Two-stream Attentive CNNs for Image Retrieval

Fei Yang
Institute of Information Science, Beijing Jiaotong University, Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing, China 100044

Jia Li*
State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University International Research Institute for Multidisciplinary Science, Beihang University
jiali@buaa.edu.cn

Shikui Wei*
Institute of Information Science, Beijing Jiaotong University, Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing, China 100044
shkwei@bjtu.edu.cn

Qinjie Zheng
Institute of Information Science, Beijing Jiaotong University, Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing, China 100044

Ting Liu
Institute of Information Science, Beijing Jiaotong University, Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing, China 100044

Yao Zhao
Institute of Information Science, Beijing Jiaotong University, Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing, China 100044

ABSTRACT

In content-based image retrieval, the most challenging (and ambiguous) part is to define the similarity between images. For the human-being, such similarity can be defined with respect to where they pay attention to and what semantic attributes they understand. Inspired by this fact, this paper presents two-stream attentive CNNs for image retrieval. As the human-being does, the proposed network has two streams that simultaneously handle two tasks. The Main stream focuses on extracting discriminative visual features that are tightly correlated with semantic attributes. Meanwhile, the Auxiliary stream aims to facilitate the main stream by redirecting the feature extraction operation mainly to the image content that human may pay attention to. By fusing these two streams into the Main and Auxiliary CNNs (MAC), image similarity can be computed as the human-being does by reserving the conspicuous content and suppressing the irrelevant regions. Extensive experiments show that the proposed model achieves impressive performance in image retrieval on four public datasets.

KEYWORDS

Visual Attention; Image Retrieval; Two-stream CNNs

*Corresponding authors: Shikui Wei and Jia Li

MM'17, October 23–27, 2017, Mountain View, CA, USA.
 ISBN 978-1-4503-4906-2/17/10...$15.00
DOI: https://doi.org/10.1145/3123266.3123396

ACM Reference format:
Fei Yang, Jia Li*, Shikui Wei, Qinjie Zheng, Ting Liu, and Yao Zhao. 2017. Two-stream Attentive CNNs for Image Retrieval. In *Proceedings of MM'17, October 23–27, 2017, Mountain View, CA, USA., ,* 9 pages.
DOI: https://doi.org/10.1145/3123266.3123396

1 INTRODUCTION

With the booming of smart phones and digital cameras, the amount of images grows surprisingly fast in our daily life. To maximize the value of such big visual data, it is necessary to develop an image search approach that is capable of retrieving images with the desired content. For such a content-based image retrieval (CBIR) approach, one of the key challenges is to infer the inherent query intention expressed by a query image. As shown in Fig. 1, confusion may arise in determining what is the desired content [26, 54], while the similarity between images may be defined in visual [4, 15, 59] and/or semantic [3, 10, 39] levels. Actually, the ambiguity in capturing the *inherent query intention* acts as a major obstacle in CBIR.

In the past decades, hundreds of approaches have been proposed for fast and reliable CBIR [46, 47]. For example, many hashing methods [27, 37, 44] based on SIFT [40, 57] and GIST [27, 43] features have been proposed to make the similarity computation faster and more semantic, while other cues like emotion [20] have been explored as well. In particular, recent advances in deep learning [13, 14, 21, 38] provide an opportunity to overcome the well-known semantic gap in CBIR [5, 6, 12, 34, 35]. For example, Razavian *et al.* [34] first extracted sub-patches from different locations in an image and characterized them with deep features. Such features are then compressed to compute patch-based similarity. Gong *et al.* [12] extracted deep features from patches at different scales and locations by

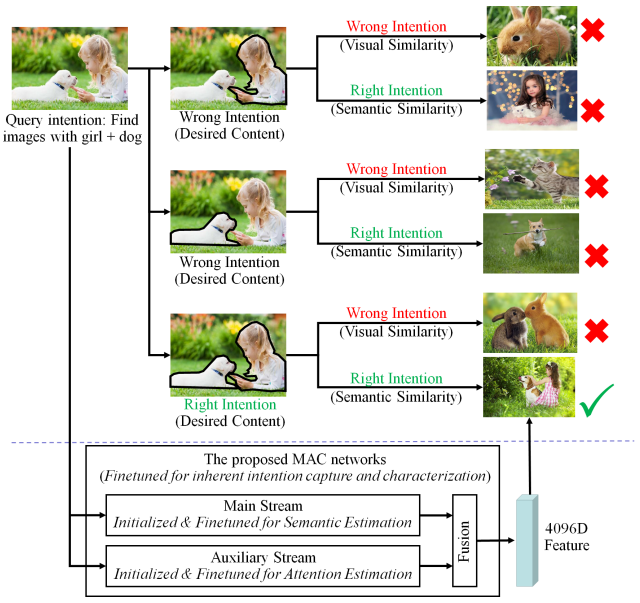

Figure 1: Capturing the inherent query intention plays an important role in correctly retrieving the desired target images. Toward this end, we propose a two-stream attentive CNNs that start from a Main semantic stream and an Auxiliary attention stream, which are fused and simultaneously fine-tuned so as to capture and characterize such inherent intention.

using Convolutional Neural Networks (CNNs) as well as orderless pooling strategies. In [5], local deep features were aggregated to produce compact global descriptors for image retrieval. Typically, such CNN-based approaches can outperform classic SIFT- or GIST-based approaches since the feature extracted by CNNs are generally considered to be closer to the semantic attributes of images. However, such features are extracted from the whole image, making them somehow inaccurate to capture and characterize the inherent query intention (*e.g.*, the desired content).

