Learning Deep Features For MSR-bing Information Retrieval Challenge

Qiang Song[1], Sixie Yu[2], Cong Leng[1]
JiaXiang Wu[1], Qinghao Hu[1], Jian Cheng[1*]
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
Beijing, China
{qiang.song, cong.leng, jiaxiang.wu, qinghao.hu, jcheng}@nlpr.ia.ac.cn[1]
bit.yusixie@gmail.com[2]

ABSTRACT

Two tasks have been put forward in the MSR-bing Grand Challenge 2015. To address the information retrieval task, we raise and integrate a series of methods with visual features obtained by convolution neural network (CNN) models. In our experiments, we discover that the ranking strategies of Hierarchical clustering and PageRank methods are mutually complementary. Another task is fine-grained classification. In contrast to basic-level recognition, fine-grained classification aims to distinguish between different breeds or species or product models, and often requires distinctions that must be conditioned on the object pose for reliable identification. Current state-of-the-art techniques rely heavily upon the use of part annotations, while the bing datasets suffer both abundance of part annotations and dirty background. In this paper, we propose a CNN-based feature representation for visual recognition only using image-level information. Our CNN model is pre-trained on a collection of clean datasets and fine-tuned on the bing datasets. Furthermore, a multi-scale training strategy is adopted by simply resizing the input images into different scales and then merging the soft-max posteriors. We then implement our method into a unified visual recognition system on Microsoft cloud service. Finally, our solution achieved top performance in both tasks of the contest

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models

Keywords

Information retrieval, visual recognition

*corresponding author

MM'15, October 26–30, 2015, Brisbane, Australia.

DOI: http://dx.doi.org/10.1145/2733373.2809928.

1. INTRODUCTION

MSR-bing Grand Challenge 2015 tries to deal with the following 2 tasks: 1) how can we build a system to assess the relevance of a text query and its returned image list? 2) how to develop an image recognition system based on the bing datasets provided by the Challenge to recognize a wide range of dog breeds.

As for the first task: the key point lies in choosing a better search engine which can successfully meet users' requirements. The dataset of MSR-bing Grand Challenge contains 11.7 million queries and 1 million images which were collected from the user click logs of bing image search in the EN-US market[3]. Participants are asked to produce a floating point score on each image-query pair that can reflect how relevant the image matches the given query, with higher numbers indicating higher relevance.

A large number of methods [10, 11, 12] have been put forward in the previous Challenges. In general, two feasible proposals have been carried out in these methods: image content based model and text based model. The former[11] firstly retrieves images and their associated queries by computing the visual similarities between the images, and then calculates the final relevance scores based on the textual similarities between the corresponding queries. However, the proposal is hard to retrieval reliable visually similar images in such a massive number of image collection with a crowd of noisy samples. The latter [10, 12] concentrates on firstly retrieving semantically similar queries and then computing the visual similarity between the corresponding images. Specifically, for instance, Wu [11] proposed a modified PageRank model for ranking images, with an assumption that the majority of images under a same query are relevant to the query and the higher similarity to other images, the higher relevance score the image should be obtained. This method worked pretty well and achieved the first place in the competition. The assumption is also widely followed by later participants [10, 12].

In this paper, we raise and integrate a series of methods to deal with the Challenge, with visual features obtained by convolution neural network (CNN) models. We discover that the ranking strategies of Hierarchical clustering and PageRank methods are mutually complementary.

The second task is a new task of the Grand Challenges this year. Contestants are asked to develop an image recognition system based on the bing datasets provided by the Challenge to recognize a wide range of dog breeds. Gen-

erally, the problem of visual fine-grained classification can be extremely challenging due to the subtle differences in the appearance of certain parts across related categories. Localizing the parts in an object is therefore central to establishing correspondence between object instances and discounting object pose variations and camera view position. Previous work [2, 8] has investigated part-based approaches to this problem. Farrell et al.[2] proposed a pose-normalized representation using poselets. Liu et al.[8] put forward an exemplar-based geometric method to detect dog faces and extract highly localized image features from keypoints to differentiate dog breeds. However, the bottleneck for many pose-normalized representations is indeed accurate part localization. Without any ground truth information, it's difficult to adopt these part-based strategies.

Thus, in this paper, we propose a method only using image-level information. Our CNN model is pre-trained on a collection of clean datasets and fine-tuned on the bing datasets. Furthermore, a multi-scale learning strategy is adopted by simply resizing the input images into different scales. We then implement our method into a unified visual recognition system on Microsoft cloud service. Fortunately, our team achieved the third prize in task 1 and the first prize in task 2.

