석 사 학 위 논 문
Master's Thesis

이미지 검색을 위한 영역 관심 기반의 깊은 특징

Regional Attention Based Deep Feature for Image Retrieval

2019

김 재 윤 (金 載 潤 Kim, Jaeyoon)

한 국 과 학 기 술 원

Korea Advanced Institute of Science and Technology

석 사 학 위 논 문

이미지 검색을 위한 영역 관심 기반의 깊은 특징

2019

김 재 윤

한 국 과 학 기 술 원

전산학부

이미지 검색을 위한 영역 관심 기반의 깊은 특징

김 재 윤

위 논문은 한국과학기술원 석사학위논문으로
학위논문 심사위원회의 심사를 통과하였음

2018년 12월 10일

심사위원장 윤 성 의 (인)

심 사 위 원 김 민 혁 (인)

심 사 위 원 조 성 호 (인)

Regional Attention Based Deep Feature for Image Retrieval

Jaeyoon Kim

Advisor: Sung-eui Yoon

A dissertation submitted to the faculty of
Korea Advanced Institute of Science and Technology in
partial fulfillment of the requirements for the degree of
Master of Science in Computer Science

Daejeon, Korea
December 10, 2018

Approved by

Sung-eui Yoon
Professor, Department of Computer Science

The study was conducted in accordance with Code of Research Ethics[1].

[1] Declaration of Ethical Conduct in Research: I, as a graduate student of Korea Advanced Institute of Science and Technology, hereby declare that I have not committed any act that may damage the credibility of my research. This includes, but is not limited to, falsification, thesis written by someone else, distortion of research findings, and plagiarism. I confirm that my thesis contains honest conclusions based on my own careful research under the guidance of my advisor.

MCS
20173130

김재윤. 이미지 검색을 위한 영역 관심 기반의 깊은 특징. 전산학부 . 2019년. 17+iii 쪽. 지도교수: 윤성의. (영문 논문)
Jaeyoon Kim. Regional Attention Based Deep Feature for Image Retrieval. School of Computing . 2019. 17+iii pages. Advisor: Sung-eui Yoon. (Text in English)

초 록

효율적인 이미지 검색을 위해 합성곱신경망(CNN, Convolutional Neural Network)을 사용하는 많은 방법들은 최근 몇 년 동안 특징 매김(Feature Embedding)보다는 특징 집합(Feature Aggregation)에 집중하여 개발 되어 왔다. 왜냐하면 합성 곱 특징(Convolutional Feature) 자체가 상당히 구별을 잘 하는것으로 알려져 있기 때문이다. 그럼에도 불구하고 우리는 이미지 검색에서 잘 알려진 영역 기반의 특징 집합 방법인 R-MAC(Regional-Maximum Activation of Convolutions)이 배경 혼잡과 영역들간의 다른 중요성으로 인해 어려움을 겪고 있음을 발견했다. 본 논문에서 전역적인 중요성(Global Attentiveness)을 고려하여 각 영역들에게 중요성을 나타내는 점수를 가지고 무게를 가하는 간단하고 효과적인 Context-Aware Regional Attention Network 를 제안하고 이것을 통해 위에서 언급한 문제를 해결한다. 우리는 이미지 검색에서 잘 알려진 데이터 셋에 대한 다양한 실험을 수행했고, 우리의 방법이 R-MAC 베이스 라인을 크게 향상시킬뿐만 아니라 "pre-trained single-pass" 범주에서 가장 높은 정확도를 보였습니다. 또한, 본 논문의 방법은 질의 확장 방법(Query Expansion)과 결합 할 때 이전 방법보다 더 높은 정확도 향상을 보여 주었다. 이러한 결과는 R-MAC과 통합 된 우리가 제안한 Context-Aware Regional Attention Network에 기인한다.

Abstract

Many approaches using Convolutional Neural Network (CNN) for efficient image retrieval have concentrated on feature aggregation rather than feature embedding over recent years, since convolutional features have been found to be reasonably discriminative. Nonetheless, we found that a well-known region-based feature aggregation method, R-MAC, for image retrieval is suffered from the background clutter and varying importance of regions. In this work, we tackle these problems with a simple and effective, context-aware regional attention network that weights an attentive score of a region considering global attentiveness. We conduct various experiments on well-known retrieval datasets, and confirm that our method does not only improve the R-MAC baseline significantly, but also present new state-of-the-art results in the category of "pre-trained single-pass". Furthermore, we show that our method shows higher accuracy improvement combined over prior methods, when combined with the query expansion method. These results are attributed by our novel regional-attention network integrated with R-MAC.

Contents

List of Tables

List of Figures

Chapter 1. INTRODUCTION

Content-based image retrieval has been actively growing over recent years, since it can be directly applied to various computer vision applications such as visual place recognition, web-scale image retrieval, face retrieval, and product recognition. Recently, this task has been addressed using Convolutional Neural Network (CNN) with significant improvements. One of early works by Babenko *et al* [1] pointed out that the CNN feature itself is sufficiently discriminative without any embedding and complex aggregation techniques commonly adopted in manually crafted features (e.g., SIFT). Since then, most following techniques [14, 27, 13] have constructed feature vectors from CNN through a simple aggregation method without complex embedding techniques.

