Automatic tagging and retrieval of E-Commerce products based on visual features

Vasu Sharma and **Harish Karnick**
Dept. Of Computer Science and Engineering
Indian Institute of Technology, Kanpur
sharma.vasu55@gmail.com hk@cse.iitk.ac.in

Abstract

This paper proposes an automatic tag assignment approach to various e-commerce products where tag allotment is done solely based on the visual features in the image. It then builds a tag based product retrieval system upon these allotted tags.

The explosive growth of e-commerce products being sold online has made manual annotation infeasible. Without such tags it's impossible for customers to be able to find these products. Hence a scalable approach catering to such large number of product images and allocating meaningful tags is essential and could be used to make an efficient tag based product retrieval system.

In this paper we propose one such approach based on feature extraction using Deep Convolutional Neural Networks to learn descriptive semantic features from product images. Then we use inverse distance weighted K-nearest neighbours classifiers along with several other multi-label classification approaches to assign appropriate tags to our images. We demonstrate the functioning of our algorithm for the Amazon product dataset for various categories of products like clothing and apparel, electronics, sports equipment etc.

Keywords: Content based image retrieval, Multi-Modal data embeddings and search, Automatic Image Annotation, E-commerce product categorization

1 Introduction

In the present day world of mass internet penetration and the advent of the e-commerce era the number of products being bought and sold online has increased exponentially in the past few years. In 2012, Business to Consumer (B2C) e-commerce sales grew 21.1% to top $1 trillion for the first time[1]. This is expected to grow steadily at the rate of around 20% and is estimated to hit $2.5 trillion by 2018 [2].

Given the explosive growth in the number of products being sold online and the relative heterogeneity in the categories these products could be allotted to, it has become physically impossible and infeasible to manually tag these products. Besides not everyone will tag the same images with the same tags. This leads to discrepancy in the kinds of tags allotted to the products. Search engines looking for products based on customers query heavily rely on these tags allotted to each image to return accurate and meaningful results to customers queries but mainly only the product images are available which is impossible for the search engine to make sense of. Besides the discrepancy in tagging leads to a lot of useful search results to get excluded.

An automatic tagging system can help take care of both of these problems and will be able to build an efficient product database querying system even if the database consists solely of visual information about the products. Such an automated systems will bring about tagging homogeneity so that similar products are tagged with the same tags. This will

[1] http://www.emarketer.com/Article/Ecommerce-Sales-Topped-1-Trillion-First-Time-2012/1009649

[2] http://www.emarketer.com/Article/Global-B2C-Ecommerce-Sales-Hit-15-Trillion-This-Year-Driven-by-Growth-Emerging-Markets/1010575

Proceedings of NAACL-HLT 2016, pages 22–28,
San Diego, California, June 12-17, 2016. ©2016 Association for Computational Linguistics

also eliminate the need for the laborious process of manually tagging such products.

The e-commerce marketplace is a truly multimodal space with visual features co-existing with product descriptions and feature specifications. To be truly effective, such a marketplace must allow the user to be able to find products based on it's visual features as well as product descriptions. This paper proposes an efficient approach to create such a Multi-Modal visual feature based product information retrieval system.
This is achieved in a 2 step process:

1. (Image to Tags) Visual features are extracted from product images and are used to automatically annotate these product images with meaningful tags.

2. (Tags to Images) Now these tags are used to query a hash table indexed on these tags and used to retrieve all images corresponding to this tag.

The rest of the paper is organized as follows. Related literature is reviewed in Section 2. Section 3 presents our proposed approach along with details of techniques used. Sections 4 presents a detailed description of obtained results for various datasets. Section 5 and 6 present Conclusion and Future Work respectively.

2 Literature Review

The e-commerce boom we have seen in the past few years has created a lot of interest in this particular area of research as such a system which can automatically tag product images is of great commercial and economic value. We saw a lot of recent work on this suggesting that this a very active area of research and is quickly gaining popularity.

