model, but without a message passing module. The following metrics have been used used in the evaluation: precision, recall and F1-score. Comparison results can be seen in Table 2.4 (category level), Table 2.5 (sub-category level) and Table 2.6 (attribute level).

Table 2.4: Evaluation results for category level; reproduced [7]

Method	Precision	Recall	F1-score
Baseline	80.01	79.43	78.73
ResNet independent	82.77	82.65	82.65
Unified, no message passing	81.66	82.57	81.47
Hierarchical, message passing	**83.53**	**84.16**	**83.35**

Table 2.5: Evaluation results for sub-category level; reproduced [7]

Method	Precision	Recall	F1-score
ResNet independent	**45.74**	34.90	**29.60**
Unified, no message passing	42.03	34.21	29.20
Hierarchical, message passing	42.68	**37.00**	29.39

Table 2.6: Evaluation results for attribute level; reproduced [7]

Method	Precision	Recall	F1-score
ResNet independent	47.55	85.16	58.51
Unified, no message passing	47.17	84.51	58.04
Hierarchical, message passing	**49.22**	**86.75**	**60.60**

As we can see in presented tables, for category level and attribute level predictions, the described network outperforms the rival solutions. For sub-category level, however, the results are comparable to other models or slightly worse. In summary, it can be concluded that the unified model, which learns from a hierarchy tree of concepts and models relations between them, produces better results for typical categorization VFU tasks, than a standard approach of separate modeling for each concept level [7].

2.3 Localization tasks

Development of visual fashion analysis field is linked to availability of high-quality fashion datasets [19, 13], which opened up research possibilities in the field for tasks, such as clothes recognition [12, 13] or retrieval [17, 19]. However, high-variance of clothes representations in these datasets led to the challenge of precise garment localization, which stems from pose variations and scaling issues [20] in available datasets. Researchers in the field tried different approaches to capture clothing localizations in a more discriminative way by introducing additional regional annotations, such as bounding boxes [13, 12], general object proposals [13] or masks [15, 29].

2.3.1 Localization schemes

Liu et al. [20] performed an experiment where multiple models were trained on a subset of DeepFashion [19] for tasks of attribute prediction and clothes retrieval. In all of the models, foundational visual features were learned with the same off-the-shelf CNN, as described in [25]. The only difference between models was the use of different garment localization schemes. Authors compared different localization schemes popular in the scientific literature. First presented scheme is based on full image and has no garment localization annotation. Second scheme is bounding box [13, 12], with minimal rectangle-sized area, which contains whole garment. Main challenge of bounding box is related to false positive area allocation close to its corners and edges. Third scheme is based on fashion landmarks. Fourth scheme is based on analysis of human body joints. This scheme is similar to landmarks but the key points are defined as points localized on body joints of a person wearing the analyzed garment.

Table 2.7: Top-5 attribute prediction recall rates for different garment localization schemes [20]

	Recall rate
full image	0.27
bounding boxes	0.53
human body joints	0.65
fashion landmarks	0.73

For comparison between different localization schemes, the experimentation performed by Liu [20] uses top-5 recall rate for attribute prediction task (results can be seen in Table 2.7) and top-k retrieval accuracy for clothing retrieval task. The results of experiment show that leveraging available landmarks helps with improving performance on other VFU tasks: attribute prediction or clothing retrieval. It also demonstrates that fashion landmarks are a better garment localization representation than bounding boxes or human body joint estimations.

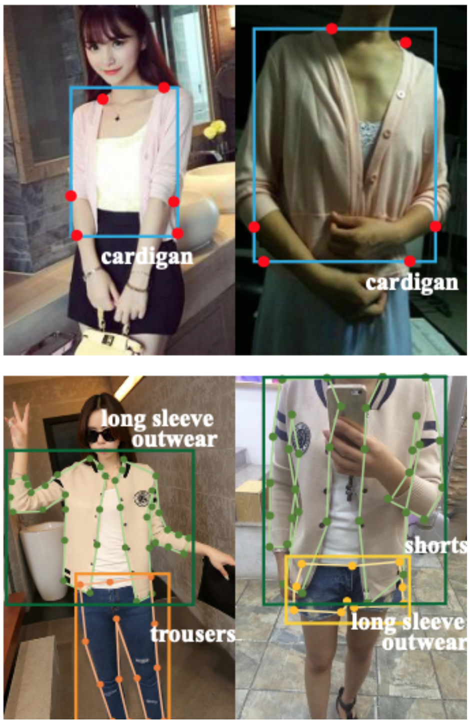


Figure 2.15: Difference between bounding boxes and landmarks, as shown in DeepFashion2 [8]

Landmarks (or fashion landmarks) are key location points situated in functional regions of clothes, such as waistline, ankle, elbow or hem. Liu et al. [20] presented landmarks, as a novel, more discriminative representation for localization properties of fashion garments. Improvement in discriminative abilities is possible thanks to landmarks' key features. Firstly, landmarks capture localization more precisely than bounding boxes.

Difference between bounding boxes and landmarks, as shown in DeepFashion2 [8], is depicted in Figure 2.15. As we can deduct from the Figure, landmarks improve on problem of bounding boxes related to false positive area allocation. Secondly, landmarks, as regional key points on clothing, introduce additional information on key points. Analysis of key points provides improvements when learning features for VFU tasks, as many essential visual patterns are located in regions around these key points.

2.3.2 Fashion grammars for landmark prediction

Many papers treat visual fashion analysis as just another possible use case for computer vision methods, and therefore do not try to introduce additional human knowledge about visual clothing problem characteristics. Wang [27] introduces a new model with high-level knowledge in the domain, called a fashion grammar model, aimed at improving performance in fashion landmark prediction sub-problem.

Grammar models aim to inject domain-specific knowledge, such as dependencies between elements, which typically would not be included in the training dataset itself. Use of grammar models is convenient when we try to model complex structures with rich annotations, and try to tackle the problem of local ambiguities. Fashion grammar describes kinematic and symmetric relations between landmarks on clothes in pictures. These relations are used to predict a confidence map of positional distribution (heatmap) for each landmark. Wang [27] argues that use of these heatmaps, which are essentially probability distributions of a landmark location, give a more detailed landmark estimation than a direct vector prediction, because of non-linear nature of pose estimation.

Kinematics grammar

Kinematics grammar describes kinematic relations between different parts of a garment, so that particular clothing landmarks are connected in a way that satisfies human anatomy constraints. There are four kinematic relations in the described grammar [27]: 1) Left collar - left waistline - left hem, 2) Left collar - left sleeve, 3) Right collar - right waistline - right hem, 4) Right collar - right sleeve.

Symmetry grammar

Symmetry grammar describes symmetry property of garments, as left and right side of a clothing piece ought to mirror each other. There are four symmetrical relations in the described grammar [27]: 1) Left collar - right collar, 2) Left sleeve - right sleeve, 3) Left waistline - right waistline, 4) Left hem - right hem.

Performance evaluation

Grammars' influence on prediction model performance can be evaluated using a difference of cost function values between model baseline and models with grammar modules. Model baseline does not incorporate any high-level knowledge and has no grammar module. Wang [27] uses normalized error to evaluate effectiveness of introducing different combinations of grammar modules.

Table 2.8: Fashion grammar effectiveness; reproduced [27]

Method	Normalized error (NE) on DeepFashion [19]	NE on FLD [20]
Without any fashion grammar	0.0615	0.0681
Just with kinematics grammar	0.0538	0.0641
Just with symmetry grammar	0.0525	0.0659
Symmetry and kinematics grammars	**0.0484**	**0.0583**

Results of performing this effectiveness evaluation can be seen in Table 2.8. As we can see in the Table, both kinematic and symmetry grammars positively contribute to model's overall performance, while being tested on DeepFashion [19] and Fashion Landmark Detection [20] datasets. Therefore, we can conclude that including these grammar modules improves CNN-based network's prediction performance for fashion landmark prediction task [27].

2.3.3 Segmentation and detection in VFU

While introduction of new, massive, highly granular fashion-dedicated datasets, such as DeepFashion [19], drastically improved fashion understanding in tasks such as attribute prediction or retrieval, those datasets only consisted of images with image-level annotations. Lack of pixel-level annotations (masks) makes it impossible to achieve high performance on tasks of clothing segmentation and detection. In order to address challenges of automatic detection and segmentation of clothing garments, subsequent published work in the field [12, 29] made contributions of pixel-annotated datasets.

Localization-based tasks of segmentation and detection in the VFU domain differ from classification-type (item category prediction, item attribute prediction) and retrieval-type (in-shop, consumer-to-shop) tasks, in terms of applicability of universal state-of-the-art computer vision algorithms. Specificity of the fashion domain makes categorization tasks in the VFU field a much more difficult problem than typical image classification problems that can be trained on ImageNet [6]. That is not the case in segmentation and detection tasks. In fashion item detection and segmentation, universally acclaimed algorithms for those tasks can be used as out-of-the-box solutions and still provide great performance [33] in VFU. State-of-the-art algorithms for object detection, such as Faster RCNN, YOLO or SSD [23, 22, 18] can be applied to clothing item detection. State-of-the-art methods designed for semantic segmentation, such as DeepLab or CRFasRNN [4, 32] can be applied to clothing item segmentation. It is outside of the scope of this work to describe the mentioned algorithms in detail.

2.4 Retrieval tasks

Retrieval in VFU focuses on finding similar pieces of clothing in a database, based on a visual query. Such search can be a task in itself or it can be used as an internal module in solutions dedicated to other tasks. In this chapter, I describe issues related to retrieval tasks in VFU.

As an example, one possible application of an VFU retrieval algorithm is finding exact garments. That example is portrayed in Figure 2.16. In this example, algorithm has been trained on Exact Street2Shop dataset, which was described in Section 2.1.9.

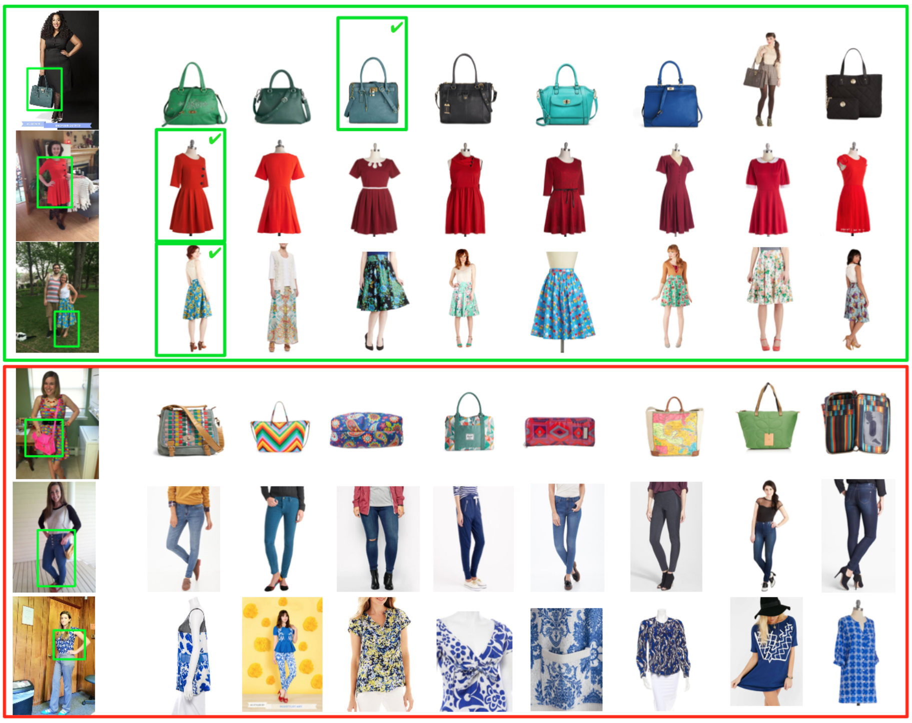


Figure 2.16: Possible retrieval results from Exact Street2Shop [13]

2.4.1 Bridging the gap between street and shop domains

Clothing companies and online retailers are actively looking to provide new services to their clients that would change the way they shop for clothes. One such use case relates to the client's need to find a garment in an online clothing store that is similar or even exact to the garment he/she saw on the picture. Technical solution to this problem requires a system that would take a fashion image from the domain of unprofessional photographs (street pictures, selfies, social media images) as input and output the image of the most relevant searched-for clothing item in the domain of professional pictures from online stores. Such problem of finding cross-domain mappings is an example of retrieval in VFU. Retrieval methods are a popular problem in the industry and one of the most researched ones in computer vision. However, traditional CV retrieval methods such as Fisher Vectors, cannot be used in VFU, because of large discrepancy in the fashion data [13].

