Multi-Modal Fashion Product Retrieval

A. Rubio
Institut de Robòtica i Informàtica Industrial (CSIC-UPC)
Wide Eyes Technologies
arubio@iri.upc.edu

LongLong Yu
Wide Eyes Technologies
longyu@wide-eyes.it

E. Simo-Serra
Waseda University
esimo@aoni.waseda.jp

F. Moreno-Noguer
Institut de Robòtica i Informàtica Industrial (CSIC-UPC)
fmoreno@iri.upc.edu

Abstract

Finding a product in the fashion world can be a daunting task. Everyday, e-commerce sites are updating with thousands of images and their associated metadata (textual information), deepening the problem. In this paper, we leverage both the images and textual metadata and propose a joint multi-modal embedding that maps both the text and images into a common latent space. Distances in the latent space correspond to similarity between products, allowing us to effectively perform retrieval in this latent space. We compare against existing approaches and show significant improvements in retrieval tasks on a large-scale e-commerce dataset.

1 Introduction

The level of traffic of modern e-commerce is growing fast. U.S. retail e-commerce, for instance, was expected to grow 16.6% on 2016 Christmas holidays (after a 15.3% increase in 2014) (Walton, 2016). In order to adapt to these trend, sellers must provide a good experience with easy to find and well classified products. In this work, we consider the problem of multi-modal retrieval, in which a user searches for either text or images given a text or image query. Existing approaches for retrieval focus image-only and require hard to obtain datasets for training (Hadi Kiapour et al., 2015). Instead, we opt to leverage easily obtained metadata for training our model, and learning a mapping from text and images to a common latent space, in which distances correspond to similarity.

We evaluate our approach in the retrieval and classification tasks and it outperforms KCCA (Bach and Jordan, 2002) and Bag-of-word features on a large e-commerce dataset.

Figure 1: Example of a text and nearest images from the test set. Our embedding produces low distances between texts and images referring to similar objects.

2 Method

Our joint multi-modal embedding approach consists of a neural network with two branches: one for image and one for text. The image branch is based on the Alexnet (Krizhevsky et al., 2012) Convolutional Neural Network (CNN) which converts a 227×227 pixel image into a fixed-size 128-dimensional vector. The text branch is based on a multi-layer neural network and uses as an input features extracted by a pre-trained *word2vec* network which are converted into a fixed-size 128-dimensional vector. Both branches are trained jointly such that the 128-dimensional output space becomes a joint embedding by minimizing the distance between related image-text pairs and maximizing the distance between unrelated image-text pairs using the contrastive loss function (Hadsell et al., 2006) shown in 1, where v_I and v_T are two embedded vectors corresponding to the image and the text respectively, y is a label that indicates whether or not the two vectors are compatible ($y = 0$) or dissimilar ($y = 1$), and m is a margin for the negatives. Two auxiliary classification