One of the best work in this field is from Zoghbi et al. (2016) who similar to us use visual features to assign textual descriptors to the image and use these descriptors for query based image retrieval. In this paper they use two latent variable models to bridge between textual and visual data: bilingual latent Dirichlet allocation and canonical correlation analysis. They report their results on a self created image dataset and work only for apparel and do not generalise to other categories.

Another promising work on this problem was done by Mason et al (2013). In this work they use SIFT (Scale Invariant Feature Transform) features Lowe (2004), along with colour features, Gabor filters Fogel and Sagi (1989) and bag of HOG (Histogram of Gradient) features Dalal and Triggs (2005) as the visual features. Gibbs sampling is used to sample topic assignments for visual terms in the test image. Then they apply the technique of Latent Dirichlet Allocation (LDA) Blei et al. (2003) to generate probable captions for a given set of visual features.

Feng et al (2010), proposed a system called MixLDA which tried to automatically generate captions for news images which is similar to our task. Their model works with a bag-of-words representation and treats each article-image-caption tuple as a single document. LDA is then used to infer latent topics which generated these documents. It now uses the distribution over these latent topics to estimate the multimodal word distributions over topics for new images which in turn is used to estimate posterior of the topic proportions over visual documents. This is used to annotate the news images.

Style Finder Di et al. (2013), is another system which tries to identify visual features present in an image and uses those to build a visual feature vocabulary which could then be used for query based product retrieval. They use the women's coats dataset and extract SIFT, HoG and GIST features from these images and train binary linear SVM's to detect the presence or absence of each feature. We feel that such an approach would not be able to scale up to a very large number of labels as it tries to train a classifier for each label .

Another work by Zhan et al (2015) is also relevant to this problem. They try to automatically tag image features for shoe images by first identifying the viewpoint and then use view-specific part localization model based on the prior knowledge of the shoe structures under different viewpoints. Finally, they use a SVM classifier on low level features extracted from these localized shoe parts, which is ultimately used for attribute prediction. Here their approach is restricted to only shoe images and cant scale to large number of images or to images of other product categories.

Whittlesearch approach Kovashka et al. (2015) was another interesting work where the authors try to cater to user queries based on relative strengths of visual attributes in a product image. First they learn a ranking function to predict the relative strength of certain visual attributes. This is achieved by casting estimation of such a ranking function as a large-margin formulation of Joachims Tsochantaridis et al. (2005) with suitable constraints, which could then easily be solved. The users can then query for products which have less or more of a certain attribute than a given image and are shown a ranked list of such products based on this ranking function. Based on the user feedback, they update their scoring function using what they call as the Relative Attribute Feedback approach.

Several other SVM based approaches on a variety of visual features like SIFT, HoG, SURF etc have also been proposed Rothe et al. (2015), Li (2010), Gangopadhyay (2001), Zeng et al. (2014), Tadse et al. (2014). However we note that none of these approaches attempt to generalize to the classification of all varieties of products and restrict themselves to certain product classes (mainly apparel and clothing). Besides, they do not operate on a multi label setting and hence need one-vs-all type classifiers to detect multiple labels which do not scale up for large number for possible tags. Our approach on the other hand is extremely adaptive and is able to identify tags for multiple categories of products. Besides we directly deal with tag annotation as a multi label problem allowing our approach to scale up to a large number of tag categories.

3 Approach

In this section we elaborate upon the approach we used to build the automatic annotation system and then how we use these tags to build an efficient tag based product retrieval system.

3.1 Dataset

Lack of open availability of a dataset is one of the biggest problems which hinders the development of effective automatic tagging systems for e-commerce products. Most of the data present on the web is highly unstructured and lacks proper labels and hence cant be effectively used. Even when the labels are there, they are extremely noisy and cant be relied upon. In our paper we present the results on the Amazon e-commerce product dataset McAuley et al. (2015b), McAuley et al. (2015a) which contains images of various product categories and their meta data which we parse to obtain the tags associated with each image. For this paper we demonstrate our approach for apparels and clothing, electronics and sports equipment categories and show that the approach scales up to large number of tags and performs well on a wide category of products. Images of these categories are tagged with a total of 1664, 886 and 2224 possible tags respectively.