To develop a CBIR approach that is capable of capturing inherent query intention, we first turn to a fundamental question: how does the human-being compute visual similarity during their cognitive visual processes? With this question in mind, we first explore the cognitive mechanisms like visual attention and object recognition in human vision system and then propose two-stream attentive CNNs for image retrieval. As shown in Fig. 1, the Main and Auxiliary CNNs, denoted as MAC, start from two separate streams that handle different cognitive tasks. The Main stream is initialized with VGG16 [38], while the Auxiliary stream is initialized with DeepFixNet [30]. In other words, the Main and Auxiliary streams start from the tasks of semantic attribute prediction and visual attention prediction, respectively. Considering that such semantic and attentive cues are tightly correlated with but not equivalent to the inherent intention in a query image, we further fuse them and fine-tune the entire networks on existing image retrieval datasets. In this manner, the semantic and attentive cues can be gradually modulated to reflect the inherent query intention. As a result, we can obtain reliable similarity scores between a query image and all candidate images by using the output features of MAC, even with a very simple $\ell 2$ distance measure. Extensive experiments show that our approach achieves impressive performance on 4 public datasets. In particular, our approach further validates its effectiveness in many (synthesized) challenging scenarios like rain/snow and low-resolution/low-quality, implying that it can be even suitable for many real-world applications.

Our main contributions can be summarized as follows: 1) we propose a two-stream attentive CNNs to capture the inherent intention in image retrieval; 2) we conduct extensive experiments on 4 public datasets to validate the performance of MAC from various perspectives. Moreover, we synthesize many challenging scenarios that further validate the scalability of MAC in real-world scenarios.

2 RELATED WORK

The proposed attentive CNNs are tightly correlated with CNN-based or attention-guided CBIR approaches, which will be briefly reviewed in this Section.

2.1 CNN-based CBIR

Due to the remarkable success of deep learning, they have been introduced into image retrieval in many different ways like local feature extraction [5, 12, 22, 31, 34, 61], global feature extraction [25, 33, 55, 60], hashing [23, 36, 53, 58], semantic annotations [49, 51], semantic segmentation [48, 50] or similarity computation [2, 7, 9, 56]. For example, Paulin *et al.* [31] proposed to learn patch descriptors without supervision. In their approach, the convolutional kernel networks were adopted to extract patch features for matching and instance-level retrieval.

In global feature extraction, Razavian *et al.* [34] used Structure-from-Motion (SfM) method to get 3D models, which can guide the selection of deep features. Zheng *et al.* [60] fused various features by extracting the output of pooling layers in VGG and Alexnet for image retrieval. Zhou *et al.* [61] used the match function to integrate SIFT and CNN features. A threshold exponential match kernel method was proposed to calculate the scores of similar images.

In CNN-based hasing, Xia *et al.* [53] proposed a CNN-based Hashing method. They broke down similarity matrix and generated the binary encoding results. Zhao *et al.* [58] proposed a Deep Semantic Ranking Hashing method. They used CNNs to learn the ranking of retrieval results and optimize the evaluation index.

In similarity computation, Zagoruyko *et al.* [56] proposed to directly learn visual similarity from image pairs by using two-stream networks. Bontar *et al.* [9] learned the similarity measure on small image patches by using CNNs.

2.2 Attention-guided CBIR

As visual attention can depict the most conspicuous content in images, they have been incorporated into CBIR. Generally speaking, the most straightforward way to use visual attention is to detect attention regions for subsequent feature extraction stage. For example, Wen *et al.* [52] extracted SIFT and color features from attention regions to retrieve images. Giouvanakis and Kotropoulos [11] combined a classic attention model with the Bag-of-Words (BOW) model by extracting SIFT features only from attention regions. Wang *et al.* [8] used attention regions to select a subset of SIFT points for image retrieval. Actually, many such attention-guided models have been proposed [1, 8, 17, 42], but the performance gained from such straightforward usage of visual attention is usually not as high as expected. It is still not clear how to correctly use visual attention in CBIR. Moreover, existing attention models, which are mainly developed for predicting human fixations in free-viewing conditions, may be not suitable for the CBIR task if they are directly used without being fine-tuned. Furthermore, most of such attentive CBIR models are developed based on classic features (*e.g.*, SIFT), which often perform much worse than deep features in depicting semantic attributes of the desired image content.

To sum up, CNN-based CBIR approaches can extract features or learn similarity measures that are closer to semantic. However, a key challenge for these approaches is: how to extract features only from the desired image content so as to avoid the influence of irrelevant regions. In other words, existing CNN-based approaches can extract powerful features with unexpected noise beyond the desired image content. On the contrary, attentive CBIR approaches can filter out irrelevant regions. But the classic features used by most attentive models are relatively weak. Moreover, most attention models are developed for fixation prediction in free-viewing conditions, and it may be inappropriate to directly use them in CBIR without revision. Inspired by these facts, we propose two-stream attentive networks for CBIR, in which the two streams are initialized for fixation prediction and semantic recognition, respectively. These two streams are then fused and fine-tuned together on image retrieval datasets so that the extracted attention cues and semantic features become more suitable for the CBIR task.

3 MAC: TWO-STREAM ATTENTIVE NETWORKS FOR CBIR

To capture the inherent query intention, we propose two-stream attentive networks for CBIR. In this section, we first introduce the architecture of the proposed Main and Auxiliary CNNs (MAC), followed by the details that describe how to train such a model and how to use MAC for CBIR.

3.1 System Framework

As shown in Fig. 2, the core of the proposed CBIR approach is the two-stream attentive networks, which can be denoted as Main and Auxiliary CNNs (MAC). Different from previous works, MAC has two separate streams that are initialized for different cognitive tasks. Both streams take a 224×224 image with three channels as the input.

The Main stream is initialized as the first five major convolutional and pooling groups (*i.e.*, from `CONV1_1` to `CONV5_3` and `POOL5`). Finally, the major stream will output a 7×7 map with 512 channels, while such a map, denoted as a 3D matrix $\mathbf{X}_{main}$, contains high-level cues extracted from both the desired image content and the irrelevant regions. As a result, such feature maps need to be further refined to obtain cleaner *semantic* features that can characterize the inherent query intention.