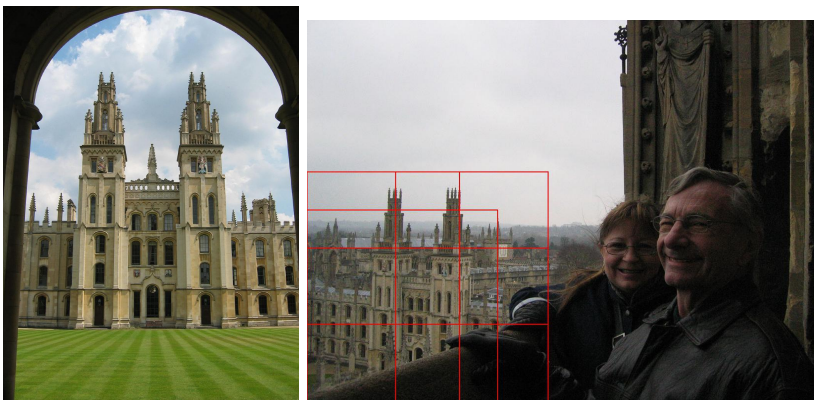
Although many image retrieval techniques have employed an off-the-shelf CNN pre-trained on ILSVRC ImageNet [24], some of recent approaches [22, 5, 19] have tried to fine-tune the CNN with training datasets related to test datasets such as the retrieval-SfM dataset [22]. While these approaches showed improved results in particular datasets, these fine-tuning methods have drawbacks of requiring training datasets with expensive annotations and newly re-training the network with different training datasets depending on a category of a test dataset.

To avoid newly fine-tuning the network depending on target categories of images, we encode images into compact feature vectors using an off-the-shelf CNN, commonly known as a general feature extractor. In this context, we set a target type of this work to the "pre-trained single-pass" category of CNN-based approaches, defined by Zheng *et al* [30] for efficient and accurate image retrieval.

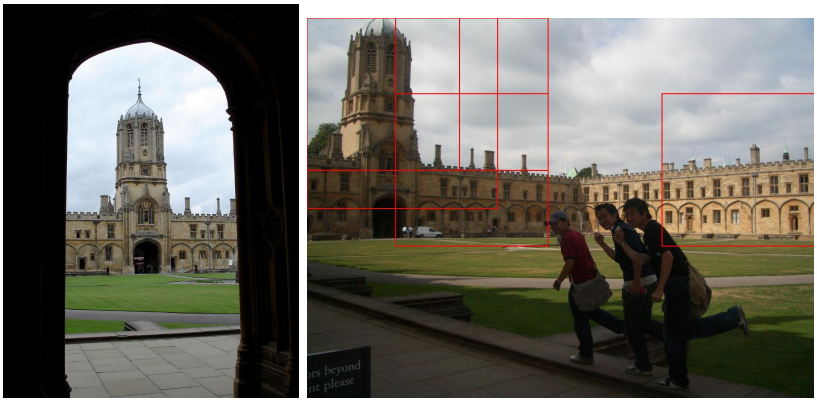
R-MAC (Regional-Maximum Activation of Convolutions) [27], a prominent method in such retrieval category, has been well known in the task of the image retrieval thanks to its attractive properties of efficiency and high accuracy, while maintaining the simplicity, resulting in a method-of-choice for image retrieval in practice. As a result, there have been many methods [5, 6, 25] utilizing R-MAC. We, however, found that R-MAC considers many regions without considering their varying importance. Some of regions, especially in small regions generated by high scales, contain meaningless backgrounds (Fig. 1.1). Furthermore, in such regions, we tend to lose its context, leading to lower retrieval accuracy.

Main contributions. To address these issues, we propose to use context-aware, regional attention module with R-MAC. The main challenge is to treat all regions with global attentiveness within a whole image, especially when there are many salient objects in the image. For tackling this problem, we propose a simple, yet effective regional attention network, which weights an attentive score of a region considering the global context (Sec. 3).

To demonstrate benefits of our method, we have applied our method to well-known image retrieval datasets, and compared it to the state-of-the-art techniques [14, 27, 13, 8] that are in the category of "pre-trained single-pass". Overall, we achieve meaningful accuracy improvement up to 10% to our baseline of R-MAC, and observe robust improvement across all tested cases, resulting in a new state-of-the-art accuracy in our tested category (Sec. 4). Furthermore, we have demonstrated that our method achieves higher accuracy improvement over the other prior methods, even when combined with the query expansion. This result is achieved by better matching results thanks to our context-aware regional attention module. Additionally, we show Region Proposal Network(RPN) [23] can cooperate with our regional attention module (Sec. 4.2) but, R-MAC is more efficient and accurate for the image retrieval.



(a)



(b)

Figure 1.1: This figure shows two challenging examples with backgrounds and clutters in Oxford5k. In each example, the left is a query, while the right is its corresponding positive image. We mark top five attentive regions of our regional attention network in the positive images as red boxes.

Chapter 2. RELATED WORK

Recent approaches [18, 28, 8, 27, 13, 1, 14] used a convolutional layer of an off-the-shelf CNN as a feature extractor for utilizing the spatial information rather than a fully-connected layer. SPoC [1] used a global sum pooling with a centering priority to a convolutional feature map. CroW [14] proposed a non-parametric spatial and channel-wise weighting method for preventing the visual burstiness [10] and focusing on salient regions. Similar to the spatial weighting of CroW, Hoang *et al* [8] proposed a spatial mask for reducing the number of local convolutional features. They effectively applied recent embedding and aggregating techniques with the reduced local convolutional features for further enhancement, but they acquired such competitive results by using a high dimensionality that is higher than the original feature's dimensionality. Jimenez *et al* [13] employed Class Activation Maps (CAMs) [31] for calculating semantic-aware spatial weights of a convolutional feature map. SCDA [28] proposed an unsupervised method for localizing the representative object while removing the noisy background, resulting in the improvement of fine-grained image retrieval. However, this method is structurally based and optimized on the VGG16 architecture.

