Cross-Modal Attribute Recognition in Fashion

Susana Zoghbi[*]
Department of Computer Science
KU Leuven
susana.zoghbi@cs.kuleuven.be

Geert Heyman
Department of Computer Science
KU Leuven
geert.heyman@cs.kuleuven.be

Juan Carlos Gomez
Department of Computer Science
KU Leuven
juancarlos.gomezcarranza@cs.kuleuven.be

Marie-Francine Moens
Department of Computer Science
KU Leuven
sien.moens@cs.kuleuven.be

Abstract

In this paper we focus on cross-modal (visual and textual) attribute recognition 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 visual characteristics, we retrieve relevant images that exhibit the required visual attributes. To this end, we collected a dataset that consists of 53,689 images coupled with textual descriptions in natural language. The images contain fashion garments that display a great variety of visual attributes. Examples of visual attributes in fashion include colors (e.g. green, purple), shapes (e.g., v-neck, a-line, one-shoulder, floor-length), patterns (e.g., floral, striped), textures (e.g., silk, chiffon, beading, sequin), and even occasions (e.g., wedding, evening, wear-to-work). Unlike previous work, 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 attribute recognition. We investigate two latent variable models to bridge between textual and visual data: bilingual latent Dirichlet allocation and canonical correlation analysis. We use visual and textual features and report promising results[1].

1 Introduction

Humans have the remarkable ability to not only recognize objects in an image, but also to describe them according to their visual attributes. A quick glance at an image allows us to identify its visual properties: shape, color, texture, etc. While computer vision systems have become very successful at recognizing objects, visual attribute recognition still remains relatively under explored. In this work we focus on jointly recognizing attributes in both images and text within the fashion domain.

Fashion products present a great deal of challenges for computer vision and natural language processing (NLP). The large visual variations in style, poses, shapes, and textures make them very

[*] http://people.cs.kuleuven.be/$\sim$ $susana.zoghbi/$.

[1] Examples of the data and results may be found in http://roshi.cs.kuleuven.be/multimodal_search [20]

interesting for attribute recognition tasks. Textual descriptions of fashion products as found in the wild Web are also very noisy, containing misspellings, improper grammar and punctuation, and incomplete sentences. Here we propose a complete system that performs two truly cross-modal search in the fashion domain. Examples are shown in Figures 2 and 4.

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.

To study this problem, 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. Unlike popular general-domain image annotation datasets [9, 5, 19], the text that accompanies the image in our dataset has not been created specifically for our tasks.

We study Scale Invariant Feature Transform (SIFT) as a visual representation (also used in [12]) and compare it against Convolutional Neural Networks (which have not been extensively explored in this domain). We investigate two different vocabularies: 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, where one modality does not rely on the other one:
    - Generate terms that visually describe the properties in a given set of query images;
    - Find images that display sought visual attributes as expressed in a textual query.
- A real-world benchmark dataset for cross-modal attribute recognition in fashion.
- An empirical evaluation of factors that influence the performance of our system, such as the choice of vocabulary, visual representation and models.

2 Methodology

2.1 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 [11]: 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 [14, 15]. All words are converterd to lowercase, but no stemming or lemmatization is applied. For the second method we have used the online glossary of *Zappos*, an online shoe and clothing shop[2] 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.

2.2 Image Representations

We investigate two different image representations: SIFT-based Visual Words [17, 1] and Convolutional Neural Networks (CNNs) [7].

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 [10]. 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.

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

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 [18].

Convolutional Neural Networks. CNN is a type of artificial neural network where the individual neurons are connected to respond to overlapping regions in the image [8]. They have been extremely successful in image recognition tasks in computer vision. The training process consists of learning a set of weights, which aim to maximize the probability of the correct class for each of the training instances. The CNN was pre-trained on the famous ImageNet [2] computer vision classification task. The activation weights corresponding to the last fully connected layer of CNNs may be used as image features. Under this framework, images may be represented as 4096-dimensional real-valued vectors. 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. For a full description of this representation, we refer the reader to the study of [7, 16]. We used the Caffe implementation of CNNs [6].

2.3 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). We focus on two successful models for multimodal retrieval: Canonical Correlation Analysis (CCA) and Bilingual Latent Dirichlet Allocation (BiLDA).