In fashion context, our images mainly come from 2 distinct domains. First domain is street/consumer/unprofessional domain, where fashion pictures being taken outside of professional studio setting. These pictures can be considered as real-world examples, taken in an uncontrolled setting. Second domain is shop/commercial/professional domain, where fashion pictures are being taken inside the professional studio. These pictures typically have clean backgrounds, more neutral human body poses and better lightening than those from the street domain [13].

Bridging the gap between those two domains is the main challenge for retrieval methods in VFU. Challenges stem from domain differences, such as significant appearance variance of garments between domains. Visual clothing features in images are deformable and highly variant between examples, which is another major challenge in retrieval tasks. One approach to bridging this gap is a retrieval method based on body parts allignment [17]. However, a much more accurate method is based on finding a similarity measure between garments in different domains. A deep similarity measure is used by Kiapour et al. [13] which introduces a VFU task in which the aim is to find not just a similar, but exact clothing item in the other domain.

In order to find the universal, domain-independent representations for clothing items, it is not enough that we train models using examples from both domains, but we also need to learn the cross-domain representation. Otherwise, our models would only learn domain-specific features which would not be transferable to different domains, therefore decreasing algorithm's performance. This is the problem of **domain adaptation**, in which the algorithms is trained to learn a transformation function that maps examples from different domains into one feature space.

2.5 Multi-task learning approaches

In many research papers, scientists use the same neural architecture to predict multiple VFU tasks at once [26, 16, 33, 13]. This approach is called multi-task learning and is based on learning clothing representations jointly across many sub-problems.

One example of using such approach to learning clothing features is FashionNet algorithm [19], which uses landmark-based features to help with prediction of attributes and categories. The algorithm simultaneously optimizes loss functions for landmark localization, landmark visibility, garment category and pairwise metric. Sharing features between tasks helps with better prediction performance, e.g. information about landmark location can be used to create better local representations and therefore, also improve categories prediction.

2.5.1 Landmark pooling

Introduction of massive and fine-grained datasets for visual fashion, such as DeepFashion [19] opened possibilities for researchers to try to predict values across different sub-problems jointly. One such approach, called FashionNet, is proposed by same authors [19]. Landmark pooling is a key mechanism described in FashionNet model [19]. The general idea of the algorithm is to utilize information about landmarks in order to learn more discriminative features for other tasks, such as category and attribute prediction.

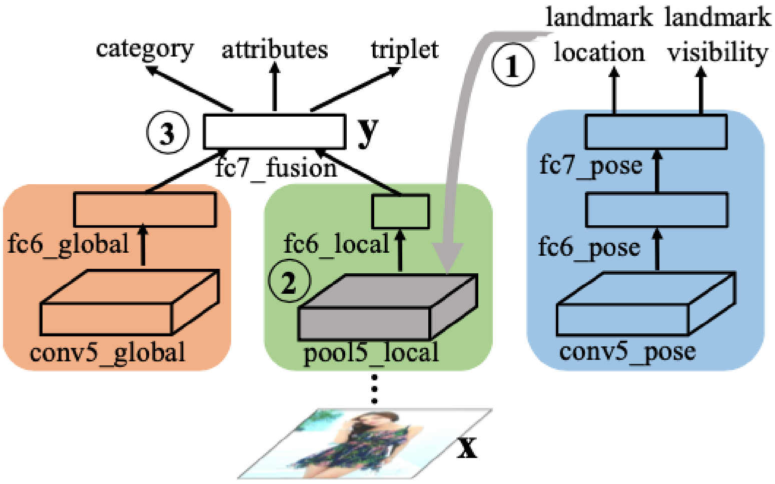


Figure 2.17: Branch structure in FashionNet [19]

FashionNet is built on top of VGG-16 neural network, a popular CNN network for automatic visual features learning in computer vision. Features learned by a VGG-16 output layer can be used as inputs for three-branched neural structure, designed specifically for joint learning in clothes domain. Outputs of these three branches, which focus on different tasks, are then combined to jointly learn features across tasks through passing information between branches. Three branches, whose structure can be seen in Figure 2.17, account for learning different tasks. Firstly, blue branch learns landmarks' feature maps: location and visibility. Secondly, orange branch learns global features, through additional convolution of features learned with VGG-16. Thirdly, green branch learns local features, through pooling over the previously-learned landmark from the blue branch [19].

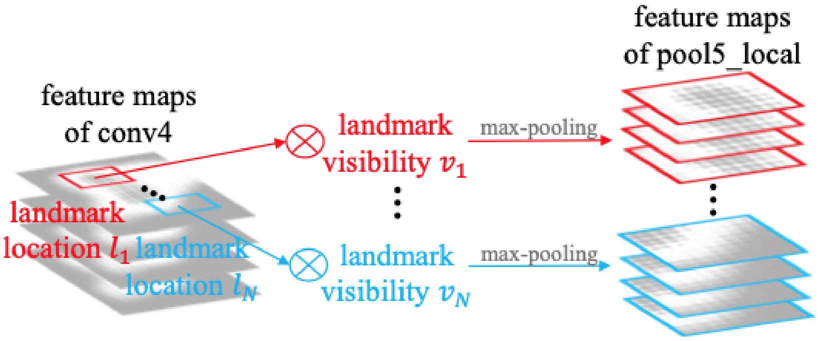


Figure 2.18: Landmark pooling layer in FashionNet [19]

First, the blue branch predicts landmarks' location and visibility. Then, inside the green branch the max-pooling aggregation process is performed around predicted landmarks, using feature maps learned previously by VGG-16 CNN. This pooling operation is depicted in Figure 2.18 and it leads to creation of new cross-task influenced feature maps. Landmark pooling layer performs max-pooling of local features around predicted landmarks to learn local feature maps, which are later used in final prediction tasks [19]. Weights of the landmarks that were not predicted are gated. Concatenation of local features allows for modeling the interaction between landmark points. These additionally learned local features are crucial, as they allow for more discriminative model representation, when the training dataset includes pictures with deformations or occlusions, which typically is the case in real-world scenarios.

Orange and green branch are combined afterwards, in order to jointly predict category, attributes and cross-domain metric (consumer-shop relationship for same clothing piece on pictures in different domains) [19]. These predictions are affected by cross-task information passing, thanks to which landmarks and attributes are learned jointly. This approach is viable as tasks of landmark recognition and category/attribute/cross-domain metric are correlated. Therefore, using localization-based landmark keypoints can help in more accurate prediction in categorization tasks as well.

In practice, joint learning comes down to simultaneous optimization of multiple loss functions at once. FashionNet [19] also uses that concept. The solution network first calculates L2 loss in order to find landmarks. Using predicted values, one-versus-all softmax loss is calculated in order to classify categories. Next, cross-entropy loss is used for attributes prediction. Lastly, the cross-domain triplet loss aims to learn clothing pairs relationship.

Performance evaluation

The effectiveness of landmark pooling method described in FashionNet [19], can be evaluated by comparison to other neural solutions, that showed good results for the same visual fashion tasks, namely WTBI [13] and DARN [12]. Both of them have been trained on the same dataset as FashionNet - DeepFashion [19]. WTBI and DARN do not employ cross-task landmark pooling technique.

Category prediction can be evaluated using top-k accuracy metric. As we can see in Table 2.9, FashionNet with landmark pooling layer outperforms both DARN and WTBI significantly in category prediction. We can also notice that fine-grained, rich attributes are essential in driving the model's accuracy [19].

Table 2.9: Category and attribute prediction accuracy comparison between FashionNet [19], WTBI [13] and DARN [12]; reproduced [19]

	FashionNet [19]	WTBI [13]	DARN [12]
Top-3 category	82.58	43.73	59.48
Top-3 attribute Texture	37.46	24.21	36.15
Top-3 attribute Fabric	39.30	25.38	36.64
Top-3 attribute Shape	39.47	23.39	35.89
Top-3 attribute Part	44.13	26.31	39.17
Top-3 attribute Style	66.43	49.85	66.11
Top-3 attribute (all)	45.52	27.46	42.35

Table 2.9 also shows comparison results for attribute prediction. FashionNet's performance on this task is also significantly better compared to other solutions. Especially great results are obtained for attributes in groups describing "shape" and "part". Attributes from those groups are usually described by information around landmarks. Therefore, we can conclude that local features, found thanks to landmark pooling layer, directly contribute to more accurate attribute prediction. FashionNet also outperforms DARN [12] and WTBI [13] in retrieval for both consumer-to-shop and in-shop scenarios [19]. We can conclude that FashionNet achieves higher accuracy than its competitors thanks to advantages of using landmarks-based feature pooling, which boosts performance across tasks.

2.5.2 Attention mechanism in VFU

Attention mechanism in neural network aims at selectively focusing on a few important aspects, while ignoring others. It was one of the biggest breakthroughs in deep learning research in the last decade. Attention is mainly used in natural language processing, but there were a few successful applications in computer vision as well, such as Visual Question Answering (VQA) or image captioning [27]. In the context of classification tasks, attention modules are effective in helping the network to learn which are the relevant picture regions "it should look at" and which regions should be ignored.

Wang [27] claims that no attention mechanisms had been used for solving any visual fashion understanding problem before his proposal. Attention mechanisms are introduced in order to help improve fashion category and attribute classification, through making clothes representations more robust as well as to leave out information which are not relevant to the prediction task. Wang introduces two types of attention mechanisms in his model: fashion landmark-aware attention and clothing category-driven attention.

Landmark-aware attention

First introduced attention mechanism is a landmark-aware attention, which uses landmarks property of having strong representation capability. Landmark-aware attention enforces that network focuses on functional parts of clothes. Attention is learned in a supervised way. That attention generates landmark-aligned features, which allows for the model to capture the significance of regions which are semantically richer [27].

Category-driven attention

Taking into account functional regions may not be enough to correctly classify attributes if their granularity is fine. That is why another attention mechanism was introduced: a category-driven attention. That attention is goal-driven and can be trained to enhance features related to the classification task at hand, which might lead to better performance [27]. Category-driven attention mechanism focuses on picture areas relevant to predicting particular category or attribute.

Performance evaluation

Influence of attention mechanisms on classification model performance can be evaluated using a difference of accuracy and recall between model baseline and models with attention mechanisms that enhance landmark-aligned and category-related features. Model baseline does not include any attention mechanism [27].

Table 2.10: Attention mechanisms effectiveness; reproduced [27]

Method	Top-3 cat. acc.	Top-5 cat. acc.	Top-3 attr. acc.	Top-5 attr. acc.
Baseline, without any attention	83.23	89.51	43.28	53.54
Just with landmark-aware attention	87.75	93.67	49.93	58.78
Just with category-driven attention	85.27	91.32	48.29	56.65
With both attention modules	90.99	95.78	51.53	60.95

Results of performing this effectiveness evaluation can be seen in Table 2.10. Metrics are shown for top-3 and top-5 accuracy for both category and attribute prediction.

