Multi-Scale Orderless Pooling of Deep Convolutional Activation Features

Yunchao Gong[1], Liwei Wang[2], Ruiqi Guo[2], and Svetlana Lazebnik[2]

[1]University of North Carolina at Chapel Hill
yunchao@cs.unc.edu
[2]University of Illinois at Urbana-Champaign
{lwang97,guo29,slazebni}@illinois.edu

Abstract. Deep convolutional neural networks (CNN) have shown their promise as a universal representation for recognition. However, global CNN activations lack geometric invariance, which limits their robustness for classification and matching of highly variable scenes. To improve the invariance of CNN activations without degrading their discriminative power, this paper presents a simple but effective scheme called *multi-scale orderless pooling* (MOP-CNN). This scheme extracts CNN activations for local patches at multiple scale levels, performs orderless VLAD pooling of these activations at each level separately, and concatenates the result. The resulting MOP-CNN representation can be used as a generic feature for either supervised or unsupervised recognition tasks, from image classification to instance-level retrieval; it consistently outperforms global CNN activations without requiring any joint training of prediction layers for a particular target dataset. In absolute terms, it achieves state-of-the-art results on the challenging SUN397 and MIT Indoor Scenes classification datasets, and competitive results on ILSVRC2012/2013 classification and INRIA Holidays retrieval datasets.

1 Introduction

Recently, deep convolutional neural networks (CNN) [1] have demonstrated breakthrough accuracies for image classification [2]. This has spurred a flurry of activity on further improving CNN architectures and training algorithms [3,4,5,6,7], as well as on using CNN features as a universal representation for recognition. A number of recent works [8,9,10,11,12] show that CNN features trained on sufficiently large and diverse datasets such as ImageNet [13] can be successfully transferred to other visual recognition tasks, e.g., scene classification and object localization, with a only limited amount of task-specific training data. Our work also relies on reusing CNN activations as off-the-shelf features for whole-image tasks like scene classification and retrieval. But, instead of simply computing the CNN activation vector over the entire image, we ask whether we can get improved performance by combining activations extracted at multiple *local* image windows. Inspired by previous work on spatial and feature space pooling of local descriptors [14,15,16], we propose a novel and simple pooling

scheme that significantly outperforms global CNN activations for both supervised tasks like image classification and unsupervised tasks like retrieval, even without any fine-tuning on the target datasets.

Image representation has been a driving motivation for research in computer vision for many years. For much of the past decade, orderless bag-of-features (BoF) methods [15,17,18,19,20] were considered to be the state of the art. Especially when built on top of locally invariant features like SIFT [21], BoF can be, to some extent, robust to image scaling, translation, occlusion, and so on. However, they do not encode global spatial information, motivating the incorporation of loose spatial information in the BoF vectors through spatial pyramid matching (SPM) [14]. Deep CNN, as exemplified by the system of Krizhevsky et al. [2], is a completely different architecture. Raw image pixels are first sent through five convolutional layers, each of which filters the feature maps and then max-pools the output within local neighborhoods. At this point, the representation still preserves a great deal of global spatial information. For example, as shown by Zeiler and Fergus [22], the activations from the fifth max-pooling layer can be reconstructed to form an image that looks similar to the original one. Though max-pooling within each feature map helps to improve invariance to small-scale deformations [23], invariance to larger-scale, more global deformations might be undermined by the preserved spatial information. After the filtering and max-pooling layers follow several fully connected layers, finally producing an activation of 4096 dimensions. While it becomes more difficult to reason about the invariance properties of the output of the fully connected layers, we will present an empirical analysis in Section 3 indicating that the final CNN representation is still fairly sensitive to global translation, rotation, and scaling. Even if one does not care about this lack of invariance for its own sake, we show that it directly translates into a loss of accuracy for classification tasks.