Toward this end, we incorporate the Auxiliary stream to filter out the unexpected *noise* from the original features extracted by the Main stream. To maintain the features that may be useful for the task of image retrieval, we initialize this stream with the DeepFixNet [30], the CNNs that are developed for fixation prediction. In the initialization, we select the first eight convolution layers together with the related pooling layers and generate a 56×56 map with 32 channels. After that, such a feature map enters a pooling layer and is then reshaped to 7×7 maps with 512 channels. Similar to the Main stream, the output map is denoted as $\mathbf{X}_{aux}$.

Compared with heuristic attention models, DeepFixNet gains impressive performance in predicting attention (see Fig. 3). With this stream, we can filter out the features from regions that are irrelevant to the inherent query intention. However, DeepFixNet is trained with eye-tracking data from free-viewing experiments. As a result, it may not perfectly meet the specific requirement of image retrieval. Therefore, the parameters of this stream, as well as the Main stream, need to be further fine-tuned on image retrieval datasets. Toward this end, we first conduct element-wise fusion of the output maps from the Main and Auxiliary streams:

$$\mathbf{X}_{\text{fuse}} = \lambda \mathbf{X}_{\text{main}} + (1 - \lambda)\mathbf{X}_{\text{aux}}, \tag{1}$$

where λ is empirically set to 0.6 to balance the output features from the two streams. In this manner, the fused feature map contains both semantic and attention cues, which is converted to a lower dimensional feature vector via three consecutive Fully Connected layers, denoted as `FC6`, `FC7` and `FC8`, respectively. Note that both `FC6` and `FC7` output 4096D feature vectors, while `FC8` outputs a vector with N components. For an image retrieval dataset, N denotes the number of categories formed by aggregating training images with similar contents (such similarity is manually annotated by the human-being). By applying a softmax layer after `FC8` to turn its output to a probability vector, we can train a classification network on image retrieval datasets by solving the minimization problem:

$$\mathbf{W}^* = \arg\min_{\mathbf{W}} \sum_{k=1}^{K} \sum_{n=1}^{N} log(p_{kn}, y_{kn}) + \beta \|\mathbf{W}\|_2^2, \tag{2}$$

where $\mathbf{W}$ denote the set of network parameters of MAC. p_{kn} is the nth component of the probability vector generated by

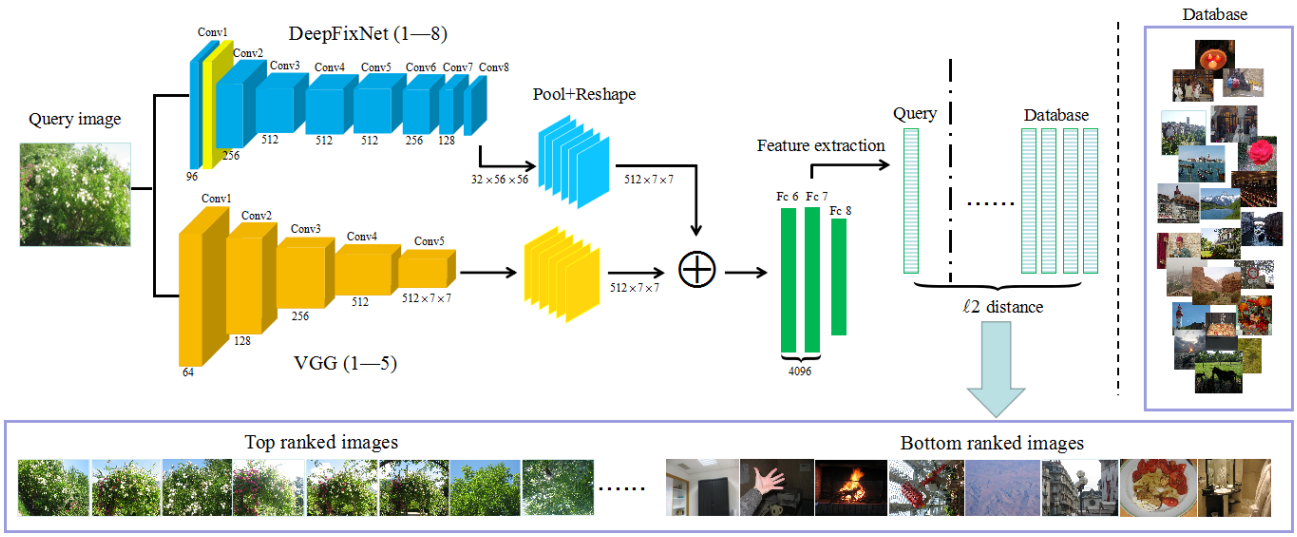

Figure 2: The framework of the proposed CBIR system. The core of this framework is the two-stream attention CNNs, which contain two separate streams. The Main stream is initialized with the semantic prediction networks VGG16, while the Auxiliary stream is initialized with the fixation prediction networks DeepFixNet. These two streams are then fused and simultaneously fine-tuned on image retrieval datasets so as to extract features that can well capture and characterize the inherent query intention. Finally, such features are used to measure the similarity (computed as the $\ell 2$ distance) between a query image and all images in the database.

the final softmax layer of MAC in processing the kth training image. y_{kn} is a binary indicator which equals to 1 only if the kth training image belongs to the nth category of similar training images. β is a constant that controls the norm of parameters in MAC.

By minimizing (2), the MAC network gains the capability to aggregate similar images and separate dissimilar images, while such similarity is defined from the perspective of image retrieval. In this manner, the features generated by MAC can well capture and characterize the inherent query intention for CBIR. In training MAC, we adopt the Caffe platform [19] and utilize a batch size of 16. The learning rate is initialized as 10^{-6}, which will decrease twice, by a factor of 10, after the 33% and 66% of the maximum iteration number have been reached, respectively. Moreover, a weight decay of 0.0005 and momentum of 0.9 are used.

