Multi-Scale Context Attention Network for Image Retrieval

Yihang Lou[1], Yan Bai[1], Shiqi Wang[2], Ling-Yu Duan[1*]
National Engineering Lab for Video Technology, Peking University, Beijing, China[1]
Computer Science, City University of Hong Kong, Hong Kong[2]
{yihanglou,yanbai,lingyu}@pku.edu.cn,shiqwang@cityu.edu.hk

ABSTRACT

Recent attempts on the Convolutional Neural Network (CNN) based image retrieval usually adopt the output of a specific convolutional or fully connected layer as feature representation. Though superior representation capability has yielded better retrieval performance, the scale variation and clutter distracting remain to be two challenging problems in CNN based image retrieval. In this work, we propose a Multi-Scale Context Attention Network (MSCAN) to generate global descriptors, which is able to selectively focus on the informative regions with the assistance of multi-scale context information. We model the multi-scale context information by an improved Long Short-Term Memory (LSTM) network across different layers. As such, the proposed global descriptor is equipped with the scale aware attention capability. Experimental results show that our proposed method can effectively capture the informative regions in images and retain reliable attention responses when encountering scale variation and clutter distracting. Moreover, we compare the performance of the proposed scheme with the state-of-the-art global descriptors, and extensive results verify that the proposed MSCAN can achieve superior performance on several image retrieval benchmarks.

KEYWORDS

Image Retrieval, Multi-Scale Context, Attention Network

ACM Reference Format:
Yihang Lou, Yan Bai, Shiqi Wang, Ling-Yu Duan. 2018. Multi-Scale Context Attention Network for Image Retrieval. In 2018 ACM Multimedia Conference (MM '18), October 22-26, 2018, Seoul, Republic of Korea. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3240508.3240602

1 INTRODUCTION

Instance image retrieval aims to retrieve an image depicting a particular object in a query from an image database, which has received a lot of research focus. The success of Convolutional Neural Network (CNN) in recent years has greatly facilitated the advance of image retrieval owing to its discriminative power and compact representation. Though significant improvements have been achieved by deep learning based descriptors, two challenges presented by clutter distracting and scale changes still exist in real applications, as shown in Figure 1. First, clutter distracting as irrelevant information greatly affects the feature representation on the informative regions for image retrieval. Second, the interest/target objects in the query and reference images are often different in terms of scale. In this work, we mainly focus on exploiting the multi-scale feature representation of informative regions in images.

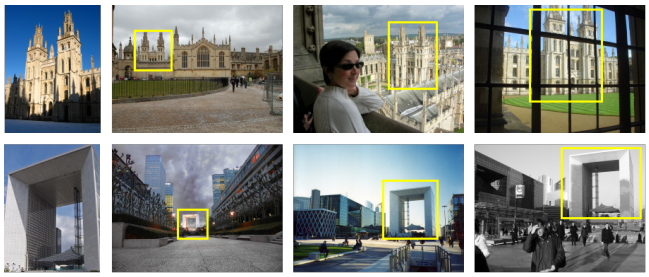

Figure 1: Examples from Oxford5K (first row) and Paris6K dataset (second row). The first column indicates query images, and the three right columns denote reference images.

In real application scenarios, the clutter distracting will severely hinder the feature matching procedure. As for image retrieval, focusing on informative regions in images is beneficial to generate discriminative feature. Recently, the CNN based features are mostly trained as global descriptors using siamese or triplet network [11] [2] [29]. As these global features are directly extracted from the output of the last convolutional layer followed by a max or average pooling [35] [15] [16], it is difficult to handle the complex scenes since the target objects in images are mostly unaligned and even take up only a small portion in some extreme cases. Therefore, it is beneficial to selectively focus on the informative regions and ignore the irrelevant ones. Such selectively focusing scheme is also termed as attention, which has been demonstrated to be effective in various research areas, such as machine translation [5], speech recognition [7] and image caption [39]. A typical attention mechanism applied in CNNs is to predict an attention map in which the value of each patch indicates the informativeness in corresponding locations.

Scale is a predominant factor affecting feature representation in image retrieval. In different scales, the attentive regions would be different. A representative work is scale invariant feature transform (SIFT) [19], which finds extreme responses at multi-scale gaussian pyramid as feature points for image matching. However, in existing deep learing based methods, the multi-scale context that is the relevance between the attentive regions in different scales has not been fully explored. Current network used to generate scale robust feature is commonly equipped in training stage with data

*Ling-Yu Duan is the corresponding author.

MM '18, October 22–26, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5665-7/18/10…$15.00
https://doi.org/10.1145/3240508.3240602

augmentation (*i.e.*, randomly resize or crop the training images, etc.), or the concatenation feature of input images with different scales is obtained as the final feature. In some extreme cases, when the interest objects take small portions of the input images, in the network forward stage, it is hard to retain the responses as the size of feature map consistently decreases. In order to perform reliable attention across different scales, intuitively, we need to acquire multi-scale context information. However, few attempts have been made to explore the context between different scales attentions.

In this paper, we propose a **M**ulti-**S**cale **C**ontext **A**ttention **N**etwork (MSCAN) which performs selective attention across multiple layers of different scales. The attentions from multiple scales constitute attention sequence in which we can explore the context information. Specifically, such context within the attention sequence is modeled by a two layer Long Short-Term Memory (LSTM) network. The first layer encodes the attention maps at different scales and generates initial multi-scale context memory. Then this context memory is fed to the second LSTM layer to assist the network to selectively focus on the informative attention and further produce a multi-scale aware attention. That is to say, if attention responses at a specific scale are informative regarding the multi-scale context, the LSTM network will import more information. To the best of our knowledge, our scheme is the first work to approach image retrieval utilizing attention mechanism, coupled with recurrent memory network to model multi-scale context in feature representation.