Intuitively, bags of features and deep CNN activations lie towards opposite ends of the "orderless" to "globally ordered" spectrum for visual representations. SPM [14] is based on realizing that BoF has insufficient spatial information for many recognition tasks and adding just enough such information. Inspired by this, we observe that CNN activations preserve too much spatial information, and study the question of whether we can build a more orderless representation on top of CNN activations to improve recognition performance. We present a simple but effective framework for doing this, which we refer to as *multi-scale orderless pooling* (MOP-CNN). The idea is summarized in Figure 1. Briefly, we begin by extracting deep activation features from local patches at multiple scales. Our coarsest scale is the whole image, so global spatial layout is still preserved, and our finer scales allow us to capture more local, fine-grained details of the image. Then we aggregate local patch responses at the finer scales via VLAD encoding [16]. The orderless nature of VLAD helps to build a more invariant representation. Finally, we concatenatenate the original global deep activations with the VLAD features for the finer scales to form our new image representation.

Section 2 will introduce our multi-scale orderless pooling approach. Section 3 will present a small-scale study suggesting that CNN activations extracted

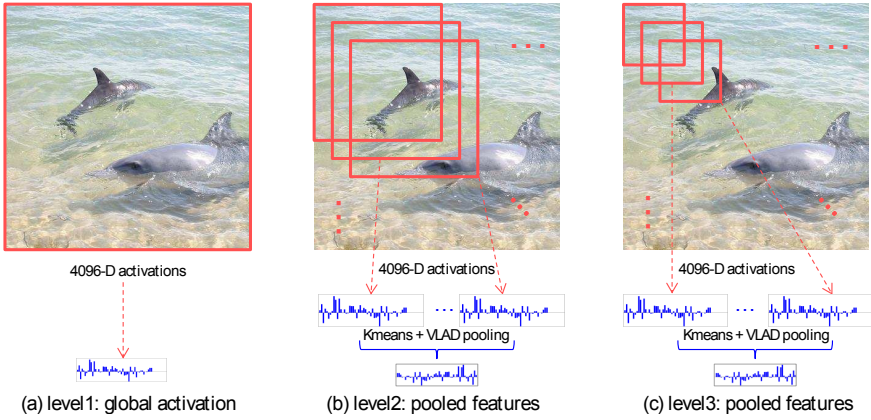


Fig. 1. Overview of multi-scale orderless pooling for CNN activations (MOP-CNN). Our proposed feature is a concatenation of the feature vectors from three levels: (a) Level 1, corresponding to the 4096-dimensional CNN activation for the entire 256×256 image; (b) Level 2, formed by extracting activations from 128×128 patches and VLAD pooling them with a codebook of 100 centers; (c) Level 3, formed in the same way as level 2 but with 64×64 patches.

at sub-image windows can provide more robust and discriminative information than whole-image activations, and confirming that MOP-CNN is more robust in the presence of geometric deformations than global CNN. Next, Section 4 will report comprehensive experiments results for classification on three image datasets (SUN397, MIT Indoor Scenes, and ILSVRC2012/2013) and retrieval on the Holidays dataset. A sizable boost in performance across these popular benchmarks confirms the promise of our method. Section 5 will conclude with a discussion of future work directions.

2 The Proposed Method

Inspired by SPM [14], which extracts local patches at a single scale but then pools them over regions of increasing scale, ending with the whole image, we propose a kind of "reverse SPM" idea, where we extract patches at multiple scales, starting with the whole image, and then pool each scale without regard to spatial information. The basic idea is illustrated in Figure 1.

Our representation has three scale levels, corresponding to CNN activations of the global 256×256 image and 128×128 and 64×64 patches, respectively. To extract these activations, we use the Caffe CPU implementation [24] pre-trained on ImageNet [13]. Given an input image or a patch, we resample it to 256×256 pixels, subtract the mean of the pixel values, and feed the patch through the network. Then we take the 4096-dimensional output of the seventh (fully connected) layer, after the rectified linear unit (ReLU) transformation, so that all the values are non-negative (we have also tested the activations before ReLU and found worse performance).

