Deep learning for visual understanding

Guo, Y.; Guo Y.

Chapter 3

Convolutional Neural Networks Features: Principal Pyramidal Convolution

The features extracted from convolutional neural networks (CNNs) are able to capture the discriminative part of an image and have shown superior performance in visual recognition. Furthermore, it has been verified that the CNN activations trained from large and diverse datasets can act as generic features and be transferred to other visual recognition tasks. In this chapter, we aim to learn more from an image and present an effective method called Principal Pyramidal Convolution (PPC). The scheme first partitions the image into two levels, and extracts CNN activations for each sub-region along with the whole image, and then aggregates them together. The concatenated feature is later reduced to the standard dimension using Principal Component Analysis (PCA) algorithm, generating the refined CNN feature. When applied in image classification and retrieval tasks, the PPC feature consistently outperforms the conventional CNN feature, regardless of the network type where they derive from.

3.1 Introduction

Convolutional Neural Networks (CNN) have achieved breakthrough achievements in various visual recognition tasks and have been extensively studied in recent years [22, 44, 45, 213]. There are several brilliant properties for the CNN feature: 1)It is highly discriminative. Related research has analyzed the behavior of the intermediate layers of CNN and demonstrated that it can capture the most obvious features [22], thus could achieve considerably better results in a number of applications [22, 213]; 2) Unlike the hand-crafted features such as SIFT [1], HOG [3], the CNN feature is generated from end-to-end, which eliminates the human intervention; 3) it can be achieved efficiently. In contrast to the standard feedforward neural networks with similarly-sized layers, CNN has fewer connections and parameters, which reduces the time cost of the feature extraction; 4) it is transferrable. Some works [45, 62] have demonstrated that CNN features trained on large and diverse datasets, such as ImageNet [149] and Places [216], could be transferred to other visual recognition tasks, even there are substantial differences between the datasets.

Owing to those notable characters, our research focuses on reusing of the off-the-shelf CNN feature. But, instead of computing the CNN feature over the full image, we ask whether we could get more information from an image and achieve a refined version of the CNN feature?

An intuitive way to achieve more knowledge is to extract multiple CNN features from one image and organize them in a proper way. In recent years, there are a number of works attempt to extract multiple features from one image, either in region proposals [44] or sliding windows [170]. But most of those methods are used for object detection, not for the refinement of CNN features. Besides, the extraction of numerous features from overlapping regions is quite inefficient.

Related works have been done in the past [170, 217]. In the work by Gong et al. [170], they extract CNN activations at multiple scale levels, perform orderless VLAD pooling separately, and concatenate them together, forming a high dimensional feature vector which is more robust to the global deformations. Koskela et al. [217] splits one image into nine regions and averages their CNN activations,

concatenating with the activation of the entire image. The resulting spatial pyramid features are certificated to be more effective in scene recognition.

Different from previous works, we show that the concatenation of the CNN features from one image could also improve the performance, without further calculation or other time-consuming processes. To avoid increasing the complexity during the test phase and keep the key components at the meanwhile, we compress the dimension to the normal one (4096-D) using PCA scheme after the concatenation and get the refined feature: Principal Pyramidal Convolution (PPC).

The idea of concatenating features has ever been done in the literatures. The most representative one is the spatial pyramid matching (SPM) [135] algorithm, which concatenates the BOF vectors of the sub-regions as well as the whole image to import the global spatial information. SPM achieves a substantial improvement over the traditional BOF and has long been a key component in the competition-winning systems for visual recognition before the surge of CNN [31, 122].

In this chapter, the BOF vector of the SPM algorithm is replaced by the discriminative CNN feature. Therefore, besides preserving the discrimination of CNN, PPC also introduces some spatial information as well as preserving the most important components. In addition, the strategy is portable, experiments show that whichever network the CNN activations derive from, PPC strategy could consistently improve the performance.

3.2 Principal Pyramidal Convolution

Inspired by SPM, which extracts features at multiple levels and aggregates them together, we propose the Principal Pyramidal Convolution (PPC) method. It divides the image into two levels and generates the final feature for the image by concatenating and extracting principal components for the features at all resolutions. The basic idea is illustrated in Figure 3.1.