After training MAC, the two-stream attentive CNNs can be used for image retrieval. Considering that the feature dimension of `FC8` varies with respect to different training data, we adopt the 4096D feature vector generated by the `FC7` layer of MAC to characterize the inherent query intention of a new query image. After that, the similarity scores between this feature vector and those pre-computed for the images in the database can be computed. Since the main objectiveness is to demonstrate the powerfulness of the proposed two-stream attentive CNNs, we only use the simplest $\ell 2$ distance as the similarity measure, which can already generate impressive performance by using the powerful features from MAC.

4 EXPERIMENTS

To validate the effectiveness and scalability of MAC in CBIR, we conduct a series of experiments from multiple perspectives, including:

(1) Effectiveness test, which compares MAC with the state-of-the-art and baseline models;

(2) Scalability test, which test the performance of MAC by adding one million images into testing datasets, or synthesizing more challenging real-world scenarios like rain/snow and low-resolution/low-quality;

(3) Performance analysis, which investigates the performance variation of MAC by changing key parameters.

Detailed experimental settings, performance scores and representative results can be found as follows.

4.1 Settings

To conduct comprehensive evaluation of MAC, we adopt four datasets from the areas of image retrieval and image classification. Among these datasets, Oxford Paris [32] and Ukbench [29] are two image retrieval datasets that are widely used in the literature. Flower [28] and Bird [41] are two fine-grained image classification datasets which can be also used to benchmark the image retrieval models [16, 26]. For each dataset, we split them into a training set and a testing set. Details of the 4 datasets can be found in Tab. 1.

One problem in comparing image retrieval models is that the performance of learning-based models may vary remarkably before/after being fine-tuned on specific training

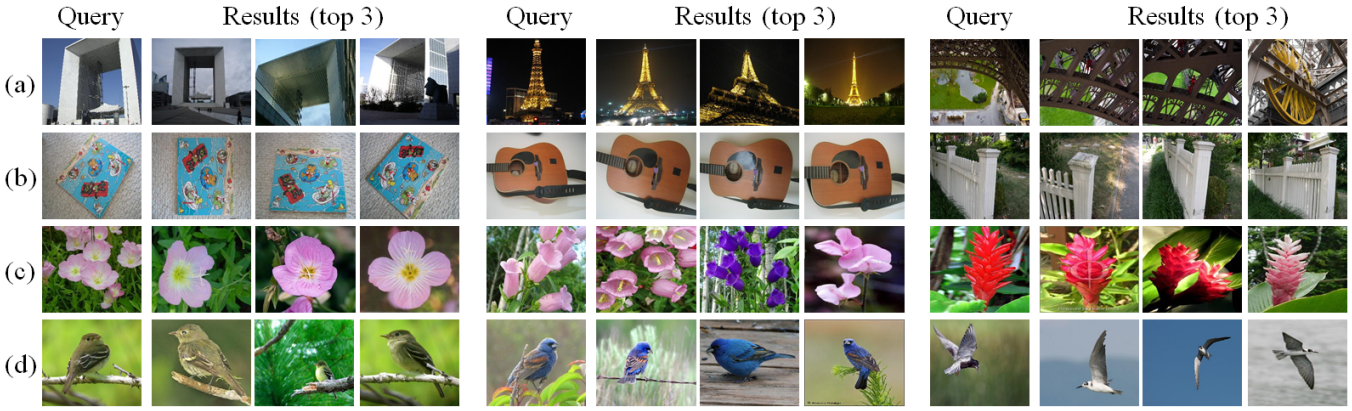

Figure 4: Representative results of MAC on 4 datasets. (a) Oxford Paris; (b) Ukbench; (c) Flower; (d) Bird.

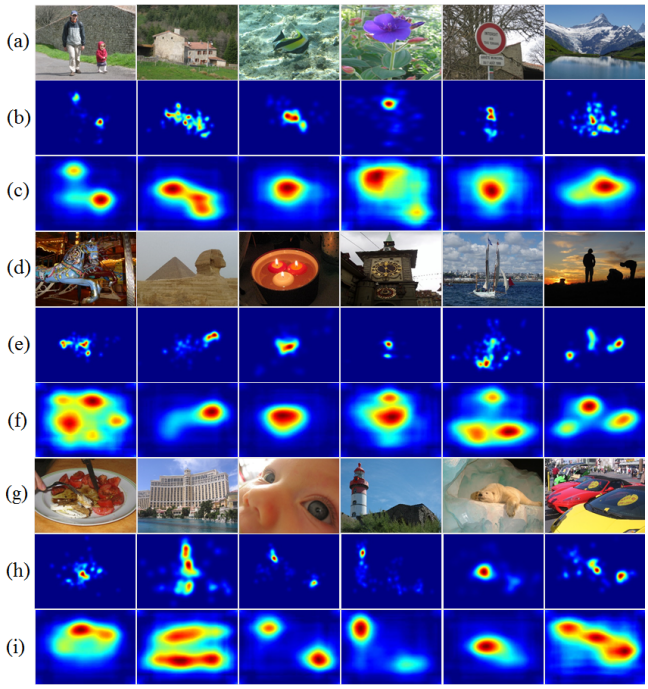

Figure 3: DeepFixNet achieves impressive performance in predicting visual attention under the free-viewing condition. However, it is still not clear whether such types of visual attention can be directly used in image retrieval task. As a result, it is necessary to fine-tune the parameters of DeepFixNet on image retrieval datasets. (a,d,g) Original images; (b,e,h) Fixation density maps captured by our eye-tracking device that show the actual human attention under free-viewing condition; (c,f,i) Attention maps predicted by the DeepFixNet.

data. Therefore, we compare MAC with two state-of-the-art models and two baselines, including:

Table 1: Details of the 4 benchmarking datasets

Datasets	Total	Training	Testing	Categories
Oxford Paris	6,392	5,192	1,200	12
Ukbench	10,200	5,100	5,100	2,550
Flower	7,169	6,149	1,020	102
Bird	11,788	6,788	5,000	200

(1) BOWE [45]: A non-deep approach that jointly optimize Bag-of-Words and embedding methods for image retrieval.

