Cross-modal Search for Fashion Attributes

Katrien Laenen
KU Leuven
Celestijnenlaan 200A
3001 Heverlee, Belgium
katrien.laenen@kuleuven.be

Susana Zoghbi
KU Leuven
Celestijnenlaan 200A
3001 Heverlee, Belgium
susana.zoghbi@kuleuven.be

Marie-Francine Moens
KU Leuven
Celestijnenlaan 200A
3001 Heverlee, Belgium
sien.moens@kuleuven.be

ABSTRACT

In this paper we develop a neural network which learns intermodal representations for fashion attributes to be utilized in a cross-modal search tool. Our neural network learns from organic e-commerce data, which is characterized by clean image material, but noisy and incomplete product descriptions. First, we experiment with techniques to segment e-commerce images and their product descriptions into respectively image and text fragments denoting fashion attributes. Here, we propose a rule-based image segmentation approach which exploits the cleanness of e-commerce images. Next, we design an objective function which encourages our model to induce a common embedding space where a semantically related image fragment and text fragment have a high inner product. This objective function incorporates similarity information of image fragments to obtain better intermodal representations. A key insight is that similar looking image fragments should be described with the same text fragments. We explicitly require this in our objective function, and as such recover information which was lost due to noise and incompleteness in the product descriptions. We evaluate the inferred intermodal representations in cross-modal search. We demonstrate that the neural network model trained with our objective function on image fragments acquired with our rule-based segmentation approach improves the results of image search with textual queries by 198% for recall@1 and by 181% for recall@5 compared to results obtained by a state-of-the-art image search system on the same benchmark dataset.

CCS Concepts

•Information systems → Image search; •Computing methodologies → Image segmentation; Neural networks; *Natural language generation; Image representations;* Cluster analysis; **•Applied computing → Online shopping;**

KDD '17 August 13–17, 2017, Halifax, Nova Scotia, Canada
© 2017 ACM. ISBN 123-4567-24-567/08/06. . . $15.00
DOI: 10.475/123_4

Keywords

Cross-modal search, fashion

1. INTRODUCTION

Fashion e-commerce is a booming business. We currently witness a shift from physical to digital retail. Therefore, applications that organize and retrieve fashion items have great economic value.

Imagine a cross-modal search tool that can be learned automatically from the organic and noisy data as found in webshops. Such a cross-modal search tool performs both tasks of image annotation, i.e., given an image, return suitable textual descriptors, and image search, i.e., given textual descriptors, retrieve images showing the visual characteristics expressed by the textual descriptors. This would not only alleviate the workload of human annotators, but also promote increased access to relevant products. Currently, products are found by matching key terms in the description of the products. Often the product description does not contain all attributes because they are visible in the accompanying image, thereby hampering the search for such attributes. The only alternative is then to navigate through the taxonomy of products that is offered by the e-retailer. Consequently, the cross-modal search tool would provide a more flexible way for searching products in webshops.

However, building such a cross-modal search tool is definitely not straightforward. On the text side, we have e-commerce product descriptions which are noisy and incomplete, making them challenging to learn from. On the image side, we observe that within a clothing category (e.g. dresses) garments share a high degree of shape similarity. They only differ in certain details, i.e., in their fashion attributes. For instance, the overall shape of two dresses is always the same, but they can have different kinds of necklines, sleeve lengths, colors, ... Over different seasons and trends we also notice that it is the fashion attributes which change appearance, while the overall garment shape remains constant. Therefore, a cross-modal search tool should operate on the level of fashion attributes. Based on its knowledge of fashion attributes, e.g., *"What does a V-neck look like?"* or *"How do we call this kind of skirt shape?"*, the cross-modal search tool will be able to search for requested attributes and to annotate fashion images.

In this paper, we obtain this knowledge of fashion attributes with a neural network which learns to align fashion attributes in images and texts by embedding them into a common, multimodal space (Figure 1). We focus on a single clothing category: dresses. The inferred intermodal repre-

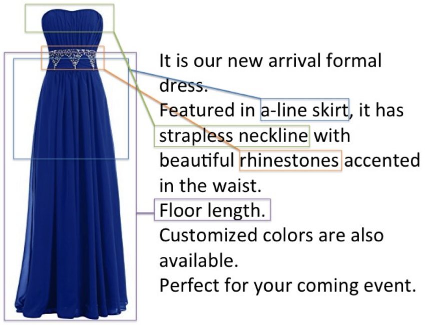
It is our new arrival formal dress.
Featured in a-line skirt, it has strapless neckline with beautiful rhinestones accented in the waist.
Floor length.
Customized colors are also available.
Perfect for your coming event.

Figure 1: Both the image and text are segmented into fragments. These fragments represent the fashion attributes. Corresponding fashion attributes in the image and text are aligned by embedding them into a common embedding space. (Image reference: www.amazon.com)

sentations are utilized in a cross-modal search tool, which we evaluate on image search and image annotation.

The contributions of our work are:

- We propose a neural network alignment model to find the latent alignment between fashion image regions and phrases by embedding them into a common, multimodal space. The obtained intermodal representations allow cross-modal search of fashion attributes.
- Our proposed model learns intermodal representations from organic and noisy data as found in webshops, and does not rely on manually curated data.
- We illustrate how similarity information of fashion image regions can be used to acquire better intermodal representations.
- We perform cross-modal search of fashion items, i.e., given an image, return suitable textual descriptors, and given textual descriptors, retrieve images exhibiting the characteristics expressed by the textual descriptors. This is realized with the inferred intermodal representations, thus without one modality relying on the other. Unlike previous work [18] that uses intermodal representations of full images and texts, our model works at a finer level and employs intermodal representations of image regions and phrases.
- We substantially outperform the results of image search obtained by a state-of-the-art fashion image search system. Compared to this state-of-the-art system, we achieve an increase of 196% on recall@1 and of 181% and recall@5 on the Amazon Dresses dataset.