We extract CNN features from two scale levels. The first level corresponds to the full image, and the second level consists of 2×2 regions by equally partitioning

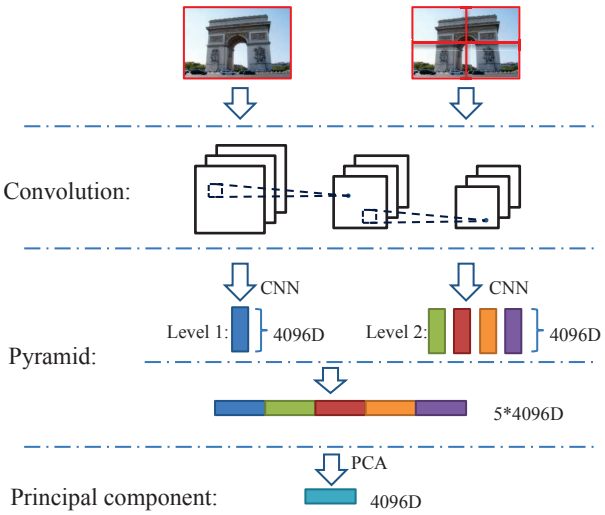

Figure 3.1: The procedure of PPC algorithm.

the full image. Therefore, we need to extract five CNN features for each image: C_0, C_1, C_2, C_3, C_4. Afterwards, we concatenate the five CNN features in an intuitive scheme: $C = [C_0, C_1, C_2, C_3, C_4]$. The resulting C is a 5*4096-dimensional vector. The CNN activations are achieved using the Caffe implementation [218]. Here, we select the 4096-dimensional output of the seventh layer (i.e. the last fully-connected layer) and L2-normalize it as the baseline CNN feature.

To eliminate the increase of computational cost, we compress the resulting feature vector to 4096-D in the last step. For the dimension reduction, we utilize the well-known PCA method [219], which could reduce the dimensionality of a data set and retain as much as possible of the variation at the same time.

In addition, we also reduce the dimension to other various sizes and compare the performance between conventional CNN and PPC for different visual tasks, including the supervised image classification and unsupervised image retrieval.

3.3 Experiment

In this part, we make some comparisons between the conventional CNN feature and PPC feature on various image classification and image retrieval databases.

3.3.1 Datasets

We present the results on four widely used datasets: Caltech-101 [220], Scene15 [135], MIT Indoor67 database [221] and INRIA Holidays [222].

The details of the datasets are summarized in Table 3.1.

Table 3.1: Details of the datasets

Datasets	Details
Caltech-101:	102 categories and a total of 9144 images, the image number per category ranges from 31 to 800. We follow the procedure of [15] and randomly select 30 images per class for training and test on up to 50 images per class;
Scene15:	4485 greyscale images assigned to 15 categories. Each category has 200 to 400 images. We use 100 images per class for training and the rest for testing;
Indoors67:	67 categories and 15620 images in total. The standard training/test split consists of 80 images for training and 20 images for testing per class;
Holidays:	1491 images corresponding to 500 image instances. Each instance has 2-3 images describing the same object or location. The images have been rectified to a natural orientation. 500 images of them are used as queries.

In the databases described above, the first three datasets are used for image classification, on which we train linear SVM classifiers (s=0, t=0) to recognize the test images, using the LIBSVM tool [223]. The last dataset is a standard benchmark for image retrieval, and the accuracy is measured by the mean Average Precision (mAP) [224].

3.3.2 Comparisons on different networks

According to which database the CNN is trained on, the CNN features can be categorized into two types: ImageNet-CNN and Places-CNN. ImageNet-CNN is the most commonly used model which is trained on the well-known database: ImageNet [149]. This database contains 1000 categories with around 1.3 million images and most of the images are object-centric. Places-CNN is another model which is trained on the recently proposed Places database and is scene-

centric [216]. This database contains about 2.5 million images assigned to 205 scene categories.

In this chapter, we utilize the off-the-shelf CNN features of ImageNet and Places respectively, and compare the performance of CNN and PPC on the four datasets. The results are shown in Table 3.2.

