Tackling Attribute Fine-grainedness in Cross-modal Fashion Search with Multi-level Features

Kenneth Goei
University of Amsterdam
Amsterdam, The Netherlands
kengoei@live.nl

Mariya Hendriksen
University of Amsterdam
Amsterdam, The Netherlands
m.hendriksen@uva.nl

Maarten de Rijke
University of Amsterdam &
Ahold Delhaize
Amsterdam, The Netherlands
m.derijke@uva.nl

ABSTRACT

Leveraging information across modalities can facilitate customers throughout their journey, especially in the fashion domain where the visual modality plays an important role. Fashion products have a variety of visual groups of attributes such as shapes, colors, patterns, etc. Every category is fine-grained, i.e., attributes within a category may be visually very similar, e.g., v-neck vs. round-neck. The fine-grainedness of fashion attributes makes cross-modal fashion retrieval more challenging. In this paper, we address the problem of attribute fine-grainedness in fashion cross-modal retrieval by leveraging multi-level feature representations. In particular, we replace the commonly used spatial segmentation approach with a multi-level feature approach. We compare our approach with state-of-the-art models in general and fashion cross-modal retrieval and evaluate it on the Fashion200K and Fashion-Gen datasets. We record a 43.4% relative increase in text-to-image retrieval and a 57.8% relative increase in image-to-text retrieval on the Fashion200K dataset and a 48.6% relative increase in text-to-image retrieval and a 67.2% relative increase in image-to-text retrieval on the Fashion-Gen dataset while reducing the number of model parameters by 70% when compared with the baselines.

ACM Reference Format:
Kenneth Goei, Mariya Hendriksen, and Maarten de Rijke. 2021. Tackling Attribute Fine-grainedness in Cross-modal Fashion Search with Multi-level Features. In *Proceedings of the 2021 SIGIR Workshop on eCommerce (SIGIR eCom'20), July 15, 2021, Virtual Event.* ACM, New York, NY, USA, 8 pages.

1 INTRODUCTION

Fashion represents a significant part of today's e-commerce [8]. Many algorithms within fashion search let the user type a textual query and retrieve the best matches. The algorithms generally take into account various types of features and input data. However, in many cases, the search engines do not leverage the rich visual information available in the product image. This solution creates a limited search experience for consumers trying to search for their desired fashion items. Hence, leveraging visual and textual information in a cross-modal fashion e-commerce platform can help better grasp the consumer's needs.

Cross-modal fashion search. In this paper, we focus on the task of cross-modal fashion search. To perform cross-modal retrieval, the model has to learn a mapping between image and text to create a shared embedding space where similar concepts from different modalities are close. In the general domain, extensive research has been performed within image-text retrieval [10, 14, 17, 18, 22, 26, 31].

Challenges. Despite the numerous advances in the general cross-modal retrieval domain, cross-modal fashion search still has many challenges [16, 33]. The first challenge is related to the fine-grainedness of fashion attributes, i.e., the difference between attributes can be quite subtle, e.g., puffed sleeve vs. butterfly sleeve. The level of granularity makes it hard to create feature representations that can represent these differences well and hence impede the cross-modal retrieval process. Furthermore, because of the fine-grainedness, it is hard to ground fashion attributes to the corresponding spatial location within an image. Ground truth for the attribute annotations is generally unavailable for e-commerce, so we only know which attributes are expected to be shown in the image but not where.

Multi-level feature approach (MLF). A common approach to address attribute fine-grainedness is *spatial segmentation*, i.e., dividing the image into spatial segments to extract image features. Examples of this approach are 64-tile segmentation [11] and 6-rule segmentation [21]. However, we argue that the spatial segmentation approach is sub-optimal due to image composition variability and the high detail level of fashion attributes. Hence, we propose to replace the spatial segmentation approach by creating features from different hidden layers within a convolutional neural network (CNN). We refer to this approach as the *multi-level feature approach* (MLF). We demonstrate that MLF allows us to create more discriminative features. Besides, we explore how the MLF approach works with different feature extractors from both the general and fashion domain, and both in end-to-end and general scenarios. We compare MLF with a spatial segmentation approach and evaluate it on the Fashion200K [12] and Fashion-Gen [30] datasets.

Additionally, we aim to improve model performance while using models with fewer parameters. This allows us to train lighter, faster, and smaller models, which is important for reproducibility, especially given limited computational resources.

In this work, we aim to answer the following research questions:

(RQ1) How can we improve upon the current state-of-the-art in cross-modal fashion search? Is it possible to perform better with fewer parameters? More specifically, what kind of architecture do we need? What is the influence of the number of parameters on the performance of the models?

(RQ2) How can we improve the image representation so that fine-

SIGIR eCom'21, July 15, 2021, Virtual Event, Montreal, Canada

grained fashion attributes are better represented? In particular, how can we use training to improve the image features? What kind of image and text encoder do we need?

