Towards MSR-Bing Challenge: Ensemble of Diverse Models for Image Retrieval

Quan Fang[1,2], Hanqiu Xu[1], Ruowei Wang[1], Shengsheng Qian[1], Ting Wang[1], Jitao Sang[1,2], Changsheng Xu[1,2]

[1]National Lab of Pattern Recognition, Institute of Automation, CAS, Beijing 100190, China
[2]China-Singapore Institute of Digital Media, Singapore, 139951, Singapore
qfang@nlpr.ia.ac.cn, {xuhangqiu.cs, younggive}@gmail.com, {qss2012, tina437213}@163.com,
{jtsang, csxu}@nlpr.ia.ac.cn

ABSTRACT

This paper describes the solution of our team NLPR_MMC for MSR-Bing Image Retrieval Grand Challenge. This challenge is to develop an image-query scoring system to assess the effectiveness of query terms in describing the images. The provided dataset includes a training set containing more than 23 million clicked image-query pairs crawled from the web, and a development set which has been manually labeled. The test set is used in the final evaluation. We employ a diverse set of models to exploit image and query features from different perspectives. And these individual models are blended to boost the performance. Our solution achieves 0.5033 in terms of DCG@25 on the test set.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Algorithms

Keywords

Image retrieval, query, image, scoring model

1. INTRODUCTION

Text-based image retrieval (TBIR) has found growing importance due to its popularity through Web image search engines[1][2]. In the task of TBIR, the input is a text query and the retrieval system outputs a ranking set of images in which the relevant images to the query should appear on the top. There are various query-by-text image retrieval approaches proposed in both industrial and academic communities. These approaches include text matching based on the textual information associated with images, automatic image annotation approaches based on Support Vector Machines (SVMs) [9, 11], boosting classifiers [10], PLSA [8], and Corr-LDA [1], and the ranking models [7, 6] which directly connect the text queries to images. Along with the increasing usage of image search engines, users generate rich retrieval records which contain the clicked image-query triads indicating certain relevance information between images and text queries. This provides huge possibility to train and evaluate the image retrieval models by leveraging the user click logs. In this context, we are engaged in the MSR-Bing Image Retrieval Challenge to develop models exploiting user click logs from Bing image search engine for image retrieval.

The task of MSR-Bing Image Retrieval Challenge is to produce a floating-point score on each image-query pair that reflects how relevant the query could be used to describe the given image. The dataset includes a training set and a development set. The training set contains 23,094,592 triads that are sampled from Bing image search engine. Each triads is a clicked image-query record. The development set is a manually labeled dataset which contains 1,000 queries and 79,665 images for evaluation. The test set is available for the final evaluation. The relevance between images and queries is defined with three levels: excellent, good, and bad. The evaluation metric is Discounted Cumulated Gain (DCG)[3]. Each query associates with multiple images. To compute DCG, for each query, we sort the images based on the floating point scores produced by the scoring system. DCG for each query is calculated as

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

where $rel_i = \{Excellent = 3, Good = 2, Bad = 0\}$ is the manually 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.

For the challenge task, we develop an ensemble system containing diverse image retrieval models towards image-query pair scoring. A rich set of models are first employed to measure the relevance of an image-query pair by capturing different information from images and queries. We then combine the individual models to generate the final

[1]http://images.bing.com/
[2]http://images.google.com/

[3]http://web-ngram.research.microsoft.com/GrandChallenge/Rules.aspx

WOODSTOCK '97 El Paso, Texas USA

Figure 1: Overview of the proposed system

relevance score. The individual models are trained on the training set. To train the ensemble model, we use half of the development set as the validation set to learn the model parameters. The remained development set is used as test set for evaluation. Experimental results on the development set show competitive performance, which validate the effectiveness of our proposed system.

2. SYSTEM OVERVIEW

We first provide an overview of our proposed system for image-query pair scoring. As shown in Figure 1, the scoring system can be divided into four stages: data preprocessing, image and query features extraction, individual models training, and ensemble models learning. In the first stage, we perform data preprocessing on the training set to generate data annotation and build a lexicon for correcting the query spelling errors. In the second stage, a rich set of image and query features are extracted for scoring models learning. In the third stage, we apply several approaches to capture different properties of the clicked image-query dataset and learn a diverse set of models. In the final stage, the learned individual models are aggregated to generate the final scoring result.

3. DATA PREPROCESSING

3.1 Dataset

The data provided by Microsoft includes two parts: (1) the training set which is a sample of Bing user click log, (2) the development dataset which is a manually labeled set, and (3) the test set which is used for final evaluation. Each triad in the training set is a clicked image-query component which contains image, query, and click count. Each record in the development set includes query, image and judgment. The judgment has three relevance levels: excellent, good, and bad. Table 1 provides summary statistics of the dataset used in the experiments.

