Bag-of-Words Based Deep Neural Network for Image Retrieval

Yalong Bai[1]
ylbai@mtlab.hit.edu.cn

Wei Yu[1]
w.yu@hit.edu.cn

Tianjun Xiao[2]
xiaotianjun@pku.edu.cn

Chang Xu[3]
changxu@nkjl.nankai.edu.cn

Kuiyuan Yang,
Wei-Ying Ma[4]
{kuyang,
wyma}@microsoft.com

Tiejun Zhao[1]
tjzhao@hit.edu.cn

[1]Harbin Institute of Technology, Harbin, China [2] Peking University, Beijing, China
[3]Nankai University, Tianjin, China [4]Microsoft Research Asia, Beijing, China

ABSTRACT

This work targets image retrieval task hold by MSR-Bing Grand Challenge. Image retrieval is considered as a challenge task because of the gap between low-level image representation and high-level textual query representation. Recently further developed deep neural network sheds light on narrowing the gap by learning high-level image representation from raw pixels. In this paper, we proposed a bag-of-words based deep neural network for image retrieval task, which learns high-level image representation and maps images into bag-of-words space. The DNN model is trained on the large scale clickthrough data, and the relevance between query and image is measured by the cosine similarity of query's bag-of-words representation and image's bag-of-words representation predicted by DNN, the visual similarity of images is computed by high-level image representation extracted via the DNN model too. Finally, PageRank algorithm is used to further improve the ranking list by considering visual similarity of images for each query. The experimental results achieved state-of-the-art performance and verified the effectiveness of our proposed method.

Categories and Subject Descriptors

I.4 [**IMAGE PROCESSING AND COMPUTER VISION**]: Miscellaneous

General Terms

Algorithms; Experimentation

Keywords

Image Retrieval; Deep Neural Network; Click Data

MM'14, November 3–7, 2014, Orlando, Florida, USA.

ACM 978-1-4503-3063-3/14/11 ...$15.00.
http://dx.doi.org/10.1145/2647868.2656402 .

1. INTRODUCTION

According to settings of MSR-Bing Grand Challenge, in this work, we developed a system to assess the relevance between image and query pair for image retrieval. That is, given a pair of query and image, the system could produce a floating-point score that reflects how relevant the images could describe the query. The database of MSR-Bing Grand Challenge contains 11.7 million of queries and 1 million of images which were collected from the user click log of Bing image Search in the EN-US market [4].

Bridging the semantic gap between visual representation and textual representation is the core issue of getting better image retrieval. The traditional solutions compute the relevance between images and textual descriptions based on features extracted from the surrounding text of images or some low-level image representation features (such as SIFT, HOG, and LBP), where the information from the image itself is still far from fully used. With large number of training data, deep neural network has demonstrated its great success in learning high-level image representation from raw pixels. In this task, large scale clickthrough is available as training data, thus we can learn high-level image representations by deep neural networks.

Deep neural network has been used for image retrieval task in [1], which trained a set of classifiers for different queries based on high-level image features learned by a transfer learning DNN architecture. But this kind of solution is hard to scale up when dealing with a huge number of queries considering that training classifier for each query is impractical. In this work, we proposed a new deep neural network architecture which can handle large scale of queries and be directly used for computing the relevance between image and query. Our model aims two main objectives: 1) query-image relevance assessment; 2) high-level visual features based image-image similarity assessment.

For query-image relevance assessment, our DNN maps the image from raw pixels into a bag of words space, the relevance of a query-image pair is computed via their cosine similarity in the bag-of-words space. The high-level visual features of each image can also be extracted by our DNN model, the image-image similarity is computed based on these high-level visual features. Finally, the scores of query-image

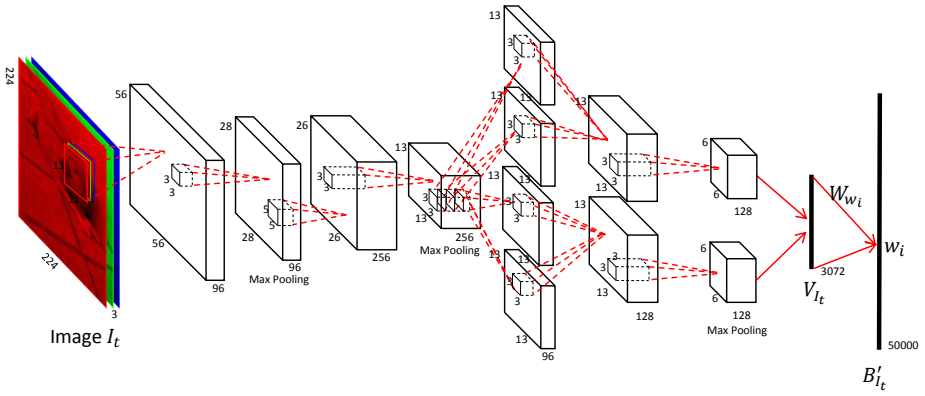

Figure 1: The architecture of bag-of-words based deep neural network.

relevance and image-image similarity are both considered to get final ranking score.