3.2 Feature Extraction from Images

Researchers in the computer vision domain have frequently used Scale Invariant Feature Transform (SIFT) vectors, Histogram of Gradients (HoG) vectors, SURF vectors, Gabor filters etc. to extract useful features from images. However in our project we use features extracted from higher layers of a very Deep Convolutional Neural Network to serve as our visual features. Here we use the 19 layer deep VGG network Simonyan and Zisserman (2014) and it is trained on the Imagenet dataset first. We then use this trained network on the Imagenet dataset Russakovsky et al. (2015) and use the concept of 'Transfer Learning' Yosinski et al. (2014) to train it on our Amazon product dataset next. We also experiment with the VGG-16 and Googlenet networks, but VGG-19 features give us the best performance and hence we use them for our paper. The network structure is presented in Image 1.

It is observed that Deep Convolutional Neural networks have the ability to learn useful feature representations by non-linearly projecting the input images into a discriminative subspace where similar images tend to have similar intermediate feature representations while non-similar images are projected far from each other. This trend is independent of the dataset it is trained on and hence a network trained on Imagenet network too is able to learn useful feature representations for the Amazon product dataset when trained on it. Oquab et al. (2014), Bengio (2012)

Hence we could use VGG-19 models pre-trained on Imagenet and then adapt them for our dataset. This cuts down our training time tremendously as training

Deep CNN models from scratch is computationally very expensive. We use the 4096 Dimensional features from the last fully connected hidden layer of the VGG net as features to represent a given visual image.

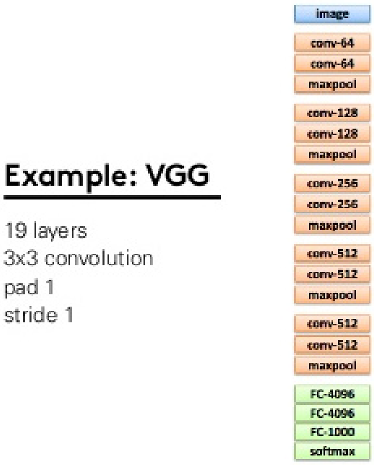


Figure 1: The VGG-19 Network
The conv layers represent convolution layers, Maxpool are Max Pooling layers and FC are Fully connected layers

3.3 Automatic tag assignment to images

After having obtained the extracted feature vectors for each images from the Deep CNN's, our next task to automatically assign all relevant tags to each image. This is essentially a multi label problem which is still a very active area of research till date. We avoid using one-vs-all classifiers like SVM's etc as such classifiers can't be scaled up for a very large number of tag categories which we observe in the case of e-commerce products. In this paper we use Weighted K-Nearest Neighbours approach with weights for each neighbour as inverse of the distance of the neighbour from the query point. For each query point we evaluate it's K nearest neighbours from the training set and evaluate the weighted average of the presence of each particular tag from among these K neighbours. Ultimately, we allot the D tags with the highest weighted average to the query image.

The 'presence score' of the occurrence of the tag t_i in the query image x could be calculated as:

$S(t_i|x) = \sum_{j=1}^{K} \frac{I(i,j)}{d_{x,n_j}} / \sum_{j=1}^{K} \frac{1}{d_{x,n_j}}$

Where d_{x,n_j} is the distance between query point and the j^{th} neighbour, $I(i,j)$ is an indicator variable which is defined as follows:

$$I(i,j) = \begin{cases} 1 & \text{if } n_j \text{ has tag } t_i \\ 0 & \text{Otherwise} \end{cases}$$

This is how we do tag allocation to the various images. K-Nearest Neighbour approach is computationally very efficient and can be scaled up to a very large number of tags without a very large increase in computational complexity.