While some of these prior methods [1, 14, 13] aggregated a feature map considering each point of a convolutional feature map, R-MAC [27] uniformly sampled and aggregated local regions in a convolutional feature map for considering region-wise information. This approach did not consider varying importance among regions when aggregating regional feature vectors. In contrast to this, we consider different importance among regions with our regional attention network.

Instead of sampling regions in the grid manner adopted by R-MAC, Gordo *et al* [5] presented a region proposal network, generating bounding boxes from a feature map of VGGnet [26]. Interestingly, Gordo *et al* [6] showed there is no substantial improvement of using the fine-tuned region proposal network with Resnet [7], compared to sampling regions in the R-MAC manner. In addition to this, the region proposal network requires high computational costs due to a large number of proposed regions; for example, Gordo *et al* [6] used 256 regions per image whereas R-MAC used 20 to 30 regions. Since our work aims for an efficient image retrieval, we decide to mainly investigate a regional attention network with the uniformly sampled regions in this work, instead of using the region proposal network for sampling regions.

Some of prior image retrieval method [14, 19], utilizing the attention mechanism, have been proposed in both of "pre-trained" and "fine-tuned" categories. DELF [19] adopted a learning-based attention network and used the attention network for densely weighting all points of a feature map, similar to CroW [14] which used the non-parametric model for calculating attention weights. These methods, however, did not utilize the global context, and used pixel-based attention in feature map space. Departing from these techniques, we adopted region-based attention, resulting in a smaller amount of attention weights and thus efficiency.

Chapter 3. OUR APPROACH

This work was published in BMVC(British Machine Vision Conference) 2018. In this section, we first review R-MAC [27] that our method is built upon. We then present our new method that efficiently suppresses backgrounds and less important regions. In addition, we employ a context-awareness on the region-wise attention method for further improvement, inspired by recent saliency detection methods [17, 15, 29]. We use ResNet101 [7] as an off-the-shelf CNN, since it was identified [6] that ResNet101 generates higher quality features for image retrieval than that of VGG16, especially when combined with the R-MAC descriptor, and works quite well with uniformly sampled regions. Some notations of Sec. 3.2 are summed up in Table 3.1.

3.1 R-MAC(Regional-Maximum Activation of Convolutions)

R-MAC (Regional-Maximum Activation of Convolutions) [27] was presented as an aggregator of local features in an image as a discriminative global image representation. The pipeline of R-MAC is summarized as follows. With a convolutional feature map, we sample square regions with a region size, R_s, of a specific scale s in a sliding window manner of 40% overlap between neighbor windows, for all $s = 1, ..., S$. Refer to Fig. 3.1 for an example of region sampling of R-MAC. The region size at a specific scale can be calculated as: $R_s = 2\min(W, H)/(s+1)$, where W and H are width and height of the feature map, respectively. After sampling the regional feature maps, R-MAC performs a max pooling for all regional feature maps and a standard post-processing such as ℓ_2-normalization and PCA-whitening [9]. It then calculates a global feature vector with a sum pooling, followed by ℓ_2-normalization. For our work, we replace the sum pooling with a mean pooling. We found that the mean pooling makes training more stable when we have many regions created by having higher scales and varying aspect ratios, because it can adjust summed gradients according to the number of sampled regions.

R-MAC has been known for effective and efficient performance in the image retrieval. Nonetheless, we would like to point out that R-MAC uniformly treats all regions of an image when aggregating their regional feature vectors, even though only specific regions would be helpful to construct a discriminative global feature. This can become a major issue, when we consider more scales (and more regions) for attempting to achieve higher accuracy. Empirically, we also found that R-MAC shows degrading performance as we adopt higher scale values due to the aforementioned issue (Table 4.1a). We aim to address this issue, by proposing a novel region-wise attention and seamlessly integrating it with the R-MAC pipeline.

3.2 Context-Aware Regional Attention

For formulating our regional attention, we suppose that $\mathbf{V}_I$ are a convolutional feature map extracted from an image I through an off-the-shelf CNN, i.e., $\mathbf{V}_I = CNN(I)$. Let Ω to be a set of regional feature maps generated by a region sampler C_S, R-MAC in this work, i.e., $\Omega = C_S(\mathbf{V}_I)$. Our global feature vector, $\mathbf{f}_I$, of the image I is then obtained by the weighted mean on regional feature vectors, as follows:

$$\mathbf{f}_I = [\mathrm{f}_{I,1}, ..., \mathrm{f}_{I,k}]^{\mathrm{T}} = \frac{\sum_{\mathbf{R}\in\Omega} \Phi(\mathbf{k})\mathbf{P}(\mathbf{M}(\mathbf{R}))}{|\Omega|}, \tag{3.1}$$