As we can see in the Table, models with attention produce better results than baseline models. Therefore, we can conclude that including these attention mechanism modules improves CNN-based network's prediction performance for category classification and attribute prediction.

2.5.3 Upsampling feature maps for improved attention

This Section builds up on the information included in Section 2.5.2. The method described here aims to improve landmark-driven attention, above the performance of standard attention, described earlier. Liu et al. [16] argues that majority of methods aiming to improve prediction accuracy of landmarks (more about landmarks in Section 2.3.1) do not succeed, because of the fact that resolution of predicted landmark heatmaps is too low. A reason for that is the usage of full CNNs with pooling layers, which downsample the output. That causes the issue of decreased accuracy for related VFU tasks, because landmarks are commonly located in "difficult" areas of the picture, such as corners, which is another reason of the mentioned problem. One example of such situation where there are multiple pooling operations involved, which downsample landmark heatmaps, is the solution described in Section 2.5.2 [27].

Authors propose a new solution [16] based on transposed convolutional operations, which upsample feature maps. Those operations lead to heatmaps of higher resolution, which have the same size as input training pictures, and that subsequently leads to better performance for the VFU task of landmark localization. The predicted landmark heatmap can be used in order to help with another task of category and attribute classification, by introducing landmark-driven attention mechanisms. Such attention mechanism enables the network to selectively focus only on the most important regions of the picture (more about landmark-based attention in Section 2.5.2), from the perspective of the category/attribute prediction problem. Element of an attention mechanism, which allows for that is the attention map, which combines both local-oriented landmark location properties with global-oriented visual features learned by convolutional layers. All the most task-related information is leveraged, while non-important features are discarded, which eventually leads to higher prediction accuracy for tasks.

The proposed solution [16], which aims to improve category and attribute classification accuracy in visual fashion understanding, consists of two parts. First part is upsampling of landmark feature maps, through transposed convolutional layers. Second part is using generated landmark heatmaps in an attention mechanism, which strengthens or limits learned visual features. There is only one, global attention mechanism, as opposed to two separate attention mechanisms in solution described in Section 2.5.2.

Separate attention branches lead to putting hard constraints on feature selection, while unified attention is considered a soft constraint, which boosts important features in a more natural way and can be learned more efficiently.

Network description

As we can see in Figure 2.19, the network consists of a VGG-16-based foundation, which is a standard CNN network, used here for extracting visual features. On top of those layers, there are two separate branches, designed to focus on their own tasks: landmark localization branch and attention branch. First, in a landmark localization branch, multiple transposed convolution operations perform the upsampling, until the point where heatmaps for all landmarks are of the input size [16]. At the end of the branch, L2 loss is calculated and landmark predictions are generated.

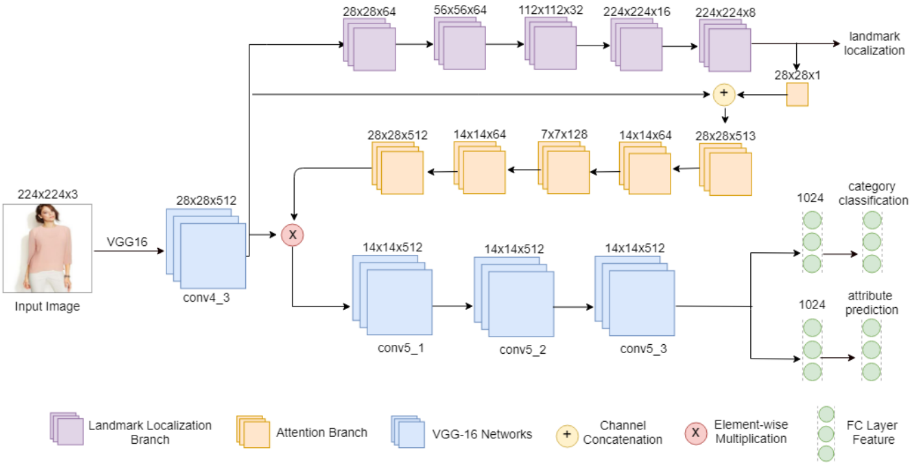


Figure 2.19: Upsampling-boosted attention network [16]

Next, output of the landmark localization branch is concatenated with features from the foundational VGG-16 network, and passed to the attention branch, where data is passed through two sets of convolutions and pooling layers [16]. Afterwards, a transposed convolution is performed and the output of the attention branch is then multiplied element-wise again, with the output of the first VGG-16 net. Landmark-driven attention allows for new way of understanding the VGG-16 features, where if the output of the concatenation is below 1 - features will be reduced according to the calculated weight, while if the concatenation value is above 1 - features will be enhanced, in order to account for importance of those features around the landmark. Next, the information is passed through multiple VGG-16 networks and lastly, spread into two final branches for category and attribute tasks, where cross entropy loss is calculated and final predictions are performed.

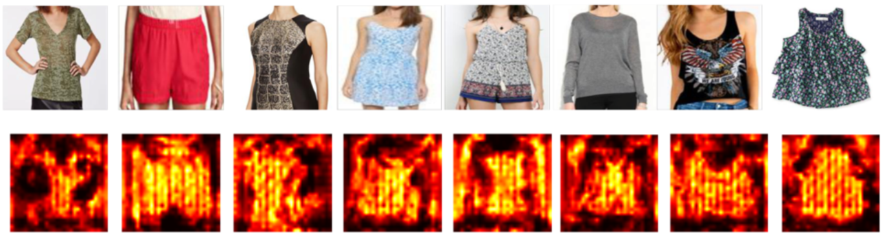


Figure 2.20: Attention maps visualization [16]

Figure 2.20 depicts attention maps visualization with pixel-wise activations for example pictures from DeepFashion [19] dataset. We can observe how areas of pictures around landmarks are enhanced, while data in useless regions is filtered out.

Performance evaluation

Performance of the above-described network [16] was evaluated on DeepFashion benchmark for category and attribute prediction, which is described in detail in Section 2.1.2. Metrics for evaluation were chosen as follows: top-k accuracy for category classification, top-k recall for attribute prediction, normalized distance between landmarks for landmark localization. The results of the upsampling-boosted landmark-driven attention network were compared to recent deep learning-based solutions for the same VFU problems. For both categorization and localization tasks, the network proposed by Liu et al. [16] outperforms its rivals and achieves what can be considered state-of-the-art prediction performance.

2.5.4 Attribute-aware retrieval feature learning

Huang et al. [12] proposed a solution to the problem of cross-domain image retrieval, by introducing a DARN network (not to be confused with DARN dataset described in 2.1.7). The specific addressed problem, tackled by DARN is, being given a fashion photograph in the street domain, to retrieve garments, which are the same or closely similar in the shop domain. Such definition of a VFU retrieval task resembles practical applications, as usually shoppers would look to find a concrete item in store, while knowing already how it looks in real-life street picture. The system was trained on DARN dataset (described in Section 2.1.7) which is a large cross-domain garment dataset. Depiction of how the system works is presented in Figure 2.21, which shows examples of street-to-shop domain retrieval.

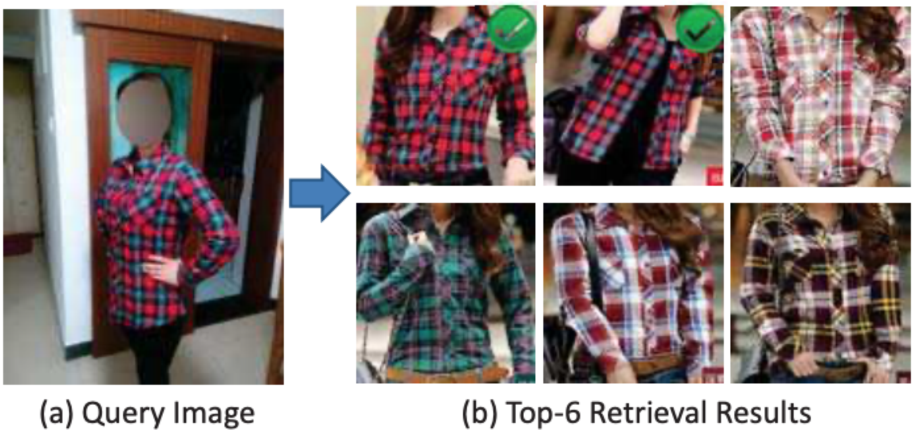


Figure 2.21: Cross-domain retrieval examples [12]

DARN, which stands for Dual Attribute-aware Ranking Network, is a neural network, which is designed for retrieval feature learning [12], based on visual features describing semantic attributes of clothing items. DARN network consists of two sub-networks, which are very similar in their design. Each of the sub-networks is designed to handle data coming from one, specific domain, either online (shop/commercial) or offline (street/consumer). Pair data from different domains is processed simultaneously in both sub-networks. Separate processing of data from different domains allows for learning characteristics of each domain and therefore decrease the discrepancy between them. Both sub-networks are attribute-aware, which means that human-level knowledge about attributes is leveraged in order to improve algorithm's visual understanding. This simultaneous contribution of fine-grained attribute knowledge and visual similarity constraints results in powerful network, which combines both semantic and visual understanding.

Network description

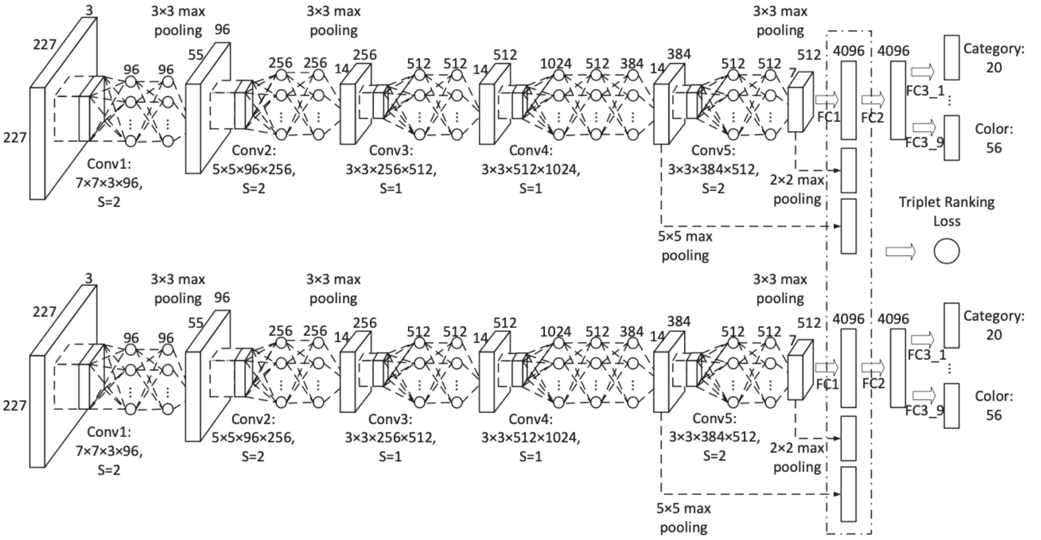


Figure 2.22: DARN network structure [12]

Figure 2.22 presents the structure of the DARN network. As we can see, there are two separate sub-networks in DARN [12]. Each network has, at its bottom, five convolutional layers, with some of the layers using additional max pooling. Weights of these low-level layers are shared across all attribute categories in each domain, and used in order to learn visual features characteristic to each domain. These low-level layers are followed by two dense layers. Then, on top of that network, tree-structure layers for modeling attributes are placed, where each branch is responsible for one specific attribute. These high-level layers aim to capture semantic meaning of attributes. The FC2 output (see Figure 2.22) is sent to all the following branches, which learn semantic attributes on their own. This weight sharing to tree-structure layers from lower layers is what allows DARN to effectively combine both visual and semantic features. Both sub-networks learn to produce outputs, which can be compared between each other, and therefore used as a measure of cross-domain similarity.

