Fashion Meets Computer Vision and NLP at e-Commerce Search

Susana Zoghbi*, Geert Heyman, Juan Carlos Gomez, Marie-Francine Moens
Department of Computer Science, KU Leuven, Belgium.

* Corresponding author. Tel.: +32 16373 930; email: susana.zoghbi@cs.kuleuven.be
Manuscript submitted October 5, 2015; accepted February 2, 2016.
doi: 10.17706/ijcee.2016.8.1.31-43

Abstract: In this paper, we focus on cross-modal (visual and textual) e-commerce search within the fashion domain. Particularly, we investigate two tasks: 1) given a query image, we retrieve textual descriptions that correspond to the visual attributes in the query; and 2) given a textual query that may express an interest in specific visual product characteristics, we retrieve relevant images that exhibit the required visual attributes. To this end, we introduce a new dataset that consists of 53,689 images coupled with textual descriptions. The images contain fashion garments that display a great variety of visual attributes, such as different shapes, colors and textures in natural language. Unlike previous datasets, the text provides a rough and noisy description of the item in the image. We extensively analyze this dataset in the context of cross-modal e-commerce search. We investigate two state-of-the-art latent variable models to bridge between textual and visual data: bilingual latent Dirichlet allocation and canonical correlation analysis. We use state-of-the-art visual and textual features and report promising results.

Key words: Content-based multimedia retrieval, internet search technologies, information retrieval.

1. Introduction

The Web is a multi-modal space. It is flooded with visual and textual information. The ability to natively organize and mine its heterogeneous content is crucial for a seamless user experience. This is true in e-commerce search, where product images and textual descriptions play key roles, because it is not feasible to physically inspect the goods. In particular, this is true for items that are predominantly visual, such as fashion products.

Automatically mining fashion products, while considering their multi-modal nature, has a large potential impact on Web technologies. Allowing users to query in one modality and obtain results in another is greatly beneficial for providing relevant content. For example, users may write textual queries indicating the visual attributes they wish to find, and the algorithm would retrieve product images that display such attributes, without having to rely on textual meta-data on the image to match against the textual query. Additionally, an e-commerce site may wish to automatically organize an image collection according to its visual attributes. In this case, automatically annotating the images with visual properties would allow it.

In this paper, we propose a complete system that performs two truly cross-modal search tasks in the fashion domain. An example is shown in Fig. 1.

Task 1 (Img2Txt): Given a query image without any surrounding text, our system generates text that describes the visual properties in the image.

Task 2 (Txt2Img): Given a textual query without any visual information, our system finds images that display the visual properties in the query.

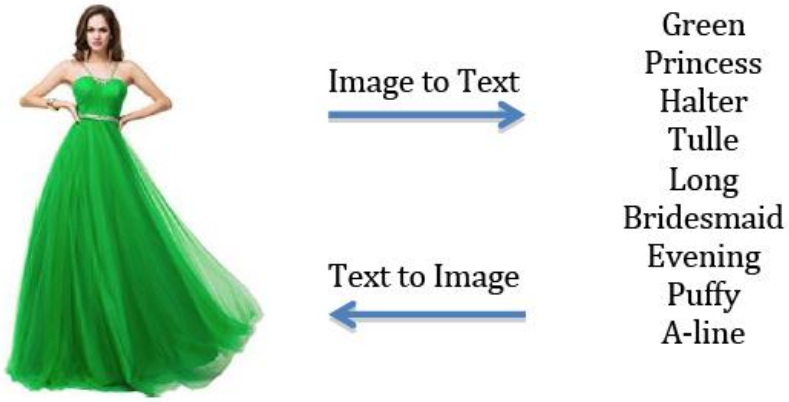


Fig. 1. Our system performs two cross-modal tasks: image to text and text to image.

Cross-modal e-commerce fashion search has not received much attention in the literature. Existing works [1]-[3] mostly focus on classifying fashion visual content into a set of predefined categories. One exception is [4], where the authors focus on automatically annotating images of shoes and bags, but fashion garments and the Txt2Img task are not investigated.

Fashion products present a great deal of challenges for computer vision and natural language processing (NLP). There exist large visual variations in style, poses, shapes, and textures. The textual descriptions in natural language are also very noisy, containing misspellings, improper grammar and punctuation, and incomplete sentences. It is also not trivial to identify which words actually serve to describe the visual properties in the image.