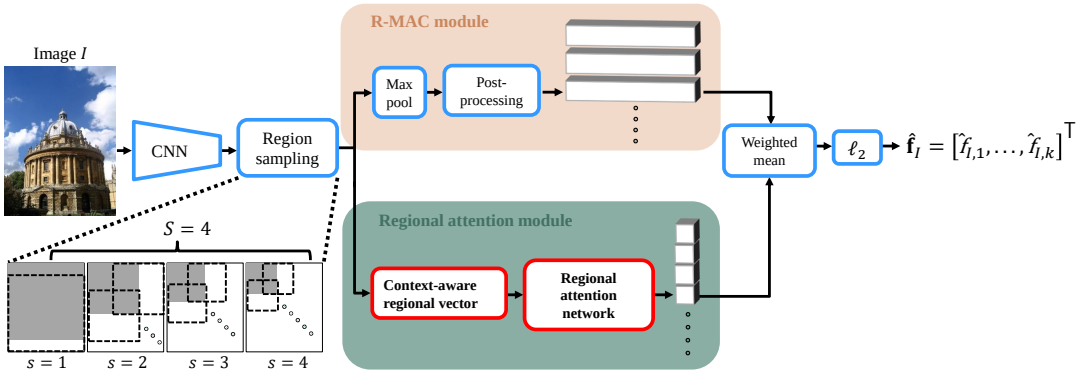


Figure 3.1: This shows our overall encoding sequence of computing feature vector $\widehat{\mathbf{f}}_I$ from a given image I, based on the R-MAC module and our novel regional attention module.

Notation	Description
$\mathbf{V}_I$	Convolutional feature map from an image I
Ω	Set of regional feature maps of $\mathbf{V}_I$
$\mathbf{R}$	A regional feature map that belongs to Ω
$P(M(\mathbf{R}))$	Regional feature vector through max pooling and post-processing
$\mathbf{k}$	Regional feature vector of $\mathbf{R}$ through mean pooling $J(\cdot)$

Table 3.1: Notations for Sec. 3.2.

where $|\cdot|$ is a cardinality of a set and $\mathbf{R}$ is a regional feature map. In this formulation, we calculate $\mathbf{k} = J(\mathbf{R})$ by performing a mean pooling $J(\cdot)$ with the regional feature map $\mathbf{R}$ for obtaining its regional attention weight with a regional attention function, $\Phi(\cdot)$. At the same time, we also execute a max pooling, $M(\mathbf{R})$, followed by post-processing, $P(\cdot)$, for obtaining a regional feature vector, $P(M(\mathbf{R}))$, as R-MAC did (Sec. 3.1). We then get $\mathbf{f}_I$ by modulating regional feature vectors with regional attention weights, which are calculated as follows:

$$\begin{aligned} \Phi(\mathbf{k}) &= \text{softplus}(\mathbf{W_c}\pi(\mathbf{k}) + \mathbf{b_c}), \\ \pi(\mathbf{k}) &= \tanh(\mathbf{W_r}\mathbf{k} + \mathbf{b_r}). \end{aligned} \tag{3.2}$$

where $\mathbf{W}_r \in \mathbb{R}^{d\times k}$ and $\mathbf{W}_c \in \mathbb{R}^{1\times d}$ are linear transformation matrices, and $b_r \in \mathbb{R}^d$ and $b_c \in \mathbb{R}^1$ are bias vector and scalar respectively. $\pi(\cdot)$ performs a linear transformation with reduction in the dimension of a vector space, followed by a non-linear activation function of tanh. Subsequently, $\Phi(\cdot)$ can calculate a final regional attention weight by performing a linear transformation with reduction to a scalar and then a softplus [4], as π did.

While this global feature considered varying importance for regions, it may be insufficient in some cases for fully understanding the global attentiveness of a region without a global context. As a result, it is critical for our region-wise attention module to consider both local regional context and global context. For this reason, we present a context-aware, regional attention network below that considers a global context for calculating a regional attention weight of a specific region.

Context Awareness. Many CNN based saliency detection methods [17, 15, 29] consider a pixel (or region) and its neighboring or whole image as the context, to calculate a saliency score of the pixel (or

region). Our regional attention model can be also regarded as one instance of saliency detection. Based on this view, we also compute a context-aware global feature vector, $\widehat{\mathbf{f}}_I$, with conditionally calculated regional attention weights, as follows:

$$\widehat{\mathbf{f}}_I = \left[\widehat{\mathrm{f}}_{I,1}, ..., \widehat{\mathrm{f}}_{I,k}\right]^{\mathrm{T}} = \frac{\sum_{\mathbf{R}\in\Omega} \Phi(\mathbf{k} \oplus \mathbf{J}(\mathbf{V_I}))\mathbf{P}(\mathbf{M}(\mathbf{R}))}{|\Omega|}, \tag{3.3}$$

where $\oplus$ represents a vector concatenation in the channel space. For conditionally considering a regional attention weight, we use concatenation of a regional feature vector $\mathbf{k}$ and the whole feature vector, $J(\mathbf{V}_I)$, as an input of $\Phi(\cdot)$.

Training Region-wise Attention. To obtain region-wise attention weights without losing the generality of the feature representation, we use the ILSVRC ImageNet dataset [24] for training our regional attention network, different from fine-tuning methods [2, 5]. For training the regional attention network, we use a classification loss and a slightly different sequence for extracting the global feature vector. An output logit vector, $\widehat{\mathbf{y}}_I$, from an image I is expressed as follows:

$$\widehat{\mathbf{y}}_I = L\left(\frac{\sum_{\mathbf{R}\in\Omega} \Phi(\mathbf{k} \oplus \mathbf{J}(\mathbf{V_I}))\mathbf{M}(\mathbf{R})}{|\Omega|}\right), \tag{3.4}$$

where $L(\cdot)$ represents the final fully connected layer of the off-the-shelf CNN to get a class prediction vector.