(2) Siamese [56]: Two-stream CNNs that take a pair of images as the input and output the similarity scores.

(3) Base-VGG: A baseline model formed by directly using the 4096D features from the original VGG16 networks and the same retrieval settings with MAC.

(4) Base-VGG-F: Different from Base-VGG, Base-VGG-F is further fine-tuned on the same training data used by MAC in all experiments so that the 4096D features it generated is refined for the retrieval tasks.

To compare different models, we adopt the mean Average Precision (mAP) as the evaluation metric, which is one of the most widely used metrics in image retrieval.

4.2 Effectiveness Test

In the effectiveness test, we fine-tune MAC and Base-VGG-F on the training set of each dataset and compare them with other models on the testing set. Performance of all approaches can be found in Table 2. Some representative retrieval results of MAC can be found in Fig. 4.

From Table 2, we can see that the proposed MAC model achieves impressive performances on all the four datasets. In particular, the MAC network outperforms Base-VGG-F, even when they are fine-tuned on the same training data. This may be caused by the fact that, after incorporating the Auxiliary stream, the semantic features from irrelevant regions can be removed, and the retrieval process will mainly

Table 2: Effectiveness test of 5 models on 4 datasets

	Oxford Paris	Ukbench	Flower	Bird
BOWE [45]	0.12	0.81	0.05	0.004
Siamese [56]	0.11	0.26	0.03	0.01
Base-VGG	0.29	0.87	0.31	0.17
Base-VGG-F	0.46	**0.92**	0.60	0.26
MAC	**0.48**	**0.92**	**0.64**	**0.28**

Table 3: Scalability test on 4 datasets after adding a one-million confusion images from Flickr1M

	Oxford Paris	Ukbench	Flower	Bird
Base-VGG-F	0.10	**0.88**	0.20	**0.12**
MAC	**0.11**	**0.88**	**0.25**	**0.12**

focus on comparing the "desired content" shared by query and target images. In other words, with the assistance of feature maps from the Auxiliary attention stream, the Main semantic stream perform better in distinguishing images from different categories by focusing on the right regions. Moreover, both the Main semantic stream and the Auxiliary attention stream are fine-tuned on image retrieval datasets. In this manner, we can assume that both the semantic features and the attentive cues extracted become more suitable for the task of image retrieval. That also explains the remarkable performance enhancement from Base-VGG to Base-VGG-F after fine-tuning the original VGG model on image retrieval datasets.

Moreover, from Table 2 we find that the proposed MAC network is not only suitable for the classic image retrieval datasets but also fits for the fine-grained image classification datasets. As shown in Fig. 4, MAC can successfully retrieve images with fine-grained birds and flowers. This is an interesting findings, implying that the usage of the Auxiliary attention stream also helps maintain the unique attributes of the desired objects while refining the noisy features. Actually, the fine-grained classification/retrieval is much more sensitive to the noise from irrelevant regions, while visual attention cues can help to "neglect" such regions in feature retrieval. In other words, the semantic stream mainly learns about what is a bird, while the attention stream may help in learning where is the right place to extract such features. From these results, we can safely conclude that incorporating an additional attention stream is effective for the task of image retrieval.

4.3 Scalability Test

Beyond effectiveness test, we also conduct several experiments to validate the scalability of MAC (and the baseline models). Toward this end, we first incorporate the one million images from the Flickr1M dataset [18] into the testing sets of each dataset and run the retrieval experiments again. Experimental results are shown in Table 3. By comparing the results in Table 3 and Table 2, we can see that even with so many confusion images the retrieval performance of both MAC and Base-VGG-F drop sharply on most datasets. However, in such a challenging setting the performance on UKbench still reaches 0.88. Considering that Flickr1M contains many objects like flower and bird, the performance of MAC on the fine-grained datasets Flower and Bird are still acceptable, implying that the MAC is a scalable network.

Table 4: Performance of four models on synthesized low-resolution/low-quality scenarios

Models	Oxford Paris	Ukbench	Flower	Bird
Base-VGG	0.24	0.84	0.26	0.07
Base-VGG-F	0.38	0.87	0.49	0.15
MAC	**0.41**	**0.88**	**0.53**	**0.17**

Table 5: Performance of four models on synthesized rain/snow scenarios

Models	Oxford Paris	Ukbench	Flower	Bird
Base-VGG	0.17	0.35	0.10	0.02
Base-VGG-F	0.36	0.47	0.38	0.09
MAC	**0.40**	**0.48**	**0.46**	**0.11**

Beyond adding confusion images, in actual life many images uploaded to the Internet are low-quality/low-resolution ones. To further validate the effectiveness of our approach, from the four datasets we generate their low-resolution version and test the performance of MAC and baseline models Base-VGG and Base-VGG-F. The performance scores are shown in Table 4, while some representative results are shown in Fig. 5. By comparing Table 4 and Table 2, we can see that the performance only slightly decreases, while the results in Fig. 5 validates that the proposed approach is scalable to low-quality and low-resolution scenarios. Moreover, in such scenarios MAC still outperforms Base-VGG and Base-VGG-F. This may be caused by the fact that the attention maps are less sensitive to resolution variation, and many attention/saliency models will resize the input image to an extremely low resolution (*e.g.*, 32×32 in [24]) to speed up the computation process. When the resolution decreases, the Auxiliary attention stream still outputs reliable cues that assists the localization of desired content, making the whole network more reliable.

