Image Retrieval with Combination of Multiple Global Descriptors

Huynh Ngoc, Tran
18520385

Nguyen Ngoc, Khanh
18520901

1 Introduction

An image retrieval system is a computer system for browsing, searching and retrieving images from a large database of digital images. The purpose of an image database is to store and retrieve an image or image sequences that are relevant to a query. Although image retrieval has been extensively explored since the early 1990s [1], it still attracts lots of attention from the multimedia and computer vision communities until now due to its necessity, since image data has become excessively popular.

There are two main approaches of image retrieval methods, which are *Text-based* and *Content-based*. Traditionally, image search engines usually adopt Text-based methods, which index multimedia visual data based on the surrounding meta data information around images on the Web, such as titles and tags. However, there are several disadvantages of this approach such as textual information inconsistency with visual content (due to polysemy problem or inaccurate annotation), manual annotation impracticality for large database, etc., Content-based approach is preferred and has been witnessed to make great advance in recent years [2]. Therefore, for our image retrieval system, we follow the Content-based approach.

Content-based image retrieval methods involve three main parts in system realization: collecting data, building up feature database, searching and ranking results of the retrieval. Among those steps, the most distinguished one that impacts the performance is building up feature database, i.e., analysing the collected images and then extracting the feature information. In this project, we adopt the feature extraction methods from the paper "***Combination of Multiple Global Descriptors for Image Retrieval***" (CGD) [3]. The main scheme of feature extraction method proposed in the paper is combining multiple global descriptors generated by different global-pooling methods, which are sum pooling of convolutions (SPoC) [4], maximum activation of convolutions (MAC) [5], and generalized mean pooling (GeM) [6]. Each of the pooling methods has its own properties, e.g., SPoC activates larger regions on the image representation while MAC activates more focused regions. By combining them, we can utilise their advantages.

In this report, we represent details about the framework we adopted from the CGD paper, especially the main module: multiple global descriptors. Afterwards, the results of our experiments on our own dataset are shown.

2 Framework

The proposed framework consists of a CNN backbone network and two modules. The first module is the main module, which is a combination of multiple global descriptors that learn the input image representation with ranking loss. The other one is auxiliary module to fine-tune a CNN with classification loss. The final loss is the sum of the ranking loss and the classification loss. The framework is trained in an end-to-end manner.