Performance evaluation

Huang [12] argues that incorporation of attribute-driven learning module is the key mechanism that contributes to improvement of accuracy in clothing item retrieval task. Six examples of top-5 retrieval can be seen in Figure 2.23, with query image being shown on the left side of each row and top-5 results being shown on the right side. Best retrieved element is additionally highlighted with a green frame. As we can see from the Figure, the network has been able to successfully retrieve from the opposite domain, items which are very similar to the ones from query.

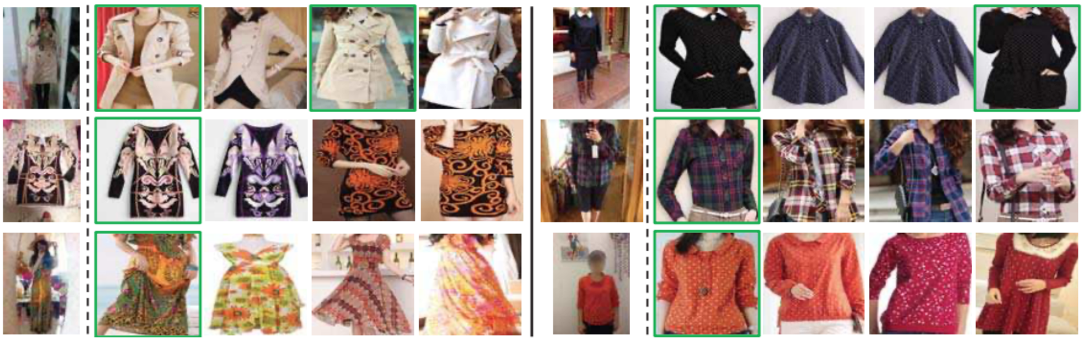


Figure 2.23: Top-5 retrieval results [12]

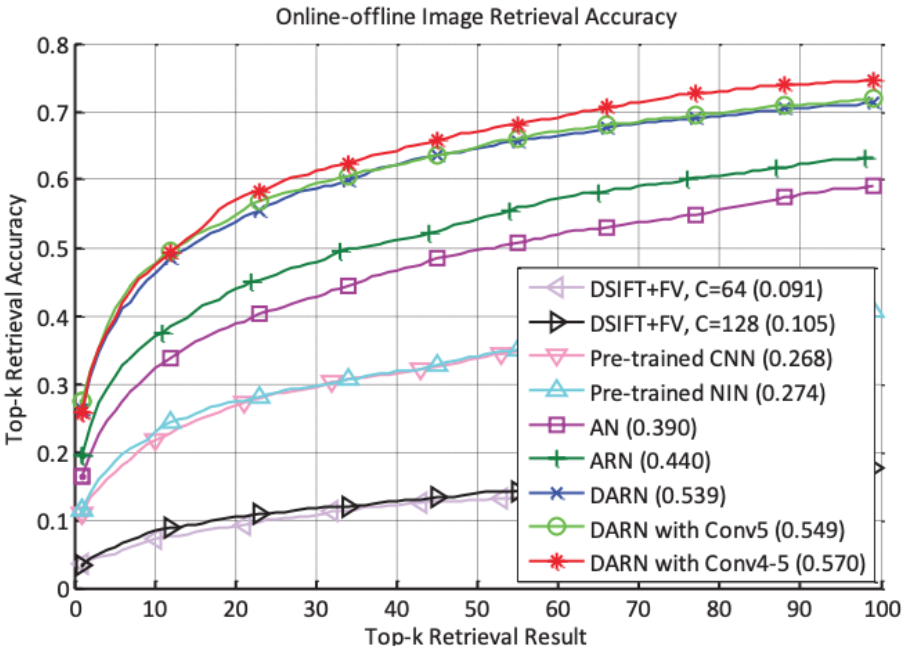


Figure 2.24: Top-k retrieval metrics [12]

A comparison study has been performed to measure and evaluate performance of DARN algorithm, as related to other popular methods in the literature. The results of the study can be seen in Figure 2.24. Top-k retrieval accuracy has been used as a comparison metric. As we can see in the Figure, DARN network outperforms all the other methods, including traditional CNN pretrained on AlexNet dataset or ARN, which is the similar attribute-aware network, which uses only 1 channel for processing images and does not have the dual nature described in this subchapter. Based on the study we can conclude that performance of an algorithm for cross-domain retrieval in visual fashion understanding increases with: introduction of sub-networks, introduction of additional semantic knowledge about attributes and separate processing of data according to the domain from which it comes from [12].

2.5.5 Localization-aided attribute representation learning

Cross-domain retrieval is one of the most important problems in VFU. While many papers have been proposed on this subject [17, 13], challenges remain for situations, where there are multiple attributes of clothes present, or where there is a need for attribute manipulation, which requires finding precise features that represent just the analyzed attribute. Some researchers tried to tackle such problems by crossing visual features of the query fashion image with the searched attribute representation [31].

Those methods, however, do not focus on leveraging localization properties for representing attributes, which is essential for determining which parts of the image are "responsible" for which attributes. DARN paper's authors [12] explored localization-aided attribute prediction and their method used bounding boxes annotations. While DARN's approach has been helpful in pushing the VFU field forward, in order to apply it to real-life solution, it would require each picture, which we are trying to make an inference for, to have fully annotated boxes for all of its attributes. However, it is very hard to annotate bounding boxes for every attribute and it is desirable for VFU computer vision algorithms to be able to deal with multiple attributes in a weakly supervised manner. FashionSearchNet takes such approach, where only meta-level image annotations are given and attribute activation maps are found by attention mechanism [2]. This method is called attribute representation learning, and enables region-specific attribute manipulation by removing some local aspects of the unwanted features.

Authors of FashionSearchNet [2] argue that in order to be able to perform extensive visual search and retrieval in the VFU area, there is a strong need for similarity learning, which would leverage information across tasks to find universal representations. Authors use weakly supervised localization-based learning and region-specific awareness, in order to find distinctive, representative features separately for each attribute. By localizing towards attributes, algorithm can find more accurate, area-specific attribute features and discard redundant data.

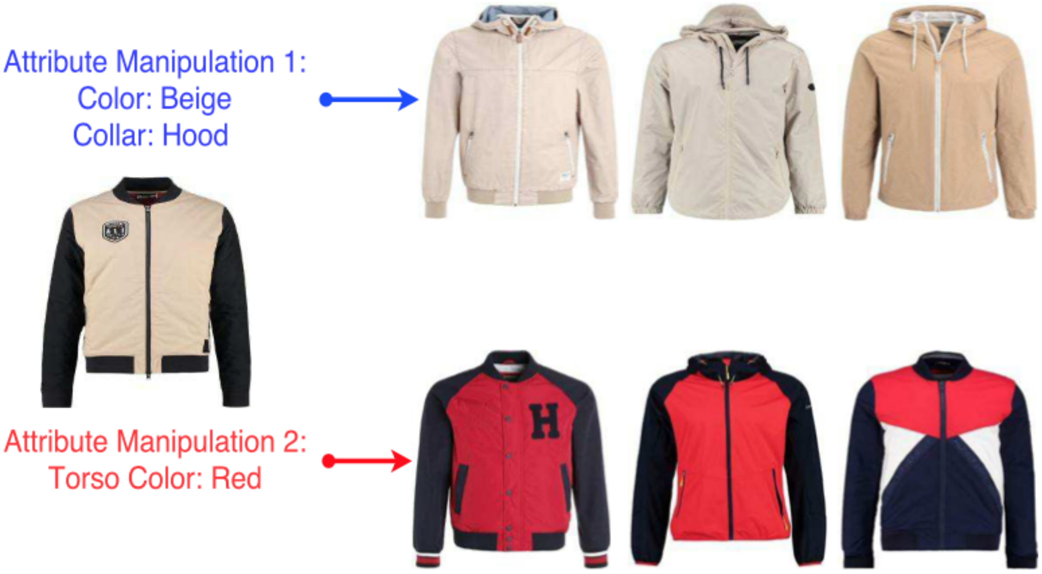


Figure 2.25: Atrribute manipulation application of FashionSearchNet [2]

Figure 2.25 shows FashionSearchNet's [2] application examples for attribute manipulation. In the first example, whole garment color was changed to beige and the collar was changed to hood. In the second example, just torso color changed. However, because of the fact that no garment with fully red torso was found in the used dataset, the most similar ones were returned. This example shows the importance of similarity learning. Authors argue that no other published method is able to perform such precise and intuitively-understood region-specific manipulations.

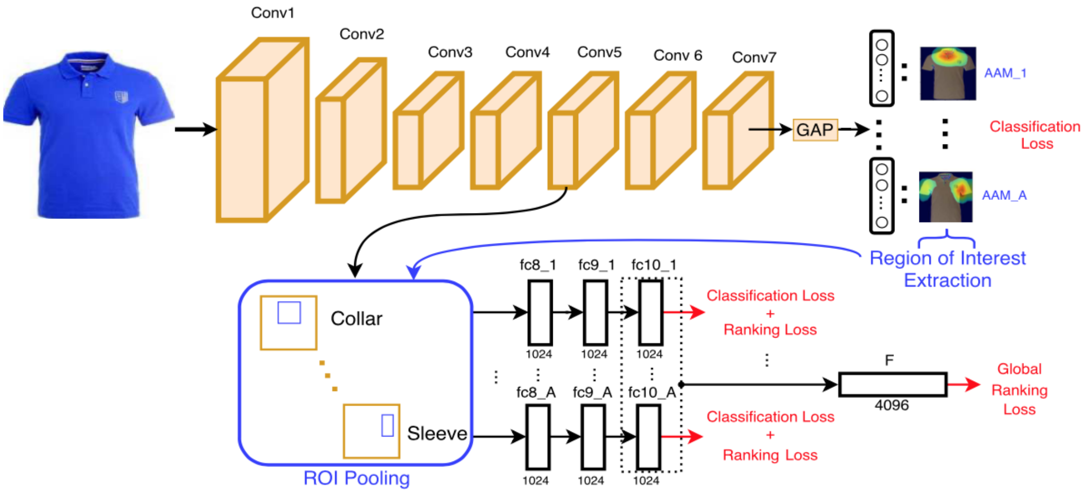
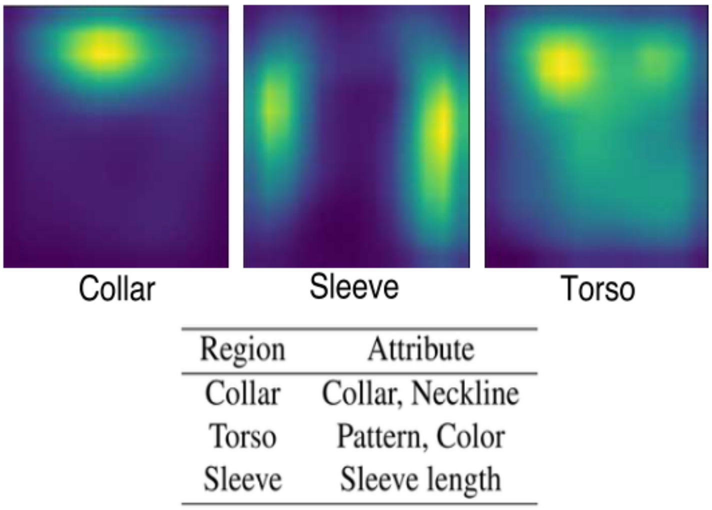


Figure 2.26: Neural architecture of FashionSearchNet [2]

Graphic depicting the neural architecture of FashionSearchNet [2] can be seen in

Figure 2.26. Firstly, input image is passed through 7 convolutional layers, which extract visual features. Then, global average pooling layer (GAP) is used as a helping layer for generating attribute activation maps (**AAMs**). These AAMs are a key mechanism, which allows for training in weakly supervised manner, just by providing meta-level attribute annotations, without any localization definitions. Additional examples of AAMs are shown in Figure 2.27 and Figure 2.28. In Figure 2.27, we can see different activation maps for different types of attributes. Also, different values of the same attribute type will produce slightly different AAMs, as can be seen in Figure 2.28.