Moreover, many images in our daily life are taken in rain or snow, and it is necessary to develop a model that can effectively retrieve such images. To test the performance of image retrieval models in such scenarios, we add synthesized rain/snow to the four datasets. As shown in Fig. 5 and Fig. 2, the performance scores of both MAC and the two baseline models decrease in rain/snow scenarios. In particular, the performance on UKBench drops remarkably due to it contains many large-scale scenes, while the other

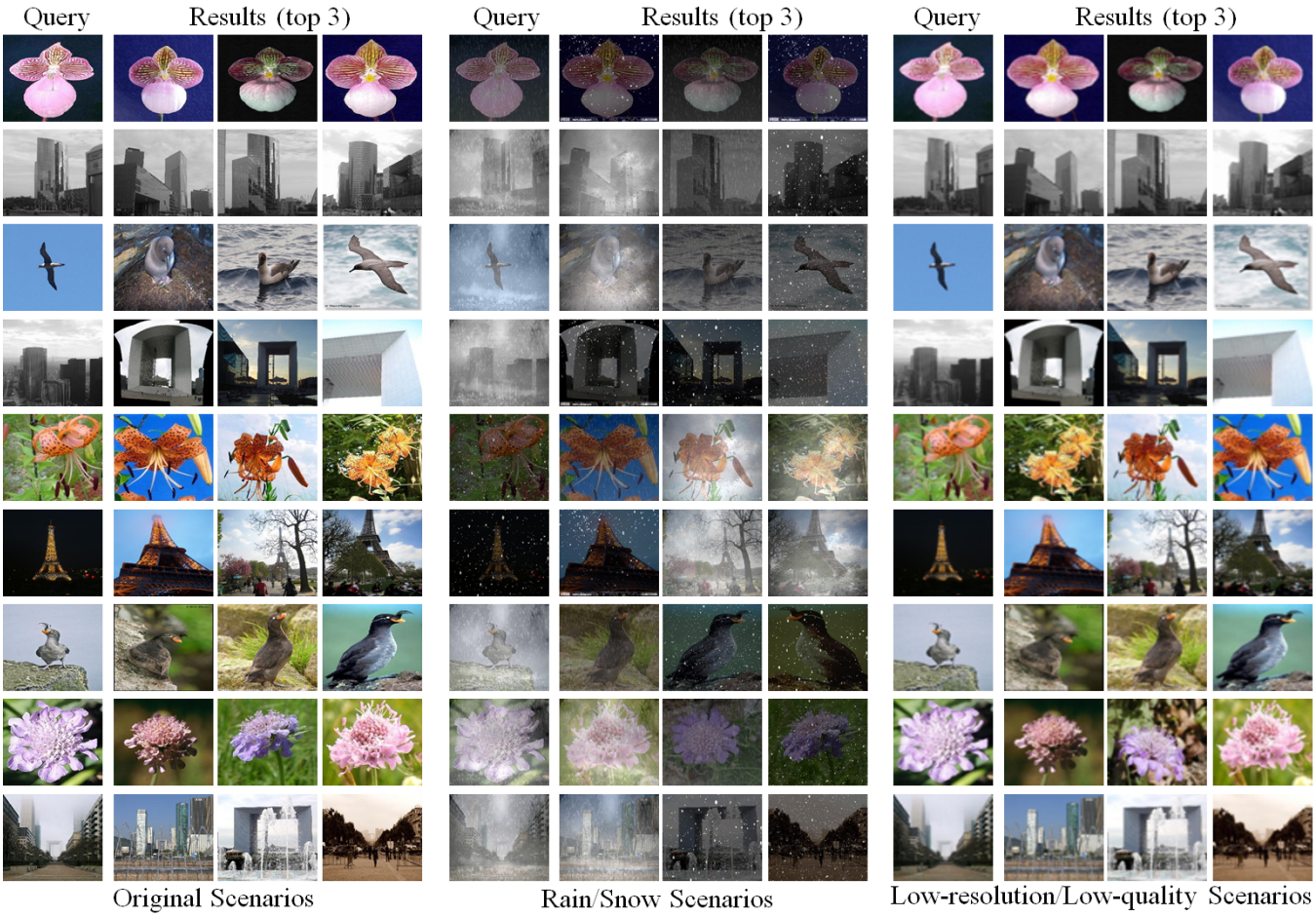

Figure 5: Representative retrieval results of MAC in rain/snow and low-resolution/low-quality scenarios. Left column: results on original datasets; Middle column: results on datasets with synthesized rain/snow; Right Column: results on datasets with degraded resolution/quality.

three datasets contain large objects that are less influenced by rain and snow, leading to smaller performance drop. Actually, rain and snow can be viewed as additive noise to the original images, while such noises can be viewed as outliers in a local region. In the convolutional operations of CNNs, such outlier will lead to unexpected local maximum or minimum, while such wrongly extracted local extremum will lead to inaccurate semantic features in the Base-VGG-F. Surprisingly, the performance decrease in MAC is often less than Base-VGG-F, which may be caused by the fact that the Auxiliary attention stream is capable to ignore such frequently appeared fake local extremum and enforces the semantic streams focus on the attractive regions. These results further validate the scalability of the proposed two-stream attentive CNNs.

4.4 Performance Analysis

Finally, we conduct an experiment to see the influence of the parameter λ, which controls the way that the two streams are fused. By varying λ from 0.0 to 1.0 with a step of 0.2, we test the performance of MAC on the Flower dataset and obtain a performance curve (as shown in Fig. 6). We find that over-emphasizing either stream will lead to degraded performance, and the best performance is achieved at λ=0.6, indicating the Main stream has weight 0.6 and the Auxiliary stream has weight 0.4.

5 CONCLUSION

In this paper, we propose two-stream attentive CNNs for image retrieval. By initializing a Main stream for semantic feature extraction and an Auxiliary stream for attention prediction, the two-streams fused and fine-tuned on image retrieval datasets. In this manner, the capability of the whole network in capturing inherent query intention can be improved. Experimental results show that the proposed approach has impressive performance on two image retrieval datasets and two fine-grained image classification datasets.