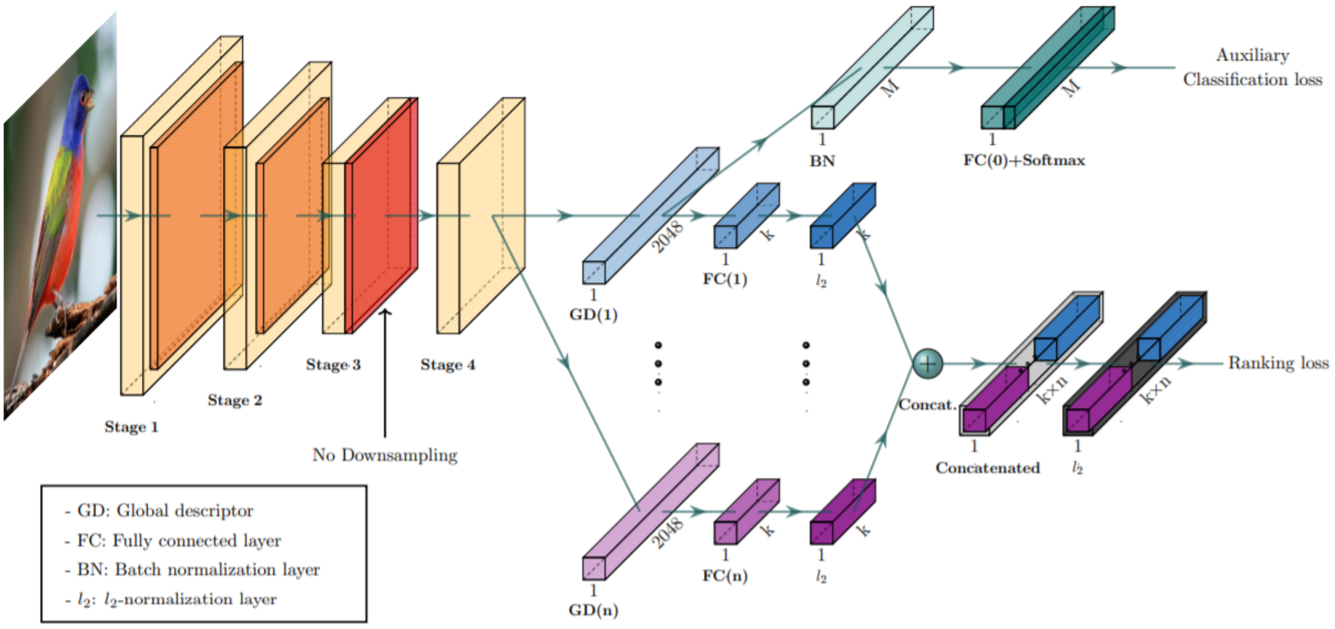

Figure 1: **The combination of multiple global descriptors (CGD) framework.** The downsampling stage is removed from the backbone. From the last feature map, each of n global descriptors produces a k-dimensional embedding vector, which is concatenated into the combined descriptor for ranking loss. The first global descriptor is used for auxiliary classification loss where M denotes the number of classes.

2.1 Backbone

The proposed framework can use any CNN backbones. In the original paper, the authors use ResNet-50 as the baseline backbone. In this project, since the framework is claimed to be flexible with different CNNs, we tries out several backbone: Wide-ResNet50, ResNet50, ResNet101, ResNext50, ResNext101, EfficientNet-B5. These CNN backbones are chosen because they have compatible number of output dimensions with the proposed framework, which is 2048.

2.2 Combination of Multiple Global Descriptors

The main module has multiple branches that output each image representation by using different global descriptors on the last convolutional layer. In the paper, the authors use three types of the most representative global descriptors on each branch, including SPoC, MAC, and GeM.

Given an input image, the output of the last convolutional layer is a 3D tensor X of $W \times H \times K$ dimensions, where K is the number of feature maps. Let X_k be the set of $W \times H$ activations for feature map $k \in \{1...K\}$. The network output consists of K such activation sets or 2D feature maps. We additionally assume that the very last layer is a Rectified Linear Unit (ReLU) such that X is non-negative.

The global descriptor takes X as input and produces a vector f as output by pooling process. We denote $f^{(s)}$, $f^{(m)}$, $f^{(g)}$ as SPoC, MAC and GeM respectively. Each pooling method is described below:

• **SPoC** (Sum Pooling and Centering Prior) [4]: This method's main scheme is similar to global average pooling, which can be generalized as follows:

$$f^{(s)} = [f_1^{(s)}...f_k^{(s)}...f_K^{(s)}]^\top, \qquad f_k^{(s)} = \frac{1}{\mid X_k \mid} \sum_{x \in X_k} x \tag{1}$$

• **MAC** (Maximum Activations of Convolution) [5]: This method builds up the result vector in a way that is similar to global max pooling:

$$f^{(m)} = [f_1^{(m)}...f_k^{(m)}...f_K^{(m)}]^\top, \qquad f_k^{(m)} = \max_{x \in X_k} x \tag{2}$$

• **GeM** (Generalized Mean Pooling) [6]: Different from average pooling and max pooling, the result of generalized mean pooling is given by:

$$f^{(g)} = [f_1^{(g)} ... f_k^{(g)} ... f_K^{(g)}]^\top, \qquad f_k^{(g)} = \left(\frac{1}{\mid X_k \mid} \sum_{x \in X_k} x^{p_k} \right)^{\frac{1}{p_k}} \tag{3}$$

In fact, pooling methods (1) and (2) are special cases of (3). In particular, when $p_k \to \infty$, (3) is equivalent to max pooling (2), while it is the same as average pooling (1) when $p_k \to 1$. For the case of GeM, p_k is differentiable and can be set manually, while we choose $p_k = 3$ as fixed value.

Afterwards, the result vector f is fed through a FC layer for dimensionality reduction, and is then normalized by l_2-normalization layer to generate output feature vector $\Phi^{(a_i)}$:

$$\Phi^{(a_i)} = \frac{W^{(i)} \cdot f^{(a_i)}}{\|W^{(i)} \cdot f^{(a_i)}\|}, \qquad a_i \in \{s, m, g\} \tag{4}$$

for $i \in \{1...n\}$ where n is the number of branches. W^i denotes the weight of FC layer. $f^{(a_i)}$ is the output vector of pooling process with the global descriptor SPoC, MAC or GeM if a_i is equal to s, m or g respectively.