For the first level, we simply take the 4096-dimensional CNN activation for the whole 256×256 image. For the remaining two levels, we extract activations for all 128×128 and 64×64 patches sampled with a stride of 32 pixels. Next, we need to pool the activations of these multiple patches to summarize the second and third levels by single feature vectors of reasonable dimensionality. For this, we adopt Vectors of Locally Aggregated Descriptors (VLAD) [16,25], which are a simplified version of Fisher Vectors (FV) [15]. At each level, we extract the 4096-dimensional activations for respective patches and, to make computation more efficient, use PCA to reduce them to 500 dimensions. We also learn a separate k-means codebook for each level with $k = 100$ centers. Given a collection of patches from an input image and a codebook of centers $\boldsymbol{c}_i,\ i = 1, \ldots, k$, the VLAD descriptor (soft assignment version from [25]) is constructed by assigning each patch $\boldsymbol{p}_j$ to its r nearest cluster centers $r\mathrm{NN}(\boldsymbol{p}_j)$ and aggregating the residuals of the patches minus the center:

$$\boldsymbol{x} = \left[\sum_{j:\, \boldsymbol{c}_1 \in r\mathrm{NN}(\boldsymbol{p}_j)} w_{j1}(\boldsymbol{p}_j - \boldsymbol{c}_1), \quad \ldots \quad , \sum_{j:\, \boldsymbol{c}_k \in r\mathrm{NN}(\boldsymbol{p}_j)} w_{jk}(\boldsymbol{p}_j - \boldsymbol{c}_k) \right],$$

where w_{jk} is the Gaussian kernel similarity between $\boldsymbol{p}_j$ and $\boldsymbol{c}_k$. For each patch, we additionally normalize its weights to its nearest r centers to have sum one. For the results reported in the paper, we use $r = 5$[1] and kernel standard deviation of 10. Following [16], we power- and L2-normalize the pooled vectors. However, the resulting vectors still have quite high dimensionality: given 500-dimensional patch activations $\boldsymbol{p}_j$ (after PCA) and 100 k-means centers, we end up with 50,000 dimensions. This is too high for many large-scale applications, so we further perform PCA on the pooled vectors and reduce them to 4096 dimensions. Note that applying PCA after the two stages (local patch activation and global pooled vector) is a standard practice in previous works [26,27]. Finally, given the original 4096-dimensional feature vector from level one and the two 4096-dimensional pooled PCA-reduced vectors from levels two and three, we rescale them to unit norm and concatenate them to form our final image representation.

3 Analysis of Invariance

We first examine the invariance properties of global CNN activations vs. MOP-CNN. As part of their paper on visualizing deep features, Zeiler and Fergus [22] analyze the transformation invariance of their model on five individual images by displaying the distance between the feature vectors of the original and transformed images, as well as the change in the probability of the correct label for the transformed version of the image (Figure 5 of [22]). These plots show very

[1] In the camera-ready version of the paper, we incorrectly reported using $r = 1$, which is equivalent to the hard assignment VLAD in [16]. However, we have experimented with different r and their accuracy on our datasets is within 1% of each other.

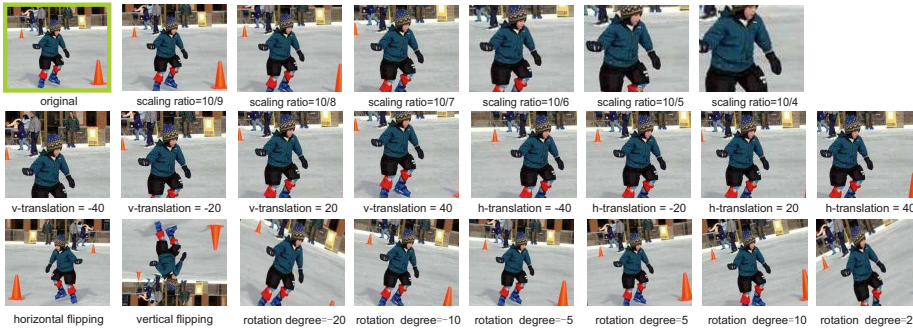


Fig. 2. Illustration of image transformations considered in our invariance study. For scaling by a factor of ρ, we take crops around the image center of $(1/\rho)$ times original size. For translation, we take crops of 0.7 times the original size and translate them by up to 40 pixels in either direction horizontally or vertically (the translation amount is relative to the normalized image size of 256×256). For rotation, we take crops from the middle of the image (so as to avoid corner artifacts) and rotate them from -20 to 20 degrees about the center. The corresponding scaling ratio, translation distance (pixels) and rotation degrees are listed below each instance.