Table 3.2: The classification accuracies of SPM and CNN, PPC on different networks

Datasets	ImageNet-CNN	ImageNet-PPC	Places-CNN	Places-PPC	SPM [135]
Caltech-101	86.44%	**87.45%**	61.07%	67.41%	64.6%
Scene15	84.49%	86.4%	89.11%	**89.88%**	81.4%
Indoor67	59.18%	64.4%	72.16%	**73.36%**	-
Holidays	73.95%	**74.9%**	71.71%	73.43%	-

From the table, we can see that the improvements of PPC over CNN vary from about 1 percent to 6 percent, depending on the network and dataset. We can further conclude that: 1) the features generated from CNN are more distinctive than SIFT in image classification, and this inherent merit brings about the improvement of PPC in contrast to SPM. 2) The features derived from ImageNet-CNN are more discriminative in classifying objects, thus perform better on the Caltech-101 dataset. In contrast, the features achieved from the Places-CNN are better at classifying scenes, and accordingly perform better on Scene15 and MIT Indoor67 datasets. On one hand, choosing a suitable network (i.e. choose ImageNet-CNN for object recognition, or choose Places-CNN for scene recognition) could bring a significant improvement in the performance. As is shown by the experiment on Caltech-101 database, the advantage of ImageNet-CNN feature over the Places-CNN feature is more than 25 percent (86.44% and 61.07% respectively). On the other hand, choosing an unsuitable network could highlight the benefit of PPC over CNN. For instance, when we utilize Places-CNN features on the Caltech-101 database, the improvement of PPC over CNN is more than 6 percent, rising from 61.07% to 67.41%. Similarly, when the ImageNet-CNN features are tested on the Indoor67 dataset, the refinement of PPC over CNN could also be more than 5 percent (from 59.18% to 64.4%). But no matter which type of networks is applied

on the datasets, PPC features consistently outperform the holistic CNN features, demonstrating the effectiveness of the strategy.

For both the CNN and PPC algorithms on the MIT Indoor67 dataset, we visualize the distance between the features of the top performing categories in 3-dimensional space using the classic multidimensional scaling technique [225]. As is shown in Figure 3.2, the axes correspond to the coordinates in the 3-dimensional space and the categories are buffet, cloister, florist, inside bus.

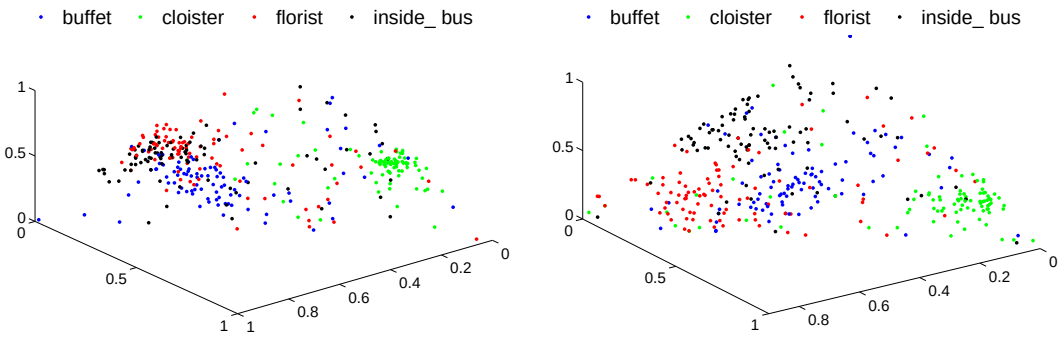

Figure 3.2: Top performing feature visualization of CNN(left) and PPC(right).

From the Figure 3.2, we can notice that the PPC features are more distinguishable than the holistic CNN features. The advantage is particularly evident on the comparisons between 'florist' and 'inside bus'.

For ImageNet-CNN model, we compare the accuracy of CNN and PPC on each category of Scene15 database, as is demonstrated in Figure 3.3. The x-axis details the categories and the y-axis corresponds to the accuracies of this category.

It can be observed that for most categories of Scene15 (ten of the fifteen categories), PPC performs better than CNN.

3.3.3 Comparisons on different dimensions

The improvement of PPC over CNN is not limited to 4096-D. To verify this, we further reduce the dimensionality to other sizes and compare the performance of

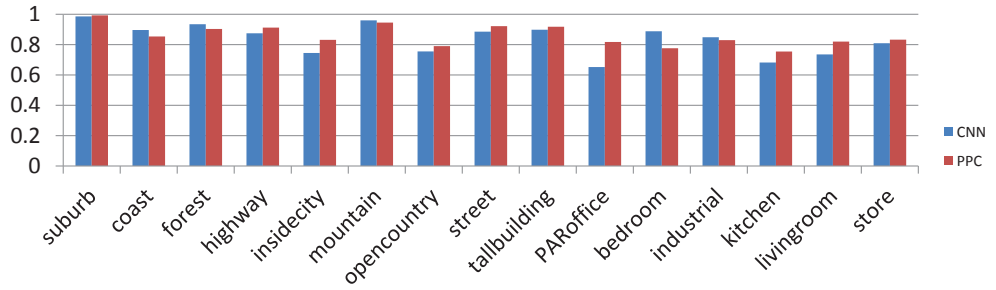

Figure 3.3: Accuracy of CNN and PPC on each category of Scene15.