Region	Attribute
Collar	Collar, Neckline
Torso	Pattern, Color
Sleeve	Sleeve length

Figure 2.27: Activation maps for different types of attributes [2]

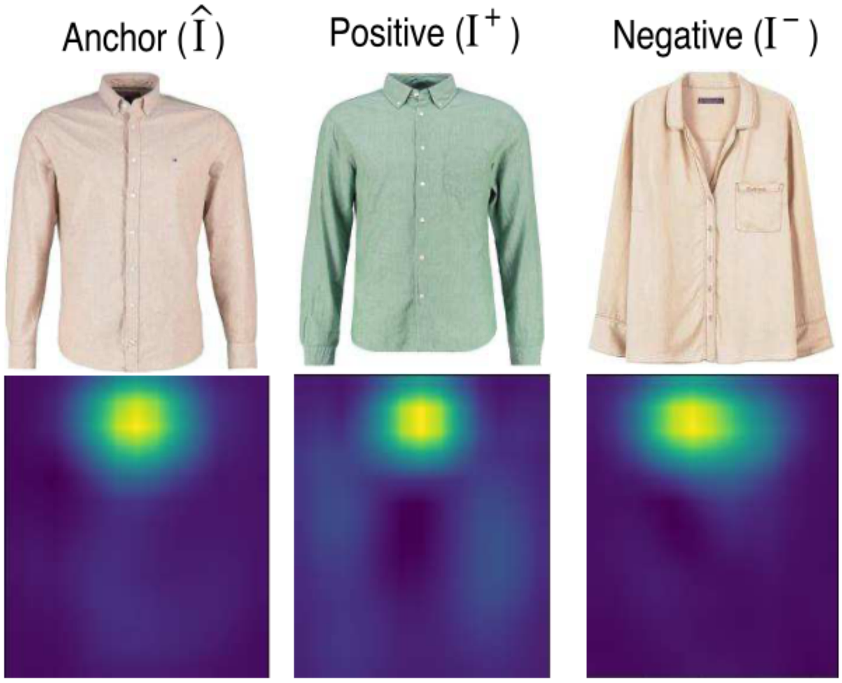


Figure 2.28: Activation maps for different values of the same attribute type [2]

AAMs are then used to estimate regions of interest, which is the strength of correlation between a particular area of image and its influence on attribute. Such regions of interest allow for representing attribute with its localization-based features and helps with more efficient representation learning, as some attributes are more heavily present only in certain areas [2]. This, in turn, leads to abandoning redundant features, which happens during the pooling operation of features from fifth convolutional layer over extracted regions of interest. Pooled features are feeded into a series of dense layers. Then, representations for features are learned, as well as similarity learning losses are calculated. At the very end, all found representations are concatenated and form a final joint representation, which can be used for search and retrieval. Search is performed through calculating global ranking loss, which penalizes features that should contribute less to the final search results. Also, having learned precise attribute representations, we can now modify certain variables to change some aspects of the attribute (such as torso color in Figure 2.25).

Performance evaluation

Authors of [2] present results of their novel FashionSearchNet by comparing its results to other algorithms in the field. FashionSearchNet's performance on DeepFashion dataset for the task of search by query is similar to other methods, as measured by top-k retrieval accuracy. However, its top-k accuracy for the task of retrieval by query and attribute manipulation, as measured on DeepFashion benchmark, outperforms other methods significantly [2].

3. My implementation

Having surveyed most important datasets and solutions in visual fashion understanding field, in this chapter I present the actual novel contribution of this work. I describe the research angle and details of implementation, as well as present results of experiments and evaluate correctness of my hypothesis.

3.1 Research angle

Multi-task learning is an approach in machine learning, where a model learns to optimize loss functions for multiple tasks at the same time. Features/representations are passed between task-specific branches, which leads to knowledge sharing between tasks. Presumption is that if tasks are related, we can benefit from sharing representations and obtain better performance, than if we tried to optimize those tasks with separate models. Basic diagram showing a general framework of multi-task architecture can be seen in Figure 3.1. After surveying major solutions in the research literature in Chapter 2, I conclude that multi-task deep learning approach is used in majority of state-of-the-art methods in the VFU field. Those methods have been described in Section 2.5.

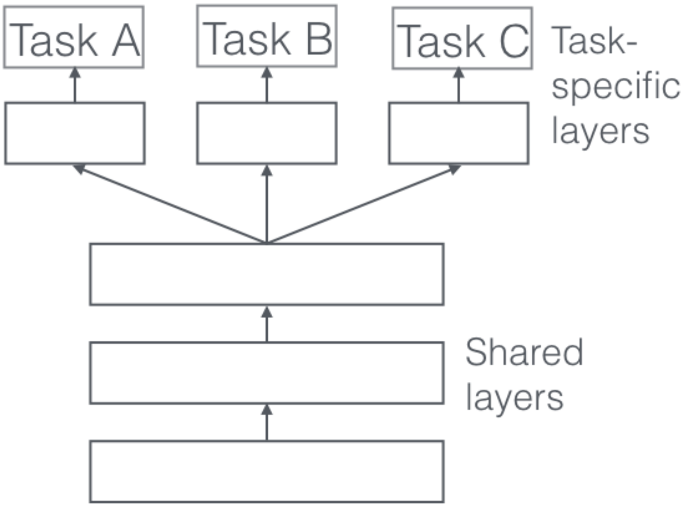


Figure 3.1: General framework of multi-task learning [1]

Traditional definition of generalization in machine learning context is that it's the model's property of being able to perform well not just on training set, but also on never-seen-before training set, so that it's able to learn important properties of the whole distribution, without learning noise in the dataset. As artificial intelligence research moves forwards, some researchers, such as Zhang et al. [30] argued for rethinking definition of generalization in ML. While, obviously, AI models cannot generalize outside of the training data distribution, ideally we would like AI models to exhibit human-like intelligence and be effective in as broad of a discipline as possible. I believe that a good way of thinking about generalization property of a model is how broad and intuitive understanding it has of the larger field of its interest domain.

I believe that VFU tasks are related to each other and building multi-task learning models can be beneficial for improving generalization across visual fashion understanding field. In order to validate my hypothesis, I performed experimental analysis that I describe in this chapter. I built two deep learning solutions designed primarily for VFU categorization tasks of category and attribute prediction. First solution uses the multi-task learning approach and shares features from localization branch, which is designed to predict landmarks, to categorization branch. Second solution is designed just for categorization tasks (only one categorization branch). I evaluate category prediction accuracy and attributes recall metrics from both solutions, and assess whether knowledge sharing between branches contributes to improved metrics and therefore generalization ability across wider visual fashion understanding domain.

3.2 Implementation details

As mentioned in the previous section, I built two deep learning based VFU solutions. First method has both localization branch (for landmark prediction) and categorization branch (for prediction of garment's category and attributes). This method will be called **MTL**, because it uses multi-task learning approach for landmark, category and attribute prediction with knowledge sharing from localization branch to categorization branch. Second solution has just categorization branch for category and attribute prediction and will be called **CAT**. As a side note, optimizing 2 loss functions for category and attribute prediction in CAT method could also be considered a multi-task learning. However, those tasks are so highly related that it would not make any sense to ever have them separated. I am interested in finding out what are results of passing knowledge between tasks, which are logically unrelated, such as categorization (described in Section 2.2) and localization (described in Section 2.3).

Both solutions were implemented in PyTorch. They were trained using DeepFashion dataset [19], which have been described thoroughly in Section 2.1.1. Each image has annotations for only one main garment in the picture. That garment's localization in a picture is annotated with a bounding box. Each garment can have only 1 out of 48 possible categories. Each garment is labeled with true/false annotation for each of 1000 possible attributes. Attributes are also assigned into 1 of 5 possible type groups: texture, fabric, shape, part, style. Garments also have landmark location labels (each landmark is a point, with x and y position) for each of 4-8 possible landmarks. Amount of landmarks depends on garment's category type. Main landmarks are: left and right collar, left and right sleeve, left and right hem, left and right waistline. Each landmark is also labeled with its visibility metric. Examples of landmark annotations can be seen in Figure 2.1 in Section 2.1.1, where I described whole DeepFashion dataset in detail. Additional operations has been applied to the dataset, such as random cropping and flipping. Before training all images have been rescaled to (224, 224) shape.

3.2.1 MTL implementation

The MTL implementation has been inspired by two novel solutions I have found in the VFU literature. First of those solutions comes from paper **Attentive Fashion Grammar Network for Fashion Landmark Detection and Clothing Category Classification** by Wang et al. [27], which has been described in Section 2.5.2. Idea of landmark-aware attention has been described in that paper, and I decided to also implement a similar concept in my solution. Second literature solution I have based my implementation on is paper **Deep Fashion Analysis with Feature Map Upsampling and Landmark-driven Attention** by Liu et al. [16], which has been described in Section 2.5.3. Ideas of passing feature representation between localization-dedicated branch and categorization-dedicated branch and feature maps upsampling have been described in that paper, and I also use those concepts in my own implementation. VGG-16 [24] CNN network has been used as a main tool for visual features extraction in both papers, and that is also the case in my solution.

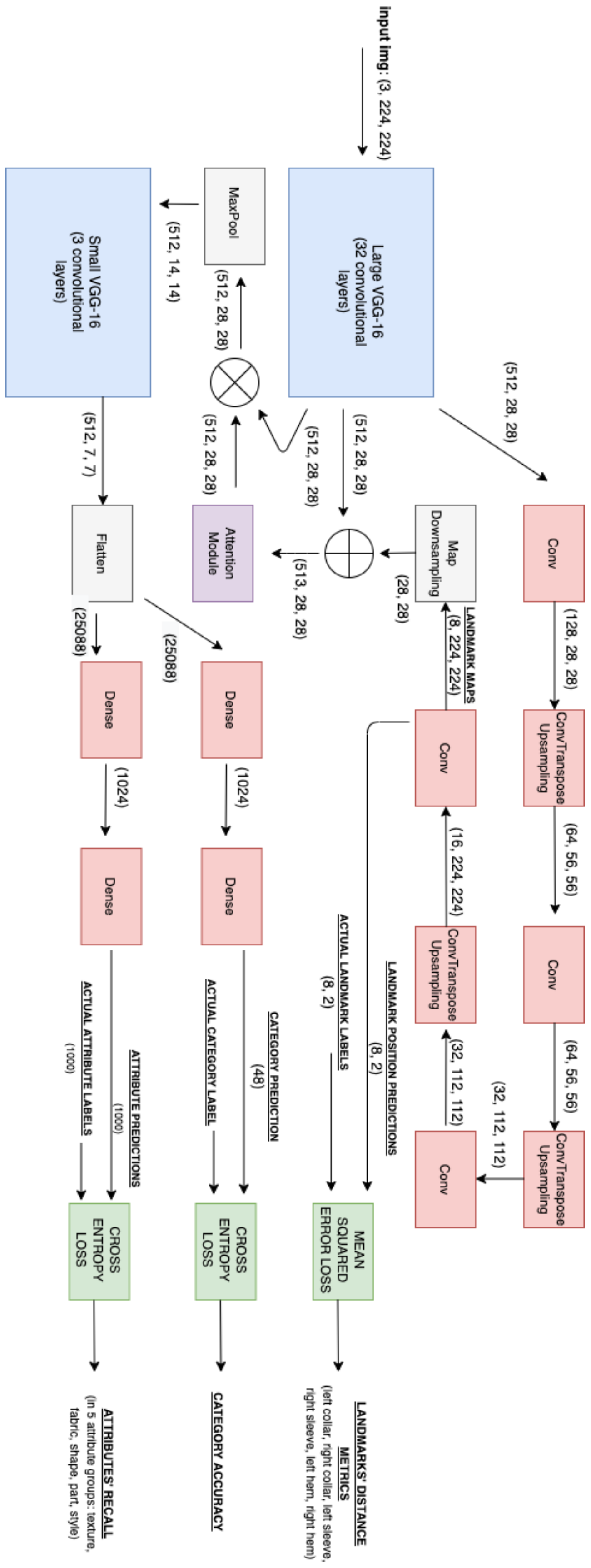


Figure 3.2: Neural architecture of MTL implementation

Neural architecture of MTL implementation can be seen in Figure 3.2. The graphic visualizes flow of data through the network in feed-forward pass. It might be helpful to follow the Figure 3.2, while I describe the network in detail in subsequent paragraphs.