The principal contributions of our research are the following:

- We achieve state-of-the-art results with a 43.4% relative increase in text-to-image retrieval and a 57.8% relative increase in image-to-text retrieval on the Fashion200K dataset and a 48.6% relative increase in text-to-image retrieval and a 67.2% relative increase in image-to-text retrieval on the Fashion-Gen dataset, while reducing the number of model parameters by 70% when compared with the baselines.
- We show that using a multi-level feature approach instead of a spatial segmentation approach allows us to create more discriminative image features for fine-grained cross-modal search as measured by cosine similarity.
- We research the requirements for successful use of the proposed multi-level feature approach and find the following:
  - Overall, the MLF approach results in a relative increase in cross-modal retrieval performance. The degree of improvement depends on network the architecture and depth.
  - Surprisingly, the use of models specifically fine-tuned on fashion data does not contribute to better retrieval performance.
  - The MLF approach in an end-to-end set-up yields a relative improvement of 18%–38% retrieval performance compared to the same model without end-to-end training.

2 RELATED WORK

Cross-modal retrieval. Early approaches to image-text mapping have focused on correlation maximization through kernelized canonical correlation analysis (kCCA) [14, 15, 31]. The problem with these approaches is that they do not scale well due to the costly kernel computation. The use of a neural approach [10, 17, 18, 22] within image-text matching became popular with the advancement of CNNs and recurrent neural networks (RNNs). Another related line of work concerns creating a universal vision-language encoder [5, 23, 24, 26]. These models are inspired by cross-language models, like BERT in the natural language processing (NLP) domain. These models achieve SOTA results on a variety of multi-modal tasks such as cross-modal retrieval and visual question answering. The downside of these models is that reproducibility of the results is difficult because of the big datasets the models are trained on and high computational costs.

Specific image-text models for the fashion domain have been researched on several occasions. Often the models from the general domain are used and tweaked to perform in this more fine-grained setting. In [32], authors have used CCA to perform cross-modal search for dresses in the fashion domain, closely resembling the work of Hodosh et al. [14] in the general domain. Improving on this work, Laenen et al. [21] used an approach very similar to Karpathy and Fei-Fei [17], embedding the images with a CNN and the sentences with a skip-gram model. The method of Laenen et al. [21] is different from Karpathy and Fei-Fei [17] in the treatment of images. Laenen et al. [21] uses the symmetry of dresses to segment the image in six regions with hand-made rules. These spatial segmentations are used because object detectors do not work in the fashion domain due to the fine-grainedness of the fashion attributes. These images are segmented in a way that the segments focus on regions that contain specific fashion attributes, for example, the neck and arms. In this way, the authors hope that the fashion attributes can be mapped to certain image segments. Another line of work suggests that attribute fine-grainedness can be tackled by using attention mechanisms, as attention is also often used to bring fine-grained attributes to the forefront in item representations [20].

Recently, the FashionBERT model has been proposed [11]. Inspired by vision-language encoders, the authors fine-tune BERT using fashion images and descriptions in combination with an adaptive loss for cross-modal search. The FashionBERT model tackles the problem of fine-grainedness similar to Laenen et al. [21], by taking a spatial approach. The image is uniformly segmented in 64 tiles. These tiles are used as "word tokens" within the FashionBERT model. The authors claim that by segmenting image in 64 tiles and feeding the tiles to the model, allows the model to learn to focus on the small details of the images, thereby addressing the fine-grainedness problem. Since the results of FashionBERT are state-of-the-art, we are using this model as one of the baselines in our work.

Most previous work in the fashion domain is focused on models that are relatively large. Unlike previous work in this domain, we aim to improve upon SOTA by using models with fewer parameters.

Combining features across multiple levels. Since the beginning of deep neural networks, research has been performed to understand the inner workings of CNN. Early work tried to visualize activations in different hidden layers to understand what happens inside a CNN. In this early work, they were able to show that early layers are activated by global structures, whereas later layers are activated by more detailed structures [1, 36, 38]. This raised the question of whether these hidden layers could be used for different tasks. Multiple works showed that different nodes within a network encode different semantic features for image classification [2, 9, 29], edge detection [35], cross-domain matching [37], and image segmentation [13]. Vittayakorn et al. [34] explored features from multiple levels in the fashion domain. In particular, the authors explored the correlation between different image attributes and neural activations across hidden layers of CNN. The authors performed experiments that suggest that global fashion attributes correspond to early layers in a CNN and more detailed attributes to later layers. This shows that the use of features from hidden layers might be interesting to use to create features that represent these fine-grained fashion attributes better.

Unlike previous work in this domain, our work focuses on leveraging multi-level features for constructing image representations for the cross-modal fashion retrieval task. We further explore how our approach helps to learn fine-grained fashion attributes.

3 EXPERIMENTS

MLF pipeline. The multi-level features approach that we propose is inspired by a line of work related to Vittayakorn et al. [34] which suggests a correlation between visual semantic attributes and features from different layers within a CNN. We propose an approach that will leverage information across multiple layers of a CNN instead of only using the features from the last layer. The intuition is that information from textual and visual modalities can be aligned better since the multi-level features contain more semantic meaning than the features obtained using the spatial segmentation approach.

Figure 1 gives a visual overview of our MLF pipeline. Our approach comprises three major components: (1) an image encoder for generating multi-level features, (2) a text encoder for generating a text representation, and (3) a text-image matching model that learns a shared multi-modal space. To obtain multi-level image representation, we sample features from evenly spaced layers of the image encoder. For 3D layers, we apply 2D average pooling. After pooling, we flatten the features to create 1D-features. We pad the features with zeros to create features of equal shape wherever required. As a text-image matching model, we use SCAN [22] because it achieves high recall scores while being a relatively small model in the number of parameters.