To study this problem, we collected a dataset of over 53,000 products, where each product contains a fashion garment image with accompanying natural language descriptions. We investigate several factors that influence the performance of our system. Specifically, we study Scale Invariant Feature Transform (SIFT) as a visual representation (also used in [4]) and compare it against Convolutional Neural Networks (which have not been explored in this domain). We investigate two different vocabulary representations: one based on part-of-speech tagging, and another one using the set of categories of a major online shop. We analyze the performance of two different modeling paradigms: one that focuses on explicitly modeling the correlations between visual and textual information, i.e., Canonical Correlation Analysis (CCA); and one that bridges the two modalities through probabilistic latent topics, i.e., Bilingual Latent Dirichlet Allocation (BiLDA). In summary the contributions of our work are:

- A complete system that performs two true cross-modal search tasks:

Generate terms that visually describe the properties in a given set of query images and Find images that display sought visual attributes as expressed in a textual query.

- A benchmark real-world dataset for cross-modal e-commerce fashion search.
- An empirical evaluation of factors that influence the performance of our system, such as the choice of vocabulary, visual representation and models.

2. Related Work

Intelligent systems for fashion analysis receive an increasing interest because of their economic and commercial value. For instance, fashion forecasting systems have received a growing attention [5]. Existing systems rely on sales data (and their metadata such as color) or other indicators such as interest of

investors to predict trends, but not on real content data such as images and natural language comments.

Current work in clothing analysis relies on hand-crafted visual features, such as SIFT and SIFT-related features. However, it seems more useful to treat the fashion attributes as latent variables, which form a latent visual pattern characterized by shape, color, and texture. In this respect [6] cluster visual words obtained with SIFT features. In this work we use Convolutional Neural Network (CNN) features to capture the latent patterns, which is novel.

Systems that combine language and visual data to build models for recognizing clothing attributes in language data or in visual data are extremely rare. Mason *et al.* [4] use a bimodal latent Dirichlet allocation model to align attributes in textual descriptions of shoes to visual words in shoe images obtained with SIFT features. Their proposed method constitutes one of our baselines to which we compare our results.

Recently we witness a large interest in automatically generating image captions trained on a large set of images with textual captions, an interest that started with aligning words with images [7], but that recently focuses on aligning content by relying on multimodal representations for instance based on Canonical Correlation Analysis (CCA) methods, e.g., [8] and dependency tree based neural networks , e.g., [9], [10]. We compare our results with the results of applying CCA, visual representations based on convolutional neural networks and textual representations based on syntactic parsing. The above methods were trained on datasets with captions for images that were fully manually annotated and curated, i.e., humans have provided several descriptions of visual scenes, which rather truthfully describe the image. Such datasets are prohibitively expensive to create. In our work, we rely on realistic and freely available data as images and their descriptions found in webshops, which are much more noisy and incomplete.

In this work, we also investigate the value of suitable representations. Text based latent Dirichlet allocation approaches have been used to represent and search products on e-commerce sites [11], [12]. Image based CNN features have been used in clothing retrieval [13]. The use of multimodal representations in an e-commerce context, which we advocate in this paper, is novel.

3. Dataset

We collected a dataset consisting of 53,689 fashion products, specifically dress-like garments. Each product contains one image and surrounding natural language text. While there are other datasets that focus on fashion apparel [1]-[3], they only contain a number of tags to indicate their properties and do not contain descriptions in natural language.

We used Amazon.com's API to query products within the Apparel category. We focused on dresses under all available categories. While all the images contain a dress-like garment, there are very large visual variations in terms of visual attributes, such as shapes, textures and colors. Examples may be found in http://roshi.cs.kuleuven.be/multimodal_search.

There are also large variations in the choice of language, i.e, the same attribute may be described using different words, e.g., *knee-length*, *hits-at-the-knee* or *short* may be used to describe roughly the same length.

Unlike popular general-domain image annotation datasets, such as MSCOCO [14], Flickr8K [8] and Flickr30K [15], the text that accompanies the image in our dataset has not been created specifically for our tasks. Previous datasets have been purposely annotated with 5 sentences using Amazon Mechanical Turk, where annotators were specifically instructed to describe the image. This is not the case in our dataset.

We used simple regular expressions to parse and clean the text. We manually removed the most obvious sizing charts, as well as very common non-descriptive comments that repeat across many products, such as “Call us if you have questions”. After this simple cleaning, the text remains very noisy.

4. Methodology

4.1. Text Representations

The textual descriptions are represented as a bag-of-words. We used two different preprocessing methods. For the first method, we replicate the setup of [16]: we use a part-of-speech (POS) tagger to retain only adjectives, adverbs, verbs and nouns (except for proper nouns and common English stopwords).We used Treetagger for tokenization and POS tagging [17], [18]. All words are converted to lowercase, but no stemming or lemmatization is applied. For the second method, we used the online glossary of *Zappos*, an online shoe and clothing shop (www.zappos.com/glossary) to index terms. That is, after lowercasing the terms, only the glossary terms are retained. The Zappos glossary contains multi-word terms (for instance *little black dress* or *one shoulder*), these will be treated as if they were a single word. Table 1 presents statistics on these representations.