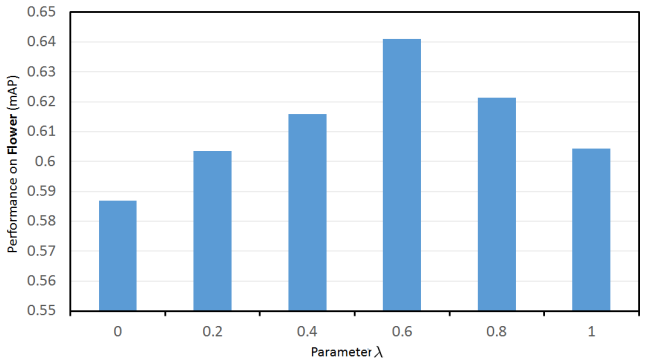

Figure 6: Influence of fusion weight λ on Flower dataset.

Moreover, its performances on retrieving low-resolution/low-quality and rain/snow images are also very promising.

In our future work, we will seek to train a network with attention cues embedded in many locations of semantic feature extraction so as to extract more discriminative features for outdoor scenes. Moreover, the hashing operations will be embedded into the network so that the retrieval process can become much faster.

ACKNOWLEDGMENTS

This work was supported in part by National Natural Science Foundation of China (No.61572065, No.61532005, No.61370113), National Key Research and Development of China (No.2016YFB0800404), Joint Fund of Ministry of Education of China and China Mobile (No.MCM20160102).

REFERENCES

[1] Satrajit Acharya and M. R. Vimala Devi. 2012. Image retrieval based on visual attention model. *Procedia Engineering* 30 (2012), 542–545.
[2] Cristhian A. Aguilera, Francisco J. Aguilera, Angel D. Sappa, Cristhian Aguilera, and Ricardo Toledo. 2016. Learning cross-spectral similarity measures with deep convolutional neural networks. In *CVPR 2015, Worshop Perception Beyond The Visible Spectrum*. 267–275.
[3] Ceyhun Burak Akgl, Daniel L. Rubin, Sandy Napel, Christopher F. Beaulieu, Hayit Greenspan, and Burak Acar. 2011. Content-based image retrieval in radiology: Current status and future directions. *Journal of Digital Imaging* 24, 2 (2011), 208–22.
[4] B Andre, T Vercauteren, A. M. Buchner, M. B. Wallace, and N Ayache. 2012. Learning semantic and visual similarity for endomicroscopy video retrieval. *IEEE Transactions on Medical Imaging* 31, 6 (2012), 1276–1288.
[5] Artem Babenko and Victor Lempitsky. 2015. Aggregating deep convolutional features for image retrieval. *Computer Science* (2015).
[6] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. 2014. Neural codes for image retrieval. 8689 (2014), 584–599.
[7] Vassileios Balntas, Edward Johns, Lilian Tang, and Krystian Mikolajczyk. 2016. PN-Net: Conjoined triple deep network for learning local image descriptors. (2016).
[8] Giulia Boato, Duc Tien Dang-Nguyen, Oleg Muratov, Naif Alajlan, and Francesco G. B. De Natale. 2015. Exploiting visual saliency for increasing diversity of image retrieval results. *Multimedia Tools. Applications* (2015), 1–22.
[9] Jure Bontar and Yann Lecun. 2015. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research* 17, 1 (2015), 2287–2318.
[10] J Faruque, C. F. Beaulieu, J Rosenberg, D. L. Rubin, D. Yao, and S Napel. 2015. Content-based image retrieval in radiology: analysis of variability in human perception of similarity. *J Med Imaging* 2, 2 (2015), 025501.
[11] Emmanouil Giouvanakis and Constantine Kotropoulos. 2014. Saliency map driven image retrieval combining the bag-of-words model and PLSA. In *International Conference on Digital Signal Processing (DSP)*. 280–285.
[12] Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik. 2014. Multi-scale orderless pooling of deep convolutional activation features. 8695 (2014), 392–407.
[13] K. He, X. Zhang, S. Ren, and J. Sun. 2014. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE TPAMI* 37, 9 (2014), 1904–16.
[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. (2015), 770–778.
[15] Xiangteng He and Yuxin Peng. 2017. Weakly supervised learning of part selection model with spatial constraints for fine-grained image classication. *AAAI Conference on Artificial Intelligence (AAAI)* (2017), 4075–4081.
[16] Ahmet Iscen, Michael Rabbat, and Teddy Furon. 2016. Efficient large-scale similarity search using matrix factorization. In *CVPR*. 2073–2081.
[17] Herve Jegou, Matthijs Douze, and Cordelia Schmid. 2008. *Hamming embedding and weak geometric consistency for large scale image search*. OAI. 304–317 pages.
[18] Herve Jegou, Matthijs Douze, and Cordelia Schmid. 2008. Hamming embedding and weak geometric consistency for large scale image search. In *European Conference on Computer Vision*. 304–317.
[19] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. *arXiv:1408.5093* (2014).
[20] Youngrae Kim, Yunhee Shin, Sojung Kim, Eun Yi Kim, and Hyoseop Shin. 2009. EBIR: Emotion-based image retrieval. In *International Conference on Consumer Electronics*. 1–2.
[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*. 1097–1105.
[22] Wei Lin Ku, Hung Chun Chou, and Wen Hsiao Peng. 2015. Discriminatively-learned global image representation using CNN as a local feature extractor for image retrieval. In *VCIP*. 1–4.
[23] Hanjiang Lai, Yan Pan, Ye Liu, and Shuicheng Yan. 2015. Simultaneous feature learning and hash coding with deep neural networks. In *CVPR*. 3270–3278.
[24] J. Li, L.-Y. Duan, X. Chen, T. Huang, and Y. Tian. 2015. Finding the secret of image saliency in the frequency domain. *IEEE TPAMI* 37, 12 (2015), 2428–2440.
[25] Yangyan Li, Hao Su, Charles Ruizhongtai Qi, Noa Fish, Daniel Cohen-Or, and Leonidas J Guibas. 2015. Joint embeddings of shapes and images via CNN image purification. *ACM TOG* 34, 6 (2015), 234.
[26] Lixin Liao, Shikui Wei, Yao Zhao, and Guanghua Gu. 2016. Improving the similarity estimation via score distribution. 1–6.
[27] W. Liu, J. Wang, R. Ji, Y. G. Jiang, and S. F. Chang. 2012. Supervised hashing with kernels. In *CVPR*. 2074–2081.
[28] Maria Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing*. 722–729.
[29] David Nister and Henrik Stewenius. 2006. Scalable recognition with a vocabulary tree. In *CVPR*. 2161–2168.
[30] Junting Pan, Elisa Sayrol, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. 2016. Shallow and deep convolutional networks for saliency prediction. In *CVPR*. 598–606.
[31] Mattis Paulin, Julien Mairal, Matthijs Douze, Zaid Harchaoui, Florent Perronnin, and Cordelia Schmid. 2016. Convolutional patch representations for image retrieval: An unsupervised approach. *IJCV* (2016), 1–20.
[32] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. 2008. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*. 1–8.
[33] Filip Radenovi, Giorgos Tolias, and Ondej Chum. 2016. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. *arXiv:1604.02426* (2016).