3.2 Query Processing

We observe that the queries contains many spelling errors such as simple typographic errors (e.g., *gril* for *girl*), and complicated cognitive errors (e.g., *camoflauge* for *camouflage*). We construct a spelling lexicon to correct the misspelled query terms. The lexicon contains correct spelling words and the corresponding misspelled words. We first obtain the correct spelling words of text queries in the training set using WordNet, and then calculate the distance between two words w_i and w_j referring to the Google distance [2]:

$$d(w_i, w_j) = \frac{\max(\log f(w_i), \log f(w_j)) - \log f(w_i, w_j)}{\log G - \min(\log f(w_i), \log f(w_j))} \quad (2)$$

where $f(w_i)$ and $f(w_j)$ are the number of images containing word w_i and word w_j respectively, $f(w_i, w_j)$ is the number of images containing both w_i and w_j, and G is the total number of images. Two words with the distance below a threshold are selected as a word pair in the lexicon for query spelling correction. In the test query, a misspelled word is detected and replaced with its corresponding correct word using the lexicon.

Table 1: Summary statistics of the datasets used in the experiments.

Dataset	Statistics	Count
Training set	#triads	23,094,592
	#images	1,000,000
	#words	867,866
Development set	#triads	79,926
	#queries	1,000
	#images	79,665
Test set	#triads	77,453
	#queries	1,000
	#images	77,453

3.3 Data Annotation

Data annotation includes two parts: (1) image annotation, and (2) tag annotation. We use the clicked image-query pair in the training set for image annotation: high click count means high relevance between image and query words. For each image in the training set, we aggregate all its associated query words with click count to obtain the ranking tags. Tag annotation is to assign relevant images for each tag. We also utilize the clicks of each image-query pair for our goal. Specifically, for each tag, we select the images associated with the tag and rank them according to the click count. Image annotation and tag annotation are used in image-query scoring models which are discussed in Section 5.

4. FEATURE EXTRACTION

This section introduces both the features of queries and the features of images used by our individual models.

4.1 Query Features

The *bag-of-words* model is used for query representation. In this context, a vocabulary V is constructed from the training set to define the set of examined words. Each query is represented as a vector $q \in \mathbb{R}^T$, where T denotes the vocabulary size. We concentrate on a vocabulary of 100,000 words by selecting the highly frequent words and removing meaningless words from the text queries in the training set. The feature weight is defined using word occurrence.

4.2 Image Features

We extract a rich set of image features including HOG [3] and multiple global features [13]. The HOG descriptors are quantized into a 21,504 dimensional sparse feature vector using LLC [12] with a codebook of the 1024 visual words and spatial pyramid matching. The 809 dimensional global features contain color moment, edge histogram, wavelet texture feature, LBP, and GIST.

5. INDIVIDUAL MODELS

In this section, we introduce several different approaches for scoring an image-query pair (x, q), and let $s(x, q)$ denote the relevance score between the image x and query q.

5.1 Concept Classification

With concept referring to a salient term or phrase, we extracted 249,527 concepts from the queries in the training set using the OpenNLP tool. Table 2 lists the statistics of our

Table 2: The statistics of our extracted concepts

#Term	132,416	#Name	30,962
#Chunk	78,860	#Location	5,289
#Query	2,000		

Table 3: Text similarity calculation

Formulations	Descriptions
$\sum_{q_i \in q \cap T_h} c(q_i, T_h)$	Term frequency (tf)
$\sum_{q_i \in q \cap T_h} \frac{c(q_i,T_h)}{\lvert T_h \rvert}$	Normalized (tf)
$\sum_{q_i \in q} \frac{IDF(q_i) \cdot TF(q_i,T_h) \cdot (k_1+1)}{TF(q_i,T_h)+k_1 \cdot (1-b+b \cdot \frac{\lvert T_h \rvert}{avdl})}$	BM25 score

extracted concepts. For each concept c, we train a binary classification model. A concept refers to a salient term or phrase. We use linear SVM in LIBLINEAR [5] combined with HOG features to train each concept classifier. The positive examples are the images with which the concept c appears in the associated clicked query, and the negative examples are randomly sampled from other irrelated images. With the trained concept classification models, in the test stage, we detect the concepts in the query and use the corresponding concept classifiers to obtain the response scores on the test image. The sum of all the response scores is used as the concept classification-based image-query relevance score. The DCG@25 of this model is 0.4937 on the test set.

5.2 Query-based Scoring Model

The idea of this model is to measure the similarity between test image and the images in the training set which are relevant to the test query. First, we use the test query to retrieve the relevant images using the tag annotation set. Second, the similarity between the test image x and the retrieved images X is calculated as

$$s(x,q) = \frac{1}{|X|} \sum_{x_k \in X} K_\sigma(x - x_k), K_\sigma(x - x_k) = \exp\left(\frac{-\|x - x_k\|^2}{\sigma^2}\right) \tag{3}$$

where σ is a scaling parameter and adaptively assigned as the median value of all pair-wise Euclidean distances between images. Here, global feature is used to represent x. We have evaluated this method on the development set. The performance of this method is inferior to the concept classification-based method and discriminative ranking method. Hence we do not evaluate this method on the test set.