Table 1. Statistics on the Training Datasets after Projection on POS- and Zappos-Based Vocabularies

Preprocessing	#tokens in corpus	Avg. #tokens per doc	|V|
POS	741 760	14.08	9 199
Zappos	317481	6.03	209

4.2. Image Representations

SIFT-based Visual Words: We compute the popular Scale-Invariant Feature Transform (SIFT) descriptors for each image. SIFT finds interest points and considers a grid of subregions around it. For each subregion it computes a gradient orientation histogram. The standard setting uses 4 by 4 subregions with 8-bin orientation histograms resulting in a 128-bin histogram. For an in-depth treatment the reader may refer to [19]. We create visual words by quantizing the descriptors into a number of clusters. We use the k-means algorithm to cluster the descriptors around centroids. These centroids are sometimes called a visual codebook. Each descriptor in each image may then be assigned a visual word and each image may be thought to contain a set of visual words. We represent the visual content of a document as a bag-of-visual-words. We compute SIFT features densely across the image using the open source library VLFeat [20].

Convolutional Neural Networks: CNN is a type of feed-forward artificial neural network where the individual neurons are connected to respond to overlapping regions in the image [21]. They have been extremely successful in image recognition tasks in computer vision. The training process may be interpreted as learning a template given by a set of weights, which aim to maximize the probability of the correct class for each of the training instances. These networks contain many layers. The activation weights corresponding to the last fully connected layer of CNNs may be used as image features. The CNN was pre-trained on over one million images from the famous ImageNet computer vision classification task [22]. For a full description of this representation, we refer the reader to the study of Krizhevsky *et al*. [23], [24]. We used the Caffe implementation of CNNs [25]. Under this framework, images may be represented as 4096-dimensional real-valued vectors. Specifically, we use a convolutional neural network from [23]. The vectors correspond to the weights of the last fully connected layer of the CNN, that is, the layer right before the softmax classifier. We may interpret each component of the CNN vectors as a visual concept or visual word. The actual value corresponding to a particular component then represents the degree to which the visual concept is present. It turns out that these visual concepts are extremely powerful to correctly classify images. Here we will show that they also outperform SIFT features for our cross-modal retrieval tasks.

4.3. Inducing a Multi-modal Space

An attribute may be seen as a latent concept that generates different representations depending on the modality, e.g., visual or textual. There are several approaches that we may use to find associations between

visual and textual words. In particular, we focus on two successful models for multimodal retrieval: Canonical Correlation Analysis (CCA) and Bilingual Latent Dirichlet Allocation (BiLDA). The former explicitly models the correlations between images and text and has been effectively used in multimodal retrieval of Wikipedia articles [26]. The latter has been used in [16] to annotate image data from shoes and bags and it constructs a multimodal space by softly clustering textual and visual features into topics.

Canonical Correlation Analysis: The set of (textual) product descriptions and the set of corresponding images can be seen as two different views on the fashion products. The objective of CCA is to find two vector $u \in R^{d_1}$ and $v \in R^{d_2}$, such that the projections of the two views as given by the *n* training examples are maximally (linearly) correlated. Let *t*, *i* be the original product description and image vectors respectively, and Σ_{ti}, Σ_{tt} and Σ_{ii} be the cross-view and the two within-view covariance matrices, then we find the projections with:

$$\max_{u,v} \frac{E[(u^T t)(v^T i)]}{\sqrt{E[(u^T t)^2]}\sqrt{E[(v^T i)^2]}} = \frac{u^T \Sigma_{ti} v}{\sqrt{(u^T \Sigma_{tt} u)}\sqrt{v^T \Sigma_{ii} v}}$$

The above equation is then extended to learn multidimensional projections by optimizing the sum of correlations in all dimensions, subject to different projected dimensions to be uncorrelated, and the resulting output are projections $U \in R^{d_1 \times d}$ and $V \in R^{d_2 \times d}$.

The dimension $d = \min\{rank(U), rank(V)\}$. After finding the projection matrices, we may project each image and each product description (even the ones that were not in our training set) into a shared multi-modal semantic space. The projection of a d1-dimensional productdescription vector t into a new d-dimensional product description vector t^p is performed as $t^p = tU$. Similarly, the projection of a d2-dimensional image vector *i* into a new d-dimensional image vector i^p is performed as $i^p = iV$. More detailed descriptions are available in the relevant literature, e.g., [8], [26]-[28].