Based on Eq. 3.4, we train the parameters ($\mathbf{W}_r$, $\mathbf{W}_c$, b_r, b_c) of the regional attention network via back-propagating gradients of the cross entropy loss of $\widehat{\mathbf{y}}_I$, while freezing parameters of the off-the-shelf CNN. This training with the classification loss can be done, since the $L(\cdot)$ originally take a mean pooled feature vector that uses a Global Average Pooling (GAP) layer [16] from a convolutional feature map, and our feature vector for training, $M(\mathbf{R})$, is also regarded as a kind of sparse mean pooled feature vector.

Note that since we fix the off-the-shelf CNN during training, our method is based on the off-the-shelf CNN, and can be thus classified to the "pre-trained single-pass" category [30]. Also, our regional attention network can be utilized on various categories of images because we use the ILSVRC ImageNet dataset as the training dataset for meeting the purpose of "pre-trained single-pass".

Chapter 4. RESULTS

As our base-network, we use Resnet101 that was released on Caffe [12] and pre-trained on the ILSVRC ImageNet dataset. We then add R-MAC [27] and our regional attention network on the base-network. For training our model, we set the R-MAC scale S to 4, and an input image for feeding the base-network is obtained by random square cropping of 800x800 resolution from an image resized to a minimum dimension of 850 with the original aspect ratio. Based on this setting, we first train our regional attention network with the classification loss from the ILSVRC ImageNet dataset by the SGD optimizer of 10^{-3} learning rate and $5 \cdot 10^{-5}$ weight decay. We change the learning rate to 10^{-4}, when a validation error of classification does not change.

4.1 Experimental Setup

We conduct experiments for testing our method with Oxford5k [20], Oxford105k, Paris6k [21], and Paris106k datasets, which are well known for the task of image retrieval. The Oxford5k dataset contains 5063 images related to particular oxford landmarks. Similar to Oxford5k, the Paris6k dataset consist of 6412 images associated with particular Paris landmarks. Both datasets provide 55 query images with a bounding box of a specific instance to conduct image retrieval, and we use cropped queries with the bounding boxes as input images for a comparison test. We also consider Oxford105k and Paris106k that are extensions of Oxford5k and Paris6k. We report mean Average Precision (mAP) as an evaluation protocol with all of these datasets.

For PCA learning, R-MAC, CroW [14], SDCF [8] and CAM [13] generally use Paris6k for testing on Oxford5k and Oxford105k, and use Oxford5k for testing on Paris6k and Paris106k. This way of learning PCA requires us to re-calculate a new PCA every time we get a new test dataset. To avoid this re-computation, we simply use the large Landmark dataset [2] to calculate the PCA parameters. Note that PCA learned from the Landmark dataset degrades the mAP performance, compared to using Oxford5k and Paris6k for calculating PCA; see the last two rows of Table 4.1c. Nonetheless, we adopted this approach of using the Landmark dataset for PCA computation, since it suits better in practice for image retrieval. For testing our method, we resize all images to a maximum dimension of 1024 as following R-MAC, and use the R-MAC scale $S = 5$, as a result of ablation study shown in Table 4.1a. We measure the accuracy of our method using mAP as following the standard evaluation protocol.

4.2 Ablation Experiments

The performance of our method can vary depending on the R-MAC scale S, since the scale controls the number of regions considered for computing the final global feature. We thus conduct an experiment for finding an optimal scale of our method. Table 4.1a displays performances of the baseline and our method according to the scales. While ours shows higher accuracy over the baseline across all the tested scales, one interesting point is that our method can use a larger scale than that of the baseline. We can interpret that our method is less affected by background or less important regions, even though smaller regions generated by using higher scales contain relatively more backgrounds.

Method	Scale (S)			
	S=3	S=4	S=5	S=6
Baseline	69.9	**70.7**	70.1	69.0
Ours	75.1	76.7	**76.8**	76.4

(a)

Method	Oxford5k	Paris6k
RPN + PCA Landmark	64.7	75.5
+ Regional attention	66.6	75.8
+ Context awareness	67.9	76.4

(b)

Method	Oxford5k	Paris6k	Time (s)
Baseline + PCA Landmark	70.1	85.4	0.095
+ Regional attention	74.9	86.0	0.115
+ Context awareness	76.8	87.5	0.123
- PCA Landmark + PCA Paris, Oxford	77.6	88.3	-

(c)

Table 4.1: a) mean Average Precision (mAP) of baseline (R-MAC) and our method with different scales. b) shows performance improvement when applying our method to RPN [23] with 256 regions, instead of R-MAC. Nonetheless, we achieve better accuracy with R-MAC. c) presents performance improvement and its computational time for an image by adding our contributions to the baseline (R-MAC).

We also experiment how much each component of our method improves the performance in Table 4.1c. We simultaneously train our regional attention network and context-aware regional attention network under the same settings such as learning rate, resolution, the number of iterations and ILSVRC ImageNet dataset. We then test the models with R-MAC scale S=5. As shown in the table, we can get a significantly improved performance with our region-wise attention as well as context-awareness, compared to the baseline, while the computational costs are not significant.