To sum up, our main contributions in this paper are as follows:

- First, we propose a multi-scale context attention network which stacks multiple attention modules in multiple layers of different scales. Thus, we are able to capture the most informative regions from multiple scales.
- Second, we explore the context among different scales attentions. Such context information is modeled by a LSTM network with context memory and attention gate to adaptively select the attentions from multiple scales.
- Third, our proposed MSCAN achieves superior performance on all the evaluated image retrieval benchmarks and the visualization results further provide evidences of the effectiveness of our method.

The rest of the paper is organized as follows. The related work is discussed in Section 2. Our network structure and learning details are presented in Section 3. In Section 4, we describe experimental setup and analyze qualitative and quantitative results. Finally, we conclude this paper in Section 5.

2 RELATED WORK

2.1 Image Retrieval

Image retrieval has drawn a quantity of research focus over the last decades. Early image retrieval methods rely on handcrafted local features, such as SIFT [19], SURF [6], ORB [30], which are all equipped with the invariance properties in terms of scale and rotation to some extent. These local descriptors are commonly combined with vocabulary trees to achieve image retrieval. With Bag of Words (BoW) [26] and geometric re-ranking, these descriptors can obtain competitive retrieval accuracy in several retrieval benchmarks. However, in the context of large scale image retrieval, the tedious matching procedure of traditional local descriptors cannot meet the practical requirements, which motivated the global descriptors such as the VLAD [14] and Fish Vector [25] based aggregated descriptor. Compared with BoW features, the aggregated global features are superior from the perspective of retrieval accuracy and feature compactness.

Recently, CNN based features have been widely adopted owing to its strong ability in semantic representation. Azizpour *et al.* [3] has shown that the max pooling of feature maps of CNNs (e.g., the output of intermediate layers) can generate more effective representations than the fully connected layers. Regional Maximum Activation of Convolutions (RMAC) was proposed in [35], which averages max pooled features over a set of multi-scale regions of interest (ROI) in feature maps. Mahedano *et al.* [22] proposed a saliency scheme to build bag of local convolutional features for efficient image representation. However, these methods mainly use off-the-shelf CNN pretrained model such as VGG16 [33] as a feature extractor, while recent works tend to fine-tuning on target dataset to get further performance improvements. In [11], Gordo *et al.* proposed an end-to-end learning framework on R-MAC feature representation with triplet network. In [2], the VLAD layer was proposed in the network to train an aggregated global feature. Filip *et al.* [29] proposed an unsupervised fine-tuning scheme using hard examples to learn global features. However, these recent methods have not fully explored the active selection of the informative responses from multi-scale context.

2.2 Attention Model

Our method is closely relevant to the attention model [5, 7, 20, 32, 38, 39] which allows the networks to selectively focus on specific information. Attention Model has been employed in the machine translation, action recognition, image caption, etc. Xu *et al.* [39] proposed soft and hard attention for image caption generation. Stollenga *et al.* [34] proposed a deep attention selective network for image classification. Yao *et al.* [40] designed a temporal attention scheme for video caption generation. Luong *et al.* [29] proposed to fuse both global and local attention for neural machine translation.

Though CNNs based methods [2, 11, 23, 29, 42] have been widely used for image retrieval, most of them have not taken attention mechanism into consideration. In [23], Noh *et al.* leveraged soft attention for keypoint selection to generate local descriptors. For deep global descriptors, the current works mainly focus on better pooling methods in learning stage such as regional pooling [11], pooling orders [18], etc. As such, the attention mechanism has not been fully explored in deep global descriptors generation. For global descriptors, scale variation would result in different attention prediction. Our method focuses on leveraging multi-scale context to attentively select informative responses on convolutional features. In particular, we use an improved LSTM to model the multi-scale context of attention maps at different scales, which can produce the more reliable attention.

2.3 Contextual Modeling

Contextual cues are important for feature representation. Many successful image retrieval systems are developed on CNNs which implicitly encode context information by cascading multiple layers.

The scale context information only flows between sibling layers. To tackle this limitation, previous work [17] attempted to build skip connections from earlier layers and ultimately aggregating the intermediate features for better semantic segmentation. Other works [8, 9] aimed to learn hierarchical and multi-scale pyramid to capture the context information. Another series of works like [43] leveraged dilated convolutions to perform contextual modeling. Recently, recurrent network like RNN and LSTM [1] are often added after convolutional layer to capture the context information between local patches in one layer. However, the multi-scale context at different layers has not been fully explored in image retrieval.

Alternatively, we leverage attention mechanism to focus on the informative regions in particular scales. As the attentive regions may not be consistent across different scales, in this work, we aim to explore the relevance between the multi-scales attentions, namely multi-scale context.

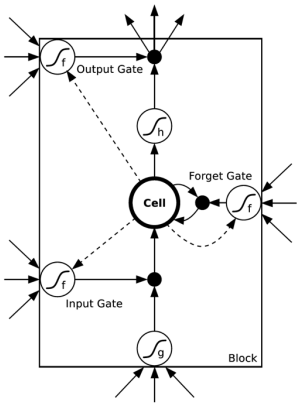

Figure 2: The visualization of LSTM unit.

3 PROPOSED METHOD

In network forward propagation, the scales of feature maps consistently decrease, and the attentive regions would correspondingly change in terms of scale. A similar example is the SIFT descriptor [19] which is collected from extreme responses from different scales. Consequently, we introduce multi-scale context in attention representation to encode the representative attention. More specifically, we introduce a two layer Long Short-Term Memory (LSTM) network to encode the multi-scale context of attention. The attention sequence composed of attention maps from multiple scales is established, and at each step of LSTM, we feed it with an attention map from a specific scale. We particularly set up a multi-scale context memory in the first layer of LSTM which memorizes the context for attention. Additionally, in the second layer of LSTM, we build attention gates that collaborate with the context-memory to selectively import attention information.

