International Journal of Remote Sensing

ISSN: 0143-1161 (Print) 1366-5901 (Online) Journal homepage: www.tandfonline.com/journals/tres20

Attention boosted bilinear pooling for remote sensing image retrieval

Yameng Wang, Shunping Ji, Meng Lu & Yongjun Zhang

To cite this article: Yameng Wang, Shunping Ji, Meng Lu & Yongjun Zhang (2020) Attention boosted bilinear pooling for remote sensing image retrieval, International Journal of Remote Sensing, 41:7, 2704-2724, DOI: 10.1080/01431161.2019.1697010

To link to this article: https://doi.org/10.1080/01431161.2019.1697010

Published online: 08 Dec 2019.

INTERNATIONAL JOURNAL OF REMOTE SENSING
2020, VOL. 41, NO. 7, 2704–2724
https://doi.org/10.1080/01431161.2019.1697010

Taylor & Francis
Taylor & Francis Group

Attention boosted bilinear pooling for remote sensing image retrieval

Yameng Wang[a], Shunping Ji[a], Meng Lu[b] and Yongjun Zhang[a]

[a]School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China; [b]Department of Physical Geography, Utrecht University, Utrecht, The Netherlands

ABSTRACT

Remote sensing image retrieval is to find the most identical or similar images to a query image in the vast archive of remote sensing images. A key process is to extract the most distinctive features. In this study, we introduce a second-order pooling named compact bilinear pooling (CBP) into convolutional neural networks (CNNs) for remote sensing image retrieval. The retrieval algorithm has three stages, pretraining, fine-tuning and retrieval. In the pretraining stage, two classic CNN structures, VGG16 and ResNet34, are pretrained respectively with the ImageNet consisting of close-range images. A CBP layer is introduced before the fully connected layers in the two networks. To extract globally consistent representations, a channel and spatial integrated attention mechanism is proposed to refine features from the last convolution layer and the features are used as the input of the CBP. In the fine-tuning stage, the new network is fine-tuned on a remote sensing dataset to train discriminable features. In the retrieval stage, the network, with fully connected layers being replaced by a PCA (principal component analysis) module, is applied to new remote sensing datasets. Our retrieval algorithm with the combination of CBP and PCA obtained the best performance and outperformed several mainstream pooling or encoding methods such as full-connected layer, IFK (Improved Fisher Kernel), BoW (Bag-of-Words) and maxpooling, etc. The channel and spatial attention mechanism contributes to the CBP based retrieval method and obtained the best performance on all the datasets, as well as outperformed several recent attention methods. Source code is available at http://study.rsgis/whu.edu.cn/pages/download.

ARTICLE HISTORY
Received 1 July 2019
Accepted 9 November 2019

1. Introduction

Content-based image retrieval (CBIR) searches and retrieves the most similar images to a query image from a large dataset. The key process of the CBIR is to extract distinctive features to represent an image labelled with a category. The CBIR can be roughly classified by the three levels of features, i.e. low, middle, and high, that are used.

A low-level representation describes features using the spectrum (Vellaikal, Kuo, and Dao 1995; Bretschneider, Cavet, and Kao 2002), texture (Du Buf, Kardan, and Spann 1990; Ma and Manjunath 1996), gradient, shape, or a combination of them. Edge histogram

CONTACT Shunping Ji jishunping@whu.edu.cn School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China

(Park, Jeon, and Won 2000), grey-level co-occurrence matrix (Haralick, Shanmugam, and Dinstein 1973), Gabor wavelet (Rangayyan et al. 2000), local binary patterns (LBP) (Ojala, Pietikainen, and Maenpaa 2002) and its variations (Zhao and Pietikäinen 2007; Zhang, Yao, and Liu 2008; Jian et al. 2008) are common texture analysis methods used in remoting sensing image retrieval. The scale-invariant feature transform (SIFT) feature (Lowe 2004) is a gradient-based feature with scale, rotation, and partial illumination invariance, many features have been developed upon it (Philbin et al. 2010; Arandjelović and Zisserman 2012). Speeded up robust features (SURF) (Bay, Tuytelaars, and Van Gool 2006), oriented FAST and rotated BRIEF (ORB) (Rublee et al. 2011), histogram of oriented gradient (HOG) (Dalal and Triggs 2005) and Haar (Papageorgiou, Oren, and Poggio 1998; Viola and Jones 2001; Lienhart and Maydt 2002) are all classic shape-based features. Shape features were also used in hyperspectral remote sensing image retrieval (Du et al. 2006). Prasad, Gupta, and Biswas (2001) proposed a combined index integrating colour and shape features, whereas Choraś, Andrysiak, and Choraś (2007) integrated colour, texture and shape information.

The mid-level representation in CBIR is typically generated from aggregating low-level local features into a global representation. Compared to low-level features, mid-level representations are commonly more robust to scale, rotation, and illumination changes. Bag-of-visual-words (BoW) (Sivic and Zisserman 2003) uses a k-means algorithm to cluster low-level feature vectors such as SIFT and SURF, to generate visual words that describe an object or image. Spatial pyramid matching (Lazebnik, Schmid, and Ponce 2006) adds a hierarchical pyramid structure to the BoW to utilize spatial information. The improved fisher kernel (IFK) (Perronnin, Sánchez, and Mensink 2010aa) uses a gaussian mixture model (GMM) to build a visual dictionary. Vector of locally aggregated descriptors (VLAD) (Jégou et al. 2010), a simplified FK (Perronnin, et al. 2010b), considers the clustering centre nearest to a feature to simplify the computation. Based on locally linear embedding (LLE) (Roweis and Saul 2000), the locality-constrained linear coding (LLC) (Wang et al. 2010) encoded SIFT features with maxpooling. Tolias, Furon, and Jégou (2014) proposed the orientation covariant embedding without codebook to obtain a single vector representation. Jégou, Douze, and Schmid (2008) utilized hamming embedding to maintain the discriminatory power of descriptors with low memory consumption. Jégou and Zisserman (2014) employed a novel embedding method to combine local descriptors such as SIFT and SURF and limited the interference between them by democratizing their contributions at the aggregation stage.

The high-level representation uses richer semantic information to identify the category an image belongs to or to detect objects. Among the machine learning methods (Wan et al. 2014; Babenko and Lempitsky 2015; Yue-Hei Ng, Yang, and Davis 2015; Gordo et al. 2016) that have been developed and applied to extract high-level features, deep learning (Lecun, Bengio, and Hinton 2015) has shown its strength and attracted a lot of interests. Convolutional neural network (CNN) is a representative architecture for image processing, widely used CNNs are AlexNet (Krizhevsky, Sutskever, and Hinton 2012), GoogleNet (Szegedy et al. 2015), VggNet (Simonyan and Zisserman 2015), and ResNet (He et al. 2016).

In CNN-based image classification or retrieval, the two-dimensional high-level features learned by the convolutional layers are commonly converted to a global vector descriptor through pooling or encoding methods for similarity measure. Compared with popular encoding methods such as fully connected (FC) layers, studies of pooling methods are

relatively few. The pooling methods can be divided into general pooling (such as the widely used maxpooling and average pooling) and overlapping pooling where the stride is smaller than the length of a sliding window. Unrestricted to the first-order pooling method, Carreira et al. (2012) proposed a second-order pooling approach and applied it to semantic segmentation. Though this approach has obtained improved results, it comes with a notably high computational burden.

Inspired by the second-order pooling, Lin, RoyChowdhury, and Maji (2015) added a bilinear pooling method to CNNs for fine-grained classification. The main idea of the bilinear CNN is to use the matrix outer product to combine two 2D feature maps derived from the CNN and output a 1D vector as a global representation. However, the outer product operation in bilinear representation squares the dimension of feature maps, which may cause huge computing and memory burden. The compact bilinear pooling (CBP) method (Gao et al. 2016) reduced the dimension of features by two orders of magnitude with little loss of precision by using a polynomial kernel function such as random Maclaurin (Kar and Karnick 2012) or tensor sketch (Pham and Pagh 2013).

Bilinear pooling calculates the outer product of outputs from different feature extractors at the same spatial location and computes sum pooling to obtain bilinear features. The outer product captures the pairwise correlation between the feature channels, which is translation invariant. Bilinear combination, therefore, provides a stronger representation than a linear model and performs comparable to or higher than the output of either of the CNN branch it encoded. The second-order pooling method may be more suitable for the retrieval of optical remote sensing data, which is complex, variable and highly nonlinear due to the changes of atmospheric conditions, illumination, viewing angles, and soil moisture. However, the bilinear pooling strategy has not yet been applied to the retrieval of remote sensing images.