different patterns for different images, making it difficult to draw general conclusions. We would like to conduct a more comprehensive analysis with an emphasis on prediction accuracy for entire categories, not just individual images. To this end, we train one-vs-all linear SVMs on the original training images for all 397 categories from the SUN dataset [28] using both global 4096-dimensional CNN activations and our proposed MOP-CNN features. At test time, we consider four possible transformations: translation, scaling, flipping and rotation (see Figure 2 for illustration and detailed explanation of transformation parameters). We apply a given transformation to all the test images, extract features from the transformed images, and perform 397-way classification using the trained SVMs. Figure 3 shows classification accuracies as a function of transformation type and parameters for four randomly selected classes: arrival gate, florist shop, volleyball court, and ice skating. In the case of CNN features, for almost all transformations, as the degree of transformation becomes more extreme, the classification accuracies keep dropping for all classes. The only exception is horizontal flipping, which does not seem to affect the classification accuracy. This may be due to the fact that the Caffe implementation adds horizontal flips of all training images to the training set (on the other hand, the Caffe training protocol also involves taking random crops of training images, yet this does not seem sufficient for building in invariance to such transformations, as our results indicate). By contrast with global CNN, our MOP-CNN features are more robust to the degree of translation, rotation, and scaling, and their absolute classification accuracies are consistently higher as well.

Figure 4 further illustrates the lack of robustness of global CNN activations by showing the predictions for a few ILSVRC2012/2013 images based on different image sub-windows. Even for sub-windows that are small translations of each other, the predicted labels can be drastically different. For example, in (f),

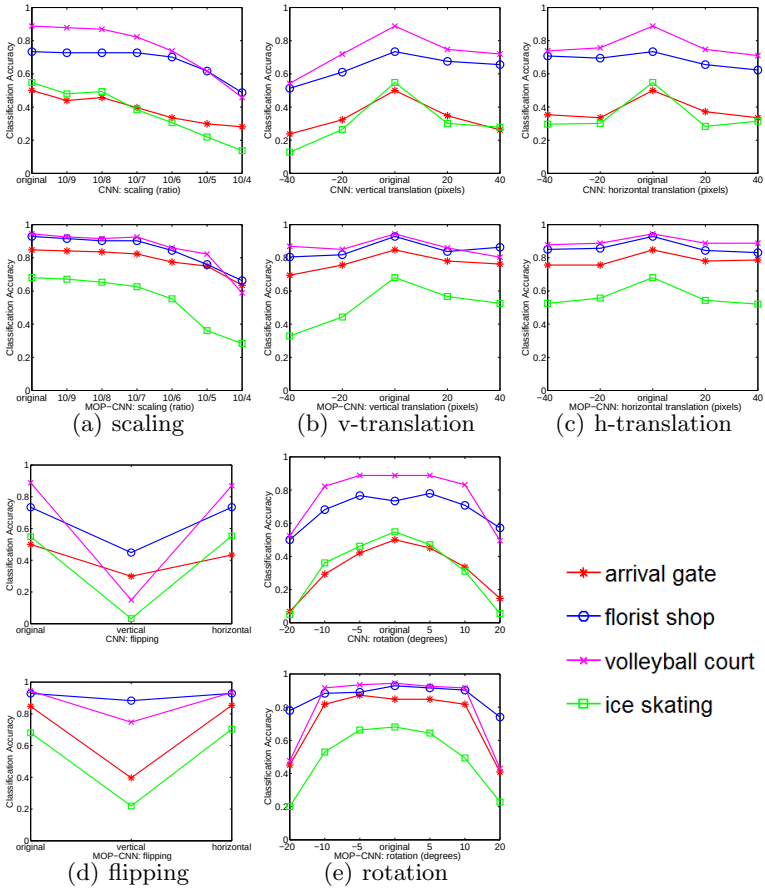
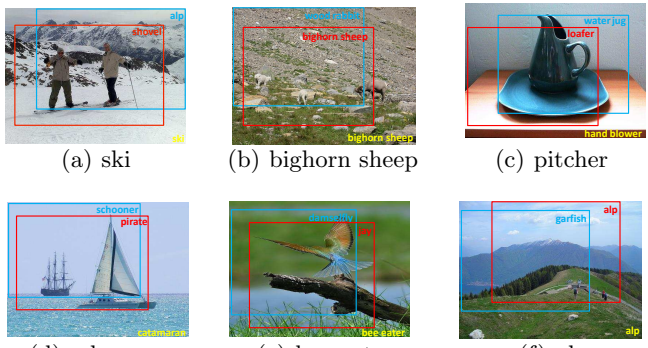