The remainder of this paper is structured as follows. In the next section we review existing work related to the subject. In Section 3 we explain our model architecture and training objective in detail. Section 4 presents the experiments conducted in this work. The results of these experiments are given and discussed in Section 5. Finally, we conclude and provide the future direction of this work in Section 6.

2. RELATED WORK

Over the past few years, several techniques have been developed to generate image regions enclosing the objects in an image. Examples are objectness [1] and selective search [15]. However, these techniques are developed for detecting straightforward objects in general image scenes, while we want to detect more fine-grained product attributes in a fashion context. Recently, there has been a lot of research on fine-grained image segmentation. Such techniques work on fine-grained classes (e.g. different bird species) and try to detect critical regions in the images that allow to discriminate between the fine-grained classes [7, 17]. Regarding the segmentation of clothing, existing techniques rather focus on generating image regions containing complete fashion items (e.g. a t-shirt) instead of fine-grained fashion attributes (e.g. short sleeves, V-neck) [4].

High-dimensional feature vectors produced by convolutional neural networks (CNNs) are currently the most popular image representations [5, 8, 16, 18]. CNNs have replaced techniques like scale-invariant feature transform (SIFT) [9, 18], and become the state-of-the-art image processing technique.

To segment product descriptions into fashion attributes, Zoghbi et al. [18] propose to filter them, either by using a part-of-speech (POS) tagger to only retain adjectives, adverbs, verbs, and nouns or alternatively by using a domain specific vocabulary and to only retain phrases present in this vocabulary.

Words and phrases are represented with low-dimensional feature vectors which capture the syntax and semantics of that word/phrase. Recently, word vectors acquired with neural networks have become very popular. In [13, 14] two simple two-layered neural networks are proposed to get word representations: the Skip-gram model and the continuous bag-of-words (CBOW) model. The context information captured by these models is rather limited though. In contrast, bidirectional recurrent neural networks (BRNNs)[5] and especially long short-term memories (LSTMs), are able to capture long-term dependencies in full sentences and discourses.

Our attribute alignment task falls into a general category of learning from multimodal data. However, current research on learning from multimodal fashion e-commerce data is still very limited. Techniques have been developed that, given a real-world fashion image as query, return fashion items on e-commerce websites which are similar [4] or identical [6] to those in the provided real-world image. While these techniques result in an image retrieval setting which is only based on visual analysis, we use both fashion images and product descriptions to learn a model which performs cross-modal retrieval of fashion items. In [11, 12] the focus is on annotating images with keywords, and learning from noisy and incomplete product descriptions, which is similar to our work. In contrast with our work, they only use handcrafted image features (e.g. SIFT), they do not address the task of image search, and their dataset is less than half the size of ours. Most closely related is the work of Zoghbi et al. [18], who learn a cross-modal search tool from multimodal fashion e-commerce data. They experiment with two different models to infer intermodal correspondences: canonical correlation analysis (CCA), which explicitly models the correlations between language and visual data, and bilingual latent Dirichlet allocation (BiLDA), a technique that bridges the two modalities through probabilistic latent topics. Their BiLDA model constitutes the state-of-the-art for cross-modal search of fashion items. However, while their model is based on the intermodal correspondences of full images and texts, we try to find the intermodal correspondences of image regions and phrases. Additionally, we will focus on neural networks to bridge the two modalities. Our objective function is inspired by the one of Karpathy et al. [5], who train a neural network which projects objects in visual indoor and outdoor scenes and their textual descriptions into a common embedding space to discover their latent alignment. This objective function uses local co-occurrence information of image regions and words and global image-text correspondence to infer the unknown alignment at the level of image regions and words, which is what we try to achieve in this work. However, while their model expects clean and complete textual annotations with general content of visual scenes composed of prominent objects, our model works with organic fashion e-commerce data as found in webshops, where fashion products are characterized by a multitude of fashion attributes, and where product descriptions barely have grammatical structure, use a sector specific vocabulary, and are often noisy and incomplete. Moreover, in order to learn a better alignment, we propose to incorporate similarity information of detected image regions across all images of our collection. In the past, including image similarity information has proved to be effective in visual relationship detection [10] and in label propagation [2].

3. METHODOLOGY

In this paper we propose a model to align image regions and phrases denoting fashion attributes (Figure 2).

First, we discuss how we detect fashion attributes in fashion images and product descriptions. Next, we elaborate on the objective function that is used to learn to align fashion attributes across different modalities. Finally, we describe how cross-modal search of fashion images and texts is achieved with the acquired intermodal representations.

3.1 Image Segmentation and Representation

3.1.1 Selective Search Segmentation

First, we use selective search [15] to get the regions of the image showing the fashion attributes. We consider all generated image regions and the full image as the image fragments. Consequently, each image has a different number of image fragments.

3.1.2 Rule-based Segmentation

Images on e-commerce websites show clothing items either on their own or worn by a model in different poses. In either case, the item is shown clearly and fully, usually on a white background. Hence, when the item is worn by a model, we assume that independently of the pose, the model always faces the camera and stands straight.