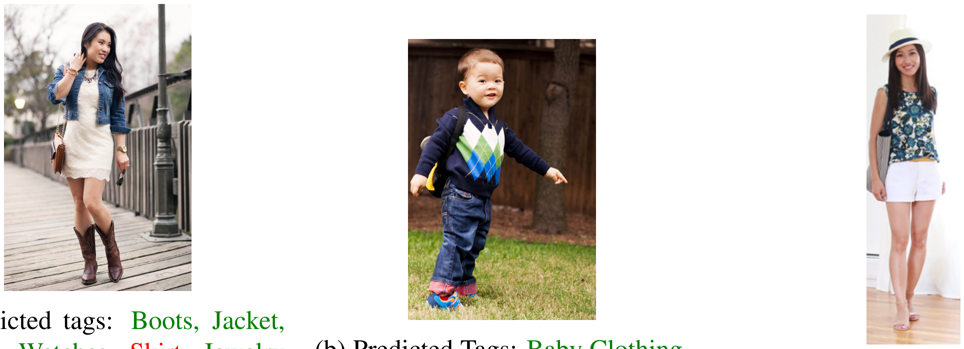
3.4 Querying product based on tagged visual features

Once we have a fully tagged dataset of product images we can easily store the tags and images which have been tagged with it in a hash table with the tag as the key. We also store the probability of the presence of this tag in each image. The image list corresponding to each tag is then sorted according to this probability score and stored and the value for the entry with the tag as the key.
Now when the user queries this database looking for products with a certain tagged feature, then the hash table is looked up for that tag and the images corresponding to those tags are returned in the order of higher probability score first. This helps us build a lightning fast retrieval system for tagged visual feature based queries for the various e-commerce products.

4 Results

4.1 Annotation Results

We present our results on 3 different categories of products i.e. apparel and clothing, electronics and sports equipment. The metrics we use to measure the performance of our method are Precision, Recall and F_1-score which are computed over all possible tag categories. These metric are defined as follows:
$Precision = TP/(TP + FP)$

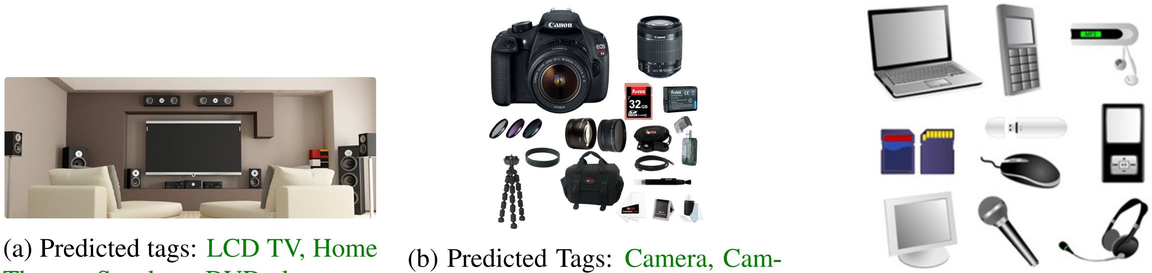


(a) Predicted tags: Boots, Jacket, Western, Watches, Skirt, Jewelry Accessories
Actual Tags:Boots, Jacket, Western, Watches, Jewelry Accessories, Sunglasses, Handbags & Purses

(b) Predicted Tags: Baby Clothing, Jeans, Shoes, Sweater
Actual Tags: Baby Clothing, Jeans, Shoes, Sweater

(c) Predicted Tags: Hats,Skirt, Handbags & Purses, Floral Top
Actual Tags: Hats, Skirt, Handbags, Floral Top, Heels

Figure 2: Examples from Apparel category

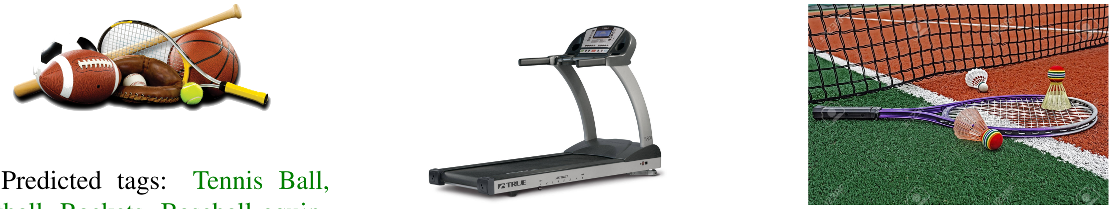


(a) Predicted tags: LCD TV, Home Theater, Speakers, DVD player
Actual Tags: LCD TV, Home Theater, Speakers, DVD player, Amplifier