Bilingual Latent Dirichlet Allocation: The BiLDA produces an elegant probabilistic representation to model our intuition that image-description pairs instantiate the same concepts (or topics) using different word modalities. To learn, the main assumptions are that each product *d*, consisting of aligned visual and textual representations $d = \{d^{\text{img}}, d^{\text{text}}\}$, may be modelled by a multinomial distribution of topics, $\theta = P(z|d)$; and each topic may be represented by two distinct word distributions Φ^{img} and Φ^{text}, which correspond to a visual-word distribution, and a textual-word distribution (both multinomial), as

$$\Phi^{\text{img}} = P(w^{\text{img}}|z), \Phi^{\text{text}} = P(w^{\text{text}}|z)$$

These may be interpreted as two views of the same entity: the topic *z*. Consequently topics become multimodal structures, and documents may effectively be projected into a shared multi-modal space via their topical distributions. We estimate the posterior probability distributions θ, Φ^{img} and Φ^{text}, using Gibbs sampling as an inference technique. We can then infer the topic distribution of an unseen image $\theta^{\text{img}} = P(z|d^{\text{img}})$, and the topic distribution of an unseen textual document $\theta^{\text{text}} = P(z|d^{\text{text}})$. We have in-house software that implements this model. We gloss over much of the details, but more information may be found in [29], [30].

4.4. Cross-Modal e-Commerce Search

Given that we can induce a multi-modal space, we may now formulate the mechanism that allows us to bridge between visual and textual content for applications in multi-modal Web search.

Canonical Correlation Analysis (CCA): For the Img2Txt task, let us project the set of m test images that we wish to annotate I_{test}, onto the d-dimensional multimodal space, such that $I^p_{test} = I_{test}V$. Likewise, let us project the set of n textual documents from the training set, $T^p_{train} = T_{train}U$. We can compute the cosine similarity between each test image and each textual document in the training set, via the multimodal space, by $F_{sim} = I^p_{test}(T^p_{train})^T$. The set of words that maximizes the similarity scores for each image constitute the top word candidates for annotation.

A similar approach can be applied in the Txt2Img task. In this case, the similarity scores are given by $F_{sim} = T^p_{test}(I^p_{target})^T$, where T^p_{test} is the projected textual query, and I^p_{target} represent the set of projected images from a target collection. The set of images that maximizes these similarity scores are the top image candidates to fulfill the textual query.

Bilingual Latent Dirichlet Allocation (BiLDA): In the case of BiLDA, we can bridge between image and textual representations in a probabilistic manner by marginalizing over all possible topics. To generate textual words w^{text}, given a query image d^{img},

$$P(w^{text} \mid d^{img}) \propto \sum_z P(w^{text} \mid z)P(z \mid d^{img})$$
$$\propto \sum \Phi^{text}_z \theta^{img}_z$$

The words with the highest probability are chosen as the top candidates to annotate the query image. To retrieve relevant images d^{img} given a textual query d^{text}, we rank the images based on the similarity of their topic distributions, and the images with the highest similarity become the top candidates to satisfy the textual query.

$$sim(d^{\text{img}}, d^{\text{text}}) = \sum_z P(z \mid d^{\text{text}})P(z \mid d^{\text{img}})$$
$$= \sum_z \theta^{\text{text}}_z \theta^{\text{img}}_z$$

5. Experiments

We conducted a set of experiments utilizing two multi-modal representations: CCA and BiLDA, two distinct visual representations: SIFT and CNN; and two different textual vocabularies: POS-based and Zappos-based.

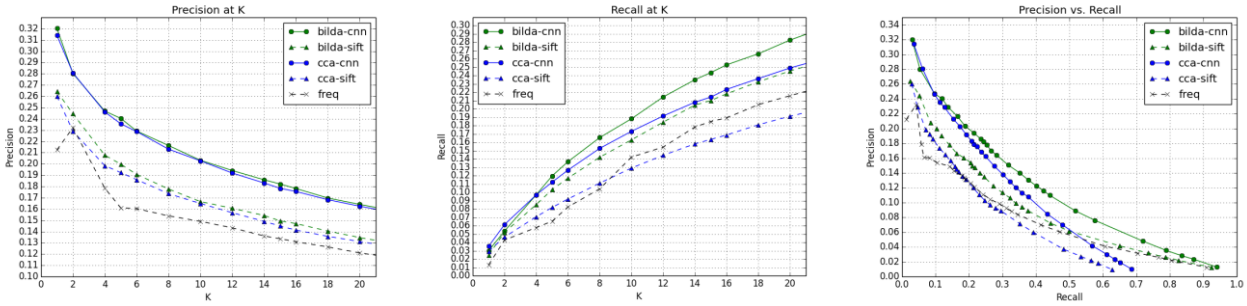
We set aside 1,000 products for testing, 4,000 for validation and the rest for training. We trained the models and chose the hyper parameters using the validation set. We first extracted the visual and textual features from the data. Using the images and corresponding text from the training set, we learned a multi-modal space using the CCA and BiLDA models.