PPC and CNN on different datasets, the results are shown in Figure 3.4.

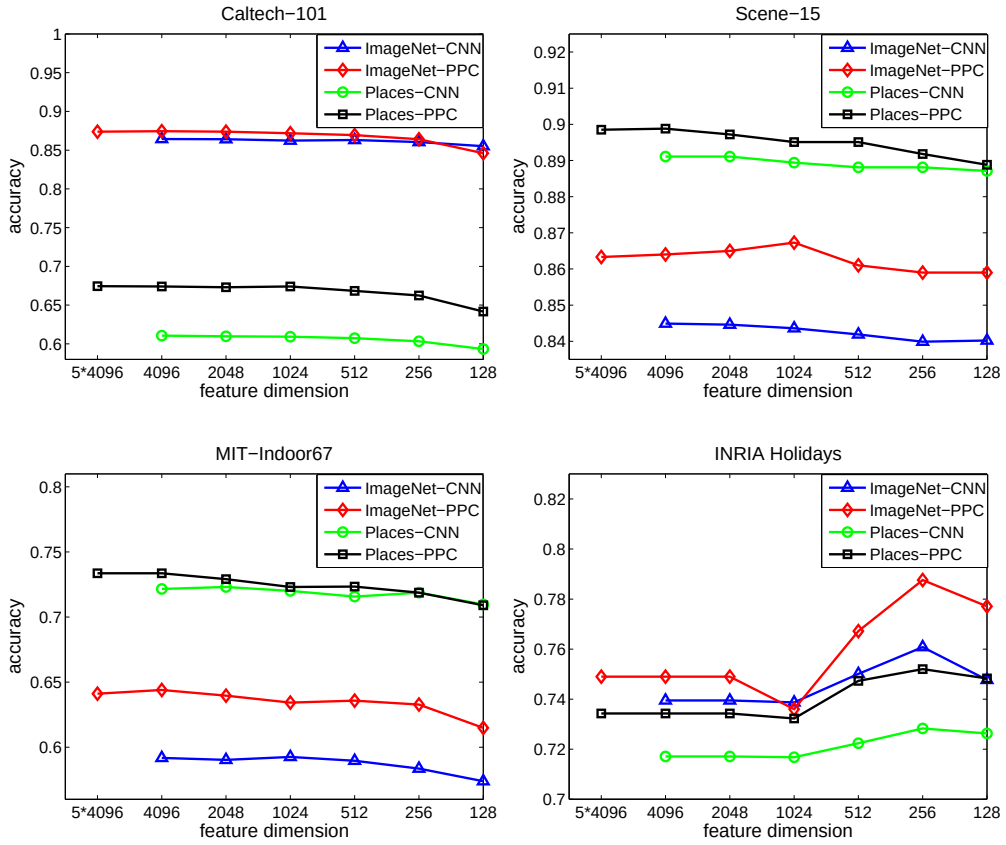

Figure 3.4: Comparisons of CNN and PPC on different dimensions

It is noticeable that the performance does not decrease even when the dimensionality of features are reduced to 128-D (most of the accuracies drift within

2 percent). On the contrary, the mAP of ImageNet-PPC on the INRIA Holidays dataset even rises to 78.76%, when the dimensionality is reduced to 256-D. This indicates that the discriminatory power of both CNN and PPC will not be greatly affected with the reduction of the dimensionality. Nevertheless, the performance of PPC is mostly better than that of CNN, indicating that PPC is more robustness than CNN.

3.4 Conclusion

CNN features have shown great promise in visual recognition. This chapter proposed the Principal Pyramidal Convolution (PPC) scheme, which aggregates the CNN features of the whole image as well as the sub-regions and then extracts the principal components. The representation from our strategy outperforms the conventional CNN feature without enlarging the feature dimensions. Furthermore, this work makes comparisons of CNN and PPC on different sizes and shows that the PPC frequently outperforms CNN.