In recent years, attention mechanism, which is inspired by the visual mechanism of human eyes, has boosted the performances of many CNN-based vision tasks (Shimaoka et al. 2016; Fu, Zheng, and Mei 2017; Zhao et al. 2017). Wang et al. (2017) designed a stackable module that merged a trunk-and-mask attention. Squeeze-and-excitation networks (SENet) (Hu, Shen, and Sun 2018) proposed a squeeze-and-excitation operation to realize feature recalibration and improved the classification accuracy. Convolutional block attention module (CBAM) (Woo et al. 2018) conducted a spatial and a channel attention successively. The most recent selective kernel networks (SKNet) (Li et al. 2019) fused attention boosted multiple-scale features to realize the approximate adaptive selection of receptive field.

Currently, the attention methods are mainly developed for refining a single feature map. As the CBP requires two feature maps as input, special attention modules should be developed to utilize the complementary advantages of two features.

In this paper, first, we extend the CBP designed for fine-grained segmentation and object recognition to remote sensing image retrieval. We introduce a CBP layer to translate the feature maps obtained by a CNN into a global vector representation, followed by a PCA (principal component analysis) to reduce dimensionality and improve the retrieval efficiency. It is found that our CBP and PCA combination outperforms all the other pooling and encoding methods such as fully connected layer, IFK, BoW, and maxpooling in various remote sensing datasets. Second, we develop a spatial-and-channel joint attention mechanism to extract attention boosted CNN features for CBP.

We utilize a spatial and a channel attention module to produce two branches of feature maps in a ResNet backbone as the bivariate inputs of the CBP. This strategy, instead of using two copies of CNN features as inputs as most related studies have done, comprehensively improves the CBP's global consistent representation capacity and obtained the best performance on all the test datasets.

In section 2, we introduce the compact bilinear pooling, the network structure, and the attention modules designed for CBP. Results of the experiments are presented in section 3 and discussed in section 4. In section 5, we summarize the paper.

2. Methods

2.1. Compact bilinear pooling

The bilinear pooling is originally proposed in (Lin, RoyChowdhury, and Maji 2015). For the feature $\boldsymbol{f}$ at the position $l \in \mathrm{L}$ of an image $I \in \mathrm{I}$, a bilinear combination with an outer product is:

$$\mathrm{bilinear}(l, I, \boldsymbol{f}) = \boldsymbol{f}(l, I)^{\mathrm{T}} \boldsymbol{f}(l, I) \tag{1}$$

If the number of channels of the original feature map is M, then the dimension of the matrix obtained after the bilinear combination is $M \times M$. A bilinear vector is obtained through summing all the $M \times M$ matrixes at all locations:

$$\boldsymbol{B}(I) = \sum_{l \in \mathrm{L}} \mathrm{bilinear}(l, I, \boldsymbol{f}) = \sum_{l \in \mathrm{L}} \boldsymbol{f}(l, I)^{\mathrm{T}} \boldsymbol{f}(l, I) \tag{2}$$

The bilinear vector $\boldsymbol{B}(I)$ is then transformed to the corresponding normalized vector $\boldsymbol{z}$ using a signed square-root, i.e. $\boldsymbol{y} = \mathrm{sign}(\boldsymbol{x})(\boldsymbol{x})^{\circ.5}$, following the L_2 normalization, i.e. $\boldsymbol{z} = \boldsymbol{y}/\|\boldsymbol{y}\|_2$.

The fully bilinear pooling has shown advantages in fine-grained segmentation, but with the problem of large memory consuming and computational overhead. The recent compact bilinear pooling (CBP, Gao et al. 2016) remarkably reduces the parameters of the bilinear pooling through a kernelized analysis. Let $\mathbf{F}$ and $\mathbf{G}$ represent two global descriptors produced from CNNs, and $\boldsymbol{f}$ and $\boldsymbol{g}$ represent the local features at the position $s \in \mathrm{S}$ and $u \in \mathrm{U}$ respectively. The inner product of the two bilinear features is:

$$\begin{aligned} \langle \boldsymbol{B}(\boldsymbol{F}), \boldsymbol{B}(\boldsymbol{G}) \rangle &= \left\langle \sum_{s \in \mathrm{S}} \boldsymbol{f}_s \boldsymbol{f}_s^{\mathrm{T}}, \sum_{u \in \mathrm{U}} \boldsymbol{g}_u \boldsymbol{g}_u^{\mathrm{T}} \right\rangle \\ &= \sum_{s \in S} \sum_{u \in \mathrm{U}} \left\langle \boldsymbol{f}_s \boldsymbol{f}_s^{\mathrm{T}}, \boldsymbol{g}_u \boldsymbol{g}_u^{\mathrm{T}} \right\rangle \\ &= \sum_{s \in \mathrm{S}} \sum_{u \in \mathrm{U}} \langle \boldsymbol{f}_s, \boldsymbol{g}_u \rangle^2 \end{aligned}$$

A kernel method attempts to represent $<\boldsymbol{f}_s, \boldsymbol{g}_u>^2$ with an approximate kernel function $k(\boldsymbol{f}, \boldsymbol{g})$, where a mapping $\Phi(\cdot)$ with much smaller dimension than $M \times M$ meets the requirement of $<\Phi(\boldsymbol{f}), \Phi(\boldsymbol{g})> \approx k(\boldsymbol{f}, \boldsymbol{g})$. Thus,

$$\begin{aligned}\langle \boldsymbol{B}(\mathbf{F}), \boldsymbol{B}(\mathbf{G})\rangle &= \sum_{s\in S}\sum_{u\in U}\langle \boldsymbol{f}_s, \boldsymbol{g}_u\rangle^2 \\ &\approx \sum_{s\in S}\sum_{u\in U}\langle \Phi(\boldsymbol{f}), \Phi(\boldsymbol{g})\rangle \\ &= \langle C(\mathbf{F}), C(\mathbf{G})\rangle,\end{aligned}$$

where

$$C(\mathbf{F}) := \sum_{s\in S}\Phi(\boldsymbol{f}_s)$$

In this paper, the Tensor Sketch (TS) (Pham and Pagh 2013) is chosen for the low dimension mapping $\Phi(\boldsymbol{f})$.

Figure 1 shows a typical process using a CBP for pooling CNN features, where the CBP takes the last features of two CNN branches as input and outputs a 1D vector representation.

2.2. Network structure

We designed a three-stage remote sensing image retrieval network based on CBP. To achieve efficient image retrieval, we use two representative light CNN structures, VGG16 and ResNet34, for feature map extraction. In the first stage, the parameters are pretrained on the ImageNet dataset (Deng et al. 2009), and the lower level features, i.e. the first fourth convolutional layers in the VGG (Figure 2(a)) and the first three building blocks in the ResNet34 (Figure 2(b)), are kept fixed.

The second stage fine-tunes the unfixed parameters in the networks with a remote sensing dataset, which translates the high semantic representations of close-range object or scenes learned from the ImageNet to the representations of specific remote sensing objects or land covers. In Figure 2, the CBP layer is right after the last convolution layer of the VGG or ResNet to obtain translation-invariant global representations. Translation invariance is helpful for remote sensing data retrieval as instances of an object usually lie in different locations in the same dataset or between training and retrieval datasets. Two fully connected layers (FCs) follow the CBP to obtain more compact features and the last FC has the dimension equal to the number of categories of the image dataset.

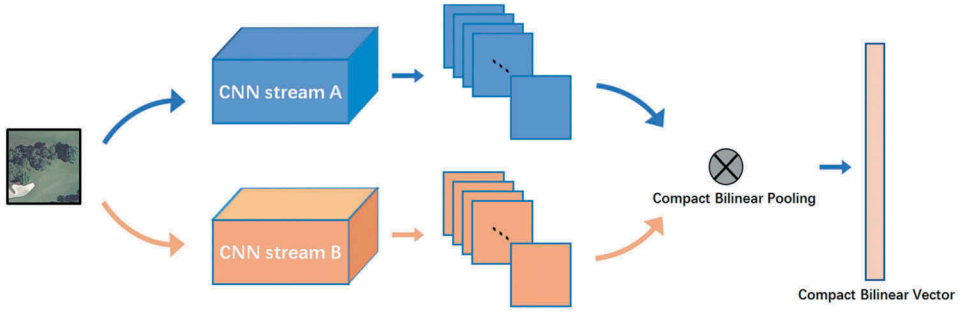


Figure 1. Compact bilinear pooling for CNN features.