In the Img2Txt task, we retrieve the top K most likely words for each image in the test set. In the Txt2Img task, we retrieve the top K most likely images for each textual description in the test set. We evaluate by computing average precision and average recall against the actual image-words pair. We compare the outputs against a random and a corpus frequency baseline.

6. Results and Discussion

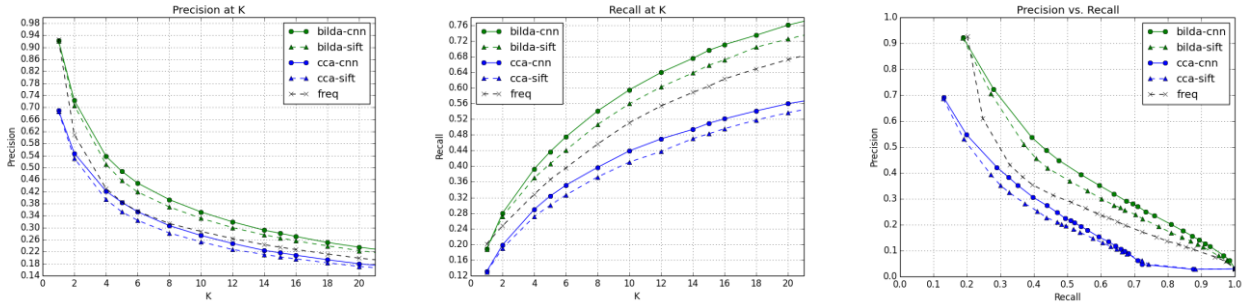
Img2Txt Results. Fig. 2 and Fig. 3 present precision and recall curves for POS- and Zappos-based vocabularies, respectively. We evaluate the contribution of different models, visual features and textual vocabularies to annotate held-out images. Regarding the choice of the latent variable model, the BiLDA model performs at least as well as, and often better than, the CCA model for both vocabularies. Furthermore, the BiLDA model also outperforms the frequency baseline.

Regarding visual features, the overall best performer is the convolutional neural network (CNN). For a given fixed triad of (model, vocabulary, K), the CNN feature always outperforms the SIFT feature in both precision and recall. For example, for the POS vocabulary, at *K*=5, BiLDA-SIFT achieves 19.97% precision and 10.34% recall; whereas BiLDA-CNN achieves 24.04% in precision (20% increase) and 11.96% in recall (15% increase). It is remarkable that CNNs perform so well compared to the SIFT counterparts because they were not trained for this particular task. Instead, as previously described, the model that generates them was trained on a large image classification task [22]. In the future, we will explore fine-tuning the CNNs to improve further performance.



(a) Precision@K POS-based (b) Recall@K POS-based (c) Precision vs. Recall POS-based

Fig. 2. Img2Txt POS-based vocabulary: Precision@K, Recall@K and Precision vs. Recall curves.



(a) Precision@KZapos-based (b) Recall@KZappos-based (c) Precision vs. Recall Zappos-based

Fig. 3. Img2Txt Zappos-based vocabulary: Precision@K, Recall@K and Precision vs. Recall curves.

Regarding the preprocessing, we observe that for this task the Zappos vocabulary performs much better than the POS-based counterpart. This makes sense, since the Zappos vocabulary is much more limited (around 200 unique tokens) than the POS-based vocabulary (over 9000 unique tokens). Having a limited, yet meaningful and complete vocabulary is beneficial for this task. In particular, the Zappos vocabulary is quite interesting because it contains the actual categories that a real-life online shop uses to manually categorize its apparel dress garments.

The results on this vocabulary are quite promising. For example at *K*=5, the frequency baseline achieves 38.52% in precision and 36.54% in recall. In contrast, our best model (BiLDA-CNN), achieves 48.75% in precision (26% increase), and 43.66% in recall (19% increase). The differences between our models and

the frequency baseline amplify for the POS-based vocabulary. At *K*=5, the frequency baseline yields 16.10% precision and 6.53% recall. For the same condition, our best overall model achieved 24.04% in precision (49% increase) and 11.96%in recall (83% increase). Similar behaviors are observed as K increases.