We additionally experiment our regional attention network on off-the-shelf RPN [23] that can be employed for region sampling, instead of R-MAC. We find that our regional attention network complementarily works well with RPN too (Table 1b); note that RPN sometimes generates outliers and noise regions, but ours can filter them out, leading to a higher accuracy.

4.3 Comparisons to State-of-the-Art

As established by Zheng *et al* [30], our target category in the task of image retrieval is "pre-trained single-pass" methods for avoiding additional training of the network depending on test image categories. The top four state-of-the-art methods [14, 27, 8, 13] were designed with VGG16 as their base-network, and we thus reproduce their approaches with Resnet101 that our work is based on. For the efficient image retrieval, we test only the offline aggregation of CAM [13] using 64 CAMs, since the online aggregation of CAM requires high storage and long query time. As following SDCF [8], we set the PCA dimension and the codebook size to 64 and 34, respectively, for final dimensionality of 2048 and use MAX-mask, T-emb and demoratic-pooling[11] while experimenting with SDCF.

Table 4.2 shows overall performance comparisons of ours and the state-of-the-art methods on different datasets. Our method outperforms all the other methods across all the test datasets, when using Resnet101. We also apply a query expansion technique [3] with top-5 retrieved images to tested

	Method	Dim.	Oxford5k	Paris6k	Oxford105k	Paris106k
VGG16	SPoC[1]	256	53.1	-	50.1	-
	BoW[18]	25k	73.9	81.9	-	-
	SDCF [8]	4096	75.3	86.7	71.4	80.6
	SDCF [8]	512	65.7	81.6	60.5	72.4
	CroW [14]	512	70.8	79.7	65.3	72.2
	R-MAC [27]	512	66.9	83.0	61.6	75.7
	CAM [13]	512	71.2	80.5	67.2	73.3
Resnet101	SDCF [8]	2048	69.1	81.7	65.4	74.3
	CroW [14]	2048	68.7	82.8	62.7	75.1
	R-MAC [27]	2048	70.1	85.4	66.9	80.8
	CAM [13]	2048	69.9	84.3	64.3	77.1
	Ours	2048	**76.8**	**87.5**	**73.6**	**82.5**
	Query expansion (QE)					
Resnet101	SDCF+QE [8]	2048	68.5	84.9	66.8	79.4
	CroW+QE [14]	2048	69.5	85.1	66.7	79.9
	R-MAC+QE [27]	2048	73.8	86.4	71.8	82.6
	CAM+QE [13]	2048	71.3	86.1	68.7	80.8
	Ours+QE	2048	**81.8**	**89.3**	**80.4**	**85.4**

Table 4.2: Performance comparisons against the state-of-the-art retrieval methods in the category of single-pass using an off-the-shelf CNN [30]. Ours includes R-MAC + regional attention + context awareness. SDCF denotes Selective Deep Convolutional Features.

approaches. As shown in Table 4.2, we can also see that our method significantly outperforms the state-of-the-arts, combined even with the query expansion technique. Specifically, the average increase, 4.1 mAP, of our method with query expansion is higher than the other methods where the average increases of CAM, R-MAC, SDCF and CROW are 2.8, 2.9, 2.3 and 3 mAP respectively. This result is acquired mainly because top-5 images initially retrieved by our method are more highly related, thanks to our context-aware, regional attention module.

4.4 Qualitative Results

Qualitative results of ours and R-MAC are shown in Fig. 4.1 on the Oxford5k dataset. In Fig. 4.1, we choose two examples out of 55 queries that have maximum AP differences between ours and R-MAC, mainly because our method surpasses the R-MAC baseline in almost every query. We visualize nine retrieved results of each example in a ranking order starting from what one of ours and R-MAC firstly fails to find correct images. For example, in the top example of Fig. 4.1, all of the tested methods report incorrect images from the second retrieved image, and we thus show images from that image.

Based on top-15 attentive regions (shown by red boxes) in Fig. 4.1, we can see that our regional attention network mostly focuses on some objects or patterns rather than backgrounds. We can also observe that our method pushes away rankings of negative images and pulls up rankings of positive images; see the first example of Fig. 4.1. Also, our method is more robust than R-MAC, even when