The rest of the paper is structured as follows. In Section 2, we introduce the architecture of Bag-of-Words based DNN and do some analysis on our proposed model. Three different ranking methods are presented on section 3. Section 4 introduces the real experimental verification of our proposed method, Finally, we conclude the paper in section 5.

2. BAG-OF-WORDS BASED DEEP NEURAL NETWORK

Bag-of-words is a simplifying representation widely used in natural language processing and information retrieval task, In this work, we regard each query as a short document built up by several individual words. Consider a query set with m queries $\{q_1, q_2, ..., q_m\}$, there are totally $|W|$ split single words $\{w_1, w_2, ..., w_{|W|}\}$, we firstly compute idf score for each single word as following

$$idf_{w_i} = log \frac{m}{|j : w_i \in q_j|} \tag{1}$$

where $|j : w_i \in q_j|$ is the number of queries that contains w_i. To decrease the computational complexity of our system, only top 50000 words with high frequency are kept as the vocabulary of bag-of-words representation. Given an image I, all of its related queries are merged as one document D_I, D_I can be regarded as the textual description for I. To represents D_I as bag-of-words B_{D_I}, the value of each word is set as the accumulation of corresponding idf value. For example, there is a puppy image clicked under query "small dog", "white dog" and "dog chihuahua", while $idf_{dog} = 0.3$, $idf_{small} = 0.1$, $idf_{white} = 0.05$, $idf_{chihuahua} = 0.9$, the bag-of-words representation of this image should be "dog:0.9; white:0.05; small:0.1; Chihuahua:0.9". The image and related bag-of-words representation vector pairs $<I, B_{D_I}>$ are used as supervised data to train our deep neural network.

2.1 Overall Architecture

As depicted in Fig. 1, the proposed bag-of-words based deep neural network (BoWDNN) contains four convolutional layers and two fully-connected layers, drop out with rate 0.5 is added to the first fully-connected layer for avoiding overfitting [3]. The output of last fully-connected layer is fed to cosine similar loss function to represent the relevance degree between image and query. The first, second and fourth convolutional layers are followed by max-pooling layers. Response-normalization layers follow the first and second convolutional layers. To accelerate the learning process, ReLU[2] non-linearity is applied as activation function in every convolutional layers and first full-connected layer.

As the input size of deep neural network is fixed, we resize each image with small dimension to 256 and crop out the central 256×256 patch. During training, 224×224 patches will be randomly extracted (four corner patches and one center patch for each image) from the 256×256 images. The first convolutional layer filters the $224 \times 224 \times 3$ input RGB image with 96 kernels of size $11 \times 11 \times 3$ with a stride of 4 pixels. The second convolution layer is with 256 kernels of size $5 \times 5 \times 96$. The third convolutional layer has 384 kernels of size $3 \times 3 \times 64$ which are split into four groups, and the fourth convolutional layer has 256 kernels of size $3 \times 3 \times 192$ which are split into two groups. The first full-connected layer has 3072 neurons, and the number of neurons in the last full-connected layer is determined by the vocabulary size of bag-of-words representation. The loss function of BoWDNN based on the cosine similarity between a predicted image's bag-of-words representation and the corresponding queries' bag-of-words representation.

2.2 Training Objective

The intent of our Bag-of-Words based deep neural network is to map images into their corresponding bag-of-words representation vectors. The objective function is to maximize the following average cosine similarly

$$J = \frac{1}{T} \sum_{t=1}^{T} Rel(B'_{I_t}, B_{D_{I_t}}) \tag{2}$$

where T is the amount of images in training data, the bag-of-words relevance of image-query pair $Rel(B'_{I_t}, B_{D_{I_t}})$ is defined as the cosine similarity in the bag-of-words space

$$Rel(B'_{I_t}, B_{D_{I_t}}) = \frac{B'_{I_t} \cdot B_{D_{I_t}}}{\|B'_{I_t}\| \|B_{D_{I_t}}\|} \tag{3}$$

where $B_{D_{I_t}}$ is the bag-of-words representation for queries of image I_t, B'_{I_t} is the bag-of-words vector predicted by

BoWDNN as following

$$B'_{I_t} =< r_{I_t,w_1}, r_{I_t,w_2}, ..., r_{I_t,w_n} > \quad (4)$$

where r_{I_t,w_i} indicates the relevance between image I_t and word w_i, given by

$$r_{I_t,w_i} = V_{I_t} \cdot W_{w_i} \quad (5)$$

where V_{I_t} is the high-level image representation features learned by BoWDNN, W_{w_i} is the weights to compute the relevance between V_{I_t} and word w_i as shown in Fig. 1.