In this section, we first introduce the attention module we construct. Then we briefly revisit the mechanism of LSTM unit and describe our proposed MSCAN in details. Finally, the model optimization procedure is further presented.

3.1 Attention Module

Selective representation is crucial for improving discriminative capability of features. Unlike the face recognition [31] and person ReID [37] tasks where the interest objects (like face or person) are cropped and aligned, in image retrieval, the target objects are often surrounded by complex background.

We design a soft attention module to actively select responses in network. In this work, we use the ResNet101[12] as base network. Attention module can be regarded as another branch to compute the importance score for each patch in feature maps. Given the input x, we obtain the output feature of the network $f(x)$, and the attention module computes attention score $s(x)$, which is used to softly weight the output features. The output attention score $s_{i,j}(x)$ can be regarded as gates for base branch $b_{i,j}(x)$, which can be formulated as:

$$b_{i,j}(x) = f_{i,j}(x) \odot s_{i,j}(x), \tag{1}$$

where i, j indicate patch positions over all feature maps. This procedure is an element-wise product, such that distracting responses can be suppressed and the responses on interest objects can be promoted. This module can be trained end-to-end with back-propagation algorithm, and the partial derivative of $b_{i,j}(\mathrm{x})$ is given by:

$$\frac{\partial b_{i,j}(x)}{\partial \theta} = \frac{\partial f_{i,j}(x) s_{i,j}(x,\theta)}{\partial \theta} = f(x)\frac{\partial s_{i,j}(x,\theta)}{\partial \theta}, \tag{2}$$

where θ are the parameters in attention module. It is worth noting that $s_{i,j}(\cdot)$ is constrained to non-negative during training.

To build up multi-scale context, we add multiple attention modules in ResNet101 after each residual block. In particular, our attention module consists of two convolutional layers with kernel size (1x1). The output of the second layer is applied with softmax function to get attention score for each patch. Directly feeding the weighted feature into next layers would affect the stability of network learning. Since the attention weights range from 0 to 1, the original identity mapping in residual block is changed. Therefore, we also establish an identity mapping for attention as in [36], which can be formulated as follows:

$$b_{i,j}(x) = f_{i,j}(x) \odot s_{i,j}(x) + f_{i,j}(x). \tag{3}$$

The motivation is similar to residual learning [12], as the identity mapping in attention modules ensures that the adding of attention would not be worse than without it.

3.2 Revisit Long Short-Term Memory Unit

In our method, the Long Short-Term Memory (LSTM) network are involved, which is constructed by stacking LSTM units. A typical LSTM unit consists of an input gate i_t, a forget gate f_t, an output gate o_t and hidden state h_t, alongside with an memory cell c_t. The visualization of LSTM unit is shown in Figure 2. The computation process can be formulated as follows:

$$\begin{pmatrix} i_t \\ f_t \\ o_t \\ u_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ tanh \end{pmatrix} \left(W \begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix} \right) \tag{4}$$

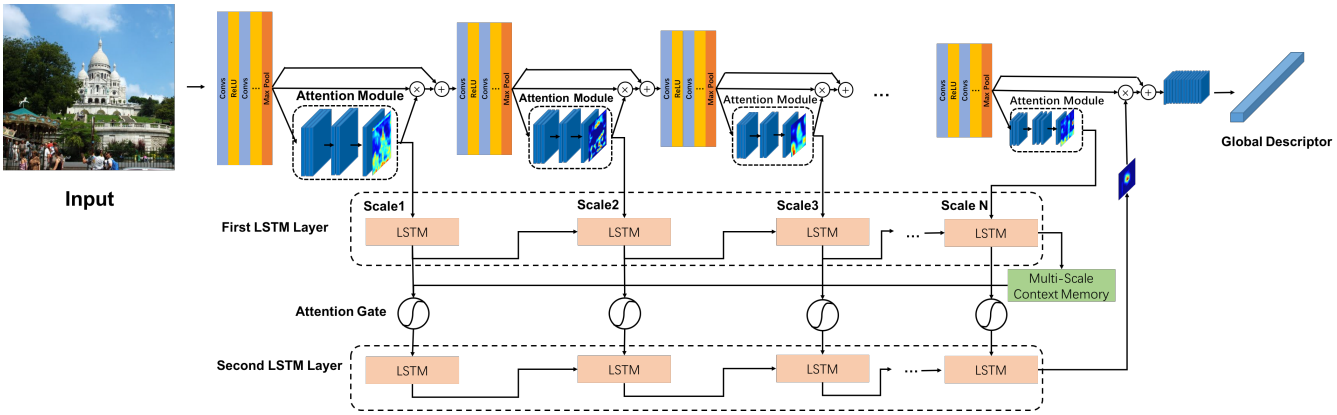

Figure 3: The visualization of the proposed MSCAN for image retrieval. Multiple attention modules are incroporated. The attention weights from different scales constitute attention sequence to be fed into LSTM. At training stage, image triplets are sampled and fed into the MSCAN and triplet loss is computed based on the distances between their global descriptors.

$$c_t = i_t \odot u_t + f_t \odot c_{t-1} \tag{5}$$

$$h_t = o_t \odot tanh(c_t), \tag{6}$$

where x_t, u_t and h_{t-1} are the input, modulated input and previous hidden state at step t. σ indicates the activation function sigmoid in LSTM unit. Operation $\odot$ denotes element-wise product. The three gates i_t, f_t and o_t are distinctive characteristics of LSTM unit for different purposes. The input gate i_t determines the degree of importing information from modulated input u_t to update c_t. Then, forget gate f_t controls importing information from previous state of the cell c_{t-1} in step t. In the last step, output gate o_t determines the degree of output from memory cell.