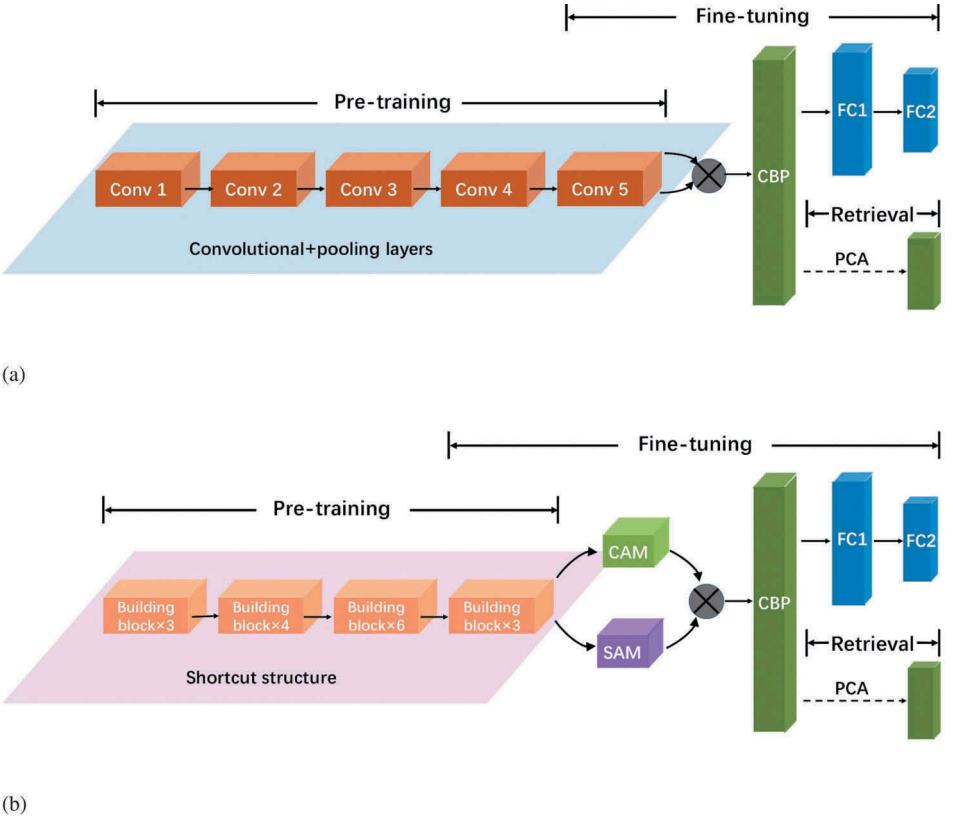


Figure 2. Three stages of remote sensing retrieval based on VGG16 (a) and ResNet34 (b) The lower level features pretrained on the ImageNet are kept fixed. In the fine-tuning stage, the rest parameters of the networks are adjusted on a remote sensing training dataset. In the retrieval stage, the FCs are replaced with PCA and the networks return the most similar images to a query image from a new dataset.

The third stage is to search a query image in a new dataset to retrieve similar images with the fine-tuned model at the second stage. We replace the FCs with a PCA module. The reason is the FC layers are heavily impacted by the original training dataset and cannot be generalized to a target dataset. PCA, which maps *n*-dimension features from the CBP to orthogonal *k*-dimension features, depends only on the more general CBP descriptors and has a better generalization ability.

2.3. Attention modules for CBP

A high-quality remote sensing image retrieval is required to recognize foreground objects from complicated and variable backgrounds of a search image. We develop an integrated spatial and channel attention mechanism to extract global consistent features of foreground objects and suppress the disturbance of backgrounds. Different from the concatenated spatial and channel attention structure (Woo et al. 2018), we perform a spatial and a channel attention module on the feature map extracted by ResNet34. The attention boosted features, namely the spatial attention boosted feature and the channel attention boosted

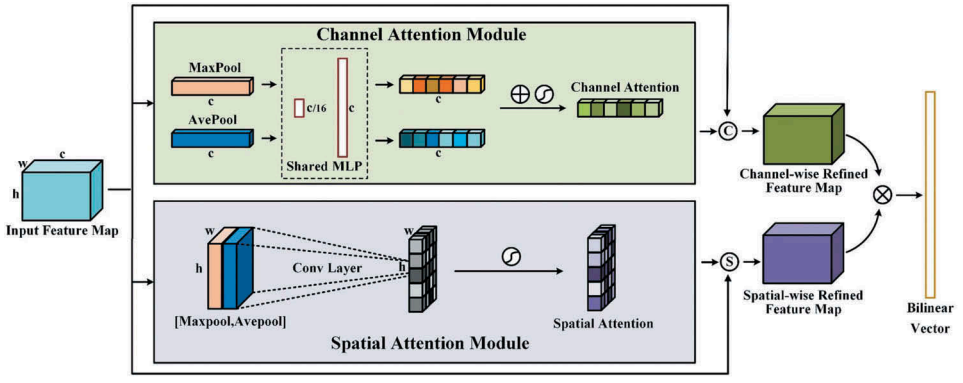


Figure 3. Attention modules for CBP. The last feature map of a CNN is processed separately by channel attention and spatial attention modules. The two attentions are each multiplied by the original feature map and the results are the input of bilinear pooling. The circled 'C' and 'S' denote the channel-wise and feature-wise multiplication of the channel attention and the spatial attention respectively. The circled '+' indicates pixel-wise addition of weight vectors. The circled curve denotes a sigmoid operation, and the circled '×' denotes compact bilinear pooling.

feature, are the bivariate inputs of the CBP. Through the CBP, a global spatial/channel consistent representation is obtained to achieve better image retrieval performance.

Figure 3 illustrates the process of using the attention modules for boosting CBP. The last feature map of the ResNet is $w \times h \times c$ dimensional, where w, h, and c are width, height and channel number, which are separately processed by a channel attention module (CAM) and a spatial attention module (SAM). In the CAM, the input feature map is respectively processed by a maxpooling and an average pooling, followed by a multi-layer perceptron (MLP). The feature dimension after pooling is $1 \times 1 \times c$, the lengths of the first and the second MLP layers are $1/16 \times c$ and c respectively. The two pooled featuresprocessed by the shared MLP are summed, followed by a sigmoid function to compute the weight vector of channel attention. The weight vector is then channel-wise multiplied with the input feature map to produce channel attention boosted feature. In the SAM, the concatenated $w \times h \times 2$ features after maxpooling and average pooling are processed with a 7×7 convolution, followed by a sigmoid to produce the weight matrix of spatial attention. The weight matrix is feature-wise multiplied by the original input to produce spatial attention boosted feature. Finally, the two boosted features are pooled by the CBP to form a bilinear vector with global spatial and channel consistent representation.

3. Experiments and results

3.1. Datasets

The experiments are carried out on four remote sensing data sets, namely PatternNet (Zhou et al. 2018), UCM (Yang and Newsam 2010), RSSCN (Zou et al. 2015), and RS-19 (Xia et al. 2010) (see Table 1). We use the PatternNet dataset to fine-tune the VGG16 and ResNet34 networks pretrained on the ImageNet dataset, and test the image retrieval performance on the other datasets.

Table 1. Attributes of the datasets.

Dataset	Size (pixels)	No. classes	No. images per class	No. total images	Usage
PatternNet	256 × 256	38	800	30400	training
WHU-RS19	600 × 600	19	around 50	1005	test
UCM	256 × 256	21	100	2100	test
RSSCN7	400 × 400	7	400	2800	test

PatternNet (Zhou et al. 2018): This large-scale remote sensing dataset consists of 38 classes and each class contains 800 images of a size of 256 × 256 pixels. The images were collected from Google Earth imagery or via a Google Map API with 0.6–4.7 m ground resolutions.

WHU-RS19 (Xia et al. 2010): The RS19 data set was captured from Google Earth imagery with various illuminations, orientations and ground sampling distance (GSD). It contains 19 types of landforms.

UC Merced Land-Use Dataset (Yang and Newsam 2010): The UCM dataset contains 2100 remote sensing images from USGS (United States geological survey). There are 21 categories and each category consists of 100 images with a GSD of 1 foot.

RSSCN7 (Zou et al. 2015): The RSSCN7 dataset contains 2,800 remote sensing images and seven scene categories. For each category, 400 images were collected from Google Earth, and every 100 images have a different scale. This is a challenging dataset with diverse scenes, scales, seasons and weather.

3.2. Compared methods

In order to explore the effectiveness and advantages of the bilinear pooling method in remote sensing image retrieval, we compare it to several other encoding and pooling methods which are applied to the features extracted from a series of convolution layers.

Bag-of-Visual-Words (BoW) (Sivic and Zisserman 2003): BoW was originally used in the field of text information retrieval. BoW can be combined with the most feature extraction methods to cluster extracted features into given groups. The centres of these groups can be regarded as 'words', all of which compose a 'dictionary'. In the end, the frequency of the 'words' in the 'dictionary' appeared in each image can be used as output vectors for subsequent classification or retrieval operations. Specifically, we compare our methods with Xia et al. (2017) who combined BoW and PCA to retrieve remote sensing images.