(b) Predicted Tags: Camera, Camera Lens, Memory card, USB drive, Camera Stand
Actual Tags: Camera, Camera Lens, Memory card, USB drive, Camera Stand, Stylus, USB cable,Camera Bag

(c) Predicted Tags: Mobile, Laptop, Mouse, Mike, Memory Card, Headphones, Monitor, MP3 Player
Actual Tags: Laptop, Mouse, Mike, Memory Card, Headphones, Monitor, MP3 Player, Calculator

Figure 3: Examples from Electronics category



(d) Predicted tags: Tennis Ball, Football, Rackets, Baseball equipment, Basketball, Rugby
Actual Tags: Tennis Ball, Football, Rackets, Baseball equipment, Basketball, Rugby ,Glove

(e) Predicted Tags: Gym equipment, Treadmill, Fitness equipment
Actual Tags: Gym equipment, Treadmill, Fitness equipment

(f) Predicted Tags: Racket, Shuttlecock, Net, Tennis, Cage
Actual Tags: Racket, Shuttlecock, Net, Badminton equipment

Figure 3: Examples from Sports category

$$Recall = TP/(TP + FN)$$

$$F_1 Score = \frac{2 \cdot Precision \cdot Recall}{(Precision + Recall)}$$

Where TP stands for True positives, FP stands for False Positives and FN stands for False Negatives.

We present our results for 4 different values of K, applying the inverse distance weighed KNN in each case. Our results are shown in Table 1

Product Category	K	F_1 score	Precision	Recall
Apparel and Clothing	1	0.282	0.284	0.283
	3	0.343	0.536	**0.252**
	5	**0.345**	0.603	0.242
	7	0.340	**0.635**	0.232
Electronics	1	0.317	0.317	0.316
	3	0.396	0.621	**0.291**
	5	**0.407**	0.706	0.286
	7	0.406	**0.743**	0.280
Sports	1	0.251	0.252	0.250
	3	0.329	0.626	**0.223**
	5	**0.336**	0.765	0.215
	7	0.335	**0.819**	0.210

Table 1: Tag Annotation results

Some sample images and tags allotted to them for each category are shown in images 2, 3, 3. To the best of our knowledge, these results are the best on this given dataset. The tagged images clearly show that the approach works very well and is able to identify tags correctly despite the large amount of noise in the data and the large number of possible tags.

4.2 Retrieval Results

Once the Tag annotation was completed we set up the hash table based indexing system with the tags as keys and a list of images relevant to that tag sorted in order of 'presence score' of occurrence of that tag. We use this to perform our retrieval. We create a list of 1000 tag queries for each category and use this retrieval system to obtain the relevant images. The retrieval accuracy depends on the accuracy of tagging. We note that by retrieval times were blazingly fast with all 1000 queries for each product category. The retrieval times are presented in Table 2. Clearly the time complexity remains more or less constant for each of the categories despite the varying number of labels denoting that the retrieval times are constant with respect to increasing dataset size and number of labels.

Table 2: Performance of the Content Based Image Retrieval System for a list of 1000 query tags

Product Category	Retrieval Time(s)
Apparel and Clothing	0.083
Electronics	0.095
Sports	0.081

5 Conclusions

In this paper we proposed an Automatic tagging approach for e-commerce products by making use of it's visual features and using these tags to build an efficient query based product retrieval system. We demonstrated that the system performs extremely well and to the best of our knowledge, it outperforms all other systems for automatic e-commerce product tagging on this dataset. Besides the approach is highly scalable, catering to a very large number of tags and products and could easily generalize to multiple product categories and performs well on each category.

The retrieval system built on top of this is also extremely fast and is able to obtain meaningful results at lightning fast speeds.