For an image I and query q, their relevance is measured via Eq. 3.

2.3 Visual Similarity Measurement

Figure 2: Visual similarity images measured by the learned high-level image feature. The first row shows some query images, the left rows are their visual similarity images ordered by the similarity degree.

The high-level image features learned by BoWDNN are employed to measure the visual similarity. Given an image pair I_a and I_b, feed them into the BoWDNN architecture as Fig. 1 shown, the outputs of first full-connected layer V_{I_a} and V_{I_b} are used as visual features of I_a and I_b respectively. The visual similarity between I_a and I_b is define as:

$$Rel(V_{I_a}, V_{I_b}) = \frac{V_{I_a} \cdot V_{I_b}}{\|V_{I_a}\| \, \|V_{I_b}\|} \quad (6)$$

Fig 2 shows the images which are similar to the query image, all of these images were retrieval from whole training dataset by using the above visual similarity measurement.

3. RANKING METHODS

Our proposed ranking method as illustrated in Fig. 3 which contains two main parts: bag-of-words based query-image relevance assessment and extracting high-level image features for computing image-image visual similarity. The outputs of these two parts will be used as the inputs of our ranking methods.

In this section, we introduce three different approaches for compute the score $s(I, q)$ an image-query pairs (I, q).

1) BoWDNN-R (Bag-of-Words similarity based ranking method). The idea of this method is to measure the image-query relevance based on the similarity between the query bag-of-words representation and image bag-of-words representation learned by BoWDNN. The relevance scores could be formulated as $s(I, q) = Rel(B'_I, B_q)$, where B'_I is the bag-of-words vector computed by BoWDNN, and B_q is the bag-of-words representation of query q.

2) BoWDNN-S (Visual similarity based ranking method). Considering that the given image-query pairs in this MSR-Bing Grand Challenge comes from users' click through, the images relevant to a query are tend to be similar. That is, an image's relevance to a query can be measured by how other candidate images similar to this image. In the other words, the visual similarity (connections) among images deserve the relevance scores. Similar to the modified PageRank model(MPM) in [6] , the relevance scores can be formulated as $s(I, q) = 1/N \sum_{j=1}^{N} sim(I_j, I) = 1/N \sum_{j=1}^{N} Rel(V_{I_j}, V_I)$, where I_j is one of the top-N most similarity images. To avoid the influence of noisy images, only the top-5 most similar images are considered.

3) BoWDNN-J (Joint ranking method). Inspired by the website ranking based on hyperlinks, we use the PageRank[5] to fuse the both bag-of-words similarity scores and visual similarity scores. Firstly, we build a visual similarity matrix P for each query q, let $P(i, j) = sim(I_i, I_j)$, where I_i and I_j is two images related to query q. Then normalize the scale of $P(i, j)$ from [-1,1] to [0,1], and iterate $s(I_i, q) = \alpha \cdot s(I_i, q) + (1 - \alpha) \sum_{n=1}^{N} s(I_n, q) \cdot s(I_i, I_m)$ for several times, where α is a parameter in [0,1), I_m is one of the N images which are most similar to I_i, $s(I_i, q)$ is initialized to the image-query relevance score of BoWDNN-R. Finally, rank each image I_i according its ranking score $s(I_i, q)$.

4. EXPERIMENT SETUP AND RESULTS

We used the entire development set including 1000 queries and the final test set to evaluate our proposed Bag-of-words based deep neural network on the image retrieval task. The evaluation metric is Discounted Cumulated Gain (DCG). To compute DCG, for each query, we sort the images based on the floating point scores produced by BoWDNN-R, BoWDNN-S and BoWDNN-J. DCG for each query is calculated as

$$DCG_{25} = 0.01757 \sum_{i=1}^{25} \frac{2^{rel_i - 1}}{log_2(i + 1)} \quad (7)$$

where $rel_i = Excellent = 3, Good = 2, Bad = 0$ is the human judged relevance for each image with respect to the query, and 0.01757 is a normalizer so that the perfect ranking has a NDCG value of 1. The final result is averaged over all queries in the test set.

Our BoWDNN was trained on single NVIDIA Tesla K20Xm 6GB GPU, limited by GPU memory size, we only used top-50000 words with high frequency as the vocabulary of bag-of-words representation.