Improved Fisher Kernel (IFK) (Perronnin, Sánchez, and Mensink 2010a): Fisher Kernel (FK) method is another classic feature encoding method. Fisher vector is essentially a gradient vector of likelihood function to express an image. First, the method uses the Gaussian Mixture Model (GMM) to obtain the probability distribution of the entire feature space. The obtained Gaussian mixture model can be regarded as a codebook. Then for each feature, the partial derivatives with respective to the parameters (i.e. the coefficients, mean, and standard deviation vectors) of every Gaussian distribution are obtained. The obtained gradient vectors are connected in series as the feature descriptor of the image. The Improved Fisher Kernel (IFK) imposes the L_2 normalization, power normalization, and spatial pyramids operations on the FK. We compare our methods with Perronnin et al. (2010a) who used the IFK for large-scale image retrieval.

Maxpooling (Krizhevsky, Sutskever, and Hinton 2012): Maxpooling down-samples a feature map and takes the maximum value of the pixels in the specified area. For the retrieval work, the maxpooling method down-samples the entire feature maps in order to obtain a one-dimensional vector, the length of which is the same as the number of the feature maps. In this study, we compare our CBP methods with Razavian et al. (2016) who used the maxpooling for visual instance retrieval.

Fully connected layer (Ge et al. 2018): The FC flattens high-dimensional feature maps obtained from the convolutional layer. Each node of the fully connected layer is connected to all nodes of the previous layer. Many studies have employed FCs for image retrieval (Babenko et al. 2014; Zhou et al. 2017).

3.3. Implementation details

In the training stage, the VGG16 and ResNet34 are pre-trained on the ImageNet dataset respectively. Then the last convolutional blocks and FCs (Figure 2) are fine-tuned on the PatternNet. The dimension of the first FC (FC1) is 4096 and the dimension of the second FC (FC2) is 38, which is the number of categories of the PatternNet dataset.

In the VGG16-based network, the training learning rate is set to 10^{-4} and the number of iterations is set to 20 epochs. The number of iterations in the ResNet34-based network is also set to 20 epochs, the learning rate of the first 15 epochs is set to 10^{-3}, and the rest is set to 10^{-4}. In all the experiments, the input images have been resized to 224×224 pixels, the batch size is set to 64 and the Adam optimizer is utilized. The graphics card used in the experiment is NVIDIA GeForce GTX 1060 with 6 GB memory.

In the channel attention module, the size of the first MLP layer is set to 1/16 of the input size. In the spatial attention module, the size of the convolution kernel is 7×7. In the BoW, the early stopping threshold is set to1×10^{-5}. The number of kernels in the IFK algorithm will affect the accuracy of the results. To optimize the performance of the hardware device, different numbers of kernels are set in each test (Table 2).

In order to maintain the uniformity of output dimension after dimension reduction, the output dimension of the PCA is set to 38. This setting is consistent to the recommendation of Xia et al. (2017) who concluded that the output dimension with the highest retrieval accuracy is between 16 and 64.

3.4. Evaluation measure

We use Precision at k (P@k) where k is the position in the list of retrieval result and mean Average Precision (mAP) to evaluate the retrieval accuracy.

The P@k is defined as,

$$\text{P@k} = \frac{\text{number of relevant images}}{k} \tag{5}$$

Table 2. The number of kernels set in each test.

Network	RS19	UCM	RSSCN
VGG16	3	1	1
ResNet34	13	8	6

The P@k only considers the number of relevant images in the retrieval results and ignores their order. The mAP is a complementarity indicator where the more relevant an image is to the query image, the higher it ranks.

Average precision (AP) is the average of the precision of the top N positions returned for a query, which is defined as:

$$\mathrm{AP} = \frac{\sum_{k=1}^{N} (\mathrm{P@k} \times \mathrm{rel}(k))}{\text{number of relevant images}} \tag{6}$$

where a binary function rel(k) is defined as,

$$\mathrm{rel}(k) = \begin{cases} 1 \text{ the } k-\text{th image is a relevant image} \\ 0 \text{ the } k-\text{th image is a irrelevant image} \end{cases} \tag{7}$$

The mean Average Precision (mAP) is defined as,

$$\mathrm{mAP} = \frac{\sum_{q=1}^{Q} \mathrm{AP}(q)}{Q} \tag{8}$$

where Q indicates the total number of queries.

In the experiment, we set $k = 5$, 10, 50, and 100, and $N = 10$. For every dataset, i.e. RS19, UCM, RSSCN, we randomly select one image as query image and the rest for retrieving. The process is repeated Q times, Q equals one-fifth of the total image number of the dataset. Large Q reduces the uncertainty from random sampling and makes our query results more conclusive.

3.5. Experimental results

We compare our method to several recent retrieval methods and strategies, including the FC layers with fine-tuning (Babenko et al. 2014; Zhou et al. 2017); the maxpooling right after convolutional layers (Razavian et al. 2016); the Bow and PCA (Xia et al. 2017); the IFK (Perronnin et al. 2010a). We use two identical CNN backbones, i.e. the VGG16 and the ResNet34, to assess the performances of different pooling or encoding algorithms at the same baseline.

Tables 3, 4 and 5 show the results of different remote sensing image retrieval methods based on the VGG16. The combination of CBP and PCA shows a higher mAP score than the other methods on the UCM and RSSCN datasets; on the RS-19 dataset, the FC1 + PCA and CBP + PCA performed almost the same. This indicates the advantage of our proposed

Table 3. The performances of different pooling and encoding methods based on the VGG16 on the RS-19 dataset.

Method	mAP	P@5	P@10	P@50	P@100
VGG16(FC1 + FC2)	0.6887	0.5980	0.5515	0.3959	0.2678
VGG16(FC1 + PCA)	**0.7074**	**0.6082**	**0.5638**	**0.4061**	0.2763
VGG16(IFK + PCA)	0.6925	0.5929	0.5301	0.3727	0.2609
VGG16(BoW + PCA)	0.6804	0.5898	0.5398	0.4019	**0.2788**
VGG16(maxpooling)	0.6721	0.5500	0.5010	0.3583	0.2537
VGG16(CBP + FC1 + FC2)	0.6719	0.5490	0.4969	0.3501	0.2563
VGG16(CBP + FC1 + PCA)	0.6928	0.5776	0.5240	0.3654	0.2657
VGG16(CBP + PCA)	0.7054	0.5990	0.5362	0.3670	0.2639

Table 4. The performances of different pooling and encoding methods based on the VGG16 on the UCM dataset.

Method	mAP	P@5	P@10	P@50	P@100
VGG16(FC1 + FC2)	0.7178	0.6171	0.5598	0.3885	0.3073
VGG16(FC1 + PCA)	0.7315	0.6257	0.5662	**0.4020**	**0.3145**
VGG16(IFK + PCA)	0.7204	0.6138	0.5417	0.3601	0.2714
VGG16(BoW + PCA)	0.6927	0.5886	0.5317	0.3716	0.2840
VGG16(maxpooling)	0.7001	0.5952	0.5364	0.3656	0.2836
VGG16(CBP + FC1 + FC2)	0.7106	0.6024	0.5395	0.3844	0.3024
VGG16(CBP + FC1 + PCA)	0.7356	**0.6357**	**0.5731**	0.3980	0.3070
VGG16(CBP + PCA)	**0.7463**	0.6319	0.5707	0.3860	0.2910

Table 5. The performances of different pooling and encoding methods based on the VGG16 on the RSSCN dataset.

Method	mAP	P@5	P@10	P@50	P@100
VGG16(FC1 + FC2)	0.7405	0.6725	0.6486	0.5731	0.5317
VGG16(FC1 + PCA)	0.7573	0.6818	0.6648	0.5868	**0.5854**
VGG16(IFK + PCA)	0.7598	0.6850	0.6586	0.5735	0.5272
VGG16(BoW + PCA)	0.7420	0.6696	0.6514	0.5881	0.5481
VGG16(maxpooling)	0.7309	0.6586	0.6359	0.5664	0.5217
VGG16(CBP + FC1 + FC2)	0.7329	0.6525	0.6368	0.5729	0.5332
VGG16(CBP + FC1 + PCA)	0.7509	0.6807	0.6630	0.5950	0.5535
VGG16(CBP + PCA)	**0.7841**	**0.7264**	**0.6987**	**0.6196**	0.5754

CBP and PCA combination. The second-best combination is FC1 + PCA. However, it only slightly exceeds the combination of CBP + FC1 + PCA. The experimental results using FC2 (FC1 + FC2 and CBP + FC1 + FC2) are worse than only using FC1, which indicates FC1 has better generalization ability and F2 is not very suitable for retrieving images across different datasets. The combination of IFK and PCA shows better results than BoW + PCA and the maxpooling, and is approaching the second-best FC1 + PCA.