3.3 Scale Context Attention Network

Our proposed MSCAN is illustrated in Figure 3. It contains three major modules, base CNN structure, attention modules and LSTM network. The CNN structure progressively forwards and outputs feature representation. Multiple attention modules at specific scale layers produce attention map on the corresponding output feature maps. The value of each location in attention map denotes informativeness of each patch in feature maps. As the dimension of input feature to LSTM at each step are required to be uniform, before being fed into the LSTM unit, these different scales attention maps are downsampled or upsampled to a uniform scale with convolutional and deconvolutional operations. Then the first LSTM layer encodes the attention weights of multiple scales and forms the multi-scale context memory. The second LSTM layer performs attention over the output of the first layer with the assistance of multi-scale context memory, and the last step output of the second layer are used as multi-scale aware attention weights to generate attentive global descriptors.

Multi-Scale Context Memory Module: Since we attempt to selectively import attention from different scales, the multi-scale context memory should be obtained. We consider to leverage the output of the first layer in LSTM to generate a multi-scale context memory. In particular, we employ the averaged hidden status h_t from all steps T in the first LSTM layer to obtain this context memory M. It can be formulated as follows:

$$M = \frac{1}{T}\sum_{t=1}^{T} h_t. \tag{7}$$

At each step t, the LSTM receives the attention map from a specific scale. Besides, we also consider to feed all the hidden states of the first layer to another forward network. For simplicity and model compactness, we adopt averaging methods since adding another subnetwork would involve more parameters.

Attention in the Second LSTM Layer: We further assess the informativeness degree of the input in the second layer where an attention gate g_t is set up to selectively control the attention information imported from each scale. It receives the input h_t and multi-scale context memory M, which can be formulated as follows:

$$e_t = tanh\left(W_{e1}\begin{pmatrix} h_t \\ M \end{pmatrix}\right) \tag{8}$$

$$g_t = \frac{exp(e_t)}{\sum_{u=1}^{T} exp(e_u)}, \tag{9}$$

where g_t is the normalized attention gate for the input at t step. After adding learnt attention gate g_t, the cell state updating rule in second LSTM layer is also changed as follows:

$$c_t = g_t \odot i_t \odot u_t + (1 - g_t) \odot f_t \odot c_{t-1}, \tag{10}$$

It indicates that if input attention h_t is important regarding multi-scale context, the cell in the second layer will import more attention information from it; while it is less informative we tend to block it and make better use of history information in LSTM units.

Global Descriptor Generation: The aim of our MSCAN is to generate more discriminative global descriptors which are able to be aware of the multi-scale context when modeling attention. The outputs of the second layer LSTM at the last step are treated as the final attention weights. The output of the last convolutional layer $\mathbb{F}(x)$ is then applied with the final attention as follows:

$$\mathbb{F}'(x) = \mathbb{F}(x) + \mathbb{F}(x) \odot S(x), \tag{11}$$

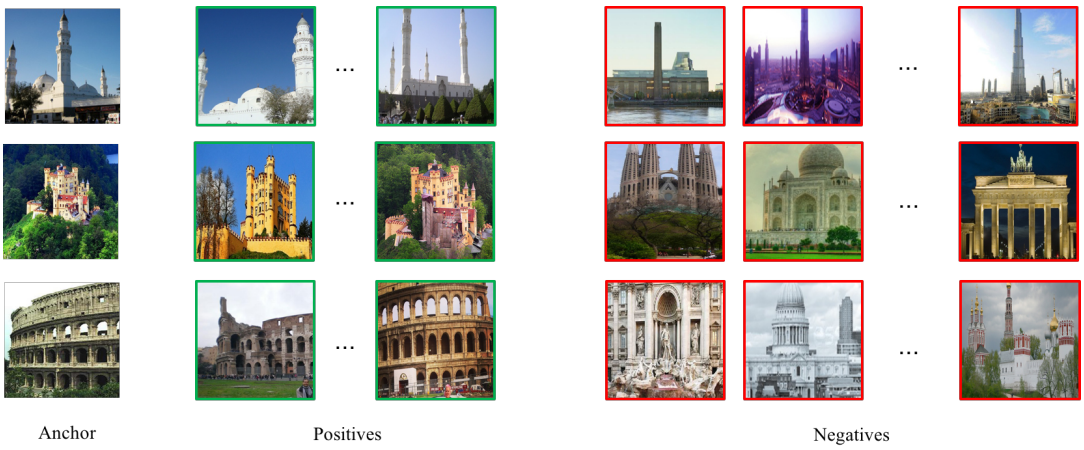

Figure 4: The training samples for building triplet units. In our experimental setting, each triplet group contains an anchor sample, two positive samples and five negative samples. In each group, the farthest positive sample and nearest negative samples regarding anchor sample are selected to compute triplet loss.

where $\mathbb{F}'(x)$ is the ultimate feature representation. Subsequently, we perform global max-pooling over $\mathbb{F}'(x)$ to generate deep global descriptor.

3.4 Model Optimization

We adopt a triplet ranking loss to train our proposed model. The triplet network aims to project samples into an embedding space where those samples belonging to the same class are closer than those from different classes. Let $< x, x^p, x^n >$ denote a triplet unit, where x is an anchor sample, x^p belongs to the same class with x, and x^n belongs to another class. The constraint can be formulated as:

$$d(x, x^p) + \alpha \leq d(x, x^n), \tag{12}$$

where α is a scalar that controls the margin between positive and negative samples. The loss function can be defined as:

$$L(x, x^p, x^n) = \frac{1}{2}\max\{\|f(x) - f(x^p)\|_2^2 + \alpha - \|f(x) - f(x^n)\|_2^2, 0\}. \tag{13}$$

The optimization process of triplet loss is less efficient due to the dramatic data expansion and the sensitivity to the selection of triplet units. As for computing triplet loss, each iteration takes dozens of triplet units, but only a minority may violate the constraints. As such, the improper triplet units can seriously degrade the performance of trained models. Therefore, we perform online hard example minings to make training more efficient. We define the hard triplets as the triplet units breaking margin constraints. More specifically, we forward randomly selected triplet units to compute triplet loss. These triplets violating constraints are recorded and collected. Then we further feed these "filtered in" hard triplet units into the network again to compute loss and perform back propagation. The sampled triplet units are visualized in Figure 4.