Canonical Correlation Analysis (CCA). The set of (textual) product descriptions and the set of corresponding images can be seen as two different views of the same product. The objective of CCA is to find two vectors $\mathbf{u} \in R^{d1}$ and $\mathbf{v} \in R^{d2}$, such that the projections of the two views as given by the n training examples are maximally (linearly) correlated [4, 13, 5, 3]. Let $\mathbf{t}$, $\mathbf{i}$ 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:

$$\max_{\mathbf{u},\mathbf{v}} \frac{E[(\mathbf{u}^\top \mathbf{t})(\mathbf{v}^\top \mathbf{i})]}{\sqrt{E[(\mathbf{u}^\top \mathbf{t})^2]}\sqrt{E[(\mathbf{v}^\top \mathbf{i})^2]}} = \frac{\mathbf{u}^\top \Sigma_{ti} \mathbf{v}}{\sqrt{\mathbf{u}^\top \Sigma_{tt} \mathbf{u}}\sqrt{\mathbf{v}^\top \Sigma_{ii} \mathbf{v}}} \tag{1}$$

The above equation is then extended to learn multi-dimensional projections by optimizing the sum of correlations in all dimensions, subject to different projected dimensions to be uncorrelated, and the resulting output are two projection matrices $\mathbf{U} \in R^{d1\times d}$ and $\mathbf{V} \in R^{d2\times d}$. The dimension $d = \min\{rank(\mathbf{U}), rank(\mathbf{V})\}$. 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 product description vector $\mathbf{t}$ into a new d-dimensional product description vector $\mathbf{t^p}$ is performed as $\mathbf{t^p} = \mathbf{tU}$. Similarly, the projection of a $d2$-dimensional image vector $\mathbf{i}$ into a new d-dimensional image vector $\mathbf{i^p}$ is performed as:[3] $\mathbf{i^p} = \mathbf{iV}$.

Bilingual Latent Dirichlet Allocation (BiLDA). 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

$$\phi^{img} = P(w^{img}|z), \qquad \phi^{txt} = P(w^{txt}|z). \tag{2}$$

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.

[3] To be more precise: since the projected vectors are coordinates in two isometric d-dimensional subspaces, they can be thought of as belonging to a single shared space, so that they can be used for cross-modal retrieval.

2.4 Cross-Modal Attribute Recognition

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 $\mathbf{I_{test}}$, onto the d-dimensional multimodal space, such that $\mathbf{I^p_{test}} = \mathbf{I_{test}V}$. Likewise, let us project the set of n textual documents from the training set, $\mathbf{T^p_{train}} = \mathbf{T_{train}U}$. We can compute the cosine similarity between each test image and each textual document in the training set, via the multimodal space, by

$$\mathbf{F_{sim}} = \mathbf{I^p_{test}}(\mathbf{T^p_{train}})^\top \tag{3}$$

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

$$\mathbf{F_{sim}} = \mathbf{T^p_{test}}(\mathbf{I^p_{target}})^\top \tag{4}$$

where $\mathbf{T^p_{test}}$ is the projected textual queries, $\mathbf{I^p_{target}}$ represents 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.

2.4.1 Bilingual Latent Dirichlet Allocation (BiLDA)

With 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},

$$\begin{aligned} P(w^{text}|d^{img}) &\propto \sum_z P(w^{text}|z)P(z|d^{img}) &(5)\\ &\propto \sum_z \phi_z^{txt}\theta_z^{img} &(6)\end{aligned}$$

To retrieve relevant images d^{img} given a textual query d^{txt}, 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.

$$\begin{aligned} sim(d^{img}, d^{txt}) &= \sum_z P(z|d^{text})P(z|d^{img}) &(7)\\ &= \sum_z \theta_z^{txt}\theta_z^{img} &(8)\end{aligned}$$

3 Experiments and Evaluation

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 hyperparameters 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.

4 Results

Img2Txt Results. Figure 1 presents precision and recall curves for the Zappos-based vocabulary. We report results only on the Zappos-based vocabulary because POS-based results are much lower.

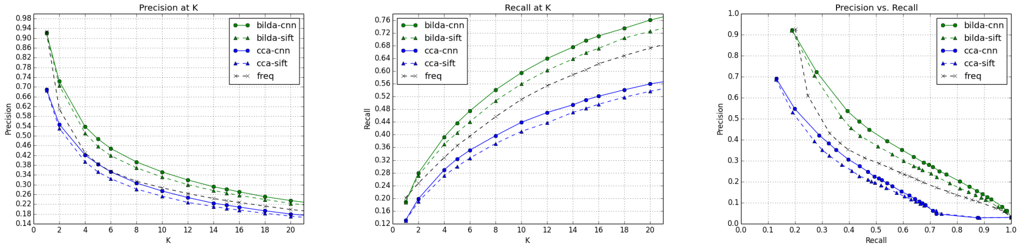

Figure 1: Img2Txt Zappos-based vocabulary: Precision@K, Recall@K and Precision vs Recall curves.