Figure 4 shows three query examples each on a dataset using the top two combinations, i.e. CBA + PCA and FC1 + PCA respectively. In the RS19, the first three retrievals with both combinations are correct. However, in the UCM, FC1 + PCA exhibits obvious errors where many structures (buildings, roads and courts) are mistaken with the queried golf course. Figure 3(c) is a challenging case where the query image labelled parking areas should be discriminated from similar images containing industrial and residential area. The CBP + PCA confused it with three industrial areas and one residential area. The FC1 + PCA made more serious mistakes where the first two retrieved images are incorrect.

Tables 6, 7 and 8 show the results of different retrieval methods based on the ResNet34 backbone on the three test datasets. First, the overall retrieval accuracy of the ResNet34 is better than the VGG16. Using CBP + PCA, the mAP on the RS-19 and UCM both improved 15%; on the RSSCN the two networks performed almost the same. This performance improvement may be due to the ResNet34 has more parameters and more suitable structures for image retrieval. Second, similar to the results of the VGG16 backbone, our CBP + PCA combination outperformed all the other pooling and encoding methods. When the attention mechanism (ATT) is introduced, the performance is comprehensively improved in terms of all indices. For example, using the ATT + CBP + PCA, the mAP improved 4% and P@5 improved 5% compared to using the CBP + PCA on the RS-19 dataset. This shows the effectiveness of the integrated spatial and channel attention

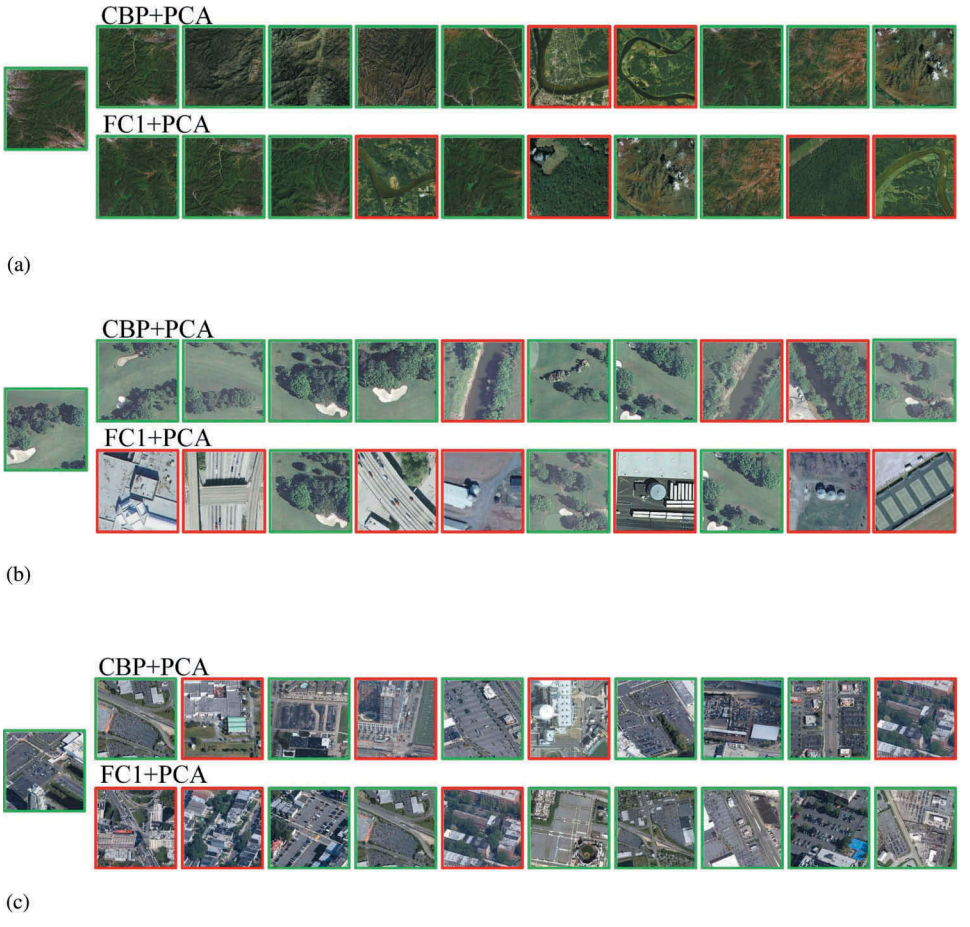


Figure 4. Image retrieval samples using VGG16-based networks. Images with green boxes are correctly retrieved images, while red boxes are incorrectly retrieved images. The first column is the query images. (a) shows the results from the RS19 dataset, and the query image is labelled with mountains. (b) shows the results from the UCM dataset, and the query images are labelled with a golf course. (c) shows the results from the RSSCN dataset, and the query image belongs to parking category.

Table 6. The performances of different pooling and encoding methods based on the ResNet34 on the RS-19 dataset.

Method	mAP	P@5	P@10	P@50	P@100
ResNet34(FC1 + FC2)	0.8101	0.7337	0.6786	0.4878	0.3277
ResNet34(FC1 + PCA)	0.8342	0.7714	0.7219	0.5113	0.3333
ResNet34(IFK + PCA)	0.7978	0.7214	0.6612	0.4091	0.2669
ResNet34(BoW + PCA)	0.7578	0.6765	0.6337	0.4811	0.3294
ResNet(maxpooling)	0.8551	0.8071	0.7628	0.5313	0.3533
ResNet34(CBP + FC1 + FC2)	0.7905	0.6980	0.6429	0.4440	0.3049
ResNet34(CBP + FC1 + PCA)	0.8010	0.7296	0.6847	0.5071	0.3405
ResNet34(CBP + PCA)	0.8568	0.8000	0.7362	0.5403	0.3534
ResNet34(ATT + CBP + PCA)	**0.8951**	**0.8490**	**0.7974**	**0.6026**	**0.3901**

Table 7. The performances of different pooling and encoding methods based on the ResNet34 on the UCM dataset.

Method	mAP	P@5	P@10	P@50	P@100
ResNet34(FC1 + FC2)	0.8714	0.8195	0.7769	0.6292	0.5032
ResNet34(FC1 + PCA)	0.8859	0.8390	0.7983	0.6466	0.5097
ResNet34(IFK + PCA)	0.8232	0.7510	0.7126	0.5266	0.3839
ResNet34(BoW + PCA)	0.7716	0.7014	0.6664	0.5361	0.4303
ResNet(maxpooling)	0.8498	0.8033	0.7738	0.6047	0.4677
ResNet34(CBP + FC1 + FC2)	0.8504	0.7995	0.7733	0.6591	0.5435
ResNet34(CBP + FC1 + PCA)	0.8739	0.8129	0.7860	0.6816	0.5719
ResNet34(CBP + PCA)	0.8890	0.8362	0.8040	0.6761	0.5484
ResNet34(ATT + CBP + PCA)	**0.9056**	**0.8638**	**0.8367**	**0.7227**	**0.5939**

Table 8. The performances of different pooling and encoding methods based on the ResNet34 on the RSSCN dataset.

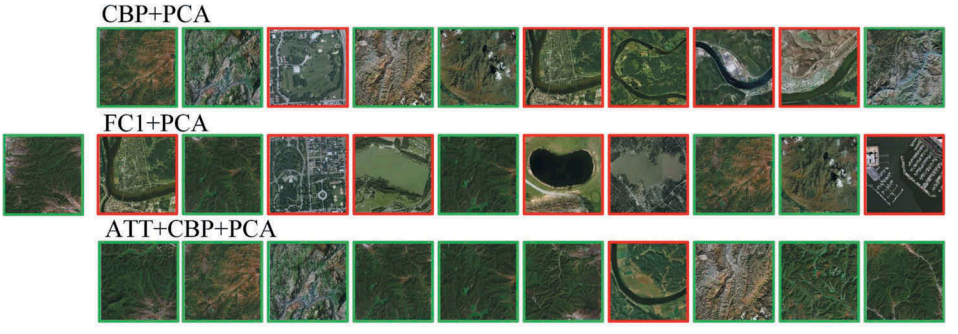
Method	mAP	P@5	P@10	P@50	P@100
ResNet34(FC1 + FC2)	0.7585	0.6829	0.6588	0.5814	0.5403
ResNet34(FC1 + PCA)	0.7811	0.7136	0.6877	0.6016	0.5529
ResNet34(IFK + PCA)	0.7467	0.6693	0.6409	0.5320	0.4654
ResNet34(BoW + PCA)	0.6955	0.6154	0.6020	0.5569	0.5203
ResNet34(maxpooling)	0.7578	0.6907	0.6805	0.6001	0.5544
ResNet34(CBP + FC1 + FC2)	0.7508	0.6729	0.6346	0.5501	0.5050
ResNet34(CBP + FC1 + PCA)	0.7524	0.6796	0.6441	0.5651	0.5217
ResNet34(CBP + PCA)	0.7777	0.7118	0.6850	0.5990	0.5464
ResNet34(ATT + CBP + PCA)	**0.8132**	**0.7550**	**0.7312**	**0.6542**	**0.6045**

mechanism. Third, the FC layers performed relatively worse for cross-dataset image retrieval. For example, the CBP + PCA outperformed the CBP + FC1 + FC2 and the CBP + FC1 + PCA more than 5% on mAP on the RS-19. Maxpooling performed worse than the FC1 + PCA but exceeded the IFK and BoW on all the datasets.