We experiment with a rule-based segmentation approach based on the geometry of garments in a clothing category. We start from the insight that the geometry of garments in a certain clothing category (e.g. dresses) gives us information about where to find the fashion attributes. More precisely, when we have the information that the garment is displayed in a straight position and in frontal view, we know the approximate location of each garment part and thus of each fashion attribute. Therefore, to find the fashion attributes in an image of a dress, we first find a bounding box enclosing the full dress. We do this by transforming the image to the Lab color space and thresholding the image to see which pixels belong to the dress and which to the background. Then, the bounding box is spanned by the leftmost, upmost, rightmost and downmost pixel belonging to the dress. If the area of this bounding box is more than 5 times smaller than the area of the complete image, we take the bounding box enclosing the complete image instead, since in this case usually something has gone wrong when thresholding. The region inside the bounding box serves as the first image fragment. Next, we use our knowledge about the geometry of a dress to segment the region inside the bounding box in 6 more image fragments containing respectively the top, the full skirt, the part of the skirt above the knee, the neckline, the left sleeve and the right sleeve (Table 1). With this rule-based segmentation approach each image has 7 image fragments corresponding to locations where fashion attributes are likely to be found.

Table 1: Expected locations of the parts and fashion attributes of a dress. A location is a rectangle represented as *(x, y), w, h* with (x, y) the coordinates of the upper left corner, w the width and h the height of this rectangle. W and H refer to respectively the width and the height of the bounding box surrounding the full dress.

Dress part	Approximate location	Expected fashion attributes
top	(0, 0), W, 0.35H	neckline, sleeve length, top shape, color/print, accessories
full skirt	(0, 0.30H), W, 0.70H	skirt shape, skirt length, color/print, accessories
skirt above knee	(0, 0.25H), W, 0.40H	skirt shape, skirt length, color/print, accessories
neckline	(0, 0), W, 0.20H	kind of neckline, color/print, accessories
left sleeve	(0, 0), 0.50W, 0.50H	sleeve length, sleeve form, color/print, accessories
right sleeve	(0.50W, 0), 0.50W, 0.50H	sleeve length, sleeve form, color/print, accessories

3.1.3 Image Representation

We represent the image fragments with the BVLC CaffeNet CNN Model[1] [3]. This CNN is pre-trained on ImageNet and only differs from the AlexNet model [8] in that it is not trained with relighting data-augmentation and that the order of the pooling and normalization layers is switched. The image fragment representations are acquired as the activation weights of the last fully connected layer before the softmax layer, which have dimension 4096 in the CNN architecture.

3.2 Text Segmentation and Representation

We train word embeddings on the product descriptions in the Amazon Dresses dataset using the Skip-gram model [13]. These word embeddings allow us to learn a single word embedding for multiword fashion expressions (e.g. dropped waist) and to better capture the syntax and semantics of fashion-related phrases.

Then, to acquire the text fragments, we first convert all words to lowercase, and remove all non-alphanumeric characters and words which occur less than 5 times in the training set. Next, we filter the product descriptions to only

[1] https://github.com/BVLC/caffe/tree/master/models/bvlc_reference_caffenet

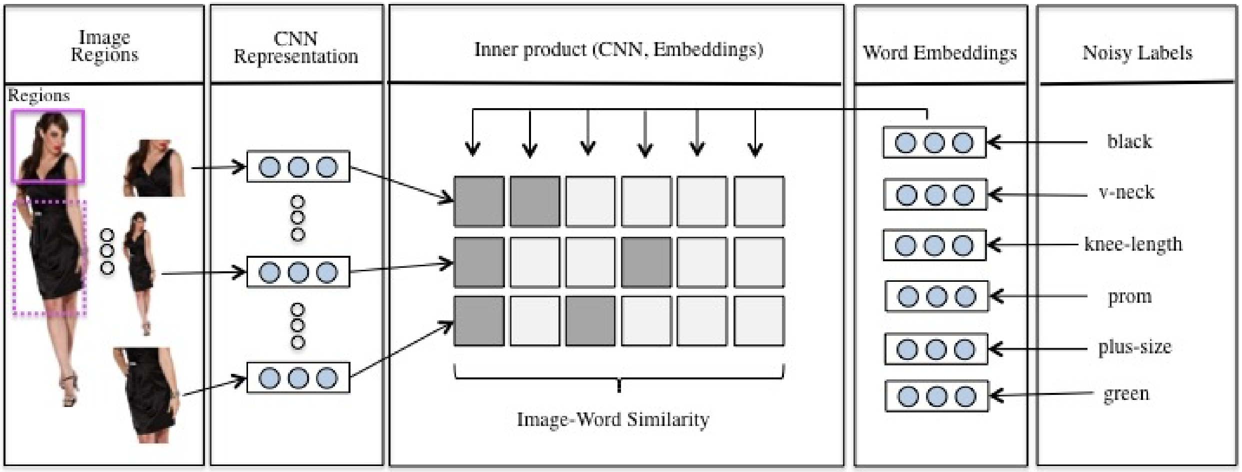

Figure 2: Model overview. Left: An image is segmented into regions, which together with the full image function as the image fragments. The image fragments are embedded through a CNN. Right: A product description is filtered with the Zappos glossary. Each Zappos phrase is a text fragment and is represented with a word embedding. Middle: The alignment model learns to project semantically related image and text fragments to vectors into a common embedding space which have a high inner product, as depicted by dark shades of grey. The resulting intermodal representations form the core building blocks for a cross-modal search tool.

retain fashion-related phrases. Following the approach of Zoghbi et al. [18] we use the glossary of the online clothing shop Zappos[2], which contains both single-word (e.g. strapless) and multiword expressions (e.g. little black dress) related to fashion. Although this removes much noise from the product descriptions, they still remain quite noisy. Remaining phrases might still refer to parts of the garment which are not visible in the image (e.g. the back or side) or describe properties of the garment which are not displayed (e.g. all possible colors). Afterwards, we consider each Zappos phrase as a text fragment. Hence, the number of text fragments differs for different product descriptions, and some product descriptions might even have no text fragments.