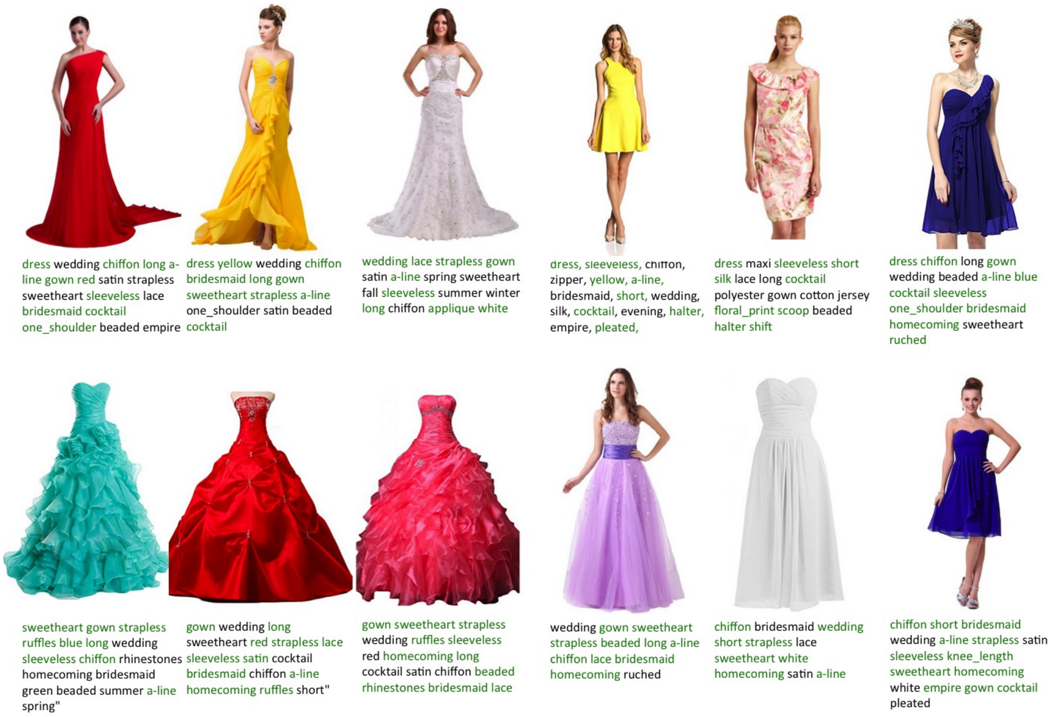
dress wedding chiffon long a-line gown red satin strapless sweetheart sleeveless lace bridesmaid cocktail one_shoulder beaded empire
dress yellow wedding chiffon bridesmaid long gown sweetheart strapless a-line one_shoulder satin beaded cocktail
wedding lace strapless gown satin a-line spring sweetheart fall sleeveless summer winter long chiffon applique white
dress, sleeveless, chiffon, zipper, yellow, a-line, bridesmaid, short, wedding, silk, cocktail, evening, halter, empire, pleated,
dress maxi sleeveless short silk lace long cocktail polyester gown cotton jersey floral_print scoop beaded halter shift
dress chiffon long gown wedding beaded a-line blue cocktail sleeveless one_shoulder bridesmaid homecoming sweetheart ruched
sweetheart gown strapless ruffles blue long wedding sleeveless chiffon rhinestones homecoming bridesmaid green beaded summer a-line spring"
gown wedding long sweetheart red strapless lace sleeveless satin cocktail bridesmaid chiffon a-line homecoming ruffles short"
gown sweetheart strapless wedding ruffles sleeveless red homecoming long cocktail satin chiffon beaded rhinestones bridesmaid lace
wedding gown sweetheart strapless beaded long a-line chiffon lace bridesmaid homecoming ruched
chiffon bridesmaid wedding short strapless lace sweetheart white homecoming satin a-line
chiffon short bridesmaid wedding a-line strapless satin sleeveless knee_length sweetheart homecoming white empire gown cocktail pleated

Figure 2: Img2Txt Example Results: Top words retrieved for several items. Correct words are marked in green.

A full comparison may be found in [21]. BiLDA performs at least as well as, and often better than, the CCA model. BiLDA model also outperforms the frequency baseline. Regarding visual features, the overall best performer is CNN as it always outperforms the SIFT feature in both precision and recall. It is remarkable that CNNs perform so well compared to the SIFT counterparts because they were not trained for this particular task. Instead, they trained on a large image classification task [2].

Txt2Img Results. Figure 3 presents recall@K. 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). Regarding the choice of model, the BiLDA model performs roughly as well as CCA.

Figure 4 presents some qualitative example 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. The query ‘wedding gown sleeveless scalloped ruffles’ retrieved wedding gowns in all four items.