Tensor dimensions for a single input image are provided at output of each module. I trained my network using a mini-batch approach with 8 images in each batch. All loss functions in this network are optimized with Adam algorithm [14] with adaptable learning rate.

Each input image is of (3, 224, 224) dimension and uses RGB color coding. It is passed through a large VGG-16 CNN network with 32 convolutional layers. Generated output is of (512, 28, 28) shape and contains aggregated visual features of the input coded into 512 feature maps. Then this representation is passed to a localization branch, which focuses on finding landmark maps (average position for each landmark), and making predictions about landmark location, based on those maps. In order to find those landmark maps, in the localization branch information is passed through a series of seven modules: four of them are convolutional layers and three of them are transposed convolutional layers. Transposed convolutional layers aim to upsample feature maps so that output maps are of the same shape as input images, and therefore can be understood in a pixel-wise manner. If a garment has all eight landmarks then a (8, 224, 224) feature map is generated. Landmark position is predicted as the most probable (x, y) pair. Then a difference between our prediction and actual landmark location is calculated by mean-squared error (MSE) loss. MSE loss calculates average squared difference between predicted and label pixel localization (landmarks' distance metrics) over both x and y axis. The MSE loss function is shown in Equation 3.1. After the feed-forward pass, error is backpropagated and modules' weights in localization branch are updated.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}((x_i - x_i')^2 + (y_i - y_i')^2) \tag{3.1}$$

Landmark maps are downsampled and channel-wise concatenated with the large VGG-16 output. This representation will be later passed down to categorization branch. The concatenation is precisely where the information sharing between localization and categorization branch happens. Concatenated tensor is passed to attention module, which uses landmark map to magnify feature values around assumed landmark positions and minify those further away from those positions. Because of the fact that landmarks are crucially important points on a garment, we assume that enhancing features around them will lead to better sharing knowledge from localization branch to categorization branch. Attention is the key mechanism that allows us to pass the knowledge more effectively. While it would still be possible to pass landmark map representations directly to the categorization branch, ability to use attention module allows network to find significant regions and that enhances network's ability to leverage localization-based knowledge in other tasks. At the same time, attention is a very powerful neural module which can have a dominant effect on algorithm's performance. Goal of my work is to investigate effect of feature sharing between tasks and using many attention modules could lead to performance improvement that cannot be attributed to feature sharing itself but rather to attention. Therefore I decided to use only one attention module, as opposed to a solution of Liu [16], which used as many as four attention modules.

In next stage, information flows into categorization branch. Output tensor from the attention module is multiplied with output of large VGG-16. Then information is max-pooled in order to reduce dimensionality. Visual features are then extracted through a new, small VGG-16 CNN with three convolutional layers. Next, multi-dimensional tensor is flattened and directed to two separate branches, each responsible for either category or attributes prediction. In both mini-branches there are two densely connected layers. First dense layer has a ReLU activation in both mini-branches, while second dense layer has softmax activation in category branch and multiple sigmoid activations in attributes branch. Afterwards category and attributes are predicted. For category output can be understood as a probability distribution over all 48 possible classes. For attributes we can observe separate 0-1 probability values for all 1000 possible attributes. Then, for both category and attribute prediction tasks, a difference is calculated between prediction and the actual label by cross-entropy (CE) loss. The CE loss function is shown in Equation 3.2. Errors are then back-propagated and weights in the branch are updated. I use accuracy as a performance metric for category prediction and recall for attributes prediction.

$$CE = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{c}(y_{ij} * log(p_{ij})) \tag{3.2}$$

MTL training results

Network has been trained over 10 epochs. After each epoch, metrics of algorithm's performance for tasks of category, attribute and landmark prediction were calculated on a testing set that had not been used in training at this particular epoch. Those metrics are shown on diagrams I present below. I decided to use top-3 category accuracy as a performance metric for category prediction, top-3 recall for attributes' prediction and landmark's average distance for landmark position prediction.

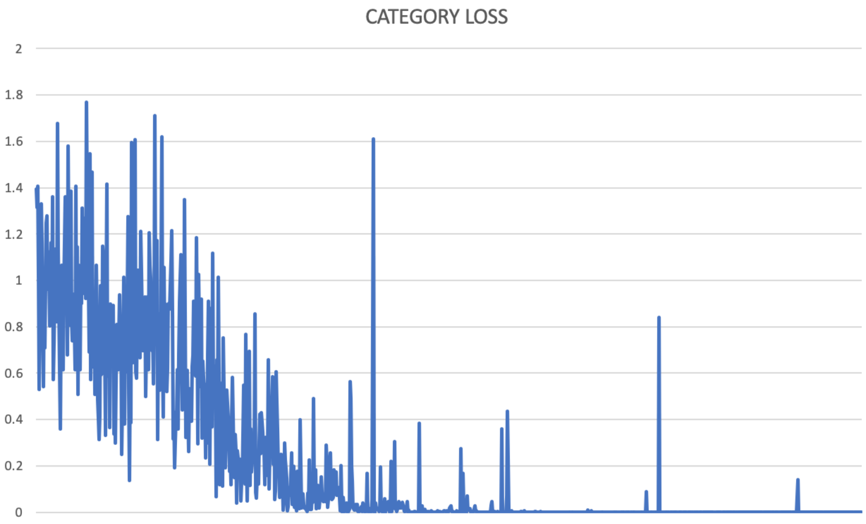


Figure 3.3: MTL's category loss over 10 epochs

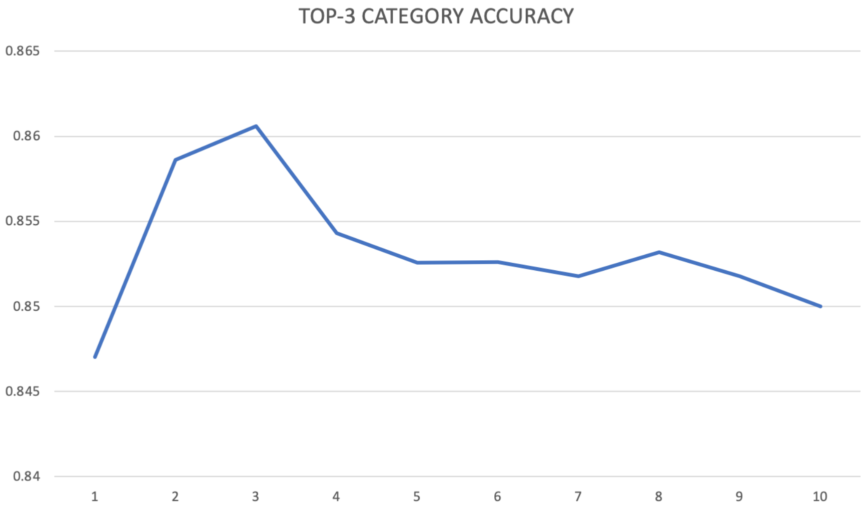


Figure 3.4: MTL's top-3 category accuracy over 10 epochs

In Figure 3.3, we can see a plot of loss for category prediction task over 10 epochs of training. Each value represents cross-entropy loss value for each mini-batch of size 8. After first 3 epochs, category loss dropped to a value of around 0. With exception of a few batches, that situation did not change much until the end of training. In Figure 3.4, we can see a plot of top-3 category accuracy over 10 epochs. During first 3 epochs, top-3 accuracy, as evaluated on a test set, was growing and reached it's maximum of **0.8605** after epoch 3. Analyzing those two plots jointly, we can conclude that only during first 3 epochs network was learning valuable general knowledge. Early stopping technique after third epoch was performed to obtain the best-possible model.

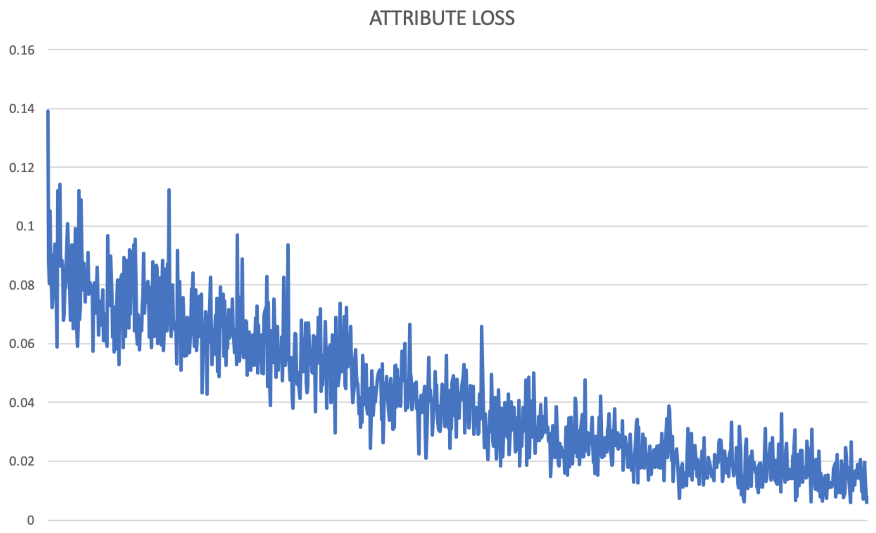


Figure 3.5: MTL's attribute loss over 10 epochs

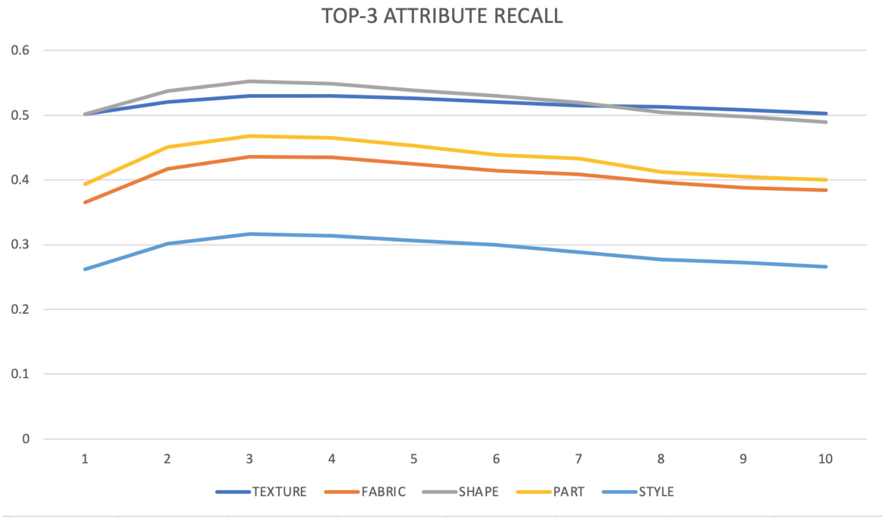


Figure 3.6: MTL's top-3 attributes' recall over 10 epochs

Table 3.1: Top-3 recall values per attribute type group

	Top-3 recall
Texture	0.5298
Fabric	0.4359
Shape	0.5528
Part	0.4677
Style	0.3169

Next, in Figure 3.5 we can see a plot of loss for attributes' prediction task over 10 epochs of training. Each value represents cross-entropy loss value for each mini-batch of size 8. Attribute loss decreases steadily over all 10 epochs. In Figure 3.6, we can see how recall metrics for different attribute groups change over 10 training epochs. Maximum recall values for all attribute groups happen after third epoch. Analyzing those two plots jointly, we can conclude that even though loss decreases throughout the whole training time, recalls on testing set starts to decrease after third epoch. That means the model starts overfitting after third epoch and model gets too closely fitted to the training data. Early stopping technique after third epoch was performed to obtain the best-possible model. Actual best recall values per attribute group can be seen in Table 3.1. As we can see, Texture and Shape are attribute type groups that MTL method was able to predict with the highest recall. Style is the attribute group with lowest average recall. Part and Fabric attribute groups have average recall values in the middle between best and worst groups.