3.5 Implementation Details

We choose ResNet101 as base network [12], and 5 attention modules after each residual blocks are incroporated. In addition, the stride of the last convolutional kernel is changed from 2 to 1 in order to obtain larger receptive field for attention. As for LSTM network, the dimension of each hidden unit equals to the squeezed dimension of output feature in the last convolutional layer. Besides, the downsampling and upsampling before LSTM is implemented by convolution and transposed convolution with kernel size (2x2) and stride 2. In order to process variable size of input images in the testing stage, we additionally inject adaptive max pooling before down and up sampling operations. The deep learning toolbox we used is Pytorch.

3.5.1 Training Details. Regarding the hyper parameters, we set $\alpha = 0.1$ as triplet margin. Since our MSCAN involves multiple modules, to make the training efficient, we adopt a multi-stage training strategy. We first train the base network as a classifier with learning rate 0.0001 with exponential decay for 30 epochs. Subsequently, we add attention modules and fix the parameters of the base network, and only fine-tune the parameters in attention module. Furthermore, we fine-tune both the attention modules and base network. Then, the down/up sampling layers and LSTM subnetwork start training with the whole network at learning rate 0.00001 for 30 epochs and staircase decay at 20th epoch with 1/10 multiplier. The weight decay for all the parameters during training is set to 0.00005.

Data augmentation is also employed to make more efficient training, which contains random crop with a factor 0.8 and probability 0.5. Moreover, horizontal flipping and color jittering are also involved.

3.5.2 Testing Details.

- **Multi-scale feature extraction.** At the testing stage, we adopt a common multi-scale feature extraction strategy as in [11] which proves that improved results can be achieved. Following [11], we resize the input images to different sizes, and then combine the generated global descriptors to a single

descriptor. Three different input image scales are used: 0.5, $\sqrt{2}$, 1. The max side of the input image is resized to 1024.

- **PCA learning.** PCA whitening is a post-processing approach usually used for image retrieval. In [35], Tolias *et al.* learned the PCA on different datasets depending on the target dataset. For example, the method in [35] tested on the Oxford5K applied PCA learned on Paris6K dataset. As both these two dataset are composed of building images, such whitening processing would sacrifice some generalization ability to obtain more competitive results. Instead, we use our training dataset to learn the PCA, which is similar to [21].

4 EXPERIMENTAL RESULTS

In this section, we first describe the dataset used for training and testing. Then we evaluate our methods from different perspectives. Finally, we compare our methods against recent baseline methods on all the benchmarked datasets.

In order to investigate our proposed MSCAN, we perform the experiments with the following different structures: (1) Attention Network (AN), (2) Multi-Scale Attention Network (MSAN), (3) Multi-Scale Context Attention Network (MSCAN):

(1)'AN'. This structure contains only one attention module after the last convolutional layer. Note that in this case, multi-scale attention and LSTM structure are removed.

(2)'MSAN'. This network is similar to our proposed MSCAN but without context information. It contains multiple scale attention modules while the LSTM network for context modeling is removed.

(3)'MSCAN'. This is the proposed MSCAN network.

4.1 Datasets

We train proposed MSCAN model on a landmark dataset [11] and then perform test evaluation on Oxford5K[27], Paris6K [28] and Holiday datasets[13].

4.1.1 *Training Datasets.*

- We use landmark dataset [11] for fine-tuning descriptors. This dataset provides full version and clean version. Due to some invalid URLs in this dataset, the full version we collect contains 135,292 from 586 landmarks and the clean version is a filtered subdataset with SIFT-based matching procedure and 34,593 images. We use the full-version dataset to train a model for initializing the fine-tuned network as performed in the training pipeline of [11]. Fine-tunned network is then trained using the clean version dataset.

4.1.2 *Testing Datasets.*

- Oxford 5K dataset [27] contains 5,062 building images captured in oxford university. It provides 55 queries with annotated region of interest (RoI). We test on this extend version Oxford105K by adding 100K distracting images from [26].
- Paris 6K dataset [28] contains 6,412 building images captured in Paris. It also provides 55 queries as Oxford5K. We also test on its extend version Paris106K.
- Holiday dataset [13] consists of 1491 images of personal holiday photos. The dataset includes a large variety of scene types (natural, man-made, water, etc) and objects as well. It contains 500 queries for retrieval evaluation.

4.2 Compared methods

We compare the proposed scheme with several recently proposed descriptors which are representative and relevant to image retrieval due to their excellent performance on retrieval benchmarks.

RMAC [35]. This is a recent deep global descriptor which is generated by performing regional max pooling over selected rigid regions on output feature maps of intermediate layers. Usually, RMAC uses the off-the-shelf CNN model to extract features.

siaMAC [29]. This is a recent global descriptor which is trained by an unsupervised fine-tuning via exploiting hard positive and negative examples.

DIR [11]. This is an extended version of RMAC, which utilizes end-to-end learning framework to learn RMAC descriptors with a triplet network.

NetVLAD [2]. NetVLAD is a recently proposed network with a learnable VLAD layer to learn aggregated features from local patches.

DELF [23]. DELF is a recently proposed local descriptors which uses attention mechanism to select keypoint in local patches. This feature achieves competitive performance on several benchmarks.

GeM [10]. GeM is generic mean pooling that is a learnable pooling layer to produce discriminative features. GeM achieves the state-of-the-art performance in several existing datasets.