3.3 Alignment Model

After segmentation, an image-text pair is represented as a set of image fragments and a set of text fragments. We know that some image fragments and text fragments in these sets correspond but it is unknown which ones. Therefore, we train a neural network to induce a common embedding space which uncovers the intermodal correspondences.

This neural network learns parameters $\theta = \{W_v, b_v, W_s, b_s\}$ to project an image fragment $\hat{v}_i$ and text fragment $\hat{s}_j$ to respectively vector

$$v_i = W_v \hat{v}_i + b_v, \tag{1}$$

and vector

$$s_j = f(W_s \hat{s}_j + b_s). \tag{2}$$

in the common embedding space, which have a high inner product if the corresponding image and text fragment are semantically similar or have a low inner product otherwise. Hence, we interpret the inner product of an image fragment and text fragment in the common embedding space as a measure of their semantic similarity. Here, W_v has dimensions $h \times 4096$ and W_s has dimensions $h \times dim$, where h is the size of the common embedding space and dim is the dimension of the word vectors. Parameters b_v and b_s are bias terms. The activation function f is set to the rectified linear unit (ReLU)[3], which computes $f(x) = \max(0, x)$.

To find the intermodal correspondences, the neural network is trained with an objective function consisting of three different objectives: the fragment alignment objective [5], the global ranking objective [5] and the image cluster consistency objective.

Following Karpathy et al. [5], we use the **fragment alignment objective** $C_F(\theta)$ which uses local co-occurrence information to infer which image fragment and text fragment should be aligned. This objective is formulated as

$$C_F(\theta) = \min_{y_{ij}} C_0(\theta) \tag{3}$$

$$C_0(\theta) = \sum_i \sum_j \max(0, 1 - y_{ij} v_i^T s_j) \tag{4}$$

$$\text{subject to } \sum_{i \in p_j} \frac{y_{ij} + 1}{2} \geq 1 \ \forall j \tag{5}$$

$$y_{ij} = -1 \ \forall i, j \text{ subject to } m_v(i) \neq m_s(j) \tag{6}$$

$$\text{and } y_{ij} \in \{-1, 1\}. \tag{7}$$

It considers all image fragments v_i and text fragments s_j in the training set. Variable y_{ij} reflects whether v_i and s_j should be aligned ($y_{ij} = 1$) or not ($y_{ij} = -1$), and consequently whether their similarity score $v_i^T s_j$ should be encouraged to be more than 1 or less than -1 (Eq. 4). To decide the value for variable y_{ij}, the fragment alignment objec-

[2] http://www.zappos.com

[3] Experiments showed that only using the ReLU activation function at the text side works best.

tive uses co-occurrence information of the fragments during training. $m_v(i)$ and $m_s(j)$ return the index ($\in \{1, ..., N\}$) of the image and sentence that the fragments v_i and s_j belong to. When v_i and s_j do not belong to the same image-text pair, they should not be aligned (Eq. 6). For the ones that do belong to the same image-text pair, the objective tries to find the variables y_{ij} which minimize Eq. 4 (Eq. 3). Here, the only constraint is that each text fragment should be aligned with at least one image fragment it occurs with (i.e., with at least one image fragment in the positive bag p_j of s_j) (Eq. 5). Since this objective benefits from a good initialization of the intermodal representations, this objective is trained with $y_{ij} = 1$ for all v_i and s_j of corresponding image-text pairs during the first 15 epochs. Later, the objective is changed to Eq. 3 to refine the fragment alignments.

The **global ranking objective** $C_G(\theta)$ [5] uses global information about fragments, and enforces that corresponding image-text pairs ($k = l$) should have a higher similarity score (by a margin Δ) than non-corresponding ones. The global ranking objective is given by the following equation

$$C_G(\theta) = \sum_k \Big[\underbrace{\sum_l \max(0, S_{kl} - S_{kk} + \Delta)}_{\text{rank images}} + \underbrace{\sum_l \max(0, S_{lk} - S_{kk} + \Delta)}_{\text{rank texts}} \Big], \quad (8)$$

where the similarity score S_{kl} of an image k and text l is computed based on the similarity scores of their respective fragments f_k and f_l :

$$S_{kl} = \frac{1}{(\mid f_l \mid + n)} \sum_{j \in f_l} \max_{i \in f_k} v_i^T s_j. \quad (9)$$

Here, n is a smoothing term to prevent shorter texts of having an advantage over longer texts.

The **image cluster consistency objective** $C_I(\theta)$ attempts to improve the intermodal representations based on image fragment similarity information. The objective tries to exploit the fact that similar image fragments should be aligned with the same text fragments. We find similar image fragments by clustering the image fragments in C clusters based on cosine distance with k-means clustering. Then, we express the image cluster consistency objective as follows:

$$C_I(\theta) = \sum_{n=1}^{N} \sum_i \sum_j (1 - \frac{\hat{v}_i^T \hat{c}_i}{\|\hat{v}_i\| \|\hat{c}_i\|}) |v_i^T s_j - c_i^T s_j|. \quad (11)$$

This objective considers all N image-text pairs in the training set, and for each pair sums over its image fragments v_i and text fragments s_j. Then, it encourages the difference between the similarity score of image fragment v_i and text fragment s_j and the similarity score of similar image fragment c_i and that same text fragment to be as small as possible. Here, we take as similar image fragment c_i the centroid of the cluster of v_i[4]. We weight the difference in similarity scores with a factor based on the cosine similarity of the two image fragments. With this weighting factor, the objective tries to prevent image fragments in the same cluster which are actually not so similar to be aligned with the same text fragments, and thus guards against introducing errors because of defects in the clustering.