Datasets. We run the experiments using Fashion200K and Fashion-Gen datasets.

Fashion200K. This dataset comprises around 200,000 images collected from webshops in five different categories: dresses, pants, tops, jackets, and skirts [12]. Every item has one or multiple images taken from different angles. Textual descriptions are filtered for words with a low frequency.

Fashion-Gen. This dataset contains 67,666 unique fashion items, every item has one to six images from different angles [30]. This results in a dataset with 293,000 image-text pairs.

3.1 Experiment 1: Logistic Regression

Setup. Firstly, we evaluate the discriminative power of MLF vs. spatial segmentation for detection of fine-grained attributes. For the experiment, we sample pairs of fashion attributes. Both attributes in every pair belong to the same category, and therefore are interchangeable. We collect all items within the dataset that contain one of the two attributes in their respective description. For every pair of attributes, we create a dataset with half of the instances containing one attribute whereas the other half of the dataset contains the second attribute. We create two independent feature sets. The first feature set is created using the spatial segmentation approach [21] and the second feature set is created using our MLF approach. In both cases, we use AlexNet as an image encoder. Afterward, we feed the obtained features into a logistic regression model and train and evaluate it on the task of attribute prediction. We perform this experiment on one group of clothing items since it creates more uniformity between the different segments. We choose the category of dresses from the Fashion200K dataset [12] since it has the most data available. We perform every run five times.

Results. The results of the logistic regression experiment can be seen in Table 1. From the F_1-scores we can see that more global attributes, like color, have higher scores than more detailed attributes. This can be explained by the prominence of color within a picture. Therefore, the features will be better in representing color. The results also show that more detailed attributes, like (silk vs. crepe) or (maxi vs. midi), are harder to discriminate. This shows that dividing the image into spatial segments is not an optimal solution because in this approach discriminative power varies per attribute. Besides, we note that the maximum scoring segment is never constant across five runs. This suggests that the model might not always look at the "real" attribute to classify the two attributes because the attribute is too fine-grained. Therefore, the model attempts to find a discriminative feature that has nothing to do with an actual attribute but can separate the two classes correctly. These results can be explained by the fact that AlexNet is trained on different data and on the task of object detection which is different from our task. Furthermore, the F_1-scores for the multi-level features are generally higher when compared to scores obtained with the spatial segmentation approach.

Overall, the MLF approach in this relatively simple setup showed an improvement over the spatial segmentation approach. Therefore, we will explore the use of the MLF approach further to see if we can use it for cross-modal fashion retrieval and under which conditions the approach works.

3.2 Experiment 2: MLF with Different Image and Text Encoders

Setup. In the second experiment, we want to investigate the relationship between MLF architecture and the resulting performance. More specificially, we explore how different combinations of image and text encoders impacts MLF performance.

3.2.1 *Image encoders.*

Differences in training techniques and networks have an influence on the interpretability of different hidden layers within a network [2]. Therefore, we experiment with a variety of SOTA image encoders trained in a supervised manner, such as AlexNet [19], ResNeSt-50 [39]. Besides, we experiment with simCLR [4], a model trained in an unsupervised way. Additionally, we experiment with MMFashion, a model trained to detect fashion attributes [25]. We incorporate the above-mentioned models in the MLF pipeline in two ways. First, we use them as out-of-the-box image encoders. Second, we fine-tune them on the fashion dataset.

3.2.2 *Text encoders.*

For text representation, we experiment with BERT [7], gated recurrent unit (GRU) [6] and continuous bag of words (CBOW) [28]. We use pre-trained BERT because of its great performance on a range of NLP tasks. Besides, we experiment with a GRU because it has a relatively small number of parameters while achieving good performance on a variety of tasks. Moreover, we experiment with CBOW because the model can be easily trained on fashion domain data. To train CBOW we create a fashion corpus that consists of unique text descriptions of the Fashion200K and Fashion-Gen datasets. The complete corpus comprises 1.3 million sentences or 5.6 million words.

We train all the models till the convergence of summation of recall scores and run the experiment three times using a different seed. We report the average recall scores.

Results. The results of this experiment are presented in Table 2. Overall, there is a significant difference in scores between the combinations of image and text encoders for both datasets. When we look at the image encoders, the ResNeSt-50 model is achieving the highest score, followed by MMFashion and AlexNet. The top score of ResNeSt-50 is not surprising since the model achieves state-of-the-art results on a range of Computer Vision tasks. However, the low scores of simCLR are somewhat surprising since the model