2. OUR APPROACH

2.1 Web Image Retrieval

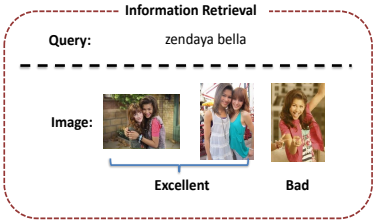

Figure 1: Web image retrieval

Shown in Figure 1, the web image retrieval task is to produce a floating point score on each image-query pair that can reflect how relevant the image matches the given query, with higher numbers indicating higher relevance. In this paper, we deal with the task by systematic strategies, including feature extraction, query matching and ranking.

2.1.1 Feature Extraction

Recently deep neural networks obtain astonishing performance at many vision tasks, outperforming the algorithms based on hand-designed features. Among variety of deep neural network architectures, Convolutional Neural Networks(CNN) have attracted much attention. A deep CNN together with large databases has achieved remarkable results[6]. Moreover, with huge amount of training data, CNN has shown to learn high-level image representation from raw pixels[13].

2.1.2 Query Match Strategy

As mentioned in [3], the original queries suffer some issues, such as meaningless words, e.g., 'picture' and 'image'. Similar to [10], we also utilize an open source tool OpenNLP[1] to remove those meaningless words and get the stems of each query. As a result, the number of distinct query triads in training dataset decreases from 23,094,502 to 8,766,023. Besides, instead of applying some traditional textual feature models in query analysis [4], e.g., 'BM25' and 'Tf-Idf', we calculate the Jaccard index to measure similarities between queries. Since query information is not long enough to extract more efficient textual features, those traditional textual feature models may not be much helpful for retrieval tasks but increase the time complexity. Furthermore, some textual expansion approaches are introduced with aims of matching possible queries. For instance, query 'hat' can be extended to 'cap' or 'chapeau'. We also discover that the queries in the dataset are written in several languages, e.g., query 'capicola' comes from Italian language, which has the same meaning as 'pork' in English. Finally, applying the above strategy in the development set obtains 450 completely matched queries and 550 partly matched queries. As for each query in the completely matched list, we sort the click count of images in descending order and select top 5 images in training dataset as the templates. As for each query in the partly matched list, we obtain top 5 similar queries in training dataset and pick images with highest click count in these 5 queries as the templates.

2.1.3 Ranking Strategies

A K-means clustering is implemented on the development set and their corresponding templates. We then sort images based on the distances to their cluster center and the number of template neighbors. Besides, we also penalize the images by the variance of their clusters. We consider that compact cluster means a higher relevance. The above method is formulated as follows:

$$score(I_i) = \frac{e^{-(D_i+\alpha Var_i)}}{1+e^{-CN_i}} \quad (1)$$

where I_i denotes the ith image, D_i, Var_i and CN_i denote the distances to the cluster center, the variance of the cluster and the number of neighbors belonging to the templates respectively. α is a trade-off parameter. We simply vary the number of the clusters and use the bagging strategy to obtain a final ranking result.

In data mining, hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters.

Besides, similar to [11, 12], we also implement a modified PageRank method to obtain the ranking of images. It is reasonable to assign the relevance score to an image based on its connections (or similarities) with others. The stronger the connections are, the higher the score should be. The relevance scores are formulated as follows:

$$score(I) = (\beta P + (1-\beta)\mathbf{1}^T)score(I) \quad (2)$$

where β denotes a trade-off parameter.

2.2 Visual Recognition

Imagine the following case: a user may have seen many lovely dogs, but can hardly tell all their names. Our target is to design a dog breeds recognition system, whose core algorithm is running on our own servers. And with the help of Microsoft cloud service, the system will be accessible to

[1]http://opennlp.apache.org/index.html

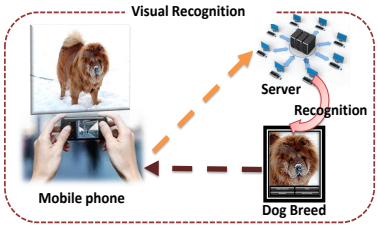

Figure 2: Visual recognition.

Table 1: Detailed descriptions of dataset.

Dataset	Category	Amount
Stanford	120	20,580
Columbia	133	8351
MM2015	344	42,886

public users.(Figure 2) In this paper, we deal with the visual recognition task by first carrying out data analysis and then building a deep learning model-based system.

2.2.1 *Data Analysis*

Figure 3: Images of distinct datasets.

