Euclidean Image Embedding in view of Similarity Ranking in Auction Search by Image

Riku Togashi Hideyuki Maeda Vibhor Kanojia Kousuke Morimoto Sumio Fujita
Yahoo Japan Corporation
{rtogashi,hidmaeda,vkanojia,kmorimot,sufujita}@yahoo-corp.jp

ABSTRACT

We propose an Euclidean embedding image representation, which serves to rank auction item images through wide range of semantic similarity spectrum, in the order of the relevance to the given query image much more effective than the baseline method in terms of a graded relevance measure. Our method uses three stream deep convolutional siamese networks to learn a distance metric and we leverage search query logs of an auction item search of the largest auction service in Japan. Unlike previous approaches, we define the inter-image relevance on the basis of user queries in the logs used to search each auction item, which enables us to acquire the image representation preserving the features concerning user intents in real e-commerce world.

Keywords

Search by image, auction search, search query logs, Euclidean embedding.

1. INTRODUCTION

Due to the wide usage of smart devices with a high definition camera, Content Based Image retrieval (CBIR), or search by image either from a SNS site or taken by the camera of user's device becomes useful means to specify their information needs in the image search. When you find a favorite celebrity wearing a pretty sexy jumpsuit on Instagram, you might want to buy something similar for yourself or your partner, by seeking an item on an e-commerce or auction site. Despite the recent breakthrough in various image processing tasks by deep learning approaches[1][4], CBIR is still a challenging task, due to the difficulties in identifying users' search intents from the image. On the above example, for example, the system would not understand, from your query image, that you are interested neither in the celebrity's scandal story nor in her hair style but in her clothes.

While a general purpose CBIR faces serious difficulties in detecting image regions relevant to search user intent, shopping/auction item search systems are able to identify user's intents by applying search log based methods frequently used by query suggestion techniques in web search. We propose a search by image system for a large scale auction service, that directly learns a mapping from auction item images to a low-dimensional Euclidean embedding by a triplet loss function[1][2][3], where the L_2 distance function represents dissimilarity between auction item images. Because of the difficulty in preparing a large number of triplets, there is no previous work addressed such issues while we leverage the bags of tokens of search queries in the logs as annotations of images, representing real search user intents. For the best of our knowledge, this is the first large scale evaluation of search by image based on the relevance to real users' search intents that adopts image indexing based on Euclidean image embedding.

WWW'17 Companion, April 3-7, 2017, Perth, Australia
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3054262

2. METHODS

At the training time, three stream deep convolutional siamese networks input triplet of 256×256 fixed size RGB images of auction items, and minimize the triplet loss defined as follows.

2.1 Triplet loss

We adopt the method proposed in [2] for Euclidean image embedding. Triplet loss is formulated as follows:

$$L = \sum_{i}^{N} \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right]_+ \quad (1)$$

The embedding is represented by $f(x) \in \mathbb{R}^d (d \in \mathbb{Z}^+)$, where x is an image and d is the dimension of the representation. Here x_i^a is the anchor image, x_i^p is the positive image and x_i^n is the negative image of the i_{th} triplet. α denotes the margin. This loss function trains d-dimensional representations to ensure that the positive is closer to the anchor than the negative.

2.2 Definition of triplet: relative P/N

Triplets are prepared by using search query logs of Yahoo! JAPAN Auction service[1]. All images are annotated by the query in response to which the auction item containing the image is clicked more than five times; if there are more than one queries, we adopt the longest query on the basis of the number of tokens, e.g. keywords separated by a whitespace. Let V be the vocabulary of query tokens and x, an image, the

[1]Unfortunately we are unable to show example images due to copyright reasons. You may see example images on the service site, http://auctions.yahoo.co.jp/

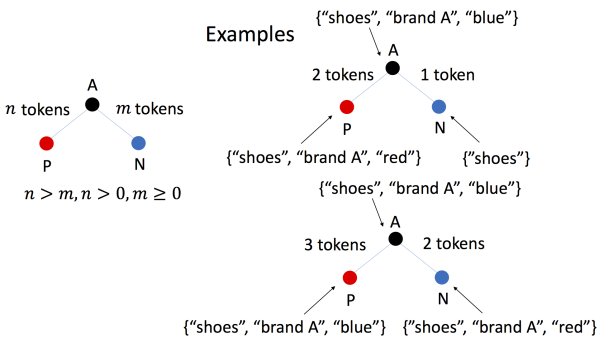

Figure 1: Triplets with relative P/N corresponding to the length of matched query tokens.

bag of tokens of thus selected annotation query, $w(x) \subseteq 2^V$ is used as the descriptor of the image.