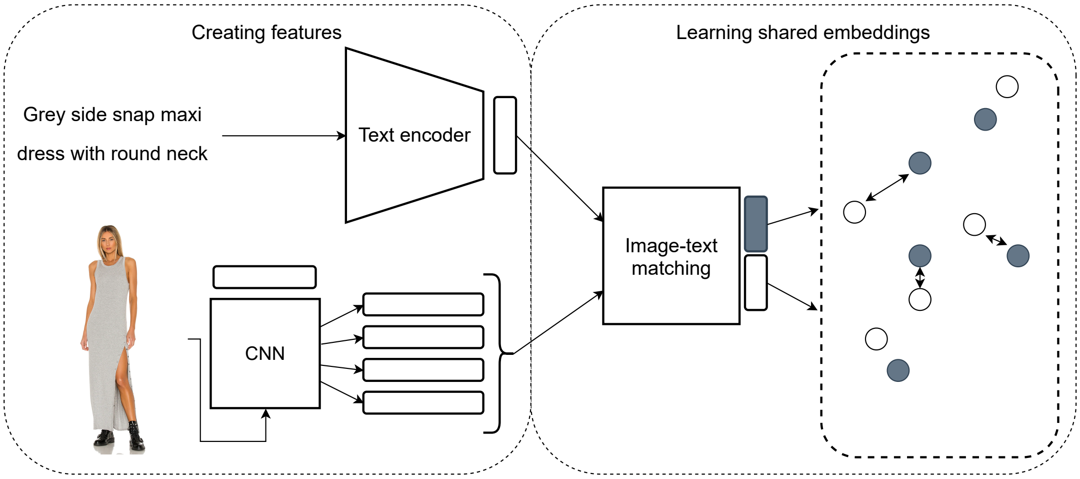

Figure 1: Overview of the MLF pipeline. We extract image and text features and feed these features to text-image matching model, where a shared embedding space is learnt

Table 1: Results of experiment 1: logistic regression

Attributes	# samples	Spatial segmentation			Multi-level features		
		F_1	Max F_1 feat.	Most freq. feat.	F_1	Max F_1 feat.	Most freq. feat.
black vs. white	2080	0.93	[4, 1, 1, 2, 2]	1,2	0.94	[0, 2, 6, 3, 0]	0
black vs. blue	5000	0.88	[1, 0, 1, 1, 6]	1	0.91	[1, 2, 1, 1, 3]	1
red vs. orange	234	0.87	[0, 6, 1, 5, 0]	0	0.87	[1, 0, 2, 0, 0]	0
yellow vs. black	246	0.98	[4, 4, 0, 6, 5]	4	0.94	[0, 6, 2, 0, 0]	0
multicolor vs. floral	1426	0.73	[4, 6, 3, 2, 4]	4	0.81	[2, 2, 3, 1, 3]	2,3
lace vs. jersey	862	0.81	[0, 6, 6, 2, 2]	2,6	0.84	[6, 4, 5, 6, 6]	6
silk vs. crepe	924	0.71	[5, 0, 5, 4, 6]	5	0.72	[3, 1, 2, 3, 1]	1,3
maxi vs. midi	2104	0.77	[1, 6, 5, 2, 2]	2	0.79	[5, 5, 5, 5, 5]	5
long vs. knee-length	880	0.87	[6, 6, 0, 1, 1]	1,6	0.87	[5, 1, 5, 2, 2]	2,5
embroidered vs. beaded	814	0.79	[4, 4, 5, 1, 2]	4	0.82	[2, 5, 2, 6, 0]	2

scores are comparable to those of top-performing networks on classification tasks. The reason for the simCLR performance could be related to the fact that the model is trained in unsupervised manner. MMFashion lags behind ResNeSt-50. It is interesting because MMFashion was specifically trained to detect fashion attributes and because ResNeSt-50 and MMFashion have comparable model architecture. The reason why ResNeSt-50 is performing better than AlexNet could be related to the depth of the network. Fine-tuning models on fashion data does not seem to improve scores, for all the models trained in supervised way. We only observe an increase in performance with simCLR, which is the model trained in unsupervised manner. Hence, we could conclude that fine-tuning models on fashion data does not necessarily create an extra advantage while learning the multimodal embedding space.

When we look at the text encoders, we can see that the score of the CBOW embeddings is the lowest for all the image encoders while the difference in performance between the GRU and BERT-embeddings differs per image encoder. The comparable performance when using the GRU and BERT-embeddings can be explained by the fact that the problem in the fashion domain is related to the image features and not the text features.

Overall we can conclude that the choice of image and text en-

Table 2: Results of experiment 2: MLF with different image and text encoders, r-sum

Model	GRU	CBOW	BERT
	Fashion 200K		
AlexNet	247.7	144.2	230.3
AlexNet (fine-tuned)	151.8	89.8	133.2
simCLR	141.8	94.3	136.7
simCLR (fine-tuned)	179.9	123.5	179.7
ResNeSt-50	**274.1**	**218.8**	**279.1**
ResNeSt-50 (fine-tuned)	239.1	172.8	254.9
MMFashion	233.3	216.1	266.9
	Fashion-Gen		
AlexNet	268.0	221.7	253.2
AlexNet (fine-tuned)	141.7	98.0	175.0
simCLR	205.2	169.8	202.1
simCLR (fine-tuned)	210.4	193.7	225.6
ResNeSt-50	**348.9**	**274.5**	**348.6**
ResNeSt-50 (fine-tuned)	261.5	230.9	276.5
MMFashion	323.1	269.4	326.9

coders is affecting model performance. We conclude that the use of a ResNeSt-50 with a GRU is the best combination since the number of parameters for the GRU is much lower than BERT while the performance is comparable. Therefore, we use this combination of image and text encoder for further experiments on the Fashion200K dataset and Fashion-Gen dataset. Additionally, we further experiment with AlexNet image encoder since this model is used in many baselines as an image encoder and thus allow for a more fair comparison.