[34] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. 2014. CNN features off-the-shelf: An astounding baseline for recognition. In *CVPR Workshops*. 512–519.
[35] Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki, and Stefan Carlsson. 2015. A baseline for visual instance retrieval with deep convolutional networks. *Computer Science* (2015).
[36] Ruslan Salakhutdinov and Geoffrey Hinton. 2009. Semantic hashing. *International Journal of Approximate Reasoning* 50, 7 (2009), 969–978.
[37] Fumin Shen, Chunhua Shen, Wei Liu, and Heng Tao Shen. 2015. Supervised discrete hashing. In *CVPR*. 37–45.
[38] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556* (2014).
[39] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. 2000. Content-based image retrieval at the end of the early years. *IEEE TPAMI* 22, 12 (2000), 1349–1380.
[40] J. Tang, Z. Li, M. Wang, and R. Zhao. 2015. Neighborhood discriminant hashing for large-scale image retrieval. *IEEE Transactions on Image Processing* 24, 9 (Sept 2015), 2827–2840. DOI:https://doi.org/10.1109/TIP.2015.2421443
[41] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. *The caltech-UCSD birds-200-2011 dataset*. Technical Report.
[42] Shouhong Wan, Peiquan Jin, and Lihua Yue. 2009. An approach for image retrieval based on visual saliency. In *International Conference on Image Analysis and Signal Processing*. 172–175.
[43] J. Wang, S. Kumar, and S. F. Chang. 2010. Semi-supervised hashing for scalable image retrieval. In *CVPR*. 3424–3431. DOI: https://doi.org/10.1109/CVPR.2010.5539994
[44] Xiaojuan Wang, Ting Zhang, Guo-Jun Qi, Jinhui Tang, and Jingdong Wang. 2016. Supervised quantization for similarity search. In *CVPR*. 2018–2026.
[45] Shikui Wei, Dong Xu, Xuelong Li, and Yao Zhao. 2013. Joint optimization toward effective and efficient image search. *IEEE Trans. on Cybernetics* 43, 6 (2013), 2216–2227.
[46] Shikui Wei, Yao Zhao, Ce Zhu, Changsheng Xu, and Zhenfeng Zhu. 2011. Frame fusion for video copy detection. *IEEE Transactions on Circuits and Systems for Video Technology* 21, 1 (2011), 15–28.
[47] Shikui Wei, Yao Zhao, Zhenfeng Zhu, and Nan Liu. 2010. Multimodal fusion for video search reranking. *IEEE Transactions on Knowledge and Data Engineering* 22, 8 (2010), 1191–1199.
[48] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. 2017. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017).
[49] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Zequn Jie, Yanhui Xiao, Yao Zhao, and Shuicheng Yan. 2016. Learning to segment with image-level annotations. *Pattern Recognition* 59 (2016), 234–244.
[50] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. 2016. STC: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2016).
[51] Yunchao Wei, Wei Xia, Min Lin, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, and Shuicheng Yan. 2016. HCP: A flexible CNN framework for multi-label image classification. *IEEE transactions on pattern analysis and machine intelligence* 38, 9 (2016), 1901–1907.
[52] Zhenkun Wen, Jinhua Gao, Ruijie Luo, and Huisi Wu. 2014. *Image retrieval based on saliency attention*. Springer Berlin Heidelberg, Berlin, Heidelberg, 177–188. DOI:https://doi.org/10.1007/978-3-642-54924-3_17
[53] R. Xia, Y. Pan, H. Lai, C. Liu, and S. Yan. 2014. Supervised hashing for image retrieval via image representation learning. (2014).
[54] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. 2015. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 842–850.
[55] Wei Yu, Kuiyuan Yang, Hongxun Yao, Xiaoshuai Sun, and Pengfei Xu. 2016. Exploiting the complementary strengths of multi-layer CNN features for image retrieval. *Neurocomputing* (2016).
[56] S. Zagoruyko and N. Komodakis. 2015. Learning to compare image patches via convolutional neural networks. In *CVPR*. 4353–4361.
[57] Xiaofan Zhang, Wei Liu, M Dundar, and S Badve. 2015. Towards large-scale histopathological image analysis: Hashing-based image retrieval. *IEEE Transactions on Medical Imaging* 34, 2 (2015), 496–506.
[58] Fang Zhao, Yongzhen Huang, Liang Wang, and Tieniu Tan. 2015. Deep semantic ranking based hashing for multi-label image retrieval. In *CVPR*. 1556–1564.
[59] Liang Zheng, Yi Yang, and Qi Tian. 2016. SIFT meets CNN: A decade survey of instance retrieval. (2016).
[60] Liang Zheng, Yali Zhao, Shengjin Wang, Jingdong Wang, and Qi Tian. 2016. Good practice in CNN feature transfer. *arXiv:1604.00133* (2016).
[61] Dan Zhou, Xue Li, and Yu Jin Zhang. 2016. A novel CNN-based match kernel for image retrieval. In *IEEE ICIP*. 2445–2449.