5.3 Image-based Scoring Model

Similar to query-based scoring model, we first use the test image to retrieve the K nearest neighbor images using the image annotation set, and then measure the text similarity between the test query and the associated tags with the K nearest neighbor set H. LSH [4] method is applied to conduct the nearest neighbor search. The image features are global features. The text similarity is calculated as

$$s(x,q) = \frac{1}{|H|} \sum_{h \in H} e^{-l_h} \Pi(q, T_h) \tag{4}$$

where l_h is the visual similarity distance between image x and h, $\Pi(q, T_h)$ is the text similarity between query q and tags T_h associated with image h. We compute text similarity using three methods summarized in Table 3.

Table 4: Performance of the individual models and ensemble models on the test set

Model	Public Leaderboard DCG@25
Concept Classification	0.4937
Query-based Scoring	-
Image-based Scoring	-
Discriminative Ranking Model	0.4962
Two Models Fusion	0.5017
Ensemble of All Models	0.5033

We have also evaluated this method on the development set. The performance of this method is inferior to the concept classification-based method and discriminative ranking method. Hence we also do not evaluate this method on the test set.

5.4 Discriminative Ranking Model

This model is referred to [7] ,where a learning criterion related to the ranking performance is adopted. The input is the features of text query q and image x. Here x is represented as HOG features. The model outputs a relevance score of the image-query pair, which is calculated as follows:

$$s(x,q) = q \cdot f(x) \tag{5}$$

where f refers to a parametric mapping from the image space to the text space. f is learnt by optimizing the supervised loss for the image-query ranking. The optimization criterion is:

$$\min_\theta \sum_{i=1}^{N^+} \sum_{j=1}^{N^-} \max(0, 1 - s(x_i,q) + s(x_j,q)) + \frac{\lambda}{2} \|\theta\|^2 \tag{6}$$

where θ represents the parameters of the model, N^+ is the number of positive samples, and N^- is the number of negative samples. For a query, we select the associated images as positive examples and use the images associated with other queries as negative images. We use the *Passive-Aggressive* (PA) algorithm [7] to train the model.

This is the best individual model we have, and can achieve 0.4962 on the test set.

6. ENSEMBLE MODEL

We employed the simple linear bending method. Referring to the performance of each kind of individual models on the development set, we empirically set the weights for the four scoring methods. The weighted sum of the four methods is taken as the final image-query relevance score. We conduct the evaluation of two fusion schemes. The first one is that we take the uniform average of the output scores of concept classification method and discriminative ranking method. The performance of this fusion scheme is 0.5017. The second one is to combine the four scoring models. Considering the individual performance of the four methods, we empirically set the fusion weights for concepts classification, query-based scoring, image-base scoring, and discriminative ranking model as 0.3, 0.1, 0.1, 0.3 respectively. The performance on the test set is 0.5033. This is our best result.

7. CONCLUSIONS

In this paper, we introduce our solution for MSR-Bing Image Retrieval Challenge. We present an ensemble framework of combining a diverse set of models to measure the relevance between a text query and an image. The results

on the provided dataset demonstrate the effectiveness of our proposed approach for image retrieval.

8. REFERENCES

[1] D. M. Blei and M. I. Jordan. Modeling annotated data. In *SIGIR*, pages 127–134, 2003.

[2] R. Cilibrasi and P. M. B. Vitányi. The google similarity distance. *IEEE Trans. Knowl. Data Eng.*, 19(3):370–383, 2007.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR (1)*, pages 886–893, 2005.

[4] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Symposium on Computational Geometry*, pages 253–262, 2004.

[5] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[6] D. Grangier and S. Bengio. A neural network to retrieve images from text queries. In *ICANN (2)*, pages 24–34, 2006.

[7] D. Grangier and S. Bengio. A discriminative kernel-based approach to rank images from text queries. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(8):1371–1384, 2008.

[8] F. Monay and D. Gatica-Perez. Plsa-based image auto-annotation: constraining the latent space. In *ACM Multimedia*, pages 348–351, 2004.

[9] M. R. Naphade. On supervision and statistical learning for semantic multimedia analysis. *Journal of Visual Communication and Image Representation*, 15(3):348–369, 2004.

[10] K. Tieu and P. A. Viola. Boosting image retrieval. *International Journal of Computer Vision*, 56(1-2):17–36, 2004.

[11] J. Vogel and B. Schiele. Natural scene retrieval based on a semantic modeling step. In *Image and Video Retrieval*, pages 207–215. Springer, 2004.

[12] J. Wang, J. Yang, K. Yu, F. Lv, T. S. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, pages 3360–3367, 2010.

[13] J. Zhu, S. C. H. Hoi, M. R. Lyu, and S. Yan. Near-duplicate keyframe retrieval by nonrigid image matching. In *ACM Multimedia*, pages 41–50, 2008.