Fig. 3. Accuracies for 397-way classification on four classes from the SUN dataset as a function of different transformations of the test images. For each transformation type (a-e), the upper (resp. lower) plot shows the classification accuracy using the global CNN representation (resp. MOP-CNN).

the red rectangle is correctly labeled "alp," while the overlapping rectangle is incorrectly labeled "garfish." But, while picking the wrong window can give a bad prediction, picking the "right" one can give a good prediction: in (d), the whole image is wrongly labeled, but one of its sub-windows can get the correct label – "schooner." This immediately suggests a sliding window protocol at test time: given a test image, extract windows at multiple scales and locations, compute their CNN activations and prediction scores, and look for the window that gives the maximum score for a given class. Figure 5 illustrates such a "scene detection" approach [29,28] on a few SUN images. In fact, it is already common for CNN implementations to sample multiple windows at test time: the systems of [2,8,24] can take five sub-image windows corresponding to the center and four corners, together with their flipped versions, and average the prediction scores over these ten windows. As will be shown in Table 4, for Caffe, this "center+corner+flip"



(a) ski (b) bighorn sheep (c) pitcher

(d) schooner (e) bee eater (f) alp

Fig. 4. Classification of CNN activations of local patches in an image. The ground truth labels are listed below each image. Labels predicted by whole-image CNN are listed in the bottom right corner.

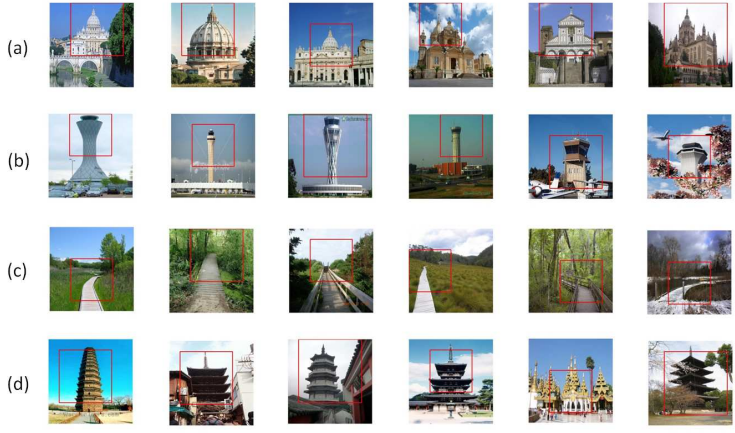


Fig. 5. Highest-response windows (in red) for (a) basilica, (b) control tower, (c) boardwalk, and (d) tower. For each test image resampled to 256×256, we search over windows with widths 224, 192, 160, and 128 and a stride of 16 pixels and display the window that gives the highest prediction score for the ground truth category. The detected windows contain similar structures: in (a), (b) and (d), the top parts of towers have been selected; in (c), the windows are all centered on the narrow walkway.

strategy gets 56.30% classification accuracy on ILSVRC2012/2013 vs. 54.34% for simply classifying global image windows. An even more recent system, OverFeat [12], incorporates a more comprehensive multi-scale voting scheme for classification, where efficient computations are used to extract class-level activations at a denser sampling of locations and scales, and the average or maximum of these activations is taken to produce the final classification results. With this scheme, OverFeat can achieve as high as 64.26% accuracy on ILSVRC2012/2013, albeit starting from a better baseline CNN with 60.72% accuracy.