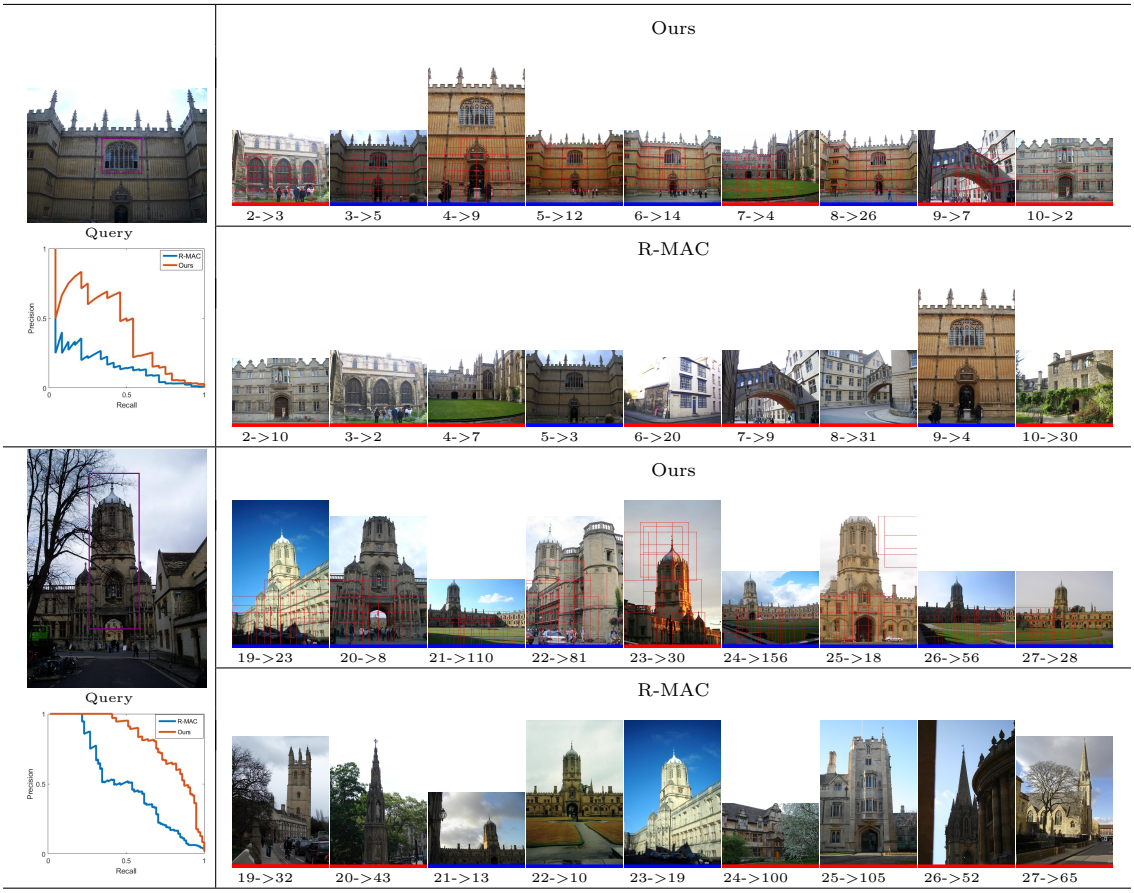


Figure 4.1: Two examples where ours outperforms R-MAC most. The first column shows query images with purple bounding boxes and precision-recall graphs of ours and R-MAC. In the second column, retrieved results are enumerated in a ranked order. Each blue and red bar of retrieved images denotes true-positive and false-positive, respectively. We also show ranking changes like "A-¿B". A is the original ranking based on each method, and B is another ranking when we use the other method. Top-15 attentive regions are shown in red boxes for our retrieved results; zoom-in view is recommended.

images include large background; see the second example of Fig. 4.1.

Fig. 4.2 shows the only single case out of 55 queries that R-MAC outperforms ours; note that ours surpasses R-MAC in the rest of all the other case. Even in this case, the gap, 3.9, of APs of ours and R-MAC is narrower than those, 24.1 and 30.0 for the top and bottom cases, shown in Fig. 4.1.

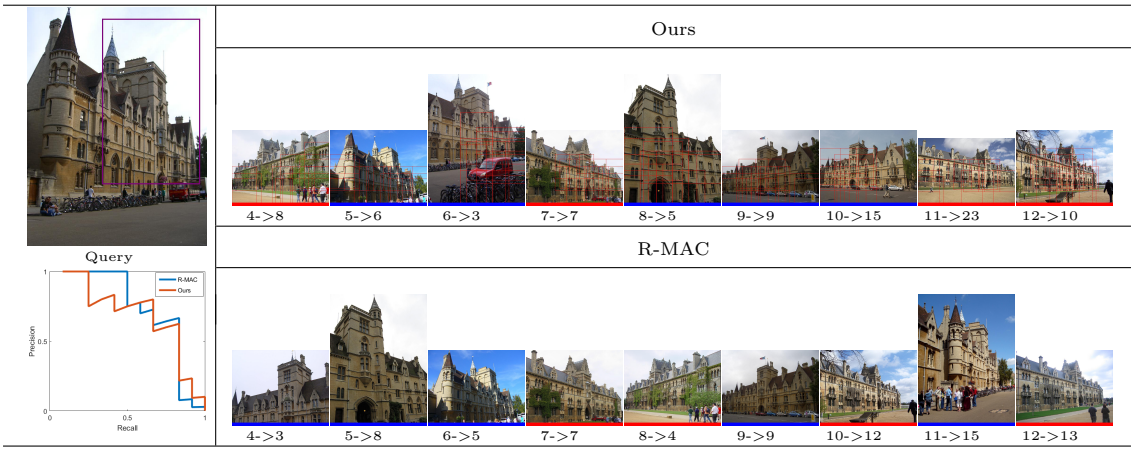


Figure 4.2: The only case where R-MAC outperforms ours, out of 55 queries in Oxford5k dataset. This figure contains the same layout as explained in Fig. 4.1.

Chapter 5. CONCLUSION

We have presented our context-aware regional attention network for tackling the problem of region-based feature aggregation, especially in R-MAC, a well-known image retrieval method. We have tested our method on different benchmarks and verified that it shows robust improvement over the prior state-of-the-art methods for the image retrieval category of "pre-trained single-pass". While we have shown RPN can be combined with ours in the image retrieval, we believe that we can take one more step for RPN coupled with our context-aware regional attention module in various fields including the image retrieval.

Bibliography

[1] A. Babenko and V. S. Lempitsky. Aggregating local deep features for image retrieval. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1269–1277, 2015.

[2] A. Babenko, A. Slesarev, A. Chigorin, and V. S. Lempitsky. Neural codes for image retrieval. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, pages 584–599, 2014.