Furthermore, when we compare our best system (BiLDA-CNN-Zappos) with the setup in [16](BiLDA-Sift-POS) we obtain remarkable improvements. For example, at *K*=5, [16] obtains 24.04% precision and 11.97% recall. In contrast our system obtains 48.75% precision (102% increase) and 43.66% recall (364% increase). Similar observations can be made for other values of *K*, as shown in Fig. 2 and Fig. 3.

Of course, our best system benefits from a more targeted vocabulary (Zapos-based). The natural question is which of these two vocabularies to use if we were to deploy this application. In general, there is a tradeoff between the size of the vocabulary and expressiveness. The larger the vocabulary, the more likely we are to capture the nuances in visual attributes. For example, we might have many more distinct words to differentiate varied shades of the same color, as opposed to one single word that encompasses the whole spectrum. We might also be able to differentiate in a more detailed way types of textures and shapes. However, as the vocabulary increases, the task becomes more difficult because we might not have enough data to instantiate and consequently learn all these nuances. This explains the large differences in performance between the two vocabularies. However, regardless of the choice of the vocabulary, the clear top choice of visual features is the CNN and not SIFT as previously used in [16].

Fig. 4. Img2Txt example results: Top words retrieved for several items. Correct words shown in green.

We further study our results qualitatively to evaluate whether our annotations are reasonable. Fig. 4 shows example annotations performed by our best model, BiLDA with CNN features using the Zappos-based vocabulary. We see that we can often recognize several colors: yellow, blue, red, white; some neckline

shapes: sweetheart, one-shoulder, halter, scoop; other patterns: floral-print, pleated, etc. Of course, there is also a degree of subjectivity when evaluating this output. In particular, the occasion to wear the garment is rather difficult to assess. Examples of occasions are homecoming, prom, evening, wedding, cocktail, bridesmaid, etc. While these are overall visual characteristics, they are not so well defined, since they depend on individual preferences or cultural norms. Nevertheless, our system attempts to categorize items for certain occasions and it does reasonably well.

Txt2Img Results. Fig. 5 presents recall@K for all conditions. In all instances, our models perform much better than random. This suggests that we have captured meaningful, useful aspects of the data. Regarding visual features, just as in the previous task, the best performer is the convolutional neural network (CNN). For any fixed (model, vocabulary, K) combination, the CNN feature always outperforms the SIFT feature. Regarding the choice of model, the BiLDA model performs roughly as well as the CCA model when using the Zappos vocabularies. However, CCA outperforms BiLDA with the POS-vocabulary. Regarding the choice of vocabulary, we see that performance is generally higher using the POS-based vocabulary. This makes sense because the POS-based vocabulary is much larger (over 9,000 unique tokens) than the Zappos-based (around 200 tokens). Having a larger vocabulary is beneficial to the Txt2Img task because it allows for more expressiveness in the required visual attributes.

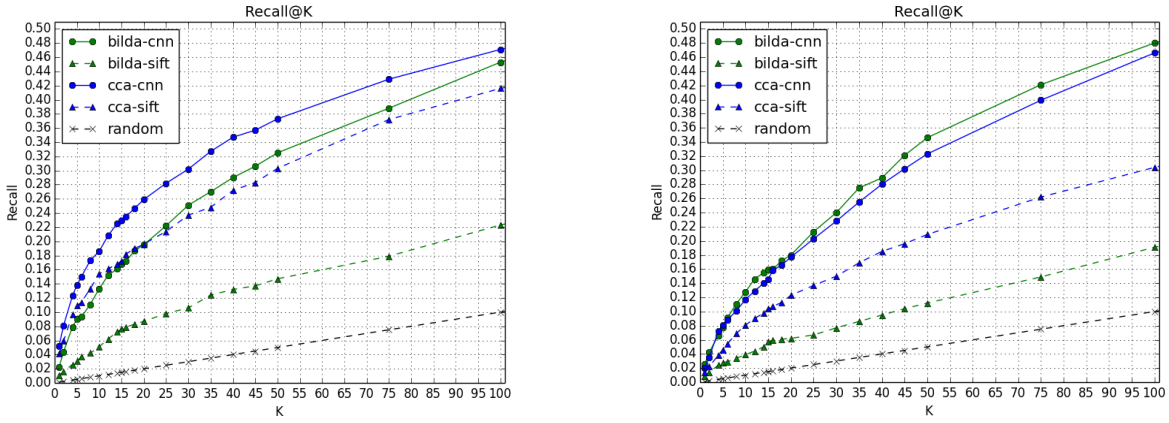


Fig. 5. Txt2Img: Recall@K for POS-based vocabulary (left) and Zappos-based vocabulary (right).