4.3 Quantitative results

We present the comparison results in Table 1 on Oxford5K, Paris6K, Holiday and their extensions Oxford105K, Paris106K, Holiday101K. For the first part of Table 1, we compare our methods with those competitive global descriptors. Our approach has yielded a significant improvement over the state-of-the-art methods GeM [10] from the results on Oxford5K and Paris6K with performance gain of 1.4% mAP and 2.4% mAP. Compared with DIR [11], we also obtain obvious performance superiority with 3.1% mAP and 0.6% mAP improvements on Oxford5K and Paris6K. These two datasets consist of building landmarks in which the scale changes are obvious due to different photo capture conditions. The performance superiority on these two datasets can provide useful evidences that the features generated by our MSCAN are more robust to scale changes and discriminative for retrieval. It should be noted that compared with the DIR [11], we also use the same dataset for training. However, the available training images in our experiment are not complete since their published dataset only provides download URLs of images and some of them are invalid now. The full and clean version of training dataset used in DIR [11] are 168,882 and 42,410, but in our case the numbers are 135,292 and 34,593 respectively. In a sense, such training dataset difference further proves the effectiveness and robustness of our descriptors.

On holiday dataset, the proposed scheme also outperforms the state-of-the-art though the performance margin is not that impressive. It can be explained that the training dataset we used is a building dataset, while the holiday dataset is mainly composed of scenes and objects. Hence, the domain difference in training

Table 1: Performance comparison with the state-of-the art methods. These methods use different networks like VGG or ResNet. To differentiate them, we show the dimensions of their descriptors. Fine-Tuned (yes/no) means whether the model is trained on another dataset or is an off-the-shelf model *i.e.*, ImageNet pretrained models. The state-of-the-art performances are in bold.

Methods	Fine-Tuned	Dim	Oxford5K	Oxford105K	Paris6K	Paris106K	Holiday	Holiday101K
MAC	No	512	56.4	47.8	72.3	58.0	79.0	66.1
CroW [15]	No	512	70.8	65.3	79.7	72.2	85.1	-
Spoc [4]	No	512	68.1	61.1	78.2	68.4	84.5	-
RMAC [35]	No	512	69.4	63.7	85.2	77.8	86.9	75.1
NIP [41]	No	512	69.3	-	-	-	88.9	-
BLCF [22]	No	336	77.8	-	83.8	-	-	-
siaMAC [29]	Yes	512	77.1	69.5	83.9	76.3	-	-
NetVLAD [2]	Yes	4096	71.6	-	79.7	-	87.5	-
FisherVector [24]	Yes	512	81.5	76.6	82.4	-	-	-
DELF [23]	Yes	128xN	83.8	82.6	85	81.7	-	-
DIR [11]	Yes	2048	86.1	82.8	94.5	90.6	89.1	-
GeM [10]	Yes	2048	87.8	84.6	92.7	86.9	89.5	87.9
AN	Yes	2048	86.0	82.7	91.1	86.2	88.8	86.9
MSAN	Yes	2048	87.4	84.5	92.8	87.3	89.3	87.6
MSCAN	Yes	2048	**89.2**	**85.8**	**95.1**	**91.0**	**90.1**	**88.2**
Methods + Query Expansion								
DELF+QE [23]	Yes	2048	90.0	88.5	95.7	92.8	-	-
DIR+QE [11]	Yes	2048	90.6	89.4	**96.0**	**93.2**	-	-
GeM+QE [10]	Yes	2048	91.0	89.5	95.5	91.9	-	-
MSCAN+QE	Yes	2048	**91.7**	**90.1**	95.8	92.9	-	-

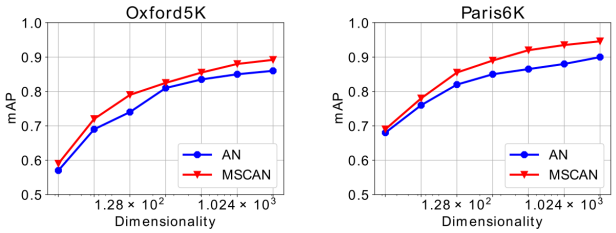

Figure 5: Performance (mAP) variations with the reduced dimensionality of descriptors on retrieval benchmarks.

dataset and testing dataset affects the representation capability of descriptors.

In particular, the MSCAN outperforms both MSAN and AN. This indicates the effectiveness of multi-scale context attention mechanism to yield more discriminative features. From the performance comparison between AN and MSAN, we find that the multi-scale attention mechanism is also beneficial for performance improvements. It can be explained that the attention module focuses on the informative regions, and the multi-scale attention mechanism tries to capture the informative responses in each scale. Moreover, the further improvements by scale context can be explained as: the multi-scale context of attention assists the network to selectively retain attention responses across a range of different scales, and generate scale context aware attention in final feature generation.

Query Expansion. We explore query expansion to further improve the performance, and the comparisons with the state-of-the-art methods involving query expansions are illustrated in the second part of Table 1. Query expansion utilizes the top k results retrieved by first query, then these results undergo a spatial verification. The remaining results are used to perform average aggregation and normalization to generate a new query. The final results are produced by the new query. Spatial verification leads to heavy memory footprint and runtime cost. The comparison methods all adopt complex strategy, for example, DLEF [23] adopts both deep global and local descriptors in query expansion; besides, GeM [10] utilizes a weighted query expansion to tune the results. In our case, we simply apply more cost friendly query-expansion, performing average query expansion on the top recall images. We achieve the best results on the Oxford5K and its extension by a substantial margin compared with others. On Paris6K and its extension, competitive results are also achieved.

Dimension reduction. To further analyze the robustness of the feature representation, we present the performance curve with the variant dimensionality via PCA whitening. Seven different operation points from (32, 64, 128, 256, 512, 1024 and 2048) are plotted in Figure 5. The performance of AN and MSCAN consistently decrease as dimensions reduce. In particular, when the dimensions are reduced to 128, the performance drop tend to be significant. It is clear that MSCAN still have superior performance margin over AN, which further demonstrates the effectiveness of our method.