While the above window sampling schemes do improve the robustness of prediction over single global CNN activations, they all combine activations (classifier

Table 1. A summary of baselines and their relationship to the MOP-CNN method.

pooling method / scale	multi-scale	concatenation
Average pooling	Avg (multi-scale)	Avg (concatenation)
Max pooling	Max (multi-scale)	Max (concatenation)
VLAD pooling	VLAD (multi-scale)	**MOP-CNN**

responses) from the final prediction layer, which means that they can only be used following training (or fine-tuning) for a particular prediction task, and do not naturally produce feature vectors for other datasets or tasks. By contrast, MOP-CNN combines activations of the last fully connected layer, so it is a more generic representation that can even work for tasks like image retrieval, which may be done in an unsupervised fashion and for which labeled training data may not be available.

4 Large-Scale Evaluation

4.1 Baselines

To validate MOP-CNN, we need to demonstrate that a simpler patch sampling and pooling scheme cannot achieve the same performance. As simpler alternatives to VLAD pooling, we consider **average pooling**, which involves computing the mean of the 4096-dimensional activations at each scale level, and **maximum pooling**, which involves computing their element-wise maximum. We did not consider standard BoF pooling because it has been demonstrated to be less accurate than VLAD [16]; to get competitive performance, we would need a codebook size much larger than 100, which would make the quantization step prohibitively expensive. As additional baselines, we need to examine alternative strategies with regards to pooling across scale levels. The **multi-scale** strategy corresponds to taking the union of all the patches from an image, regardless of scale, and pooling them together. The **concatenation** strategy refers to pooling patches from three levels separately and then concatenating the result. Finally, we separately examine the performance of individual scale levels as well as concatenations of just pairs of them. In particular, **level1** is simply the 4096-dimensional global descriptor of the entire image, which was suggested in [8] as a generic image descriptor. These baselines and their relationship to our full MOP-CNN scheme are summarized in Table 1.

4.2 Datasets

We test our approach on four well-known benchmark datasets:

SUN397 [28] is the largest dataset to date for scene recognition. It contains 397 scene categories and each has at least 100 images. The evaluation protocol involves training and testing on ten different splits and reporting the average classification accuracy. The splits are fixed and publicly available from [28]; each has 50 training and 50 test images.

MIT Indoor [30] contains 67 categories. While outdoor scenes, which comprise more than half of SUN (220 out of 397), can often be characterized by global scene statistics, indoor scenes tend to be much more variable in terms of composition and better characterized by the objects they contain. This makes the MIT Indoor dataset an interesting test case for our representation, which is designed to focus more on appearance of sub-image windows and have more invariance to global transformations. The standard training/test split for the Indoor dataset consists of 80 training and 20 test images per class.

ILSVRC2012/2013 [31,32], or ImageNet Large-Scale Visual Recognition Challenge, is the most prominent benchmark for comparing large-scale image classification methods and is the dataset on which the Caffe representation we use [24] is pre-trained. ILSVRC differs from the previous two datasets in that most of its categories focus on objects, not scenes, and the objects tend to be highly salient and centered in images. It contains 1000 classes corresponding to leaf nodes in ImageNet. Each class has more than 1000 unique training images, and there is a separate validation set with 50,000 images. The 2012 and 2013 versions of the ILSVRC competition have the same training and validation data. Classification accuracy on the validation set is used to evaluate different methods.

INRIA Holidays [33] is a standard benchmark for image retrieval. It contains 1491 images corresponding to 500 image instances. Each instance has 2-3 images describing the same object or location. A set of 500 images are used as queries, and the rest are used as the database. Mean average precision (mAP) is the evaluation metric.

4.3 Image Classification Results

In all of the following experiments, we train classifiers using the linear SVM implementation from the INRIA JSGD package [34]. We fix the regularization parameter to 10^{-5} and the learning rate to 0.2, and train for 100 epochs.