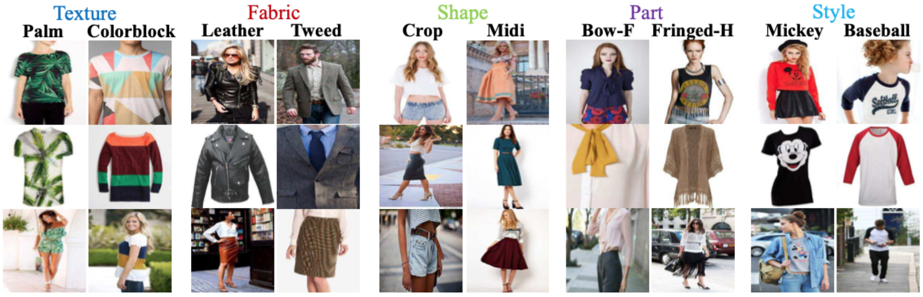


Figure 3.7: Attribute group examples [27]

In Figure 3.7 we can see image examples with attribute annotations being assigned to one of five attribute groups. As we can see in the Figure, Style attribute group is the most ambiguous of all groups and it is relatively hard for the attention module to find regions of the image which are responsible for being labeled with particular attribute, e.g. 'baseball' attribute in Style group in Figure 3.7. On the other hand, Shape attribute group has certain distinct visual appearance traits that make it easy for attention module to find regions which contribute to image having attribute from this group, e.g. 'crop' attribute in Shape group in Figure 3.7.

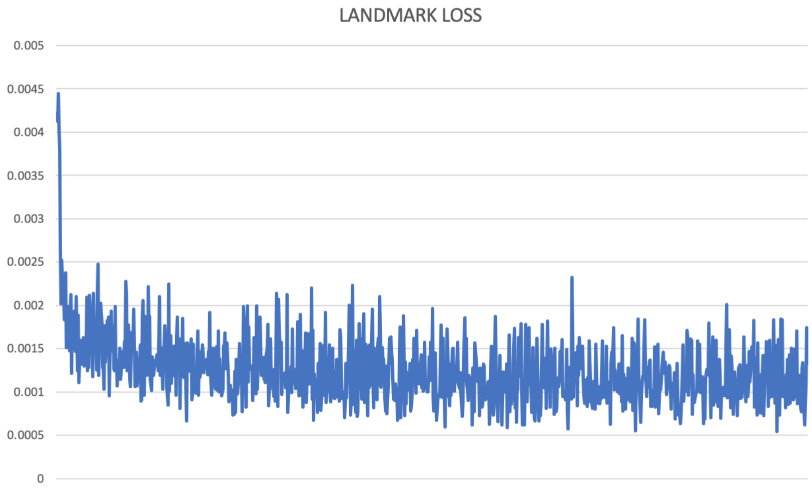


Figure 3.8: Landmark loss over 10 epochs

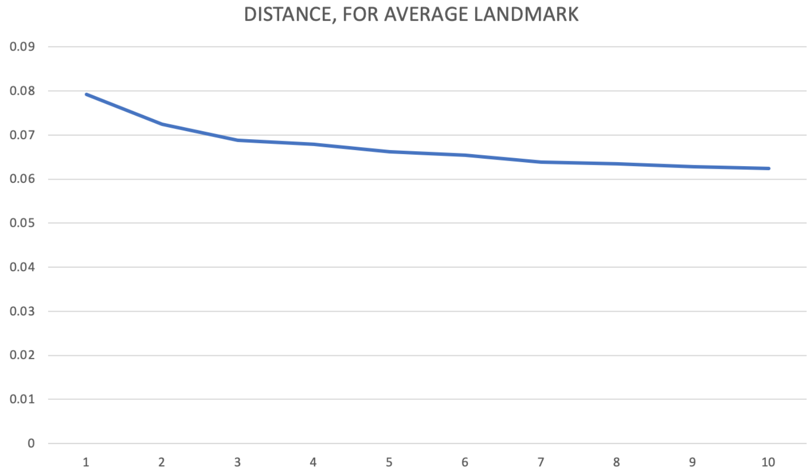


Figure 3.9: Landmark prediction performance over 10 epochs

Lastly, let's look at plots for landmarks' position prediction task in Figure 3.8 and in Figure 3.9, with former plot visualizing loss value for the task over 10 epochs and latter plot showing performance on test set for that task, as measured by average distance between prediction and actual label (lower value means better prediction). While analyzing results of that task is not essential from the perspective of my research angle, as I only try to validate whether landmarks-related features might contribute to categorization tasks described earlier, it is helpful to also see how network performs for landmark position prediction task. As we can see in the Figure 3.8, loss decreases more sharply during first two epochs, then decreases very slowly through the next six epochs and completely plateaus in the last two epochs. Average distance on plot in Figure 3.9 drops strongly during first three epochs and then establishes a slight decrease trend over the rest of time.

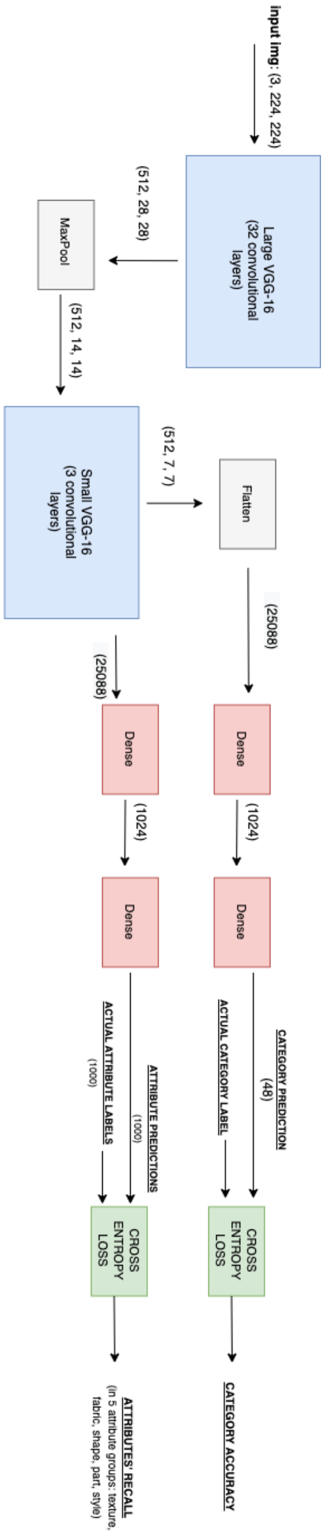


Figure 3.10: Neural architecture of CAT implementation

3.2.2 CAT implementation

The difference between CAT and MTL solutions is that in CAT there is only categorization branch, while localization branch and attention module are absent. Structure of remaining modules is exactly the same, therefore information flow in the network is analogues to the one described in the previous section. Neural architecture of CAT implementation can be seen in Figure 3.10. As we can see, output from large VGG-16 network goes directly to max-pooling module and then to small VGG-16. Same loss functions and performance metrics are used for categorization tasks, which allows for easy comparison of results. Network has been trained over 10 epochs. Because of the fact that this method was implemented primarily for comparison with the MTL method, I am not providing separate training results in this section.

3.3 Results

In this section I describe results I was able to obtain and how I evaluated my research questions, based on two implementations: MTL and CAT, which were described in previous sections. As a reminder, my research angle focuses on finding out whether using a multi-task learning approach and sharing knowledge between seemingly unrelated tasks improves generalization properties of VFU models, as understood by having a broader understanding of the VFU domain. In my particular case, I measure that generalization ability through evaluation of performance metrics (accuracy for category prediction and recall for attribute prediction) for categorization tasks in two solutions: MTL which uses a multi-task approach by passing feature representation from landmark-designed branch to categorization branch, and CAT which is designed only for categorization. If accuracy/recall improves with sharing knowledge between tasks, it would mean that being able to predict landmarks also helps with predicting category and therefore model has broader understanding of the VFU domain.

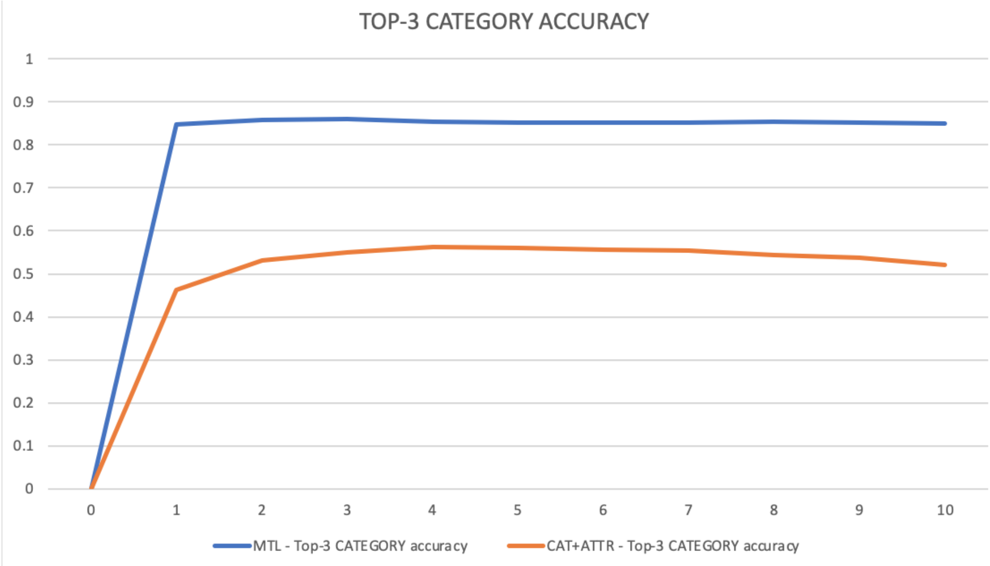


Figure 3.11: Top-3 category accuracy of MTL and CAT methods over 10 epochs

Figure 3.11 depicts top-3 accuracy for category prediction for both built methods (CAT method is marked as "CAT+ATTR" in the legend). Multi-task learning solution MTL with localization-to-categorization knowledge sharing obtains best accuracy of **0.8605** after epoch 3, while CAT solution with no knowledge sharing between tasks obtains best accuracy of **0.5621** after epoch 4.