Ultimately, the complete objective function to train our neural network is

$$C(\theta) = C_F(\theta) + \gamma C_I(\theta) + \beta C_G(\theta) + \alpha \|\theta\|_2^2, \quad (12)$$

where θ refers to the network parameters and α, β and γ are hyper parameters to be cross-validated, which we set based on a validation set.

3.4 Cross-modal Search

We can use the inferred intermodal representations for image search and image annotation. In **image search** we retrieve images that display the fashion attributes expressed in a textual query l. We compute the similarity score S_{kl} of text l with all images k and return the top K images with the highest similarity scores. In **image annotation**, we annotate an image query k with suitable fashion attributes. We compute the similarity score S_{kl} of image k with all textual descriptors l and retrieve the top K descriptors with the highest similarity scores.

4. EXPERIMENTAL SETUP

4.1 Dataset

We train our model on the Amazon Dresses dataset, which was collected from the Amazon webshop by Zoghbi et al. [18] between January and February 2015. This dataset consists of 53 689 images of dresses and their product descriptions. The images show dresses from different categories, such as bridesmaid, casual, mother of the bride, night out and cocktail, special occasion, wear to work, and wedding. The product descriptions consist of the surrounding natural language text in the webshop, like the title, features and editorial content. Hence, this dataset contains natural multimodal e-commerce data, where the product descriptions are noisy, incomplete and can contain misspellings, incorrect grammar and incorrect punctuation.

We use 48 689 image-text pairs for training, 4000 for validation and 1000 for testing. During testing, we evaluate the quality of the inferred intermodal representations in a cross-modal retrieval setting. For image search, the textual queries are the complete product descriptions of the test images. In the absence of a complete ground truth reference collection, we consider as the ground truth for each textual query the corresponding test image. Likewise for image annotation, the visual queries are the test images and the ground truth for a visual query is the complete product description of the test image. As such, we follow the exact same setup as Zoghbi et al. [18].

4.2 Experiments

First, we identify the fashion attributes in the images with two image segmentation techniques: selective search and rule-based segmentation based on garment geometry. We use k-means clustering on the resulting image fragments to find C groups of similar image fragments. In our experiments $C = 500, 2500, 5000, 7500, 10000, 12500, 15000, 17500$ and 20000, and we found that $C = 10000$ works best. Next,

[4] We also experimented with the medoid and the nearest neighbour in the same cluster.

we train 300-dimensional word embeddings with the Skip-gram model, for which code is publicly available on GitHub[5]. We concatenate the product descriptions in the training set of the Amazon Dresses dataset, convert all words to lowercase, and remove non-alphanumeric characters. We train the Skip-gram model on the resulting text, where we treat each fashion phrase as a single word. We consider a context size of 5. Then, we filter the product descriptions with the Zappos glossary and consider the remaining Zappos phrases as our text fragments. Afterwards, we input the image and text fragments in our alignment model, and train it with the fragment alignment objective and global ranking objective to induce a 1000-dimensional common embedding space. We use stochastic gradient descent with mini-batches of 100, a fixed learning rate of 10^{-5}, a momentum of 0.90, and make 20 epochs through the training data. Here, a smoothing term n in S_{kl} of 10, a margin Δ in $C_G(\theta)$ of 40, and a factor β in $C(\theta)$ of 0.50 were found to work well. Finally, we investigate the influence of image similarity information on the quality of the intermodal representations, by including the image cluster consistency objective in the objective function. We achieve the best results with γ set to 0.25. We evaluate the inferred intermodal representations in image search and image annotation, and investigate the performance of the image segmentation techniques and the proposed novel objective.

In image search, we retrieve for each textual query the top K most likely test images. We evaluate by computing recall@K for $K = 1, 5, 10, 20, 40$. Precision@K does not say much about performance, since there is only one relevant image for each textual query. To qualitatively evaluate our results, we ourselves construct realistic textual queries asking for different colors, prints, shapes, fabrics and occasions. We avoid infrequent attributes, since these might not occur in the test set. For each textual query, we retrieve the top 5 most likely test images. We consider a retrieved image as relevant if it exhibits all requested attributes.

In image annotation, we retrieve for each visual query the top K most likely product descriptions. We evaluate image annotation by computing recall@K for $K = 1, 5, 10, 20, 40$. Precision@K does not say much about performance here either, since there is only one relevant product description for each visual query. To get further insight in the annotation capabilities of our model, we also show the top 5 most likely Zappos phrases for some visual queries.

It is important to note that for both tasks, recall computed at the cut-off of K items regards a very strict evaluation. Because this evaluation relies on incomplete product descriptions, it might be that we retrieve an image or a textual description for an image, which is not present in the current incomplete ground truth reference collection, but which is relevant. Hence, we might retrieve an image, which satisfies the textual description given as query, or we might retrieve an annotation, which is not (part of) the original product description of an image, but still accurately describes it. Therefore, the actual evaluation results might be higher than those reported in this paper.

4.3 Comparison with other Models

We compare our alignment model to the CCA model and BiLDA model of Zoghbi et al. [18], the latter of which constitutes the state-of-the-art for cross-modal search of fashion items. For image annotation, our baseline is a linear support vector machine (SVM) trained in Zoghbi et al. [18] with the scikit-learn toolkit[6]. This SVM is trained on the CNN representations of the training images of the Amazon Dresses dataset using a one-vs-rest scheme, i.e., they train one classifier for each Zappos phrase.