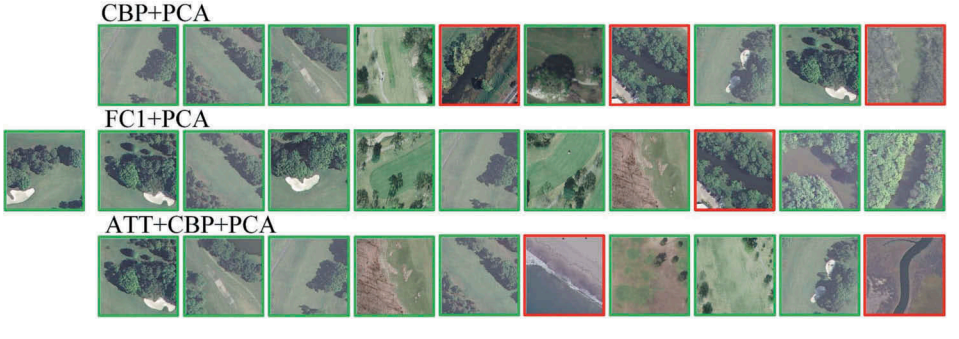
Figure 5 shows three query examples using the ResNet34 with the CBA + PCA, the FC1 + PCA, and the ATT + CBP + PCA respectively. The query images are the same as in Figure 4. In the RS19, rivers are mistaken with mountains. This phenomenon is the same as using the VGG16 but the first retrieval image is wrong using the FC1 + PCA. However, the results of the ATT + CBP + PCA is much better than the others, indicating the contribution of our specific attention mechanism. In the UCM, the CBP + PCA performed the same as in Figure 4(b) where three river images were classified to a golf course. But the FC1 + PCA performed much better than with the VGG16 building blocks, the latter mistook seven manmade constructions as a golf course. It is also observed that the last three retrieved images using the FC1 + PCA as well as the three wrongly labelled images using the CBP + PCA are very similar (covered with the river). This is also ambiguous to human vision as whether a river locates within a golf course or not determines its label. The method based on the ATT + CBP + PCA confused beach texture with grassland.

In the challenging case of Figure 5(c), the errors are caused by similar scenes, i.e. industrial areas and residential areas. The results of ATT + CBP + PCA is better than the others and the result of the FC1 + PCA is the worst.

The last experiment is designed to evaluate the performance of recent attention methods. In Tables 9, 10 and 11, the 'SE' or 'SK' indicates an SE (Hu, Shen, and Sun 2018) or SK (Li et al. 2019) module is added behind the convolution encoder as ours. 'SE-



(a)



(b)

(c)

Figure 5. Examples of remote sensing image retrieval using ResNet-based networks. Images with green boxes are correctly retrieved images, while red boxes are incorrectly retrieved images. The first column is the query images. (*a*) shows the results on the RS19 dataset, and the query image belongs to the mountain category. (*b*) shows the results retrieved from the UCM dataset, and the query images belong to the golfcourse category. (*c*) shows the results of retrieval on the RSSCN dataset, and the query image belongs to the parking category.

boosted' or 'SK-boosted' indicates the SE or SK module is added in every residual block, as has been used in the original SENet (Hu, Shen, and Sun 2018) and SKNet (Li et al. 2019).

Table 9. The performance of different attention modules on RS19.

Method	mAP	P@5	P@10	P@50	P@100
ResNet34(ATT+CBP+PCA)	**0.8951**	**0.8490**	**0.7974**	**0.6026**	**0.3901**
ResNet34(SE+CBP+PCA)	0.8472	0.7980	0.7592	0.5473	0.3628
ResNet34(SK+CBP+PCA)	0.8299	0.7704	0.7143	0.5368	0.3610
SE-boosted ResNet34 (CBP+PCA)	0.7961	0.7133	0.6684	0.4573	0.3113
SK-boosted ResNet34 (CBP+PCA)	0.7773	0.6959	0.6362	0.4350	0.3006

Table 10. The performance of different attention modules on UCM.

Method	mAP	P@5	P@10	P@50	P@100
ResNet34(ATT+CBP+PCA)	**0.9056**	**0.8638**	**0.8367**	**0.7227**	**0.5939**
ResNet34(SE+CBP+PCA)	0.8987	0.8567	0.8143	0.6923	0.5592
ResNet34(SK+CBP+PCA)	0.8768	0.8357	0.7988	0.6925	0.5844
SE-boosted ResNet34 (CBP+PCA)	0.8707	0.8148	0.7660	0.5948	0.4536
SK-boosted ResNet34 (CBP+PCA)	0.8108	0.7338	0.6724	0.4720	0.3593

Table 11. The performance of different attention modules on RSSCN.

Method	mAP	P@5	P@10	P@50	P@100
ResNet34(ATT+CBP+PCA)	**0.8132**	**0.7550**	**0.7312**	**0.6542**	**0.6045**
ResNet34(SE+CBP+PCA)	0.7845	0.7132	0.6902	0.6179	0.5700
ResNet34(SK+CBP+PCA)	0.7674	0.7064	0.6845	0.6134	0.5630
SE-boosted ResNet34 (CBP+PCA)	0.7868	0.7193	0.6818	0.5818	0.5350
SK-boosted ResNet34 (CBP+PCA)	0.7829	0.7132	0.6787	0.6002	0.5496

From Table 9–11, it is observed that our attention module (ATT) is obviously better than the SE and SK modules on all the indicators. It is due to that we designed a specific attention structure for the bivariate inputs of the CBP, on the contrary, both the SE and SK modules were developed for boosting unary input. It is also observed that putting the attention module on the last features of a CNN is better than placing it at every residual block.

From the above experiments, four main conclusions can be drawn. First, CBP, as a powerful pooling method, performed better than FC, IFK, BoW and maxpooling on the different convolutional building blocks. Second, with the integrated spatial and channel mechanism, the performance of the CBP is boosted. The combination of the ATT + CBP + PCA based on the ResNet34 backbone performed the best in all tests. Third, FC is a good pooling method but only the FC1 has enough generalization ability and the FC2 should be replaced with PCA. Finally, our spatial-channel attention mechanism designed for CBP is better than other attention methods.

4. Discussion

In this section, we discuss why the attention mechanism is particularly effective on the ResNet building block. We have applied the attention module (Section 2.3) to the VGG16 structure, but the improvement is not significant.

The difference in performance mainly comes from the specific shortcut connection in the ResNet. In each shortcut block the relationship between the output $\mathbf{y}$ of an input $\mathbf{x}$ is denoted as $\mathbf{y} = F(\mathbf{x}) + \mathbf{x}$, where $F(\mathbf{x})$ represents a series of convolutions and activations. The summation operator implies the weights of $\mathbf{x}$ and $F(\mathbf{x})$ are equal in the spatial and channel

dimensions. However, this is usually not true and may cause an imbalance between channels and spatial locations. The channel and spatial attention module after the last residual module achieved a global consistent representation and improved the retrieval results remarkably.

We visualized the improvement of representation ability from the attention modules with the tool of Grad-CAM (Selvaraju et al. 2017). Figure 6 shows the heat maps of the last feature maps of ResNet34 with and without the attention module on the PatternNet dataset. Without the attention module (original heat map), the heat maps failed to cover the entire target regions. For example, in the first column, the hot spots only cover a wing of a plane, and in the second column, the location of hot spots is biased from the baseball field. Using this kind of feature map (and its copy) as the input of the CBP cannot reach a global optimal representation. In contrast, the heat maps after the adjustment of the channel (channel-wise refined heat map) and spatial attention (spatial-wise refined heat map) can cover the entire target regions, which makes their corresponding feature maps ideal bivariate inputs of the CBP to produce a global optimal representation.

We have carried out additional experiments. Some observations are summarized below:

The number of fully connected layers. We compared the impact of a different number of fully connected layers in the fine-tuning stage on the retrieval results. It was observed that two fully connected layers can achieve the best performance while increasing or decreasing them decreases the accuracy.

Dropout operation. We attempted to add a dropout operation between the fully connected layers (remove 50% connections), but the retrieval accuracy decreased.