6 Future Work

We plan to extend this work by incorporating better multi label algorithms which could provide even better performances. We are also exploring alternate feature representation techniques which could provide us with further semantic information. One such representation we plan to explore is to use the activation values from multiple layers of the VGG network as we know that each layer of the network learns a certain kind of distinguishing feature. A combination of such features might provide superlative performance over just using the features from a single layer.

References

[Bengio2012] Y. Bengio. 2012. Deep learning of representations for unsupervised and transfer learning. *JMLR: Workshop on Unsupervised and Transfer Learning*, 27:17–37.

[Blei et al.2003] D. Blei, A. Ng, and M. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning*, 3:993–1022.

[Dalal and Triggs2005] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. pages 886–893.

[Di et al.2013] W. Di, C. Wah, A. Bhardwaj, R. Piramuthu, and N. Sundaresan. 2013. Style finder: Fine-grained clothing style detection and retrieval. pages 8–13.

[Feng and Lapata2010] Y. Feng and M. Lapata. 2010. How many words is a picture worth? automatic caption generation for news images. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1239–1249.

[Fogel and Sagi1989] I. Fogel and D. Sagi. 1989. Gabor filters as texture discriminator. *Biological Cybernetics*, 61(1).

[Gangopadhyay2001] A. Gangopadhyay. 2001. An image-based system for electronic retailing. *Decision Support Systems*, 32.

[Kovashka et al.2015] A. Kovashka, D. Parikh, and K. Grauman. 2015. Whittlesearch: Interactive image search with relative attribute feedback. *Int. J. Comput. Vision*, 115(2):185–210.

[Li2010] Jing Li. 2010. The application of cbir-based system for the product in electronic retailing. *IEEE 11th International Conference on Computer-Aided Industrial Design Conceptual Design*, 2:1327–1330.

[Lowe2004] David Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110.

[Mason and Charniak2013] Rebecca Mason and Eugene Charniak. 2013. Annotation of online shopping images without labeled training examples. *Proceedings of the NAACL HLT Workshop on Vision and Language (WVL 13)*, pages 1–9.

[McAuley et al.2015a] J. McAuley, R. Pandey, and J. Leskovec. 2015a. Inferring networks of substitutable and complementary products. *Knowledge Discovery and Data Mining*.

[McAuley et al.2015b] J. McAuley, C. Targett, J. Shi, and A. van den Hengel. 2015b. Image-based recommendations on styles and substitutes. *SIGIR*.

[Oquab et al.2014] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. 2014. Learning and transferring mid-level image representations using convolutional neural networks. *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1717–1724.

[Rothe et al.2015] R. Rothe, M. Ristin, M. Dantone, and L. Van Gool. 2015. Discriminative learning of apparel features. pages 5–9.

[Russakovsky et al.2015] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and Li Fei-Fei. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

[Simonyan and Zisserman2014] K. Simonyan and A. Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

[Tadse et al.2014] R. Tadse, L. Patil, and C. Chauhan. 2014. Review on content based image retrieval for digital library using text document image. *International Journal of Computer Science and Mobile Computing*, 4:211–214.

[Tsochantaridis et al.2005] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. 2005. Large margin methods for structured and interdependent output variables. *J. Mach. Learn. Res.*, 6:1453–1484.

[Yosinski et al.2014] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. 2014. How transferable are features in deep neural networks? *CoRR*, abs/1411.1792.

[Zeng et al.2014] K. Zeng, N. Wu, and K. Yen. 2014. A color boosted local feature extraction method for mobile product search. *Int. J. on Recent Trends in Engineering and Technology*, 10.

[Zhan et al.2015] Huijing Zhan, Sheng Li, and A. C. Kot. 2015. Tagging the shoe images by semantic attributes. pages 892–895.

[Zoghbi et al.2016] Susana Zoghbi, Geert Heyman, Juan Carlos Gomez, and Marie-Francine Moens. 2016. Fashion meets computer vision and nlp and e-commerce search. *International Journal of Computer and Electrical Engineering*.