Table 2 reports our results on the SUN397 dataset. From the results for baseline pooling methods in (a), we can see that VLAD works better than average and max pooling and that pooling scale levels separately works better than pooling them together (which is not altogether surprising, since the latter strategy raises the feature dimensionality by a factor of three). From (b), we can see that concatenating all three scale levels gives a significant improvement over any subset. For reference, Part (c) of Table 2 gives published state-of-the-art results from the literature. Xiao et al. [28], who have collected the SUN dataset, have also published a baseline accuracy of 38% using a combination of standard features like GIST, color histograms, and BoF. This baseline is slightly exceeded by the level1 method, i.e., global 4096-dimensional Caffe activations pre-trained on ImageNet. The Caffe accuracy of 39.57% is also comparable to the 40.94% with an analogous setup for DeCAF [8].[2] However, these numbers are still worse than

[2] DeCAF is an earlier implementation from the same research group and Caffe is its "little brother." The two implementations are similar, but Caffe is faster, includes support for both CPU and GPU, and is easier to modify.

Table 2. Scene recognition on SUN397. (a) Alternative pooling baselines (see Section 4.1 and Table 1); (b) Different combinations of scale levels – in particular, "level1" corresponds to the global CNN representation and "level1+level2+level3" corresponds to the proposed MOP-CNN method. (c) Published numbers for state-of-the-art methods.

	method	feature dimension	accuracy
(a)	Avg (Multi-Scale)	4,096	39.62
	Avg (Concatenation)	12,288	47.50
	Max (Multi-Scale)	4,096	43.51
	Max (Concatenation)	12,288	48.50
	VLAD (Multi-Scale)	4,096	47.32
(b)	level1	4,096	39.57
	level2	4,096	45.34
	level3	4,096	40.21
	level1 + level2	8,192	49.91
	level1 + level3	8,192	49.52
	level2 + level3	8,192	49.66
	level1 + level2 + level3 (MOP-CNN)	12,288	**51.98**
(c)	Xiao et al. [28]	–	38.00
	DeCAF [8]	4,096	40.94
	FV (SIFT + Local Color Statistic) [35]	256,000	47.20

the 47.2% achieved by high-dimensional Fisher Vectors [35] – to our knowledge, the state of the art on this dataset to date. With our MOP-CNN pooling scheme, we are able to achieve 51.98% accuracy with feature dimensionality that is an order of magnitude lower than that of [35]. Figure 6 shows six classes on which MOP-CNN gives the biggest improvement over level1, and six on which it has the biggest drop. For classes having an object in the center, MOP-CNN usually cannot improve too much, or might hurt performance. However, for classes that have high spatial variability, or do not have a clear focal object, it can give a substantial improvement.

Table 3 reports results on the MIT Indoor dataset. Overall, the trends are consistent with those on SUN, in that VLAD pooling outperforms average and max pooling and combining all three levels yields the best performance. There is one interesting difference from Table 2, though: namely, level2 and level3 features work much better than level1 on the Indoor dataset, whereas the difference was much less pronounced on SUN. This is probably because indoor scenes are better described by local patches that have highly distinctive appearance but can vary greatly in terms of location. In fact, several recent methods achieving state-of-the-art results on this dataset are based on the idea of finding such patches [38,37,36]. Our MOP-CNN scheme outperforms all of them – 68.88% vs. 64.03% for the method of Doersch et al. [38].

Table 4 reports results on ILSVRC2012/2013. The trends for alternative pooling methods in (a) are the same as before. Interestingly, in (b) we can see that, unlike on SUN and MIT Indoor, level2 and level3 features do not work as well as level1. This is likely because the level1 feature was specifically trained on ILSVRC, and this dataset has limited geometric variability. Nevertheless, by