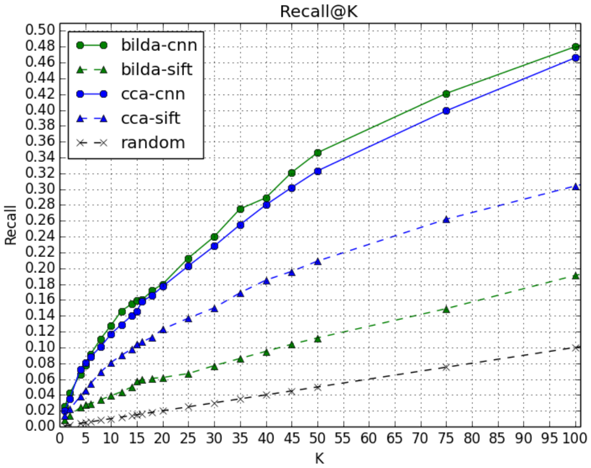

Figure 3: Txt2Img: Recall@K for POS-based vocabulary (left) and Zappos-based vocabulary (right)

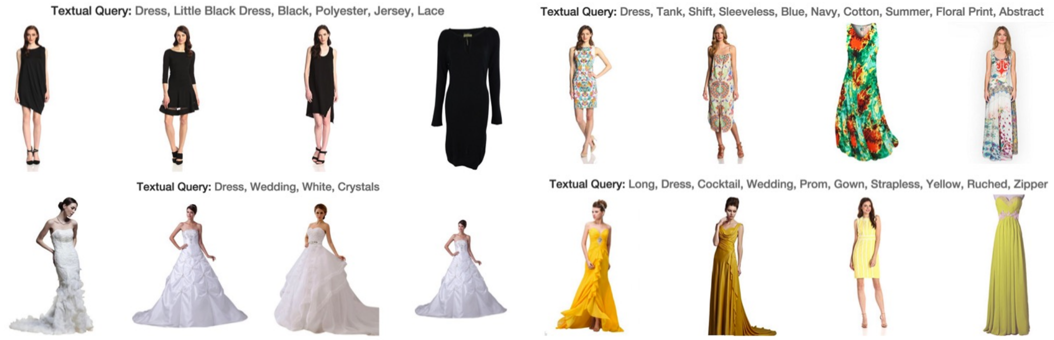
Textual Query: Dress, Little Black Dress, Black, Polyester, Jersey, Lace
Textual Query: Dress, Tank, Shift, Sleeveless, Blue, Navy, Cotton, Summer, Floral Print, Abstract
Textual Query: Dress, Wedding, White, Crystals
Textual Query: Long, Dress, Cocktail, Wedding, Prom, Gown, Strapless, Yellow, Ruched, Zipper

Figure 4: Txt2Img: Example results. Given a textual query, we show the top retrieved images

Two of them are sleeveless and three of them contain ‘scalloped ruffles’. Overall, these results are highly impressive given the difficulty of the task. We are able to correctly identify different lengths, shapes, colors and textures. We show this both quantitatively and qualitatively[4]. Full technical details and evaluation may be found in [21].

5 Conclusions

We investigated cross-modal attribute recognition in fashion items. Given a textual query composed of visual attributes, our system retrieves relevant product images, and given a product image as query, the system describes its attributes. We have implemented and compared algebraic and probabilistic graphical models to learn latent components that bridge the visual and textual features. We have experimented with different visual 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 [11], which uses SIFT-based features and a vocabulary based on part-of-speech. Inspecting the annotations our system generates, we find reasonable descriptions that capture different garment lengths, colors and textures. These results are extremely useful for navigating the multi-modal fashion data and jointly recognizing attributes in images and text.

References

[1] G. Csurka and C. Dance. Visual categorization with bags of keypoints. *Workshop on statistical learning in computer vision, ECCV*, 1(1-22):1–2, 2004.

[4] More results may be found in `http://roshi.cs.kuleuven.be/multimodal_search`

[2] 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.

[3] M. Faruqui and C. Dyer. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 462–471, 2014.

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

[5] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

[6] 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.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[8] 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.

[9] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

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

[11] 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.

[12] R. Mason and E. Charniak. Domain-specific image captioning. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 11–20, Ann Arbor, Michigan, 2014. ACL.

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

[14] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, volume 12, pages 44–49. Citeseer, 1994.

[15] H. Schmid. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop*. Citeseer, 1995.

[16] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[17] J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 1470–1477 vol.2. IEEE, 2003.

[18] A. Vedaldi and B. Fulkerson. VLFeat - an open and portable library of computer vision algorithms. In *ACM International Conference on Multimedia*, 2010.

[19] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[20] S. Zoghbi, G. Heyman, J. C. G. Carranza, and M.-F. Moens. Cross-Modal Fashion Search. In *Lecture Notes in Computer Science (LNCS) Vol. 9517, pp 367-373*, 2016.

[21] S. Zoghbi, G. Heyman, J. C. G. Carranza, and M.-F. Moens. Fashion Meets Computer Vision and NLP at E-Commerce Search. In *International Journal of Computer and Electrical Engineering (IJCEE). (Accepted)*, 2016.