3.3 Experiment 3: MLF and End-to-end Training

Setup. In the third experiment, we investigate the relationship between MLF performance and end-to-end (E2E) training. End-to-end training in this context implies that we do not only train the text-image matching model but the image encoders as well. We refer to this model architecture as *MLF-E2E*. As discussed in Section 3.2, we only perform end-to-end training with the best-performing combination of image and text encoders and AlexNet. We train all the models till the convergence of summation of recall scores and run the experiment three times using a different seed. We report the average recall scores.

Results. Table 3 demonstrates that end-to-end training (MLF-E2E) improves the scores compared to the models using no end-to-end training (MLF) for both image encoders on all the datasets. For Fashion200K dresses subset and complete dataset the relative gains are 35.6% and 22.3% respectively, whereas for Fashion-Gen dataset E2E training improves scores by 28.6%. Hence, we suggest that training image encoders end-to-end is an easy way improve MLF performance.

3.4 Experiment 4: MLF vs. Spatial Segmentation

Setup. In the fourth experiment, we explore how our different MLF models perform against other baseline models that are using spatial segmentation methods. As baselines, we select several state-of-the-art cross-modal retrieval models from the general and fashion domain.

FashionBERT [11]. We use the original implementation of the FashionBERT. More specifically, we divide every image into 64 equal tiles and extract features using a pre-trained ResNet-50. First, we use the pre-trained model directly in a zero-shot setting since it is already trained on fashion images. Later, we fine-tune the model on the Fashion200K dataset to see the performance improvement. However, we fine-tune the model only on the dresses subset of Fashion200K dataset. This is done due to memory constraints emerging from the fact that training FashionBERT requires that every image has to be represented in 64 tiles.

SCAN [22]. We use the original implementation of the SCAN model. However, for a fair comparison, we replace object detection mechanisms with the 6-rule segmentations approach. We create MLF features using AlexNet and use a GRU as a text encoder.

Table 3: Results of experiment 3: MLF and end-to-end training, and experiment 4: MLF vs. spatial segmentation

Model	Image-to-text			Text-to-image			
	R@1	R@5	R@10	R@1	R@5	R@10	Sum
	Fashion 200k, dresses						
Laenen (6-rule)	4.3	14.8	22.7	5.3	15.1	23.6	85.8
VilBERT (6-rule)	-	-	-	4.0	16.7	26.3	47.0
FashionBERT (64-tile) (zero-shot)	1.3	5.6	11.0	2.1	9.3	17.4	46.7
FashionBERT (64-tile) (fine-tuned)	3.6	13.0	24.9	4.2	18.9	29.9	94.5
SCAN (6-rule)	9.0	25.9	35.5	9.1	26.1	37.8	143.4
SCAN (6-tile)	9.0	24.6	35.0	10.1	25.6	35.5	139.8
MLF (AlexNet)	12.5	30.2	39.8	13.3	31.6	40.3	167.7
MLF (ResNeSt-50)	11.0	27.7	38.9	12.5	29.8	41.2	161.1
MLF-E2E (AlexNet)	14.3	35.9	48.1	14.9	36.8	48.7	198.7
MLF-E2E (ResNeSt-50)	**21.1**	**44.8**	**55.4**	**22.2**	**46.6**	**57.1**	**247.2**
	Fashion 200K, complete dataset						
Laenen (6-rule)	6.5	24.6	35.6	9.0	26.6	38.7	141.0
VilBERT (6-rule)	-	-	-	21.2	48.9	61.6	131.7
SCAN (6-rule)	17.8	43.9	56.3	22.5	45.2	57.7	243.4
SCAN (6-tile)	15.7	40.2	52.9	18.9	43.6	55.7	227.0
MLF (AlexNet)	18.4	44.9	57.5	21.5	45.9	59.1	247.3
MLF (ResNeSt-50)	23.6	49.5	61.1	25.0	51.1	62.8	273.1
MLF-E2E (AlexNet)	25.6	55.1	69.4	26.0	54.9	68.4	299.4
MLF-E2E (ResNeSt-50)	**35.3**	**69.7**	**81.2**	**38.3**	**69.8**	**81.2**	**375.5**
	Fashion-Gen, complete dataset						
Laenen (6-rule)	8.4	29.4	45.3	11.0	33.9	47.1	175.1
VilBERT (6-rule)	-	-	-	22.1	53.3	66.9	142.3
FashionBERT (64-tile) (zero-shot)	1.1	4.4	9.6	1.3	5.7	11.7	33.8
SCAN (6-rule)	18.4	47.3	60.9	21.0	48.8	62.8	259.2
SCAN (6-tile)	18.2	45.5	61.0	19.9	48.1	60.8	253.5
MLF (AlexNet)	18.1	48.8	63.9	20.9	50.6	65.7	268.0
MLF (ResNeSt-50)	30.3	64.4	76.4	32.0	67.6	78.2	348.9
MLF-E2E (AlexNet)	32.4	68.7	82.7	35.7	70.9	83.3	373.7
MLF-E2E (ResNeSt-50)	**41.6**	**77.4**	**89.1**	**42.8**	**79.2**	**89.5**	**419.6**

Laenen et al. [21]. Following the original implementation [21], we segment the image according to the 6-rule segmentations, extract the image features using a pre-trained AlexNet, and use GRU for text representations.