Multi-Scale Context. Moreover, we further test the effects of different number of scales in context modeling for our attention network and the results are given in Table 2. In our network, the downsampling ratio is $32=2^5$, and therefore we may exploit attention context from 5 different scales attention maps. From small scale to large scale (attention from deep to shallow), we progressively model attention from more scales context. From such progressive

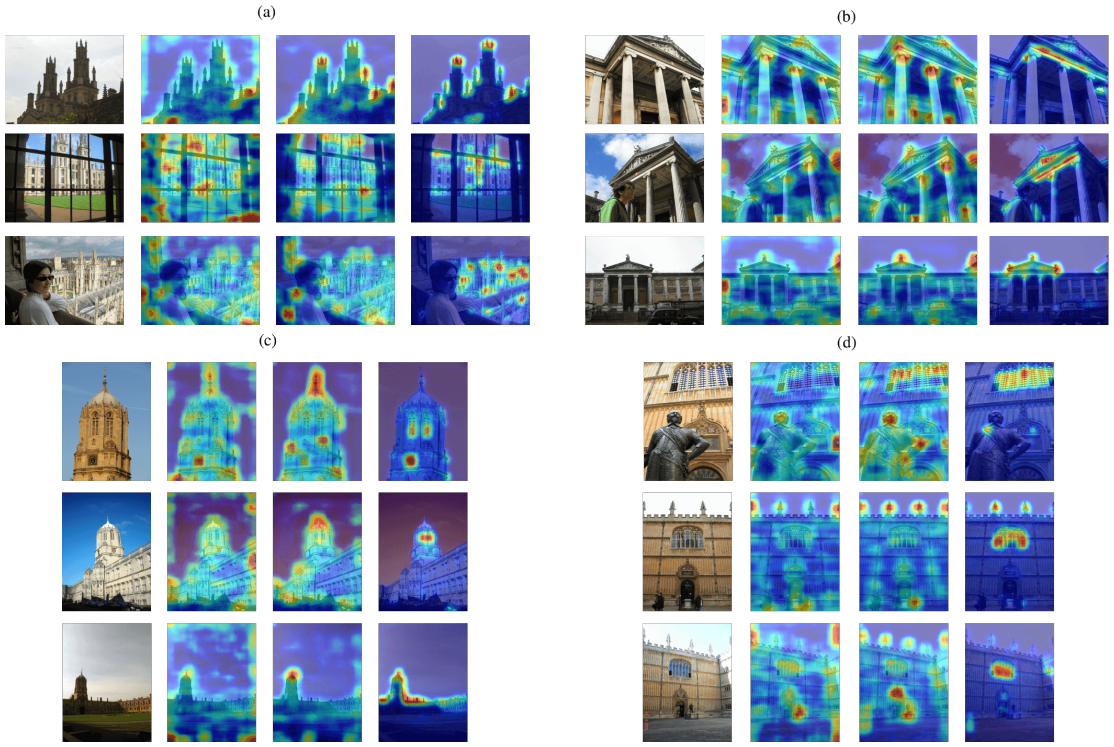

Figure 6: Response feature maps of different methods. These three columns represent the responses from pretrained model, fine-tuned model without and with our multi-scale context attention mechanism respectively. We select three images of a particular building from Oxford5K. There are four groups from (a) to (d) and in each group three images of one particular building in different scales are presented. From up to down, the scale of building in images consistently decreases. Note that these buildings are not included in our training set.

Table 2: Performance (mAP) comparison for different number of scales in context modeling on Oxford5K and Paris6K datasets.

Number of Scales	Oxford5K	Paris6K
1	87.4	92.8
2	87.9	93.3
3	88.8	94.6
4	89.0	95.1
5	89.2	95.1

results, exploiting three scales can obtain performance gains. Moreover, the performance gains consistently increase with more scales context available.

4.4 Visualization and Discussion

To better understand our model, we visualize the responses maps from final attention weighted feature maps which undergo pooling to generate deep global descriptors. The visualization comparisons are listed in Figure 6. Note that these buildings are not included in our training set. Each building is captured from near-middle-far perspectives with different scales. It is interesting to find that our method can better focus on the essential part of target buildings. Moreover, in Figure 6 (a)(b)(d), there are persons or other occlusions in images taking up large portions of content, while comparison methods still generate distinct responses on these irrelevant content. By incorporating scale aware attention mechanism our model can well capture the part of buildings and their distinctive regions.

5 CONCLUSIONS

We present MSCAN, a novel network architecture to generate global descriptors for large scale image retrieval. MSCAN is an incorporated multi-scale context attention mechanism which owns robust capability to selectively focus on the informative regions with the assistance of scale context. An improved LSTM with attention gate and context memory are proposed, which models the multi-scale context between different scales attention map and generate more informative attention for feature representation. The demonstrated experimental results validate the effectiveness of our methods by achieving superior performance on retrieval benchmarks.

6 ACKNOWLEDGEMENT

This work is supported by the National Natural Science Foundation of China (61661146005, U1611461, 61390515) and the National Key Research and Development Program of China (No. 2016YFB1001501), and in part by the PKU-NTU Joint Research Institute (JRI) sponsored by a donation from the Ng Teng Fong Charitable Foundation and Hong Kong RGC Early Career Scheme 9048122 (CityU 21211018), in part by City University of Hong Kong under Grant 7200539/cs.

REFERENCES

[1] Abrar H. Abdulnabi, Bing Shuai, Stefan Winkler, and Gang Wang. 2017. Episodic CAMN: Contextual Attention-Based Memory Networks with Iterative Feedback for Scene Labeling. In *IEEE Conference on Computer Vision and Pattern Recognition*. 6278–6287.