Memory limits for aggregation approach. The retrieval accuracy based on the BoW and IFK methods may be limited by available memory. However, in practice, the demand for resource and efficiency is also crucial. With the same resource, the CBP- based method exceeds both of them.

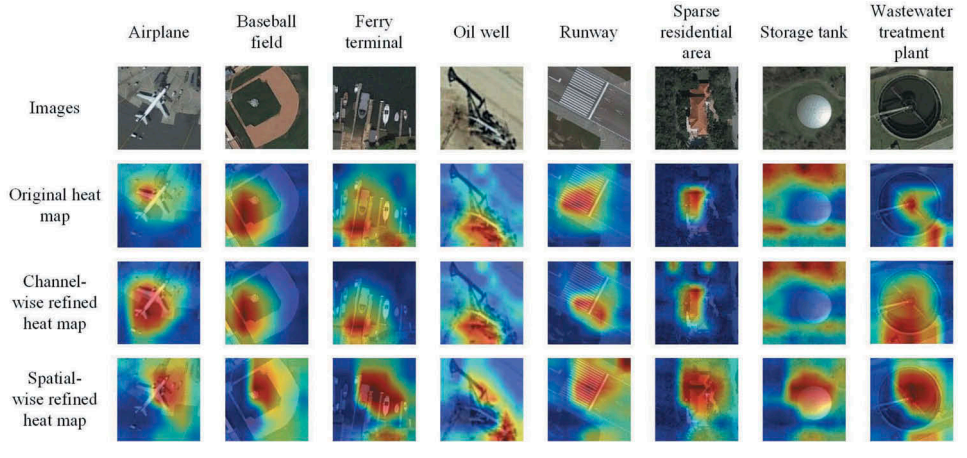


Figure 6. Heat map visualization of the last ResNet34 feature map on the PatternNet dataset. Original heat map means the results without attention mechanism.

Triplet network. We assessed alternative network structures, for example, the triplet network (Hoffer and Ailon 2015). In this structure, three images (anchor, positive, negative) are input into the triplet network at the same time, and the corresponding features are extracted by using a three-branch CNN with shared weights. The triplet loss (Schroff, Kalenichenko, and Philbin 2015) is defined to minimize the differences between relevant images as well as to maximize the distance between the anchor and negative images. However, the triplet network fine-tuned on the PatternNet did not perform as good as the single network on the three test datasets. This may be due to the differences between our datasets and the datasets used in Hoffer and Ailon (2015), the latter is much smaller. Moreover, the complexity of the network will lead to difficulties in parameter adjustment, which may cause the network to become unstable.

5. Conclusion

In this work, we integrated compact bilinear pooling into the CNN based remote sensing image retrieval. The results on various datasets have proved that the CBP is better than the fully connected layer, BoW, IFK and maxpooling, indicating the potential of CBP in remote sensing image retrieval. Using PCA has shown to be better than using FC layer as the former is only related to the more general features and the latter is more sensitive to the training samples. We also developed two attention modules in the ResNet to produce spatial boosting and channel boosting feature maps, from which the CBP learns global consistent representations and achieved comprehensive improvement in all image retrieval indices.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the National Key Research and Development Program of China, Grant No. [2018YFB0505003].

References

Arandjelović, R., and A. Zisserman. 2012. "Three Things Everyone Should Know to Improve Object Retrieval." Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16–21.

Babenko, A., A. Slesarev, A. Chigorin, and V. Lempitsky. 2014. "Neural Codes for Image Retrieval." Paper Presented at the Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, September 6–12.

Babenko, A., and V. Lempitsky. 2015. "Aggregating Deep Convolutional Features for Image Retrieval." *arXiv preprint arXiv:1510.07493*.

Bay, H., T. Tuytelaars, and L. Van Gool. 2006. "SURF: Speeded Up Robust Features." Paper Presented at the Proceedings of the European Conference on Computer Vision, Graz, Austria, May 7–13.

Bretschneider, T., R. Cavet, and O. Kao. 2002. "Retrieval of Remotely Sensed Imagery Using Spectral Information Content." Paper Presented at the Proceedings of IEEE International Geoscience and Remote Sensing Symposium, Toronto, Canada, June 24–28.

Carreira, J., R. Caseiro, J. Batista, and C. Sminchisescu. 2012. "Semantic Segmentation with Second-order Pooling." Paper Presented at the Proceedings of the European Conference on Computer Vision, Florence, Italy, October 7–13.

Choraś, R. S., T. Andrysiak, and M. Choraś. 2007. "Integrated Color, Texture and Shape Information for Content-based Image Retrieval." *Pattern Analysis and Applications* 10 (4): 333–343. doi:10.1007/s10044-007-0071-0.

Dalal, N., and B. Triggs. 2005. "Histograms of Oriented Gradients for Human Detection." Paper Presented at the Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, June 20–26.

Deng, J., W. Dong, R. Socher, L. Li-Jia, L. Kai, and L. Fei-Fei. 2009. "Imagenet: A Large-Scale Hierarchical Image Database." Paper Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, USA, June 20–25.

Du Buf, J. M. H., M. Kardan, and M. Spann. 1990. "Texture Feature Performance for Image Segmentation." *Pattern Recognition* 23: 291–309. doi:10.1016/0031-3203(90)90017-F.

Du, P., Y. Chen, H. Tang, and T. Fang. 2006. "Hyperspectral Remote Sensing Image Retrieval Based on Spectral Similarity Measure." Paper Presented at the Proceedings of SPIE Remote Sensing and Space Technology for Multidisciplinary Research and Applications, Beijing, China, May 19.

Fu, J., H. Zheng, and T. Mei. 2017. "Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition." Paper Presented at the Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii, USA, July 21–26.

Gao, Y., O. Beijbom, N. Zhang, and T. Darrell. 2016. "Compact Bilinear Pooling." Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, June 27–30.

Ge, Y., Y. Tang, S. Jiang, L. Leng, S. Xu, and F. Ye. 2018. "Region-based Cascade Pooling of Convolutional Features for HRRS Image Retrieval." *Remote Sensing Letters* 9 (10): 1002–1010. doi:10.1080/2150704X.2018.1504334.

Gordo, A., J. Almazán, J. Revaud, and D. Larlus. 2016. "Deep Image Retrieval: Learning Global Representations for Image Search." Paper Presented at the Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, October 8 –16.

Haralick, R. M., K. Shanmugam, and I. Dinstein. 1973. "Textural Features for Image Classification." *IEEE Transactions on Systems, Man and Cybernetics* 6: 610–621. doi:10.1109/TSMC.1973.4309314.

He, K., X. Zhang, S. Ren, and J. Sun. 2016. "Deep Residual Learning for Image Recognition." Paper Presented at the Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, June 27–30.

Hoffer, E., and N. Ailon. 2015. "Deep Metric Learning Using Triplet Network." *International Workshop on Similarity-Based Pattern Recognition* 84–92. doi:10.1007/978-3-319-24261-3_7.

Hu, J., L. Shen, and G. Sun. 2018. "Squeeze-and-Excitation Networks." Paper Presented at the Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Utah, USA, June 18–22.

Jégou, H., and A. Zisserman. 2014. "Triangulation Embedding and Democratic Aggregation for Image Search." Paper Presented at the Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, June 24–27.

Jégou, H., M. Douze, and C. Schmid. 2008. "Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search." Paper Presented at the Proceedings of the European Conference on Computer Vision, Marseille, France, October 12–18.

Jégou, H., M. Douze, C. Schmid, and P. Patrick. 2010. "Aggregating Local Descriptors into a Compact Image Representation." Paper Presented at the Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, USA, June 13–18.

Jian, X., D. Xiao-qing, W. Sheng-jin, and W. You-shou. 2008. "Background Subtraction Based on a Combination of Texture, Color and Intensity." Paper Presented at the Proceedings of 2008 9th International Conference on Signal Processing, Beijing, China October 26–29.

Kar, P., and H. Karnick. 2012. "Random Feature Maps for Dot Product Kernels." Paper Presented at the Proceedings of Artificial Intelligence and Statistics, La Palma, Canary Islands, Spain, April 21–23.

Krizhevsky, A., I. Sutskever, and G. E. Hinton. 2012. "Imagenet Classification with Deep Convolutional Neural Networks." Paper Presented at the Proceedings of Advances in Neural Information Processing Systems, Lake Tahoe, Nevada, December 3–8.

Lazebnik, S., C. Schmid, and J. Ponce. 2006. "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories." Paper Presented at the Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, New York, USA, June 17–22.

Lecun, Y., Y. Bengio, and G. Hinton. 2015. "Deep Learning." *Nature* 521 (7553): 436. doi:10.1038/nature14539.