ViLBERT [26]. We use the original ViLBERT implementation while adapting it to our task. More specifically, we obtain image regions by applying 6-rule segmentation instead of extracting bounding boxes. We do not use the bounding boxes method because it is not possible in the case of fashion images. However, following the original implementation, we use a pre-trained ResNet-101 from PyTorch torchvision model zoo to extract features. We fine-tune the pre-trained ViLBERT 6-layer model on our fashion datasets before testing. We only evaluate the model on text-to-image retrieval, since the model is only designed for caption-based image retrieval.

We train all the models till convergence of summation of recall scores (r-sum) and run the experiment three times using a different seed. We report the average recall scores.

Results. We start by comparing the performance of the two different spatial segmentation methods. Next, we evaluate the MLF approach on Fashion200K and Fashion-Gen datasets against the spatial segmentation baselines. The results of the experiments can be found in Table 3.

3.4.1 *6-rule segmentation vs. tile-based segmentation.*

First, we compare performance between two spatial segmentation methods, 6-rule segmentation vs. tile-based approach. For a fair comparison, we evaluate the performance of the SCAN model, i.e., SCAN (6-rule) vs SCAN (6-tile). Table 3 demonstrate that 6-rule segmentation method achieves higher r-sum scores. More specifically, the relative difference when evaluated on the Fashion200K dresses category is 2.6% whereas the relative difference for the full Fashion200K dataset is 7.2%. For the Fashion-Gen dataset, we observe a relative difference for the 6-rule segmentations of 2.2%. We find it interesting that the performance of the full dataset is relatively better compared to the dresses category alone, even though the segmentation rules are specifically made for dresses. This suggests that segmenting the dress images in certain segments adjusted to the item does not necessarily lead to the score improvement. Overall, the results show us that the type of spatial segmentation can influence the performance of the model.

3.4.2 MLF vs. Spatial segmentation. Table 3 demonstrates that MLF approach using AlexNet is performing better than the basic SCAN model using 6-rule segmentations. The relative increase in r-sum score for the Fashion200K dataset is 17.1% in the dresses category vs. 1.5% in the full dataset. For the Fashion-Gen dataset, the relative improvement of r-sum score is 3.4%. Besides, MLF with ResNeSt-50 gives better performance than MLF-SCAN with AlexNet. In particular, when comparing the MLF approach with a ResNeSt-50 to the MLF approach with an AlexNet, the relative improvement in r-sum is 10.4% for the full Fashion200K dataset and 30.1% for the Fashion-Gen dataset. Such difference in performance could be explained by that fact that ResNeSt-50 is a deeper model.

We further analyze the discriminative power of the MLF with ResNeSt-50 by collecting the average attention distribution of different fashion attributes across hidden layers. Figure 2 demonstrates that more global attributes like color are represented with early layers, whereas the more detailed attributes like type of fabric or dress are represented by later layers. The observations are in line with [34].

Overall, the results suggest that multi-level features have more discriminative power and represent the semantic fashion concepts better than the features obtained with spatial segmentation method.

3.4.3 MLF vs. Baselines. When we compare MLF with a ResNeSt-50 image encoder against the baselines, we observe that our model performs better. Compared to the model of Laenen et al. [21], we achieve a relative improvement of 97% on the two full datasets. The relative improvement when compared to ViLBERT is 5.5% on the full Fashion200K dataset and 24.9% on the Fashion-Gen dataset for image retrieval. The improvement of our model is positive since ViLBERT (252 million) is using much more parameters than our model (8 million). Lastly, if we compare our model against FashionBERT on the Fashion200K subset, we observe that we achieve a relative gain of 70.4%.

Overall, we achieve state-of-the-art results with the MLF-E2E-SCAN architecture in combination with a ResNeSt-50 image encoder and GRU text encoder. We record 43.4% relative increase in text-to-image retrieval and 57.8% relative increase in image-to-text retrieval on the Fashion200K dataset and 48.6% relative increase in text-to-image retrieval and 67.2% relative increase in image-to-text retrieval on the Fashion-Gen dataset.

3.5 Experiment 5: MLF Performance and Network Depth

Setup. The results of the experiments described earlier suggest that the use of a ResNeSt-50 leads to higher performance when compared to AlexNet. We believe that this is because ResNeSt-50 is a deeper network and therefore the features underwent more transformations, which creates more diversity. To investigate this further, we experiment with ResNeSt-50 and ResNeSt-101. We aim to explore how using a deeper network leads to improvement. We train all the models till convergence of summation of recall scores and run the experiment three times using a different seed. We report the average recall scores.