[3] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007*, pages 1–8, 2007.

[4] C. Dugas, Y. Bengio, F. Bélisle, C. Nadeau, and R. Garcia. Incorporating second-order functional knowledge for better option pricing. In *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA*, pages 472–478, 2000.

[5] A. Gordo, J. Almazán, J. Revaud, and D. Larlus. Deep image retrieval: Learning global representations for image search. In *ECCV*, 2016.

[6] A. Gordo, J. Almazán, J. Revaud, and D. Larlus. End-to-end learning of deep visual representations for image retrieval. *International Journal of Computer Vision*, 124(2):237–254, Sept. 2017.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778, 2016.

[8] T. Hoang, T. Do, D. L. Tan, and N. Cheung. Selective deep convolutional features for image retrieval. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pages 1600–1608, 2017.

[9] H. Jégou and O. Chum. Negative evidences and co-occurences in image retrieval: The benefit of PCA and whitening. In *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part II*, pages 774–787, 2012.

[10] H. Jegou, M. Douze, and C. Schmid. On the burstiness of visual elements. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 1169–1176, 2009.

[11] H. Jégou and A. Zisserman. Triangulation embedding and democratic aggregation for image search. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 3310–3317, 2014.

[12] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM*

International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014, pages 675–678, 2014.

[13] A. Jimenez, J. M. Alvarez, and X. G. i Nieto. Class-weighted convolutional features for visual instance search. In *28th British Machine Vision Conference (BMVC)*, September 2017.

[14] Y. Kalantidis, C. Mellina, and S. Osindero. Cross-dimensional weighting for aggregated deep convolutional features. In *In ECCV Workshops*, 2016.

[15] G. Li and Y. Yu. Visual saliency based on multiscale deep features. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 5455–5463, 2015.

[16] M. Lin, Q. Chen, and S. Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[17] N. Liu and J. Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 678–686, 2016.

[18] E. Mohedano, K. McGuinness, N. E. O'Connor, A. Salvador, F. Marqués, and X. Giró i Nieto. Bags of local convolutional features for scalable instance search. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, ICMR 2016, New York, New York, USA, June 6-9, 2016*, pages 327–331, 2016.

[19] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han. Large-scale image retrieval with attentive deep local features. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 3476–3485, 2017.

[20] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[21] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[22] F. Radenovic, G. Tolias, and O. Chum. CNN image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, pages 3–20, 2016.

[23] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99, 2015.

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[25] O. Seddati, S. Dupont, S. Mahmoudi, and M. Parian. Towards good practices for image retrieval based on CNN features. In *2017 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2017, Venice, Italy, October 22-29, 2017*, pages 1246–1255, 2017.

[26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[27] G. Tolias, R. Sicre, and H. Jégou. Particular object retrieval with integral max-pooling of CNN activations. *CoRR*, abs/1511.05879, 2015.

[28] X. Wei, J. Luo, J. Wu, and Z. Zhou. Selective convolutional descriptor aggregation for fine-grained image retrieval. *IEEE Trans. Image Processing*, 26(6):2868–2881, 2017.

[29] R. Zhao, W. Ouyang, H. Li, and X. Wang. Saliency detection by multi-context deep learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1265–1274, 2015.

[30] L. Zheng, Y. Yang, and Q. Tian. SIFT meets CNN: A decade survey of instance retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(5):1224–1244, 2018.

[31] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2921–2929, 2016.

Acknowledgments in Korean

석사과정 동안 연구에 집중할 수 있도록 좋은 환경을 만들어 주셨고 부족한 제가 이렇게 논문을 작성 할 수 있게 가장 많은 도움을 주신 윤성의 교수님께 가장 큰 감사 전하고 싶습니다. 또한, 연구실 생활을 위해 여러 도움을 주신 SGVR 연구실 선배, 동기 그리고 후배분들에게도 감사의 말을 전합니다. 연구 관련해서 항상 좋은 코멘트 주신 이미지팀 수민 누나, 태영 형, 우빈 형, 졸업하신 영기 형, Jared, 치완 형 그리고 석사 기간 동안 같이 생활한 SGVR 연구실 멤버 동혁 형, 용선 형, 명배 형, 인규 형, 민철 형, 희찬 형, 도헌 군, 태운 형, 훈민 군 모두에게 감사의 말을 전합니다.

마지막으로 항상 응원해 주셨던 하지만 지금은 하늘나라에 계신 아버지 그리고 현재도 열심히 응원 해 주시고 기도 해주시는 할머니, 큰 아버지, 고모, 큰 어머니 및 다른 친척 분들에게도 큰 감사 인사를 올립니다. 늦게 만났지만 도움을 주고 계시는 어머니에게도 감사 인사 올립니다.

Curriculum Vitae in Korean

이 름: 김 재 윤

생 년 월 일: 1995년 01월 05일

주 소: 대전 유성구 대학로 291 한국과학기술원 전산학과 3443호

학 력

2013. 2. – 2017. 2. 충남대학교 컴퓨터공학과 (학사)

연 구 업 적

1. **J. Kim**, S. Yoon, "Regional Attention Based Deep Feature for Image Retrieval," In *Proc. British Machine Vision Conference (BMVC)*, 2018.
2. **J. Kim**, S. Um, D. Min, "Fast 2D Complex Gabor Filter With Kernel Decomposition," *IEEE Trans. Image Processing (TIP)*, Apr. 2018.