Table 4 summarizes the DCG@25 of different approaches conducted on the DEV and TEST dataset. Both MPM (the winner of last year in MSR-Bing Grand Challenge) in [6] and BoWDNN-S are considering only top-5 similar images and compute relevance scores by the modified PageRank model, but the MPM in [6] measures image similarity based on state-of-the-art bag-of-visual-words features, while BoWDNN-S uses high-level visual features learned by BoWDNN, the DCG@25 was improved from 0.537 to 0.544 after using the learned features. It may be due to the learned high-level visual features are more suitable to describe the content of images. Meanwhile, BoWDNN-S also achieves

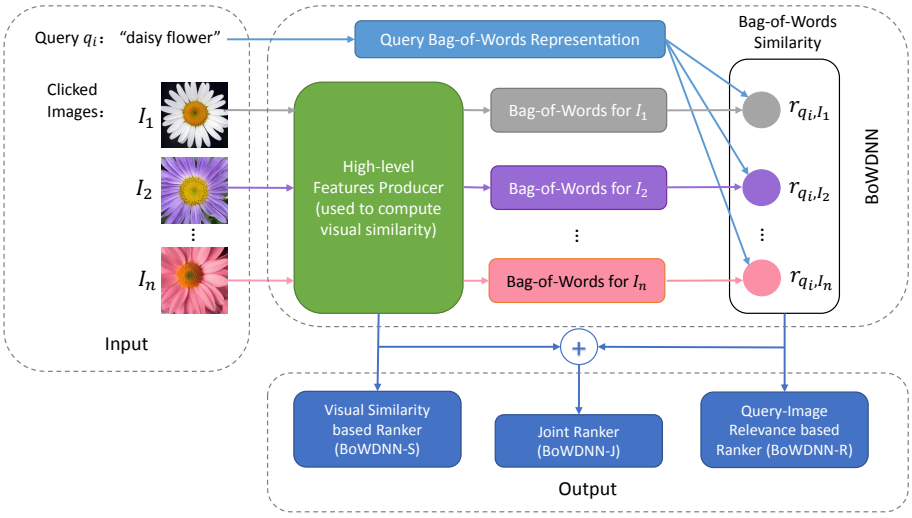

Figure 3: The system diagram of our proposed method: given query and its clicked images, the query-image relevance score will be computed according to the bag-of-words representations of image and the high-level visual features.

Table 1: Performance of three different ranking methods powered by Bag-of-words based deep neural network.

Methods	Dataset	
	Dev set	Test set
Random	0.4683	0.3906
Upper Bound	0.6840	0.5266
Ring Training	0.5021	-
BoWDNN-R	0.5089	0.4677
MPM in [6]	0.5370	-
BoWDNN-S	0.5436	0.4969
BoWDNN-J	0.5441	0.4968

better performance than BoWDNN-R, there are three possible limitations in current BoWDNN-R, 1) the size of bag-of-words vocabulary is not large enough to cover the all user intensions. There are many rare word in query log, and lots of words in queries are missed in current vocabulary. 2) the unigram representation is limited in some situations like query about people's name, each single word in a name can not represent any visual information about the related people, all of the words in people's name will be projected to human visual feature in BoWDNN, it's hard to learn some special and exact features for each person. 3) there is large numbers of queries which is hard to mining the relevance with images, like "if you really knew me" which is a title of video having no word to describe the visual content. On the other hand, the performance of visual similarity propagation methods like BoWDNN-S to a large extent depends on the quality of the initial search results, different with BoWDNN-R, this kind of method can only be used to rerank the search results. Finally, we joint these two different ranking methods and get further improvement.

5. CONCLUSIONS

In this paper, we proposed a bag-of-words based deep neural network for image retrieval task and trained on large scale clickthrough dataset from MSR-Bing Grand Challenge. By utilizing the predicted bag-of-words image representations and high-level visual features learned by BoWDNN, we propose three different methods, bag-of-words similarity based, visual similarity based, and their joint. The experiments evaluated on dev and test dataset demonstrate the effectiveness of our proposed approach for image retrieval.

6. REFERENCES

[1] Y. Bai, K. Yang, W. Yu, W.-Y. Ma, and T. Zhao. Learning high-level image representation for image retrieval via multi-task dnn using clickthrough data. *ICLR*, 2014.

[2] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier networks. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. JMLR W&CP Volume*, volume 15, pages 315–323, 2011.

[3] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[4] X.-S. Hua, L. Yang, J. Wang, J. Wang, M. Ye, K. Wang, Y. Rui, and J. Li. Clickage: Towards bridging semantic and intent gaps via mining click logs of search engines. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 243–252. ACM, 2013.

[5] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. 1999.

[6] C.-C. Wu, K.-Y. Chu, Y.-H. Kuo, Y.-Y. Chen, W.-Y. Lee, and W. H. Hsu. Search-based relevance association with auxiliary contextual cues. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 393–396. ACM, 2013.