Table 2: Image search results. *R@K* is recall@K. The reported results are for $C = 10000$.

	Image search				
Model	**R@1**	**R@5**	**R@10**	**R@20**	**R@40**
CCA model (Zoghbi et al. [18])	2.00	8.10	11.70	17.70	28.00
BiLDA model (Zoghbi et al. [18])	2.50	7.80	12.80	18.00	28.900
selective search, $C_F(\theta)$ and $C_G(\theta)$	6.40	18.60	26.40	36.90	50.10
rule-based segmentation, $C_F(\theta)$ and $C_G(\theta)$	**9.10**	20.90	29.20	42.70	**57.70**
rule-based segmentation, $C_F(\theta)$, $C_G(\theta)$ and $C_I(\theta)$	7.40	**21.90**	**32.60**	**43.10**	**57.70**

5. RESULTS AND DISCUSSION

5.1 Image Search

Our image search results are presented in Table 2. These results show that our naive rule-based segmentation technique outperforms selective search. This proves that our assumptions made about e-commerce fashion images are valid. Hence, we can rely on the geometry of the garments in a product category and the cleanness of e-commerce images to locate fashion attributes in these images.

We also observe that including image similarity information in the objective function produces improved intermodal representations. When the image cluster consistency objective is incorporated in the objective function, the alignment model outperforms the one trained with only the fragment alignment and global ranking objective on all image search metrics except recall@1. However, users usually want to see more items than one, so recall@5 or recall@10 are more relevant metrics in terms of usability.

Hence, our best model is the neural network trained with an objective function composed of the fragment alignment, global ranking and image cluster consistency objective on image fragments acquired with rule-based segmentation. Our best model outperforms both the CCA model and state-of-the-art BiLDA model of Zoghbi et al. [18]. Compared to the state-of-the-art, our best model achieves an increase of 196% on recall@1, of 181% and recall@5, of 155% on recall@10, of 139% on recall@20, and of 100% on recall@40. The BiLDA model uses the topic similarity between the textual query and the target image collection to find the most relevant images. This makes the image retrieval model rather coarse. The CCA model explicitly models the correlations between CNN features and discrete word representations. In contrast, our neural network model exploits the expressiveness of real-valued representations on both the visual and textual domains via semantic embeddings. Our results indicate that the common embedding space induced by our neural network encodes the latent semantic alignment of language and visual data in a more meaningful way than the space induced by either cross-modal topics or correlations.

In the absence of a complete ground truth reference collection, recall@K regards a rather strict evaluation. There-

[5] https://github.com/tensorflow/tensorflow

[6] http://scikit-learn.org

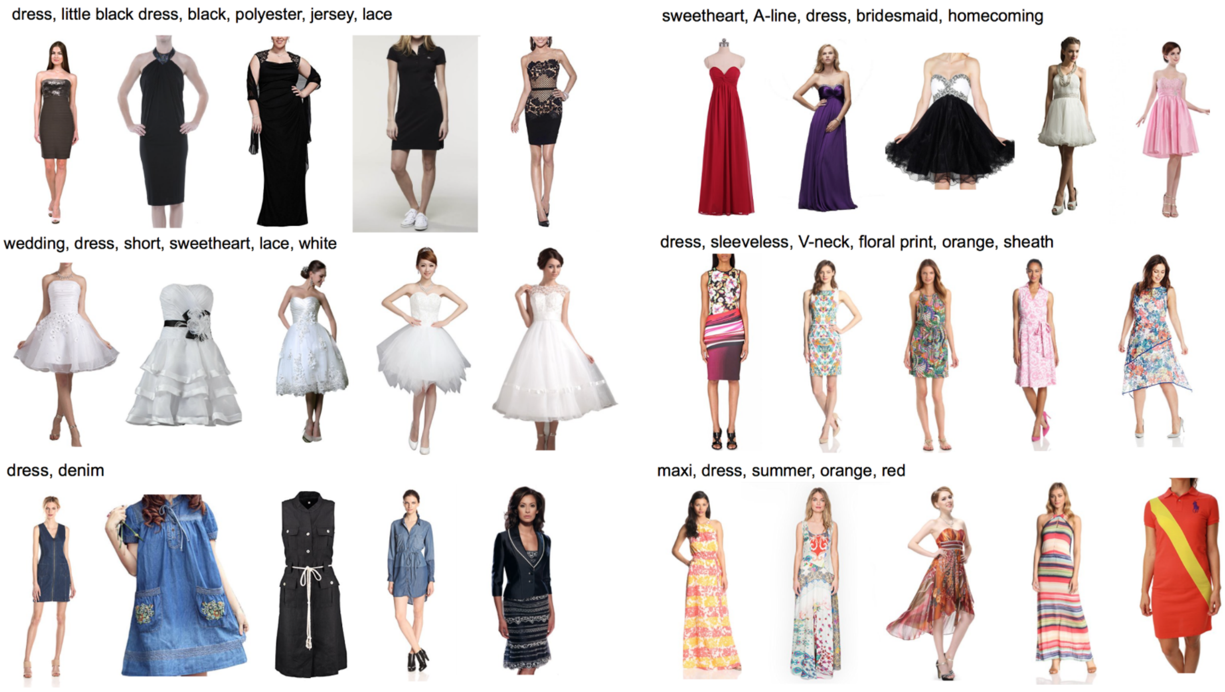
dress, little black dress, black, polyester, jersey, lace
sweetheart, A-line, dress, bridesmaid, homecoming
wedding, dress, short, sweetheart, lace, white
dress, sleeveless, V-neck, floral print, orange, sheath
dress, denim
maxi, dress, summer, orange, red