All of feature vectors $\Phi^{(a_i)}$ are concatenated sequentially, and then go through l_2-normalization to create final feature vector. The output vector of the combine descriptor is described as:

$$\Psi_{CGD} = \frac{\Phi^{(a_1)} \oplus ... \oplus \Phi^{(a_i)} \oplus ... \oplus \Phi^{(a_n)}}{\|\Phi^{(a_1)} \oplus ... \oplus \Phi^{(a_i)} \oplus ... \oplus \Phi^{(a_n)}\|}, \qquad a_i \in \{s, m, g\} \tag{5}$$

This combined descriptor is trained with batch-hard triplet loss as ranking loss.

2.3 Auxiliary Module

The auxiliary module fine-tunes the CNN backbone based on the first global descriptor of the main module by using classification loss, which is softmax cross-entropy in particular. Label smoothing and temperature scaling are also used to avoid over-confidence and over-fitting.

3 Experiments

We evaluate the framework by conducting experiments on our dataset, which includes approximately 1500 images divided into 22 classes about the topic *"Vietnamese cuisine"*.

3.1 Experiments on Combination of Global Descriptors

We first implement 3 configurations with each descriptor individually. We choose Wide-ResNet50 as our baseline network. The results are shown below (S, M G are SPoC, MAC, GeM respectively):

Config.	Dim.	Train accuracy	R@1	R@2	R@4	R@8
S	1536	99.8	81.3	86.9	92.0	94.9
G	**1536**	**100.0**	**81.3**	**87.5**	**92.0**	**94.9**
M	1536	32.3	14.7	23.3	32.9	53.4

Table 1: Results of using global descriptor individually. *Dim.* denotes the number of embedding vector dimensions. *R@k* means recall rate at rank k.

Among the descriptors, G achieves the highest result, while S result is just slightly lower. Since the result of M is significantly lower than the others, plus it takes longer to train, we decide to discard it. We then try out different combinations of S and G, using the same baseline:

Config.	Dim.	Train accuracy	R@1	R@2	R@4	R@8
GS	512	99.2	75.0	84.1	90.9	96.0
SG	**512**	**100**	**86.4**	**90.3**	**94.3**	**98.3**
GS	768	99.8	83.0	87.5	91.5	94.9
SG	768	100	85.2	88.6	93.8	97.7

Table 2: Results of combining S and G.

By using the combination of S and G, the results are improved. SG/512 achieves the highest results (86.4% R@1, 98.3% R@8), which are considerably higher than the best results of experiments on individual descriptor (81.3% R@1, 94.9% R@8). We continue using SG/512 as descriptor and try using different baseline networks. The results are as follows:

Backbone	Train accuracy	R@1	R@8
ResNet50	99.9	83.5	96.0
ResNet101	100	79.5	94.9
ResNext50	99.9	83.5	97.2
ResNext101	100	84.7	97.2
Wide-ResNet50	**100**	**86.4**	**98.3**
EfficientNet-B5	100	85.8	96.0

Table 3: Results of using different backbones.

Since Wide-ResNet50 with SG/512 achieves the best results, we adopt this model for our image retrieval system. We show the visualization of our results with some queries below:

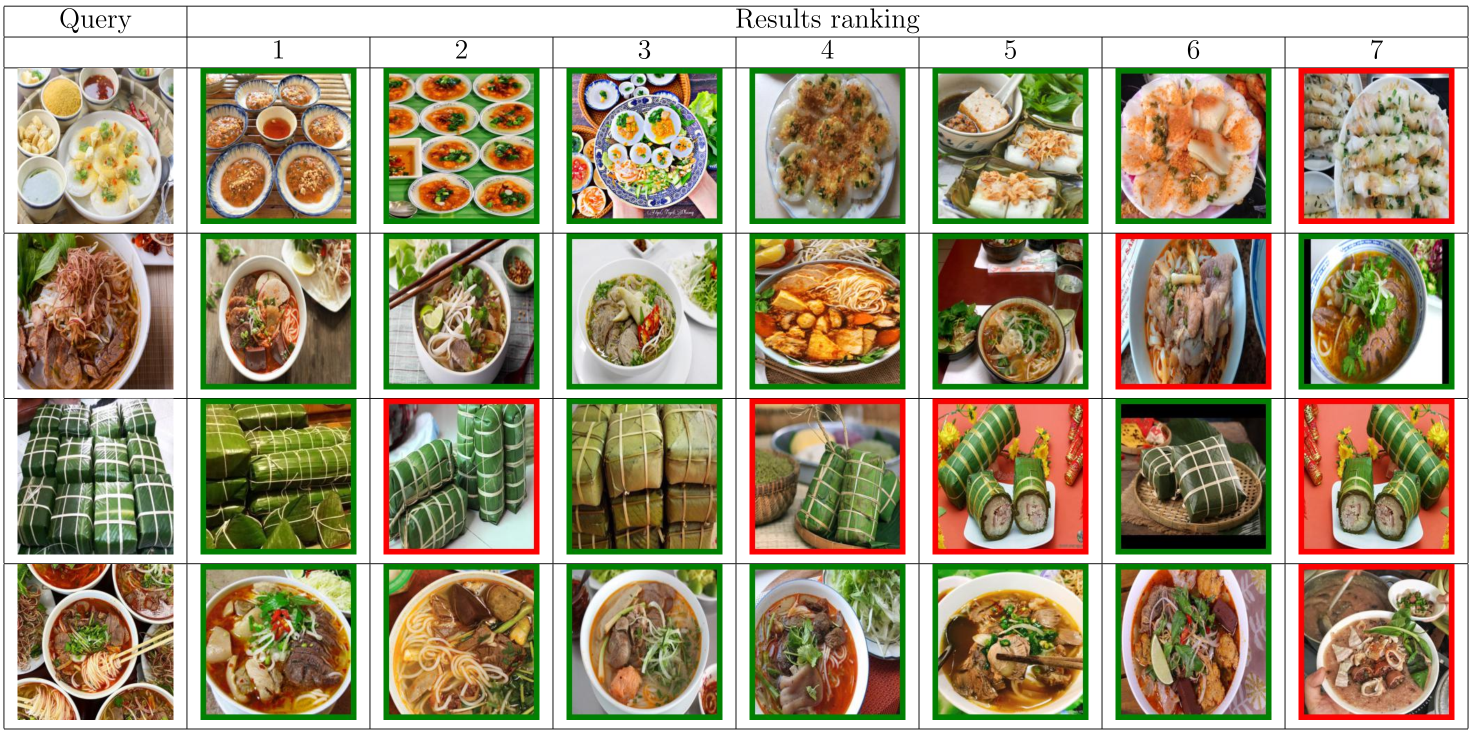

Table 4: Queries and retrieved images with their ranks. Green box indicates that the image is in the same class, while red box indicates that the image is in a different class.

4 Conclusion

In this project, after conducting a brief survey on image retrieval approaches, we have chosen *"Combination of Multiple Global Descriptors for Image Retrieval"* to research and implement beacause of its simple yet effective framework. The CGD framework exploits multiple global descriptors to get an ensemble effect. As combining descriptor can manipulate different types of feature properties, it really help achieving better result than using each descriptor individually. We have also conducted several experiments to analyse this method as well as in order to choose the best model for our image retrieval system.

References

[1] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, “Relevance feedback: a power tool for interactive content-based image retrieval”, *IEEE Transactions on Circuits and Systems for Video Technology, vol. 8, no. 5, p. 644–655,* 1998.

[2] Zhou, W., Li, H. and Tian, Q., “Recent advance in content-based image retrieval: A literature survey”, *arXiv preprint arXiv:1706.06064*, 2017.

[3] H. Jun, B. Ko, Y. Kim, I. Kim, and J. Kim, “Combination of Multiple Global Descriptors for Image Retrieval”, *arXiv preprint arXiv:1903.10663*, 2019.

[4] A. Babenko and V. Lempitsky, “Aggregating local deep features for image retrieval”, *The IEEE International Conference on Computer Vision (ICCV)*, 2015.

[5] G. Tolias, R. Sicre, and H. Jegou, “Particular object retrieval with integral max-pooling of CNN activations”, *arXiv preprint arXiv:1511.05879*, 2015.

[6]] F. Radenovic, G. Tolias, and O. Chum, “Fine-tuning CNN image retrieval with no human annotation”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.