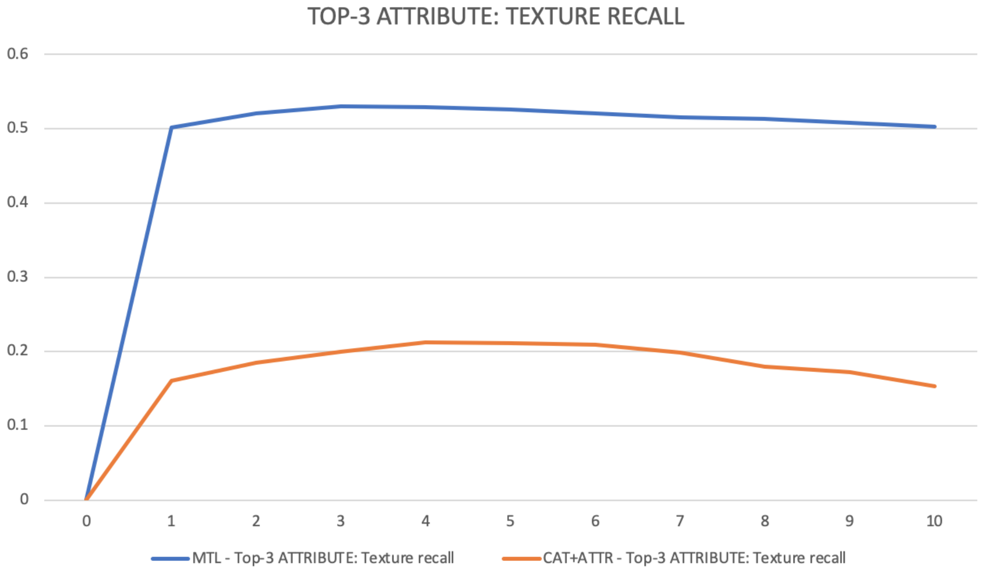


Figure 3.12: Top-3 Texture attribute recall of MTL and CAT methods over 10 epochs

Figure 3.12 depicts top-3 recall for attribute prediction of Texture group for both built methods (CAT method is marked as "CAT+ATTR" in the legend). MTL solution obtains best recall of **0.5287** after epoch 3, while CAT solution obtains best recall of

0.2124 after epoch 4. I chose Texture attribute group for comparison purposes, but plots for other groups are very similar.

Looking at best results of both models for categorization tasks, MTL solution with knowledge sharing obtained **higher accuracy by 53%** and **higher attribute recall by 149%**, when compared to CAT solution without knowledge sharing. Therefore I conclude that addition of knowledge passing from localization task to categorization tasks drastically improves performance on main tasks of category and attribute prediction. While in my implementation that knowledge sharing happens in a form of passing landmark heatmaps, the above conclusion is valid for message passing in other forms too. Performed experimentation confirms my hypothesis and indeed, **multi-task learning approach of feature sharing between tasks contributes to better generalization of VFU models**. Even though that localization task of landmark position prediction and categorization tasks of garment's category and attribute prediction are seemingly unrelated, my experiment confirmed that in fact these tasks are related and have common areas. Visual fashion understanding is a domain where tasks influence each other and those tasks should not be treated separately. VFU models gain tremendous advantages thanks to using large multi-task learning models, compared to separate, smaller single-task-dedicated models. Majority of recent state-of-the-art solutions in VFU research literature move towards multi-task learning (those solutions were described in Section 2.5) and my experiment also confirmed that it is crucial to share features between tasks, if we are aiming for the most generalizable deep learning VFU models, with as broad understanding of the domain as possible.

3.4 Successes, challenges and possible continuation steps

Main success of my research is that through experimentation with two separate implementations, I was able to confirm a more general hypothesis that multi-task learning contributes to a broader understanding of VFU domain by deep learning models.

Main challenge is related to the fact that it is not entirely possible to isolate just the issue of knowledge passing and assume that all performance improvement can be attributed to that. While inclusion of attention module and additional convolutional layers in MTL solution might be other factors that contribute to improved performance metrics, I made sure to minimize such contributions by making those mentioned modules as small as possible.

While results of my research are satisfactory as they led to confirmation of my initial hypothesis, there are certain next steps that should be taken to improve my solution. I believe that improvement areas described below might contribute to even more generalizable VFU models.

In the previous section I stated that passing landmarks' heatmap representation from localization branch to categorization branch improves results for categorization tasks. Drawing on those conclusions, I believe that passing features in the opposite direction might lead to better performance on localization task. While it is more intuitive to understand that focusing on regions around landmarks might help with attribute prediction, it is also possible that knowing what kind of attribute a garment in the image has, might contribute to better landmark prediction. Therefore passing representation in the opposite direction could be the next thing I add to my MTL implementation.

Another possible addition to the implemented method could be sharing knowledge between even more tasks. If localization-related features help with category prediction, then it's possible that adding branches related to retrieval or segmentation, and passing knowledge between all of them, could return even better results for all tasks.

Deep learning models cannot generalize outside of the training data distribution. Therefore, if we are aiming for our model to be as close to human-level understanding of the domain as possible, then we need to train it on highly-granular data. While DeepFashion [19] dataset, that I used, is a great starting point for my analysis, it has certain drawbacks, such as only one garment annotation per image or relatively low amount of landmarks (4-8). Next step could be training my model on DeepFashion2 [8] dataset, which is much more fine-grained than DeepFashion, contains multiple garments per image and as much as 23 landmarks per average clothing category.

Last aspect that I could add to my implementation is hierarchical modeling of categorization labels, similar to the method described in Section 2.2.1.

3.5 Performance comparison to other VFU solutions

Beyond answering questions related to my research angle, it is also interesting to see how the results I obtained compare to other state-of-the-art methods in the VFU field. In Table 3.2, we can see prediction metrics of my solution for top-3 category accuracy, as well as for other VFU methods I described in Chapter 2. All algorithms were evaluated on the same DeepFashion's [19] Category and attribute prediction benchmark, which was described thoroughly in Section 2.1.2.

Table 3.2: Comparison of my MTL and CAT implementations to other solutions, tested on DeepFashion's Attribute and Category benchmark [19]

	Top-3 category accuracy
My MTL solution	0.8605
My CAT solution	0.562
Liu et al. [16] (original paper), Section 2.5.3	0.9116
Wang et al. [27], Section 2.3.2	0.9099
Luo et al. [19], Section 2.5.1	0.8258
Huang et al. [12], Section 2.5.4	0.5948

As we can see, MTL implementation with top-3 accuracy of 0.8605 is only worse than solutions by Wang [27] (Section 2.3.2) and by Liu et al. [16] (Section 2.5.3), which is the original paper that inspired localization-to-categorization message passing in my solution. A possible reason for Liu et al. obtaining better results than me is having more attention modules stacked, which could lead to even more efficient leveraging of localization-related knowledge. However, I was able to obtain better results than landmark pooling-based solution by Luo et al. [19] (Section 2.5.1), which also used multi-task learning. A possible reason for that is the fact that I used landmark heatmap upsampling, which allows for retaining high resolution of maps, while landmark pooling from Luo's solution [19] downsamples maps, which leads to them being of lower resolution.

My CAT implementation with top-3 accuracy of 0.562 performs worse than any other analyzed method. It clearly shows that simple CNN-based solution is not enough to compete with novel state-of-the-art methods, which use multi-task learning approach.

4. Conclusions

Advancements in neural networks are the main reason of rapid progress in computer vision in recent times. Visual fashion understanding is a field where usage of deep learning-based computer vision algorithms is natural, as nearly all VFU-related data is in form of images. Understanding clothing garments in pictures is the fundamental problem in all of VFU tasks, such as garment category prediction, attribute prediction, landmark localization, clothes detection and segmentation or street-to-shop retrieval. In order to help the reader better understand those tasks, I made a survey of the field and described most important solutions in Chapter 2.

Based on the observation that all VFU tasks are based around visually understanding garments, I came to the conclusion that those tasks might be in fact related. I presented a hypothesis that building larger multi-task learning models dedicated to predicting multiple tasks at once might lead to better generalization of VFU models. In order to assess validity of my hypothesis, I implemented two deep learning solutions dedicated primarily to category and attribute prediction. First solution used multi-task learning concept of sharing features from additional branch dedicated to localization task of landmarks' position prediction. Second solution did not have that concept implemented, but all the remaining modules stayed the same. Comparison of those two implementations confirmed my hypothesis, as sharing knowledge between tasks increased category prediction accuracy by 53% and attributes prediction recall by 149%. Next steps for developing this solution further are incorporating even more task-dedicated branches into the network and sharing features in more directions between those branches.

After having surveyed scientific literature and having conducted my own experiments, I believe that multi-task learning improves generalization properties of deep learning-based visual fashion understanding models across tasks.

Bibliography

[1] An overview of multi-task learning in deep neural networks. https://ruder.io/multi-task/. Accessed: 2020-10-02.

[2] K. E. Ak, A. A. Kassim, J. Hwee Lim, and J. Yew Tham. Learning attribute representations with localization for flexible fashion search. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[3] E. Bell. Lyst: Working with fashion models; pydata london 2016, 2016. https://www.youtube.com/watch?v=emr2qaCQOQs.

[4] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *CoRR*, abs/1606.00915, 2016.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, 2005.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[7] B. Q. Ferreira, L. BaÃa, J. Faria, and R. G. Sousa. A unified model with structured output for fashion images classification, 2018.

[8] Y. Ge, R. Zhang, L. Wu, X. Wang, X. Tang, and P. Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images, 2019.

[9] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. The MIT Press, 2016.

[10] S. Guo, W. Huang, X. Zhang, P. Srikhanta, Y. Cui, Y. Li, M. R. Scott, H. Adam, and S. Belongie. The imaterialist fashion attribute dataset, 2019.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[12] J. Huang, R. Feris, Q. Chen, and S. Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1062–1070, Dec 2015.

[13] M. H. Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg. Where to buy it: Matching street clothing photos in online shops. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3343–3351, Dec 2015.

[14] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

[15] X. Liang, C. Xu, X. Shen, J. Yang, J. Tang, L. Lin, and S. Yan. Human parsing with contextualized convolutional neural network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(1):115–127, Jan 2017.

[16] J. Liu and H. Lu. *Deep Fashion Analysis with Feature Map Upsampling and Landmark-Driven Attention: Munich, Germany, September 8-14, 2018, Proceedings, Part III*, pages 30–36. 01 2019.

[17] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3330–3337, June 2012.

[18] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg. SSD: single shot multibox detector. *CoRR*, abs/1512.02325, 2015.

[19] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1096–1104, June 2016.

[20] Z. Liu, S. Yan, P. Luo, X. Wang, and X. Tang. Fashion landmark detection in the wild, 2016.

[21] D. G. Lowe. Distinctive image features from scale-invariant keypoints, 2004. https://www.cs.ubc.ca/~lowe/papers/ijcv04.pdf.

[22] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015.

[23] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015.

[24] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*, 09 2014.

[25] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*, 09 2014.

[26] I. Tautkute, T. Trzcinski, A. Skorupa, L. Brocki, and K. Marasek. Deepstyle: Multimodal search engine for fashion and interior design. *CoRR*, abs/1801.03002, 2018.

[27] W. Wang, Y. Xu, J. Shen, and S. Zhu. Attentive fashion grammar network for fashion landmark detection and clothing category classification. 05 2018.

[28] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang. Learning from massive noisy labeled data for image classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[29] K. Yamaguchi, M. H. Kiapour, and T. L. Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In *2013 IEEE International Conference on Computer Vision*, pages 3519–3526, Dec 2013.

[30] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *CoRR*, abs/1611.03530, 2016.

[31] B. Zhao, J. Feng, X. Wu, and S. Yan. Memory-augmented attribute manipulation networks for interactive fashion search. 07 2017.

[32] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr. Conditional random fields as recurrent neural networks. *CoRR*, abs/1502.03240, 2015.

[33] S. Zheng, F. Yang, M. H. Kiapour, and R. Piramuthu. Modanet: A large-scale street fashion dataset with polygon annotations, 2018.