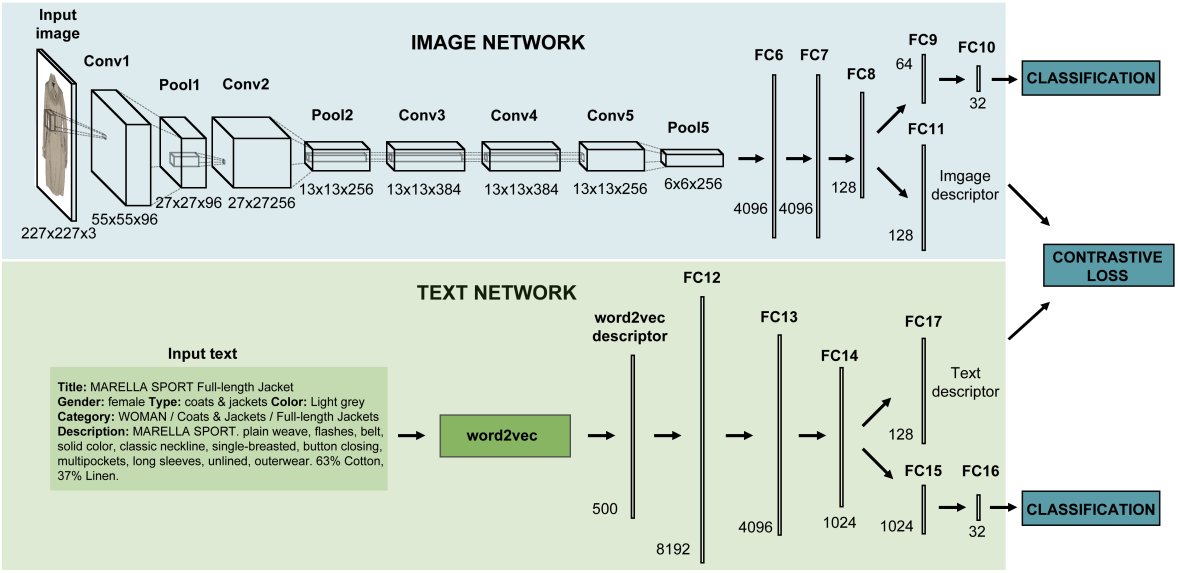

Figure 2: Architecture of the neural network used. *Conv*, *Pool* and *FC* refer to convolutional, pooling and fully connected layers, respectively. When sizes of two dimensions are equal, some of them are omitted for clarity. Fully connected layers are uni-dimensional. *Text descriptor* and *Image descriptor* are the embedded vectors describing the input text and image, respectively.

Table 1: Results of our method compared to *KCCA* and our method using *Bag of Words* for text representation.

Model	**Median rank**	
	Img v. txt	**Txt v. img**
KCCA	52.42%	46.65%
Bag of Words	4.50%	4.54%
word2vec	1.61%	1.63%

networks are also used during training that encourages the joint embedding to also encode semantic concepts. An overview can be seen in Fig. 2.

$$L_C(v_I, v_T, y) = (1-y)\frac{1}{2}\left(\|v_I - v_T\|_2\right)^2 + (y)\frac{1}{2}\left\{\max\left(0, m - \|v_I - v_T\|_2\right)\right\}^2 \quad (1)$$

3 Results

Next, we describe the results obtained by applying our method to a Fashion e-commerce dataset of $431,841$ images of fashion products with associated texts, classified in 32 categories (such as *boots*, *jewelry*, *skirt*, *shirt*, *dress*, *backpack*, *swimwear*, *glasses/sunglasses*, *shorts*, *sandals*, etc.). In order to evaluate our method, we compute all the 128-dimensional descriptors of images and texts in the testing set. Then, use the text as queries to obtain the images, and vice-versa. Looking at the position at which the exact match is, we compute the median rank for each case. The resultant values are below 2%, meaning that the exact match is usually closer than the 98% of the dataset, beating the result obtained by KCCA[1] and by our same architecture substituting the *word2vec* by a classical *Bag of Words*. We compare this metrics with two baselines: a version of our method replacing *word2vec* by *Bag of Words* and KCCA (see Table 1. We also obtained a recall value of nearly 80% for the top 5%, meaning that 80% of times the exact match for the input query is in the closest 5% results. At the same time, for the classification task we obtain accuracy values of 90% for images and 99% for texts with the *word2vec* approach.

4 Conclusions

We have presented an approach for joint multimodal embedding with neural networks with a focus on the fashion domain that is easily amenable to large existing e-commerce datasets by exploiting readily available images and their associated metadata, and can be easily used for retrieval tasks.

[1]The KCCA model has been trained with less descriptors (only 10000) due to memory errors when trying to use the whole training set

References

F. R. Bach and M. I. Jordan. 2002. Kernel independent component analysis. *JMLR*, 3(Jul):1–48.

M. Hadi Kiapour, X. Han, S. Lazebnik, Alexander C. Berg, and Tamara L. Berg. 2015. Where to buy it: Matching street clothing photos in online shops. In *CVPR*.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *CVPR*.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*.

D. Walton. 2016. *The Ultimate List of E-Commerce Stats for Holiday 2016*. `http://blog.marketingadept.com/the-ultimate-list-of-e-commerce-marketing-stats-for-holiday-2016/`. Accessed: 2017-01-23.