Cross-Modal Fashion Search

Susana Zoghbi, Geert Heyman, Juan Carlos Gomez, and Marie-Francine Moens

KU Leuven, Belgium.
{susana.zoghbi, geert.heyman, juancarlos.gomezcarranza,
sien.moens}@cs.kuleuven.be
WWW demo site: http://roshi.cs.kuleuven.be/multimodal_search/

Abstract. In this demo we focus on cross-modal (visual and textual) e-commerce search within the fashion domain. Particularly, we demonstrate two tasks: 1) given a query image (without any accompanying text), we retrieve textual descriptions that correspond to the visual attributes in the visual query; and 2) given a textual query that may express an interest in specific visual characteristics, we retrieve relevant images (without leveraging textual meta-data) that exhibit the required visual attributes. The first task is especially useful to manage image collections by online stores who might want to automatically organize and mine predominantly visual items according to their attributes without human input. The second task renders useful for users to find items with specific visual characteristics, in the case where there is no text available describing the target image. We use a state-of-the-art visual and textual features, as well as a state-of-the-art latent variable model to bridge between textual and visual data: bilingual latent Dirichlet allocation. Unlike traditional search engines, we demonstrate a truly *cross-modal* system, where we can directly bridge between visual and textual content without relying on pre-annotated meta-data.

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 especially 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 which 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. Globally, consumers spend billions of dollars on clothing every year. Therefore, applications that organize and retrieve fashion items would have great societal value. More specifically, in fashion e-commerce search, 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 our system retrieves 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 facilitate this time-consuming task.

2 Functionality Overview

In this work, we demonstrate a complete system that performs two truly cross-modal search tasks in the fashion domain: **Task 1 (Img2Txt):** Given a query image without any surrounding text, our system generates text that describes the visual properties in the image; and **Task 2 (Txt2Img):** Given a textual query without any visual information, our system finds images that display the visual properties in the query. Unlike traditional search engines, these tasks are realized in a purely cross-modal way, where the query modality is completely distinct from the target. In other words, textual queries retrieve target images that have not been previously annotated, therefore, keyword matching is not possible. Likewise, image queries without any textual annotations, retrieve words that describe the image. These are challenging tasks for both computer vision and natural language processing. Figures 1 and 2 show examples of these tasks.