Three distinctive datasets are utilized in our experiments, including two public real-world datasets, namely Stanford[1] and Columbia [8] respectively. The rest one is a collection on the Clickture-Full of MSR-bing dataset. Specifically, MM2015 is a public released dataset with our post-denoising process. Detailed descriptions are shown in Table 1. After analyzing images of these distinct datasets (Figure 3), we have discovered the following observations (where $\succ$ denotes as 'is superior to'):

- Clarity: Columbia $\succ$ Stanford $\succ$ MM2015
- Complexity: MM2015 $>$ Stanford $>$ Columbia
- Slant degree of samples: Stanford>Columbia>MM2015
- Category: MM2015>Columbia>Stanford

2.2.2 *Dog Breeds Recognition System*

Convolutional Neural Network has recently obtained great success on several visual recognition and detection Challenges. With CNN pre-trained on large amount of data, we can fine-tune the trained CNN on our own datasets so as to accomplish specific tasks. In our model, we utilized the architecture proposed by VGG group, which is a 19-layer convolutional neural network[9] pre-trained on ILSVRC dataset. In order to generalize this CNN to dog breeds classification, we pre-trained it on two datasets, Stanford Dogs Dateset[1] and Columbia Dogs with Parts[8]. Stanford Dataset contains images of 120 breeds of dogs from ImageNet and Columbia Dataset contains images of 133 breeds of dogs downloaded from Google, ImageNet and Flickr. The 120 breeds of Stanford Dogs Dataset are included in 133 breeds of Columbia Dogs with Parts. Therefore we denoted the 120 breeds as our targeted categories. For each category we selected 200 images as training data and 30 images as testing data. In pre-training stage, our model was pre-trained with dropout(dropout ratio set to 0.5) and regularized by weight decay(The L_2 penalty multiplier set to $5 \cdot 10^{-4}$). The initial learning rate is 0.01 and then decreased by a factor of 10 after each epoch, the learning rate was stopped after 100K iterations(2 epochs). In fine-tuning stage, we first detected dog breeds in MSR-bing query data. We found 344 dog breeds occurred in queries. Then we extracted corresponding images of each breed from MSR-bing training set and removed obviously unrelated images by human, which resulted in 42886 images as training data. Since we noticed that dogs in images can be of different sizes, we followed the multi-scale training proposed in [9] to fine-tune our model. We first rescaled all training images to 256×256 and then fine-tuned our model with these images. Then we rescaled all training images to 384×384, reduced initial learning rate to 10^{-3} and fine-tuned only fully-connected layers. Our data augmentation generally followed [6] and training procedure is the same as that in pre-training stage. We conducted experiments in Caffe[5]. In a word, the synthesized model is shown in Figure 4. For future convenience, we will plan to implement our CNNs model into mobile phones through hashing methods(e.g. [7]).

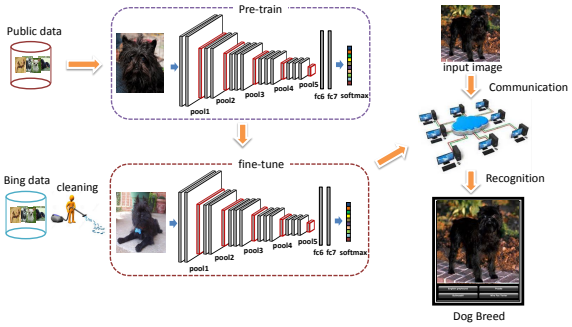

Figure 4: A flow chart of our system.

3. EXPERIMENT

3.1 Web Image Retrieval

We utilized CNNs trained on ImageNet[6] and Places[14] datasets respectively, to extract features that will be used in later stage. Both CNNs have the same architecture, with the only difference is that they are trained on two different large datasets, namely object-centric ImageNet dataset and scene-centric Places dataset[14]. The reason for using the two CNNs is that high-level features learned by them are complementary[14]. We also experimented with 19-layer VGG[9] network. Nevertheless it does not perform well in our experiments. We attribute its inferior performance to absence of fine-tuning such that in later layers it cannot learn meaningful high-level features. We use open source framework Caffe[5] to extract features from 7-th layer(fc7).

We then compare our methods with those proposals in the previous Challenges, and the results are shown in Table 2. We can observe that the $NDCG$ of completely matched queries of our methods outperforms that of partly matched ones. For instance, compared with the Ground-truth, ImageNet(P) achieves 84.34% of completely matched queries and 75.34% of partly matched queries respectively. The PageRank method still surpasses other two methods whatever the CNN features are. Furthermore, although the Hierarchical clustering is not as good as PageRank method, a model fusion with PageRank achieves a great improvement in $NDCG$, indicating that the strategies of these 2 methods do act complementary. However, combining the three models makes the performance a little bit lower, since the ranking strategy of K-means is similar to Hierarchical clustering, and the bagging may weaken the complemental benefit. Besides, we also discover that a simply re-ranking strategy on the top 25 images can improve the performance. The final competition result of our method is: 0.4715.(a little format error occurd in the submission.) We only achieved the third prize in web image retrieval task.