Figure 3: Image search examples. For each textual query, the top 5 retrieved images are shown. (Image reference: www.amazon.com)

fore, we also present qualitative results to assess our model's performance (Figure 3). The query "wedding, dress, short, sweetheart, lace, white" only returns white wedding dresses. Four of them are short, the fifth one is medium length. Three of the dresses have a sweetheart neckline, and two have lace (the third and fifth dress). The query "sweetheart, A-line, dress, bridesmaid, homecoming" retrieves four dresses with a sweetheart neckline. They could all be considered A-line, but the shape of the first and second dress is somewhat between A-line and empire. For the query "dress, sleeveless, V-neck, floral print, orange, sheath" all retrieved dresses have a floral print and are sleeveless. Two of the dresses have a sheath model, and only one has a V-neck. Retrieving dresses in the requested fabric is more difficult, but it seems that our model has some idea what a certain fabric looks like. When asked for a denim dress, our model is able to find three. Of course, denim is a fabric with clear characteristics. But also when asked for dresses in polyester, it is plausible that the returned dresses are made of this fabric. However, it is hard if not impossible to identify the exact fabric from the images, even for a human. When asked for dresses suitable for a specific occasion, like 'summer' or 'bridesmaid', the model comes up with reasonable suggestions, although this is rather subjective. Overall, when looking at the query results from left to right and top to bottom, respectively one, four, one, no, three and three dresses have all requested attributes. We conclude that our model is able to retrieve dresses with the requested color, print and shape. Retrieving dresses in the correct fabric and distinguishing between similar fashion attributes (e.g. 'empire' and 'A-line' or 'maxi' and 'high low') appear to be the main difficulties. However, it is remarkable that such nice results can be achieved without the matching of terms between the textual query and product descriptions, but instead through the use of inferred intermodal representations.

5.2 Image Annotation

On image annotation, our best model exceeds the SVM baseline, but is surpassed by both the CCA and BiLDA model of Zoghbi et al. [18]. Image annotation seems to benefit from a probabilistic topic representation. Given an unseen image, the BiLDA model infers its topic distribution and generates descriptive words via the topics, in a probabilistic fashion. In contrast, our model projects every unseen image onto an algebraic multimodal space and finds the closest text fragments. Our induced common, multimodal space has been proved quite useful for image search, as discussed in the previous section, however for image annotation, algebraically searching for close-by words might not encode enough semantic expressiveness to retrieve the relevant text fragments. In the future, we may explore how to combine these two modelling paradigms, probabilistic topics and algebraic spaces, into one framework that exploits the benefits of both.

We show examples of image annotations generated by our model in Figure 4. We see correct annotations regarding colors, prints, shapes and accessories. As for image search, finding the correct fabric and distinguishing between similar fashion attributes are things our model still struggles with. Even so, these examples demonstrate that even if our image annotation results are lower than the state-of-the-art, our model is nevertheless capable of generating meaningful annotations.

5.3 Image Segmentation

Figure 5 shows and compares the image fragments gener-

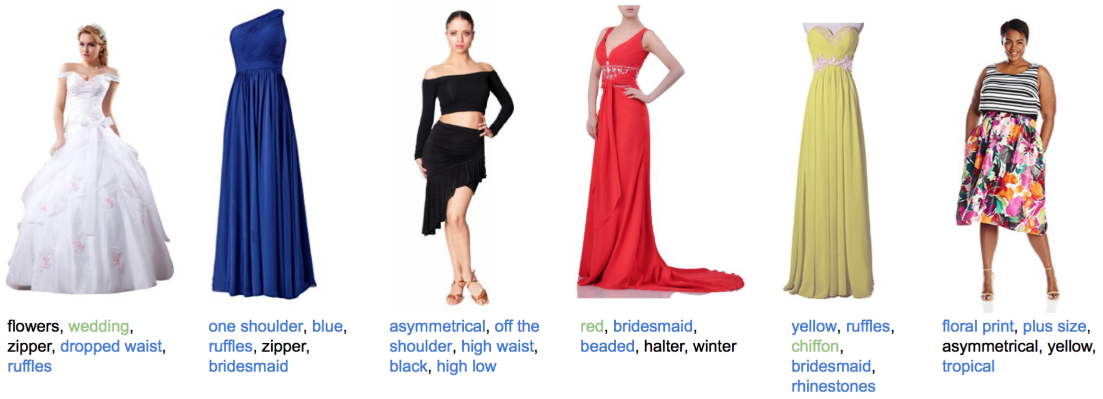
flowers, wedding, zipper, dropped waist, ruffles
one shoulder, blue, ruffles, zipper, bridesmaid
asymmetrical, off the shoulder, high waist, black, high low
red, bridesmaid, beaded, halter, winter
yellow, ruffles, chiffon, bridesmaid, rhinestones
floral print, plus size, asymmetrical, yellow, tropical

Figure 4: Image annotation examples. For each visual query, the top 5 annotations are shown. In green: annotations which are part of the original product description. In blue: annotations which are not in the original product description, but which are correct. In black: annotations which are incorrect or unknown based on what is displayed in the image. (Image reference: www.amazon.com)