Fig. 6 presents qualitative results, where given a textual query, we show the top 4 images retrieved. We see very interesting results. The query 'little black dress black polyester jersey lace' actually finds little black dresses. It can be argued that the retrieved items also display some jersey-type characteristics, especially the first and third items. For the attribute polyester, it is not clear from the image what the fabrics of the garments are, as this is not a particularly visual word. Also, we do not see any laces on the dresses, so there is room for improvement. The query 'wedding gown sleeveless scalloped ruffles' retrieved wedding gowns in all four items. Two of them are sleeveless and three of them contain 'scalloped ruffles'. The query 'casual sleeveless floral print' retrieves garments with floral patterns on them. They are sleeveless and casual. The query 'long cocktail wedding gown strapless yellow ruched' retrieves yellow items in all cases. Three of them are long and ruched. Two of them are strapless. In this case, the query might be slightly misleading, since it mentions wedding gown, and it is often assumed that that corresponds to an actual white wedding dress. However, the word wedding is often used to describe the occasion that the dress may be worn to. It is similar to the word `cocktail' as it also refers to the occasion. The occasion to which a particular garment is appropriate to wear is of course highly subjective.

Overall, these results are highly impressive given the difficulty of the task. A demonstration and workshop papers of this work be found in [31], [32]. We are able to correctly identify different lengths, shapes, colors and textures. We show this both quantitatively and qualitatively.

Query: little black dress black polyester jersey lace

Query: wedding gown sleeveless scalloped ruffles

Query: casual sleeveless floral print

Fig. 6. Txt2Img: Example results. Given a textual query, we show the top retrieved images.

7. Conclusions

We investigated cross-modal search of fashion items. Given a textual query composed of visual attributes of dresses, our system retrieves relevant images of dresses, and given a picture of a dress as query, the system describes the attributes of the dress in natural language terms. We implemented and compared several algebraic and probabilistic graphical models to learn latent components that bridge the visual and textual features. We have experimented with different types of visual and textual features. Our system was trained on real Web data found at Amazon.com composed of fashion products and their textual descriptions and was evaluated on an additional set of Amazon data. Our best approach uses CNN-based visual features and a controlled, commonly used fashion vocabulary. It obtained a remarkable performance when compared to the state-of-the-art setting of [16], which uses SIFT-based features and a vocabulary based on part-of-speech. For example, at K=5, the previous setting obtains 24.04% precision and 11.97% recall. In contrast our best system obtains 48.75% precision (102% increase) and 43.66% recall (364% increase).We find a similar behaviour for other values of K. Additionally, by visually inspecting the annotations our system generates, we find reasonable descriptions that capture different garment lengths, colors and textures.

References

[1] Chen, H., Gallagher, A., & Girod, B. (2012). Describing clothing by semantic attributes. *Proceedings of the 12th European Conference on Computer Vision*. Berlin, Heidelberg: Springer-Verlag.

[2] Yamaguchi, K., Kiapour, M. H., & Berg, T. L. (2013). Paper doll parsing: Retrieving similar styles to parse clothing items. *Proceedings of the IEEE International Conference on Computer Vision*.

[3] Bossard, L., Dantone, M., Leistner, C., Wengert, C., Quack, T., & Van Gool, L. (2013). Apparel classification with style. *Proceedings of the 11th Asian Conference on Computer Vision* (pp. 321-335). Springer-Verlag.

[4] Mason, R., & Charniak, E. (2014). Domain-specific image captioning. *Proceedings of the Eighteenth Conference on Computational Natural Language Learning* (pp. 11-20). Ann Arbor, Michigan: ACL.

[5] Choi, T-M., Hui, C-L., & Yu, Y. (2013). *Intelligent Fashion Forecasting Systems: Models and Applications.* Springer Publishing Company, Incorporated.

[6] Chen, Q., Wang, G., & Tan, C. L. (2013). Modeling fashion. *Proceedings of IEEE International Conference on Multimedia and Expo*.

[7] Mori, Y., Takahashi, H., & Oka, R. (1999). Image-to-word transformation based on dividing and vector quantizing images with words. *Proceedings of First International Workshop on Multimedia Intelligent Storage and Retrieval Management*.

[8] Hodosh, M., Young, P., & Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif Intell Res*.

[9] Socher, R., Karpathy, A., Le, Q. V., *et al*. (2014). Grounded Compositional Semantics for Finding and Describing Images with Sentences. TACL.

[10] Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3128-3137).