Table 2: Metrics evaluation of different methods on MSR-bing Dev. 'K', 'H' and 'P' are short for K-means, Hierarchical clustering and PageRank respectively. 'Comb' denotes the model fusion of ImageNet and Places. 'Rerank' denotes a re-ranking strategy for the top 25 images.

Method	Comp Match	Part Match	Total
Randomly	-	-	0.4690
Ground-truth	0.8090	0.5800	0.6840
GP[10]	0.6780	0.4330	0.5400
CrossMedia[12]	-	-	0.5608
ImageNet(K)	0.6720	0.4210	0.5340
ImageNet(H)	0.6787	0.4235	0.5383
ImageNet(P)	0.6823	0.4370	0.5474
ImageNet(H+P)	0.6884	0.4456	0.5549
ImageNet(K+H+P)	0.6880	0.4448	0.5542
Comb(H+P)	0.6912	0.4468	0.5567
Comb-Rerank(H+P)	0.6986	0.4440	0.5586

3.2 Visual Recognition

We compare our methods on the two CNN architectures and the results are shown in Table 3. We can observe that the metric of VGG is much higher than that of AlexNet. It demonstrates that deep network has better representation. Furthermore, Multi-scale training improved 1.2% performance during testing, which demonstrates our idea. The final competition results of our method are: Accuracy@1: 57% and Accuracy@5: 85%. Fortunately, we achieved the first prize in fine-grained recognition task.

Table 3: Experiment on MM2015 dataset. where S, C denote as 'Stanford' and 'Columbia' dataset respectively.

CNN Structure	Pre-train	Fine-tune	Accuracy@1(%)
AlexNet	S+C	MM	51.7
VGG	S+C	MM	63.2
Mul-AlexNet	S+C	MM	52.8
Mul-VGG	S+C	MM	64.5

4. CONCLUSION

In this paper, we propose a series of methods to accomplish the two tasks of MSR-bing Challenge. We discover that the Hierarchical clustering and PageRank methods are mutually complementary for information retrieval. Besides, we propose a method only using image-level information for fine-grained visual recognition. Our CNN model is pre-trained on a collection of clean datasets and fine-tuned on the bing datasets. We then implement our methods into a unified visual recognition system on Microsoft cloud service.

5. ACKNOWLEDGMENTS

This work was supported in part by National Natural Science Foundation of China (Grant No. 61332016,61170127), and 863 program (Grant No.2014AA015105).

6. REFERENCES

[1] E. Dataset. Novel datasets for fine-grained image categorization. In *First Workshop on Fine Grained Visual Categorization, CVPR*. Citeseer, 2011.

[2] R. Farrell, O. Oza, N. Zhang, V. Morariu, T. Darrell, L. S. Davis, et al. Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 161–168. IEEE, 2011.

[3] X.-S. Hua. Looking into" msr-bing image retrieval challenge. Technical report, Microsoft Research Technical Report MSR-TR-2013–76, 2013.

[4] Y. Hua, J. Shao, H. Tian, Z. Zhao, F. Su, and A. Cai. An output aggregation system for large scale cross-modal retrieval. In *Multimedia and Expo Workshops (ICMEW), 2014 IEEE International Conference on*, pages 1–6. IEEE, 2014.

[5] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[7] C. Leng, J. Wu, J. Cheng, X. Bai, and H. Lu. Online sketching hashing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2503–2511, 2015.

[8] J. Liu, A. Kanazawa, D. Jacobs, and P. Belhumeur. Dog breed classification using part localization. In *Computer Vision–ECCV 2012*, pages 172–185. Springer, 2012.

[9] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[10] J. Wang, C. Kang, Y. He, S. Xiang, and C. Pan. Cross modal deep model and gaussian process based model for msr-bing challenge. In *Proceedings of the ACM International Conference on Multimedia*, pages 225–228. ACM, 2014.

[11] 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.

[12] Z. Xu, Y. Yang, A. Kassim, and S. Yan. Cross-media relevance mining for evaluating text-based image search engine. In *Multimedia and Expo Workshops (ICMEW), 2014 IEEE International Conference on*, pages 1–4. IEEE, 2014.

[13] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014*, pages 818–833. Springer, 2014.

[14] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, pages 487–495, 2014.