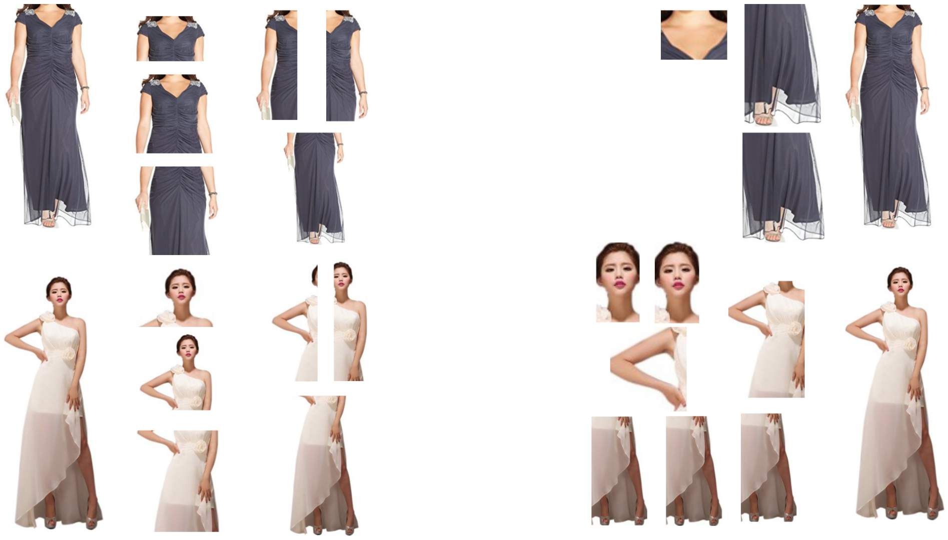

Figure 5: Image segmentation results. Left: Image segmentations acquired with rule-based segmentation. Right: Image segmentations of the same images, but when using selective search. (Image reference: www.amazon.com)

ated with selective search and our rule-based segmentation technique.

While our rule-based segmentation technique is rather naive, it works very nicely on the e-commerce fashion images. By exploiting our knowledge of garment geometry and the fact that e-commerce images are usually very clean, we are able to acquire more meaningful and complete image segmentations than those produced by the selective search algorithm. Additionally, our rule-based segmentation approach has the benefit of generating the same number of regions for each image, where we approximately know which fashion attributes can be observed in each region. In contrast, selective search [15] segments an image in regions based on color similarity, texture similarity and goodness of fit. The resulting regions can focus on smaller regions of interest than in rule-based segmentation. However, we observe that with selective search multiple fashion attributes are not enclosed in any region (e.g. no fragment of the top or neckline) and that some regions show parts of the image that are irrelevant (e.g. only the head). In addition, selective search also produces many near duplicates. Therefore, we prefer our proposed rule-based approach over selective search to segment e-commerce fashion images.

6. CONCLUSION

In this paper, we have proposed a neural network that learns intermodal representations for fashion attributes from organic e-commerce data. We illustrated how we can rely on the cleanness of e-commerce images and the geometry of garments to locate fashion attributes in fashion images. Our proposed rule-based segmentation technique to segment e-commerce images of dresses into regions outperforms selective search. Additionally, we have demonstrated how similarity information of fashion image regions can be used to acquire better intermodal representations for fashion attributes. We introduced the image cluster consistency objective, which encourages similar image fragments to be aligned with the same text fragments, and report increased results when adding this objective to our objective function. The obtained intermodal representations allow cross-modal search of fashion attributes, and we have shown the quality of these representations in image search and image annotation. In image search, our model substantially outperforms a state-of-the-art fashion image search system. Generally, we showed nice search results for colors, prints, shapes, and accessories, but the identification and retrieval of fabrics needs further refinement.

In the future, we want to experiment with other ways to optimize the intermodal representations. While including image similarity information is one way to compensate for errors caused by the noise and incompleteness of product descriptions, we also plan to investigate other ways to deal with this.

7. REFERENCES

[1] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 34(11):2189–2202, November 2012.

[2] T. D. Bui, S. Ravi, and V. Ramavajjala. Neural graph machines: Learning neural networks using graphs. *CoRR*, abs/1703.04818, 2017.

[3] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, pages 675–678, 2014.

[4] Y. Kalantidis, L. Kennedy, and L.-J. Li. Getting the look: Clothing recognition and segmentation for automatic product suggestions in everyday photos. In *Proceedings of International Conference on Multimedia Retrieval (ICMR)*, April 2013.

[5] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137, 2015.

[6] 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 *The IEEE International Conference on Computer Vision (ICCV)*, pages 3343–3351, December 2015.

[7] J. Krause, H. Jin, J. Yang, and L. Fei-Fei. Fine-grained recognition without part annotations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[8] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012.

[9] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2), November 2004.

[10] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, 2016.

[11] R. Mason and E. Charniak. Annotation of online shopping images without labeled training examples. *North American Chapter of the ACL Human Language Technologies*, 2013.

[12] R. Mason and E. Charniak. Domain-specific image captioning. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning, CoNLL 2014, Baltimore, Maryland, USA, June 26-27, 2014*, pages 11–20, 2014.

[13] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[14] T. Mikolov, I. Sutskever, K. Chen, and J. Corrado, G. S.and Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS)*, pages 3111–3119, 2013.

[15] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. *International Journal of Computer Vision (IJCV)*, 2013.

[16] M. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, pages 818–833, 2014.

[17] N. Zhang, J. Donahue, R. Girshick, and T. Darrell. Part-based r-cnns for fine-grained category detection. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I), pages = 834-849, year = 2014,*.

[18] S. Zoghbi, G. Heyman, J. C. Gomez, and M.-F. Moens. Fashion meets computer vision and nlp at e-commerce search. *International Journal of Computer and Electrical Engineering (IJCEE)*, 8(1):31–43, February 2016.