[11] Vulić, I., Zoghbi, S., & Moens, M.-F. (2014). Learning to bridge colloquial and formal language applied to linking and search of e-commerce data. *Proceedings of the 37th International ACM SIGIR Conference on Research; Development in Information Retrieval* (pp. 1195–1198). New York, NY, USA: ACM.

[12] Yu, J., Mohan, S., Putthividhya, D., & Wong, W.-K. (2014). Latent dirichlet allocation based diversified retrieval for e-commerce search. *Proceedings of the 7th ACM International Conference on Web Search and Data Mining* (pp. 463–472). New York, NY, USA: ACM.

[13] Lin K., Yang H-F., Liu K-H., Hsiao J-H., & Chen C-S. (2015). Rapid Clothing Retrieval via Deep Learning of Binary Codes and Hierarchical Search. *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval* (pp. 499–502). New York, NY, USA: ACM.

[14] Lin, T-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., *et al*. (2014). Microsoft (COCO) common objects in context. *Proceedings of European Conference on Computer Vision (ECCV)*.

[15] Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans Assoc Comput Linguist*, *2*, 67-78.

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

[17] Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of the International Conference on New Methods in Language Processing* (pp. 44-49).

[18] Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. *Proceedings of the ACL SIGDAT-Workshop* (pp. 47-50).

[19] Lowe, D. G. (Nov. 2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput Vis*. *60(2),* 91-110.

[20] Vedaldi, A., & Fulkerson, B. (2010). VLFeat — An open and portable library of computer vision

algorithms. *Proceedings of ACM International Conference on Multimedia* (pp. 1469-1472).

[21] Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE Conference* (pp. 2278–2324).

[22] Deng, J., Dong, W., Socher, R., Li, L-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *CVPR09*.

[23] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (pp. 1097-1105). Curran Associates, Inc..

[24] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv abs/1409.1556.

[25] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, *R. B.*, *et al*. (2014). Caffe: Convolutional architecture for fast feature embedding. *Proceedings of the 22Nd ACM International Conference on Multimedia* (pp. 675-678).

[26] Rasiwasia, N., Pereira, J. C., Coviello, E., Doyle, G., Lanckriet, G. R. G., Levy, R., *et al.* (2010). A new approach to cross-modal multimedia retrieval. *Proceedings of the 18th International ACM Conference on Multimedia* (pp. 251-260).

[27] Hardoon, D. R., Szedmák, S., & Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.*, *16(12),* 2639–2664.

[28] Faruqui, M., & Dyer, C. (2014). Improving vector space word representations using multilingual correlation. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 462-471).

[29] De Smet, W., & Moens, M.-F. (2009). Cross-language linking of news stories on the Web using interlingual topic modeling. *Proceedings of the CIKM 2009 Workshop on Social Web Search and Mining*.

[30] De Smet, W., Tang, J., & Moens, M.-F. (2011). Knowledge transfer across multilingual corpora via latent topics. *Proceedings of the 15th Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 549-560).

[31] Zoghbi, S., Heyman, G., Carranza, J. C. G., & Moens, M.-F. (2015). Cross-modal fashion search. *Proceedings of the 22nd International Conference on Multimedia Modelling*.

[32] Zoghbi, S., Heyman, G., Carranza, J. C. G., & Moens, M-F. (2015). Cross-modal attribute recognition in fashion. *Proceedings of NIPS Multimodal Machine Learning Workshop*.

Susana Zoghbi is a PhD student in computer science at the KU Leuven. She obtained a masters degree from the University of British Columbia in 2011. Her research interests lie at the boundary of computer vision and natural language processing, and include deep learning, topic modeling and graphical models.

Geert Heyman is a doctoral researcher in the Department of Computer Science, KU Leuven, Belgium. He completed his undergraduate studies and his master thesis at the Faculty of Engineering Science at KU Leuven in July 2014. His research interests are statistical models (such as neural networks and graphical models) for natural language processing, in particular for language modeling and machine translation.

Juan C. Gomez received a Ph.D. degree in computer science from the INAOE, Mexico, in 2007. He worked at the KU Leuven, Belgium, from 2008 to 2009, and from 2011 to 2015, he worked as a postdoctoral researcher. He is currently an associated lecturer on informatics at the UAD, Mexico. His research interests are machine learning, information retrieval, evolutionary computing and data mining, and he has published several papers in those fields.

Marie-Francine Moens is a full professor at the Department of Computer Science of the Katholieke Universiteit Leuven, Belgium. She holds a M.Sc. degree and Ph.D. degree in computer science from this university. She is the head of the Language Intelligence and Information Retrieval (LIIR) Research Group, and is a member of the Human Computer Interaction Unit. She is currenly also the head of the Informatics Section of the Department of Computer Science at KU Leuven.