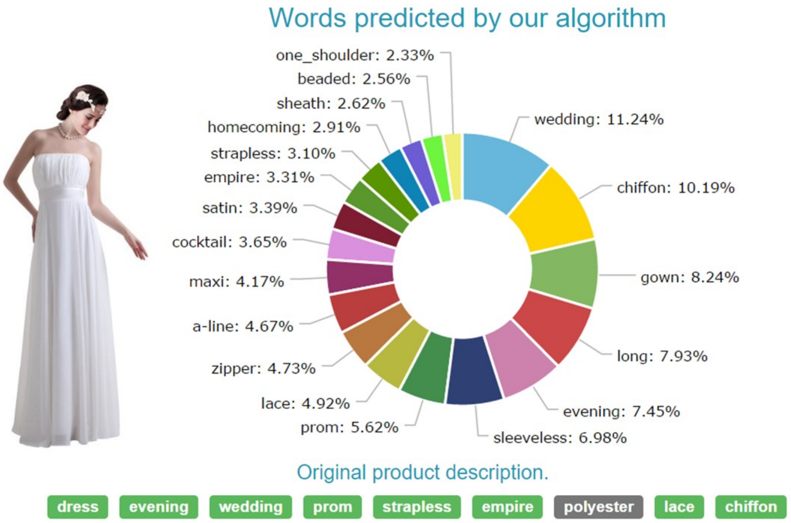
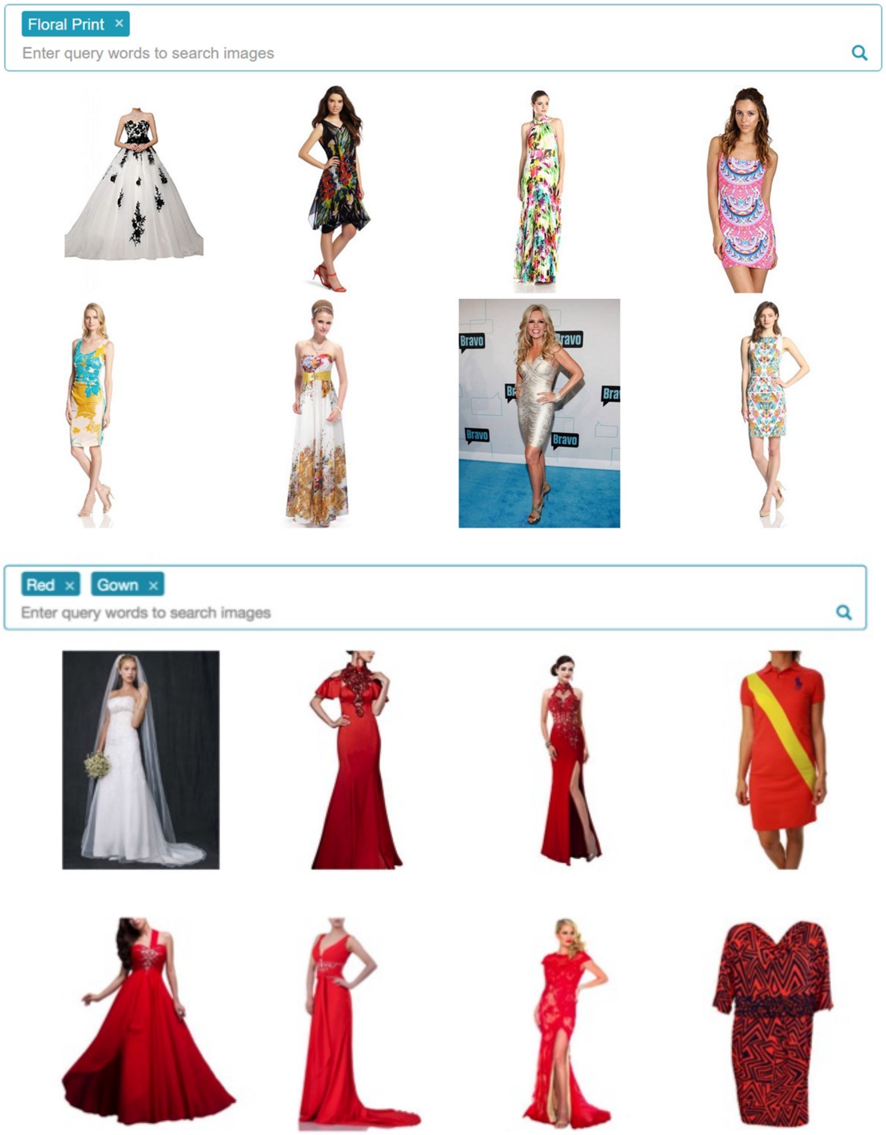


Fig. 1: Img2Txt: Given a query image (left), our system predicts words that describe the attributes of the image (right), ordered by the probability of the word occurrence. On the bottom, we show the original words from the product description and highlight in green those predicted by our algorithm.

To build this demo, we collected a dataset consisting of 53,689 fashion products, specifically dress-like garments. Each product contains one image and surrounding natural language text. We used Amazon.com's API to query products within the Apparel category. We focused on dresses under all available categories, such as Casual, Bridesmaids, Night Out & Cocktail, Wear to Work, Wedding, Mother



Floral Print ×

Enter query words to search images

Red × Gown ×

Enter query words to search images

Fig. 2: Txt2Img: The user may enter a textual query (e.g., ’floral print’ on the top half or ’red gown’ on the bottom), and our system finds images that display the attribute in the query.

of the Bride, etc. From the API's output, we concatenate all the natural text that surrounds the image, such as title, features and editorial content. While all the images contain a dress-like garment, there exist very large visual variations in terms of visual attributes, such as shapes, textures and colors[1].

3 Methodology

Text Representation. The textual descriptions are represented as a bag-of-words. We used the online glossary of www.zappos.com, an online clothing shop[2] to index terms. After lower-casing the terms, only the glossary terms are retained. The Zappos glossary contains multi-word terms (for instance 'little black dress' or 'one shoulder'), these were treated as if they were a single word.

Image Representation. We use Convolutional Neural Networks (CNNs) [5] to represent images. A CNN is a type of feed-forward artificial neural network where the individual neurons are connected to respond to overlapping regions in the image [6]. 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 [3]. For a full description of this representation, we refer the reader to the study of Krizhevsky et al. [5],[8]. We used the Caffe implementation of CNNs [4]. Under this framework, images may be represented as 4096-dimensional real-valued vectors. Specifically, we use a 16-layer convolutional neural network from [8]. 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.