Table 4: Multi-level features experiment comparing ResNeSt-50 and ResNeSt-101

Model	Image-to-text			Text-to-image		
	R@1	R@5	R@10	R@1	R@5	R@10
MLF-SCAN (ResNeSt-50)	23.6	49.5	61.1	25.0	51.1	62.8
MLF-SCAN (ResNeSt-101)	21.9	51.3	64.6	26.9	53.1	64.6
MLF-E2E-SCAN (ResNeSt-50)	**35.3**	**69.7**	81.2	38.3	69.8	81.2
MLF-E2E-SCAN (ResNeSt-101)	34.9	**69.7**	**82.1**	**38.7**	**72.4**	**82.4**

Table 5: Number of trainable parameters per model

Model	Trainable parameters
ViLBERT [26]	252,100,000
FashionBERT [11]	110,000,000
MLF-E2E (AlexNet)	65,300,000
MLF-E2E (ResNeSt-50)	33,700,000
SCAN [22]	8,300,000
Laenen et al. [21]	8,300,000

Results. Table 4 demonstrates the results of the experiment. If we compare the multi-level features, the use of ResNeSt-101 features achieves a better relative r-sum perforce performance of 3.4%. For the MLF-E2E-SCAN, the same relative improvement of r-sum is 1.3%. In general, the results further support our intuition that the using deeper networks for creating multi-level features leads to better performance.

4 ANALYSIS

Number of trainable parameters. In section 3.4.2, we showed that the number of parameters impacts the performance of the models. Therefore, we think it is important to not only look at evaluation metrics but also at how many parameters are these models using. The trend these days is to create models with more and more parameters, e.g., GPT-3 [3] and Turing-NLG [27]. While the results of these models are impressive, they are often difficult to reproduce due to the limited resources. Therefore, in this work, we focus on presenting a solution with fewer parameters that would facilitate reproducibility and be faster and cheaper when it comes to computational costs.

In Table 5 the number of trainable parameters for the different models can be seen. The best performing model in both datasets was the MLF-E2E-SCAN using a ResNeSt-50. The recall scores achieved by this model are more impressive when we compare the number of parameters. The MLF-E2E architecture is making efficient use of the parameters and that the MLF approach with end-to-end training is working. The fine-grainedness of fashion images can not be solved by creating bigger models with more parameters. There is a need to extract features in a smart way to capture the semantics of the attributes. In this work, we showed that we took a step in the right direction with the MLF approach. Furthermore, it is interesting to see that the MLF-E2E with an AlexNet backbone is

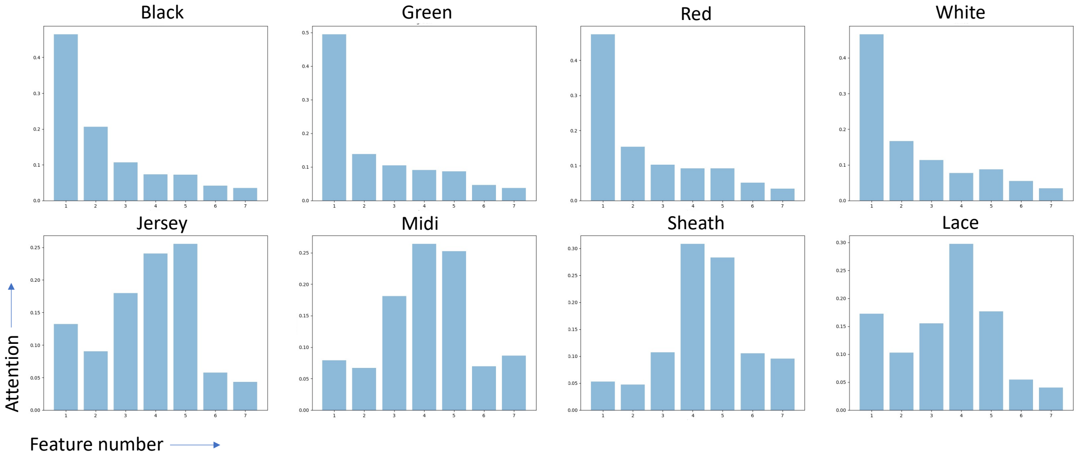

Figure 2: Experiment 4: MLF vs. spatial segmentation. The average attention distribution for different fashion attributes using the MLF with ResNeSt-50.

achieving lower recall scores, even though the number of parameters is higher. This shows again that deeper networks create better features because the difference between the features is probably bigger due to more filters. Overall, we can conclude that we created an efficient architecture that achieves state-of-the-art performance while reducing the number of parameters used by 70% compared to other high-performing models like ViLBERT and FashionBERT.

5 DISCUSSION & CONCLUSION

In this work, we addressed the challenge of fine-grainedness in fashion images for cross-modal fashion search. We proposed to use the MLF approach instead of a spatial segmentation approach to creating more discriminative features. Our architecture MLF-E2E (ResNeSt-50) achieved state-of-the-art results for cross-modal fashion search,. In particular, we gained a 43.4% relative increase in text-to-image retrieval and a 57.8% relative increase in image-to-text retrieval on the Fashion200K dataset, and a 48.6% relative increase in text-to-image retrieval and a 67.2% relative increase in image-to-text retrieval on the Fashion-Gen dataset. Additionally, we reduced the number of parameters by 70% compared to other baselines. Furthermore, we showed that the MLF approach allows us to create more discriminative image features. Besides, we researched the requirements for the MLF approach and found that end-to-end training is beneficial, the use of deeper networks improves score and that in general fine-tuning on domain data does not improve model performance.