Li, X., W. Wang, X. Hu, and J. Yang. 2019. "Selective Kernel Networks." Paper Presented at the Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Los Angeles, CA, USA, June 16–20.

Lienhart, R., and J. Maydt. 2002. "An Extended Set of Haar-like Features for Rapid Object Detection." Paper Presented at the Proceedings of the 2002 International Conference on Image Processing. New York, USA, September 22–25.

Lin, T. Y., A. RoyChowdhury, and S. Maji. 2015. "Bilinear CNN Models for Fine-grained Visual Recognition." Paper Presented at the Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, December 7–13.

Lowe, D. G. 2004. "Distinctive Image Features from Scale-invariant Keypoints." *International Journal of Computer Vision* 60: 91–110. doi:10.1023/B:VISI.0000029664.99615.94.

Ma, W. Y., and B. S. Manjunath. 1996. "Texture Features and Learning Similarity." Paper Presented at the Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, June 18–20.

Ojala, T., M. Pietikainen, and T. Maenpaa. 2002. "Multiresolution Gray-scale and Rotation Invariant Texture Classification with Local Binary Patterns." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24: 971–987. doi:10.1109/TPAMI.2002.1017623.

Papageorgiou, C., M. Oren, and T. Poggio. 1998. "A General Framework for Object." Paper Presented at the Proceedings of the IEEE International Conference on Computer Vision, Bombay, India, January 4–7.

Park, D. K., Y. S. Jeon, and C. S. Won. 2000. "Efficient Use of Local Edge Histogram Descriptor." Paper Presented at the Proceedings of the 2000 ACM Workshops on Multimedia, Los Angeles, CA, USA, October 30–November 3.

Perronnin, F., J. Sánchez, and T. Mensink. 2010a. "Improving the Fisher Kernel for Large-Scale Image Classification." Paper Presented at the Proceedings of the European Conference on Computer Vision, Crete, Greece, September 5–11.

Perronnin, F., Y. Liu, J. Sánchez, and H. Poirier. 2010b. "Large-scale Image Retrieval with Compressed Fisher Vectors." Paper Presented at the Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, USA, June 13–18.

Pham, N., and R. Pagh. 2013. "Fast and Scalable Polynomial Kernels via Explicit Feature Maps." Paper Presented at the Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, Chicago, USA, August 11 – 14.

Philbin, J., M. Isard, J. Sivic, and A. Zisserman. 2010. "Descriptor Learning for Efficient Retrieval." Paper Presented at the Proceedings of the European Conference on Computer Vision, Crete, Greece, September 5–11.

Prasad, B. G., S. K. Gupta, and K. K. Biswas. 2001. "Color and Shape Index for Region-based Image Retrieval." *International Workshop on Visual Form* 716–725. doi:10.1007/3-540-45129-3_66.

Rangayyan, R. M., R. J. Ferrari, J. L. Desautels, and A. F. Frere. 2000. "Directional Analysis of Images with Gabor Wavelets." Paper Presented at the Proceedings of the 13th Brazilian Symposium on Computer Graphics and Image Processing, Gramado, Brazil, October 17–20.

Razavian, A. S., J. Sullivan, S. Carlsson, and A. Maki. 2016. "Visual Instance Retrieval with Deep Convolutional Networks." *ITE Transactions on Media Technology and Applications* 4 (3): 251–258. doi:10.3169/mta.4.251.

Roweis, S. T., and L. K. Saul. 2000. "Nonlinear Dimensionality Reduction by Locally Linear Embedding." *science* 290 (5500): 2323–2326. doi:10.1126/science.290.5500.2323.

Rublee, E., V. Rabaud, K. Konolige, and G. Bradski. 2011. "ORB: An Efficient Alternative to SIFT or SURF." Paper Presented at the IEEE International Conference on Computer Vision, Barcelona, Spain, November 6–13.

Schroff, F., D. Kalenichenko, and J. Philbin. 2015. "Facenet: A Unified Embedding for Face Recognition and Clustering." Paper Presented at the Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, June 7 –12.

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. 2017. "Grad-cam: Visual Explanations from Deep Networks via Gradient-based Localization." Paper Presented at the Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, October 22–29.

Shimaoka, S., P. Stenetorp, K. Inui, and S. Riedel. 2016. "An Attentive Neural Architecture for Fine-grained Entity Type Classification." *arXiv preprint arXiv:1604.05525*.

Simonyan, K., and A. Zisserman. 2015. "Very Deep Convolutional Networks for Large-Scale Image Recognition." Paper Presented at the Proceedings of International Conference on Learning Representations, San Diego, CA, USA, May 7–9.

Sivic, J., and A. Zisserman. 2003. "Video Google: A Text Retrieval Approach to Object Matching in Videos." Paper Presented at the Proceedings of the IEEE International Conference on Computer Vision, Nice, France, October 13 –16.

Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2015. "Going Deeper with Convolutions." Paper Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, June 7–12.

Tolias, G., T. Furon, and H. Jégou. 2014. "Orientation Covariant Aggregation of Local Descriptors with Embeddings." Paper Presented at the Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, September 6 –12.

Vellaikal, A., C. C. J. Kuo, and S. K. Dao. 1995. "Content-based Retrieval of Remote-sensed Images Using Vector Quantization." Paper Presented at the Proceedings of Visual Information Processing IV, Orlando, FL, USA, April 17.

Viola, P., and M. Jones. 2001. "Rapid Object Detection Using a Boosted Cascade of Simple Features." Paper Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Kauai, HI, USA, December 8–14.

Wan, J., D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li. 2014. "Deep Learning for Content-based Image Retrieval: A Comprehensive Study." Paper Presented at the Proceedings of the 22nd ACM international conference on Multimedia, Orlando, Florida, November 3–7.

Wang, F., M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. 2017. "Residual Attention Network for Image Classification." Paper Presented at the Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii, USA, July 21–26.

Wang, J., J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. 2010. "Locality-Constrained Linear Coding for Image Classification." Paper Presented at the Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, USA, June 13–18.

Woo, S., J. Park, J. Y. Lee, and I. So Kweon. 2018. "CBAM: Convolutional Block Attention Module." Paper Presented at the Proceedings of the European Conference on Computer Vision, Munich, Germany, September 8–14.

Xia, G. S., W. Yang, J. Delon, Y. Gousseau, H. Sun, and H. Maître. 2010. "Structural High-resolution Satellite Image Indexing." Paper Presented at the Proceedings of ISPRS TC VII Symposium – 100 Years ISPRS, Vienna, Austria, July 5 –7.

Xia, G. S., X. Y. Tong, F. Hu, Y. Zhong, M. Datcu, and L. Zhang. 2017. "Exploiting Deep Features for Remote Sensing Image Retrieval: A Systematic Investigation." *arXiv preprint arXiv:1707.07321*.

Yang, Y., and S. Newsam. 2010. "Bag-Of-Visual-Words and Spatial Extensions for Land-Use Classification." Paper Presented at the Proceedings of the 18th ACM SIGSPATIAL international conference on advances in geographic information systems, San Jose, CA, USA, November 2–5.

Yue-Hei Ng, J., F. Yang, and L. S. Davis. 2015. "Exploiting Local Features from Deep Networks for Image Retrieval." Paper Presented at the Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, December 7–13.

Zhang, S., H. Yao, and S. Liu. 2008. "Dynamic Background Modeling and Subtraction Using Spatio-temporal Local Binary Patterns." Paper Presented at the Proceedings of 15th IEEE International Conference on Image Processing, San Diego, CA, USA, October 12–15.

Zhao, B., X. Wu, J. Feng, Q. Peng, and S. Yan. 2017. "Diversified Visual Attention Networks for Fine-grained Object Classification." *arXiv preprint arXiv: 1606.08572*.

Zhao, G., and M. Pietikäinen. 2007. "Dynamic Texture Recognition Using Volume Local Binary Patterns." *Dynamical Vision* 165–177. doi:10.1007/978-3-540-70932-9_13.

Zhou, W., S. Newsam, C. Li, and Z. Shao. 2017. "Learning Low Dimensional Convolutional Neural Networks for High-Resolution Remote Sensing Image Retrieval." *Remote Sensing* 9 (5): 489. doi:10.3390/rs9050489.

Zhou, W., S. Newsam, C. Li, and Z. Shao. 2018. "Patternnet: A Benchmark Dataset for Performance Evaluation of Remote Sensing Image Retrieval." *ISPRS Journal of Photogrammetry and Remote Sensing* 145: 197–209. doi:10.1016/j.isprsjprs.2018.01.004.

Zou, Q., L. Ni, T. Zhang, and Q. Wang. 2015. "Deep Learning Based Feature Selection for Remote Sensing Scene Classification." *IEEE Geoscience and Remote Sensing Letters* 12 (11): 2321–2325. doi:10.1109/LGRS.2015.2475299.