Inducing a Multi-Modal Space. An attribute may be seen as a latent concept that generates different representations depending on the modality, (visual or textual). For example the word *a-line* is an attribute of the object, in particular it may describe the A-shape of a skirt. This attribute is instantiated in the image by a set of pixel values; and it is instantiated in the text by the actual textual words. In our training set, we have a collection of image and textual description pairs. From these, we wish to learn associations between the visual and textual information. We may project onto a multimodal space where we enforce that image and text elements that often occur together during training, are close together in this multimodal space.

[1] Examples of our data are available in `http://glenda.cs.kuleuven.be/multimodal_search` under the 'Training Data' tab.

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

There exist several approaches that we may use to find associations between visual and textual words. In particular, we focus on a state-of-the-art multimodal retrieval model: Bilingual Latent Dirichlet Allocation (BiLDA). It has been used in [7] to annotate image data from shoes and bags and it constructs a multimodal space by softly clustering textual and visual features into topics.

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 $d = \{d^{img}, d^{text}\}$, 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^{txt}, which correspond to a visual-word distribution, and a textual-word distribution (both multinomial), respectively as

$$\phi^{img} = P(w^{img}|z), \qquad \phi^{txt} = P(w^{txt}|z). \tag{1}$$

These may be interpreted as two views of the same entity: the topic z. Consequently topics become multi-modal structures, and documents may effectively be projected into a shared multi-modal space via their topical distributions. We estimate the posterior probability distributions of θ, ϕ^{img} and ϕ^{txt}, using Gibbs sampling as an inference technique, which is a simple and easy to implement method. We can infer the topic distribution of an unseen image document $\theta^{img} = P(z|d^{img})$, and the topic distribution of a previously unseen textual document $\theta^{txt} = P(z|d^{text})$, by using again Gibbs sampling, but this time we keep the word-topic distributions fixed ϕ^{img} and ϕ^{txt}, as we have already learned them, and we need only to infer the topic assignments within each unseen document. We have in-house software that implements this model. We gloss over much of the details, but more information may be found in [1,2].

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. Using Bilingual Latent Dirichlet Allocation (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},

$$\begin{aligned} P(w^{text}|d^{img}) &\propto \sum_z P(w^{text}|z)P(z|d^{img}) & (2)\\ &\propto \sum_z \phi_z^{txt}\theta_z^{img} & (3)\end{aligned}$$

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^{txt}, we rank the images based on the similarity of their topic distributions,

$$\begin{aligned} sim(d^{img}, d^{txt}) &= \sum_z P(z|d^{text})P(z|d^{img}) & (4)\\ &= \sum_z \theta_z^{txt}\theta_z^{img} & (5)\end{aligned}$$

the images with the highest similarity become the top candidates to satisfy the textual query. Full technical details and formal evaluation can be found in [9].

4 Conclusions

In this work we have demonstrated cross-modal search of fashion items. Given a textual query, 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. 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. We used CNN-based visual features and a controlled, commonly used fashion vocabulary. By visually inspecting the annotations our system generates, we find reasonable descriptions that capture different garment lengths, colors and textures. For example an image displaying a yellow long gown, receives these actual words as descriptions. Furthermore, we were able to find relevant images given textual queries. For example, a query indicating 'casual sleeveless floral print', retrieves garments with these characteristics. These results are extremely useful for navigating the multi-modal data on fashion search.

5 Acknowledgments

We greatly thank Anirudh Tomer for building the Web interface of our demonstrator. This project is part of the SBO Program of the IWT (IWT-SBO-Nr. 110067).

References

1. W. De Smet and M.-F. Moens. Cross-language linking of news stories on the Web using interlingual topic modeling. In *Proceedings of the CIKM 2009 Workshop on Social Web Search and Mining (SWSM)*, pages 57–64, 2009.
2. W. De Smet, J. Tang, and M.-F. Moens. Knowledge transfer across multilingual corpora via latent topics. In *Proceedings of the 15th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 549–560, 2011.
3. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.
4. Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.
5. A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.
6. Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.
7. R. Mason and E. Charniak. Annotation of online shopping images without labeled training examples. In *North American Chapter of the ACL Human Language Technologies*, volume 2013, page 1, 2013.
8. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.
9. S. Zoghbi, G. Heyman, J. C. Gomez, and M.-F. Moens. Fashion meets computer vision and natural language processing. International Journal of Computer and Electrical Engineering (Accepted).