In this paper, we mainly focused on the fashion domain. However, product retrieval shares many similarities with fashion retrieval. Therefore, it would be interesting to explore how the MLF approach impacts the general cross-modal product retrieval with the MLF approach. Furthermore, in this work, we focused on the limited set of image and text encoders as well as text-image matching models. Therefore, it would be interesting to see how the MLF approach works with different text encoders, image encoders, and text-image matching models.

6 REPRODUCIBILITY

All code for the experiments can be found on our Git repository: https://github.com/jantje676/cross-modal.

7 ACKNOWLEDGEMENT

We thank Katrien Laenen, Maurits Bleeker, and Pengjie Ren for their helpful feedback. Additionally, we thank the three anonymous reviewers whose comments helped to improve and clarify the manuscript.

REFERENCES

[1] Pulkit Agrawal, Ross Girshick, and Jitendra Malik. 2014. Analyzing the performance of multilayer neural networks for object recognition. In *European conference on computer vision*. Springer, 329–344.

[2] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6541–6549.

[3] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709* (2020).

[5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740* (2019).

[6] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[8] Kinga Edwards. 2020. Key takeaways from E-commerce Region Report: Europe 2020. https://ecommercegermany.com/blog/key-takeaways-from-e-commerce-region-report-europe-2020. [Online; accessed 4-May-2021].

[9] Victor Escorcia, Juan Carlos Niebles, and Bernard Ghanem. 2015. On the relationship between visual attributes and convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1256–1264.

[10] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc' Aurelio Ranzato, and Tomas Mikolov. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. In *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 2121–2129.

[11] Dehong Gao, Linbo Jin, Ben Chen, Minghui Qiu, Peng Li, Yi Wei, Yi Hu, and Hao

Wang. 2020. Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2251–2260.

[12] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. 2017. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE International Conference on Computer Vision*. 1463–1471.

[13] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. 2015. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 447–456.

[14] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research* 47 (2013), 853–899.

[15] Harold Hotelling. 1992. Relations between two sets of variates. In *Breakthroughs in statistics*. Springer, 162–190.

[16] Dietmar Jannach, Surya Kallumadi, Tracy Holloway King, Weihua Luo, and Shervin Malmasi. 2020. ECOM'20: The SIGIR 2020 Workshop on eCommerce. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2459–2460.

[17] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3128–3137.

[18] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539* (2014).

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[20] Katrien Laenen and Marie-Francine Moens. 2020. A Comparative Study of Outfit Recommendation Methods with a Focus on Attention-based Fusion. *Information Processing & Management* 57, 6 (2020), 102316.

[21] Katrien Laenen, Susana Zoghbi, and Marie-Francine Moens. 2017. Cross-modal search for fashion attributes. In *Proceedings of the KDD 2017 Workshop on Machine Learning Meets Fashion*, Vol. 2017. ACM, 1–10.

[22] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 201–216.

[23] Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. 2019. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *arXiv preprint arXiv:1908.06066* (2019).

[24] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* (2019).

[25] Xin Liu, Jiancheng Li, Jiaqi Wang, and Ziwei Liu. 2020. MMFashion: An Open-Source Toolbox for Visual Fashion Analysis. *arXiv preprint arXiv:2005.08847* (2020).

[26] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*. 13–23.

[27] Microsoft. 2020. Turing-NLG: A 17-billion-parameter language model by Microsoft. https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/

[28] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[29] Makoto Ozeki and Takayuki Okatani. 2014. Understanding convolutional neural networks in terms of category-level attributes. In *Asian Conference on Computer Vision*. Springer, 362–375.

[30] Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. 2018. Fashion-gen: The generative fashion dataset and challenge. *arXiv preprint arXiv:1806.08317* (2018).

[31] Richard Socher and Li Fei-Fei. 2010. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 966–973.

[32] Juan Carlos Gomez Marie-Francine Moenss Susana Zoghbi, Geert Heyman. 2016. Fashion Meets Computer Vision and NLP at e-Commerce Search. In *International Journal of Computer and Electrical Engineering (IJCEE)*. 31—-43.

[33] Manos Tsagkias, Tracy Holloway King, Surya Kallumadi, Vanessa Murdock, and Maarten de Rijke. 2021. Challenges and research opportunities in ecommerce search and recommendations. In *ACM SIGIR Forum*, Vol. 54. ACM New York, NY, USA, 1–23.

[34] Sirion Vittayakorn, Takayuki Umeda, Kazuhiko Murasaki, Kyoko Sudo, Takayuki Okatani, and Kota Yamaguchi. 2016. Automatic attribute discovery with neural activations. In *European Conference on Computer Vision*. Springer, 252–268.

[35] Saining Xie and Zhuowen Tu. 2015. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*. 1395–1403.

[36] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. 2015. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579* (2015).

[37] Qian Yu, Xiaobin Chang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. 2017. The devil is in the middle: Exploiting mid-level representations for cross-domain instance matching. *arXiv preprint arXiv:1711.08106* (2017).

[38] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*. Springer, 818–833.

[39] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. 2020. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955* (2020).