Unlike classification tasks where the goal is learning a class separation hypersurface, image representations for ranked retrieval are preferably able to discriminate every images throughout the semantic space of the target domain, e.g. "red Ferragamo pumps" are closer to "blue Ferragamo pumps" rather than simple "blue pumps". Although image metadata is widely used for the preparation of image examples[3], we adopted queries used to search auction items where the focus of user interests in item search is well represented.

As shown in Figure 1, we sample a triplet based on partial matching of bags of tokens: assume an image x_i^a as anchor of the i_{th} triplet, we sample two images where the positive image x_i^p shares more common tokens with the anchor than the negative image x_i^n so that an image with the descriptor, {"$shoes$","$brandA$", "red"} is closer to that with {"$shoes$","$brandA$","$blue$"} rather than {"$shoes$"}. Depend on whether considering partial matching as positive or not, we examine two types of P/N triplet definition in the hope that the partial matching enables the ranker a smooth ranking through wider range of semantic similarity spectrum. More formally, we define :

- **Absolute P/N triplet** as (x_i^a, x_i^p, x_i^n) where $w(x_i^a) = w(x_i^p)$ and $w(x_i^a) \neq w(x_i^n)$, and
- **Relative P/N triplet** as (x_i^a, x_i^p, x_i^n) where $w(x_i^a) \supseteq w(x_i^a) \cap w(x_i^p)$ and $w(x_i^p) \supset w(x_i^a) \cap w(x_i^n)$.

3. EXPERIMENTS

Training: We leverage 29,552,528 query and clicked image pairs collected from Yahoo! JAPAN Auction search logs of October 2016, where we acquired 1,026,824 annotated images, which we partitioned into three sets, 80% for training and search target, 10% for validation and 10 % for test query images.

Evaluation: We compared three feature extraction methods as follows:

1. Intermediate representation of VGG16[4],
2. Euclidean embedding by absolute P/N triplet loss,
3. Euclidean embedding by relative P/N triplet loss(Ours).

Search by image is implemented by a nearest neighbor search on the feature spaces of the above three methods. We evaluated the retrieved image ranking against query images in the held out set, by normalized Discounted Cumulated Gain (nDCG), the measure commonly used in search evaluations. We define the grade based on descriptor matching.

$$DCG_q@k = \sum_{i=1}^{k} \frac{2^{rel_{q,i}} - 1}{log_2(1+i)} \quad (2)$$

$$NDCG_q@k = \frac{DCG_q@k}{idealDCG_q@k} \quad (3)$$

where, $rel_{q,i}$ is the grade of relevace between the query image q and the i_{th} ranked image.

4. RESULTS AND DISCUSSIONS

Table 1: Search by Image Evaluations

method	k=5	k=50	k=500	k=1000
VGG16	0.272	0.263	0.328	0.351
abs P/N	0.305	0.316	0.426	0.465
rel P/N	**0.377**	**0.440**	**0.642**	**0.717**

Table 1 shows the results of search by image evaluation in nDCG@k where the cutoff point k is 5, 50, 500 and 1000. We might attribute the poor performance of VGG16 to the absence of an appropriate distance function; the L_2 distance used here for comparison purpose is not suitable for ranking especially when k is large. With a small k, it approximately works well in the vicinity of the query. Meanwhile, the Euclidean embedding method performs much better; the relative P/N is slightly better than the absolute P/N when k is small whereas the former's performance significantly improves when k becomes larger. This endorses our hypothesis that the relative P/N triplet enables a smooth ranking through the wide range of semantic spectrum, from the most similar to the least similar, by improving the ranking at the middle to lower ranges. Although, empirically, both the relative and absolute successfully rank red shoes of the same brand as the query image, followed by red shoes with different brands, only the relative managed to promote blue shoes with the same brand among different brand red shoes.

Hence, thanks to an adequate distance metric inherent to it and the relative P/N triplet, an Euclidean embedding with triplet loss approach is a very effective way to implement search by image in auction or e-commerce services. We leveraged search logs of auction item search and proposed the relative P/N definition of the triplet, which is especially adequate for the ranking through wide range of semantic spectrum in the auction domain.

5. REFERENCES

[1] A. Gordo, J. Almazan, J. Revaud, and D. Larlus. Deep image retrieval: Learning global representations for image search. *arXiv*, 2016.

[2] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. *IEEE Conf. on CVPR*, 2015.

[3] E. Simo-Serra and H. Ishikawa. Fashion style in 128 floats:joint ranking and classification using weak data for feature extraction. *IEEE Conf. on CVPR*, 2016.

[4] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv*, 2014.