[2] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. 2016. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5297–5307.

[3] Hossein Azizpour, Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki, and Stefan Carlsson. 2015. From generic to specific deep representations for visual recognition. In *Computer Vision and Pattern Recognition Workshops*. 36–45.

[4] Artem Babenko and Victor Lempitsky. 2015. Aggregating Deep Convolutional Features for Image Retrieval. *Computer Science* (2015).

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *Computer Science* (2014).

[6] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. 2006. SURF: Speeded Up Robust Features. In *European Conference on Computer Vision*. 404–417.

[7] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-Based Models for Speech Recognition. *Computer Science* 10, 4 (2015), 429–439.

[8] Clément Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. 2012. Scene parsing with multiscale feature learning, purity trees, and optimal covers. *arXiv preprint arXiv:1202.2160* (2012).

[9] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. 2013. Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence* 35, 8 (2013), 1915–1929.

[10] Radenovi Filip, Giorgos Tolias, and Chum. 2017. Fine-tuning CNN Image Retrieval with No Human Annotation. In *arXiv:1711.02512*.

[11] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. 2016. Deep image retrieval: Learning global representations for image search. In *European Conference on Computer Vision*. Springer, 241–257.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[13] 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*. Springer, 304–317.

[14] Herve Jegou, Matthijs Douze, Cordelia Schmid, and Patrick Perez. 2010. Aggregating local descriptors into a compact image representation. In *Computer Vision and Pattern Recognition*. 3304–3311.

[15] Yannis Kalantidis, Clayton Mellina, and Simon Osindero. 2016. Cross-dimensional weighting for aggregated deep convolutional features. In *European Conference on Computer Vision*. Springer, 685–701.

[16] Jie Lin, Lingyu Duan, Shiqi Wang, Yan Bai, Yihang Lou, Vijay Chandrasekhar, Tiejun Huang, Alex Kot, and Wen Gao. 2017. HNIP: Compact Deep Invariant Representations for Video Matching, Localization and Retrieval. *IEEE Transactions on Multimedia* PP, 99 (2017), 1–1.

[17] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3431–3440.

[18] Yihang Lou, Yan Bai, Jie Lin, Shiqi Wang, Jie Chen, Vijay Chandrasekhar, Ling Yu Duan, Tiejun Huang, Alex Chichung Kot, and Wen Gao. 2017. Compact Deep Invariant Descriptors for Video Retrieval. In *Data Compression Conference*. 420–429.

[19] David G Lowe. 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* 60, 2 (2004), 91–110.

[20] Minh Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. *Computer Science* (2015).

[21] Krystian Mikolajczyk and Jiri Matas. 2008. Improving Descriptors for Fast Tree Matching by Optimal Linear Projection. In *IEEE International Conference on Computer Vision*. 1–8.

[22] Eva Mohedano, Kevin McGuinness, Xavier Giro-i Nieto, and Noel E O'Connor. 2017. Saliency Weighted Convolutional Features for Instance Search. *arXiv preprint arXiv:1711.10795* (2017).

[23] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. 2017. Large-Scale Image Retrieval with Attentive Deep Local Features. In *IEEE International Conference on Computer Vision*. 3476–3485.

[24] Eng-Jon Ong, Sameed Husain, and Miroslaw Bober. 2017. Siamese network of deep fisher-vector descriptors for image retrieval. *arXiv preprint arXiv:1702.00338* (2017).

[25] Florent Perronnin, Yan Liu, Jorge Sánchez, and Hervé Poirier. 2010. Large-scale image retrieval with compressed fisher vectors. In *Computer Vision and Pattern Recognition*. IEEE, 3384–3391.

[26] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. 2007. Object retrieval with large vocabularies and fast spatial matching. In *Computer Vision and Pattern Recognition*. 1–8.

[27] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. 2007. Object Retrieval with Large Vocabularies and Fast Spatial Matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[28] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. 2008. Lost in Quantization: Improving Particular Object Retrieval in Large Scale Image Databases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[29] Filip Radenović, Giorgos Tolias, and Ondřej Chum. 2016. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. In *European Conference on Computer Vision*. Springer, 3–20.

[30] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. 2012. ORB: An efficient alternative to SIFT or SURF. In *IEEE International Conference on Computer Vision*. 2564–2571.

[31] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 815–823.

[32] Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. 2015. Action Recognition using Visual Attention. *Computer Science* (2015).

[33] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[34] Marijn F Stollenga, Jonathan Masci, Faustino Gomez, and Jürgen Schmidhuber. 2014. Deep networks with internal selective attention through feedback connections. In *Advances in neural information processing systems*. 3545–3553.

[35] Giorgos Tolias, Ronan Sicre, and Jegou. 2015. Particular object retrieval with integral max-pooling of CNN activations. *Computer Science* (2015).

[36] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. 2017. Residual Attention Network for Image Classification. (2017), 6450–6458.

[37] Longhui Wei, Shiliang Zhang, Hantao Yao, Wen Gao, and Qi Tian. 2017. Glad: Global-local-alignment descriptor for pedestrian retrieval. In *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 420–428.

[38] Caiming Xiong, Stephen Merity, and Richard Socher. 2016. Dynamic memory networks for visual and textual question answering. In *International Conference on Machine Learning*. 2397–2406.

[39] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *Computer Science* (2015), 2048–2057.

[40] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Describing Videos by Exploiting Temporal Structure. 53 (2015), 199–211.

[41] Dong Yi, Zhen Lei, and Stan Z Li. 2014. Deep metric learning for practical person re-identification. *arXiv preprint arXiv:1407.4979* (2014).

[42] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. 2016. LIFT: Learned Invariant Feature Transform. In *European Conference on Computer Vision*. 467–483.

[43] Fisher Yu and Vladlen Koltun. 2015. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122* (2015).