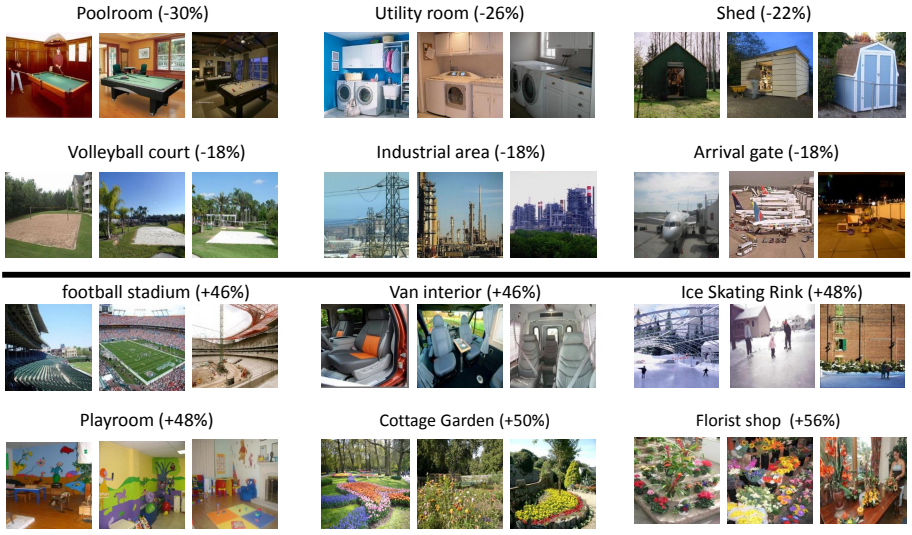


Fig. 6. SUN classes on which MOP-CNN gives the biggest decrease over the level1 global features (top), and classes on which it gives the biggest increase (bottom).

combining the three levels, we still get a significant improvement. Note that directly running the full pre-trained Caffe network on the global features from the validation set gives 54.34% accuracy (part (c) of Table 4, first line), which is higher than our level1 accuracy of 51.46%. The only difference between these two setups, "Caffe (Global)" and "level1," are the parameters of the last classifier layer – i.e., softmax and SVM, respectively. For Caffe, the softmax layer is jointly trained with all the previous network layers using multiple random windows cropped from training images, while our SVMs are trained separately using only the global image features. Nevertheless, the accuracy of our final MOP-CNN representation, at 57.93%, is higher than that of the full pre-trained Caffe CNN tested either on the global features ("Global") or on ten sub-windows ("Center+Corner+Flip").

It is important to note that in absolute terms, we do not achieve state-of-the-art results on ILSVRC. For the 2012 version of the contest, the highest results were achieved by Krizhevsky et al. [2], who have reported a top-1 classification accuracy of 59.93%. Subsequently, Zeiler and Fergus [22] have obtained 64% by refining the Krizhevsky architecture and combining six different models. For the 2013 competition, the highest reported top-1 accuracies are those of Sermanet et al. [12]: they obtained 64.26% by aggregating CNN predictions over multiple sub-window locations and scales (as discussed in Section 3), and 66.04% by combining seven such models. While our numbers are clearly lower, it is mainly because our representation is built on Caffe, whose baseline accuracy is below that of [2,22,12]. We believe that MOP-CNN can obtain much better performance when combined with these better CNN models, or by combining multiple independently trained CNNs as in [22,12].

Table 3. Classification results on MIT Indoor Scenes. (a) Alternative pooling baselines (see Section 4.1 and Table 1); (b) Different combinations of scale levels; (c) Published numbers for state-of-the-art methods.

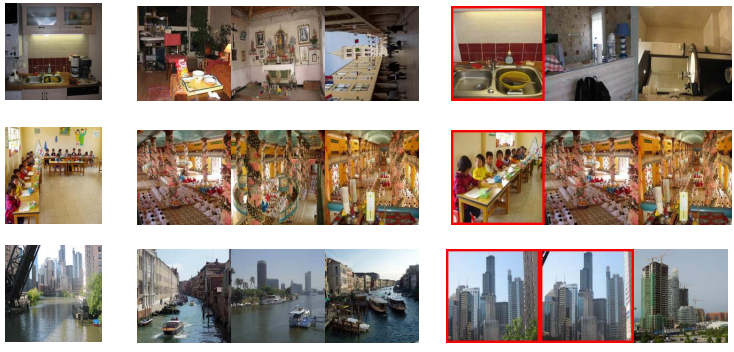
	method	feature dimension	accuracy
(a)	Avg (Multi-Scale)	4,096	56.72
	Avg (Concatenation)	12,288	65.60
	Max (Multi-Scale)	4,096	60.52
	Max (Concatenation)	12,288	64.85
	VLAD (Multi-Scale)	4,096	66.12
(b)	level1	4,096	53.73
	level2	4,096	65.52
	level3	4,096	62.24
	level1 + level2	8,192	66.64
	level1 + level3	8,192	66.87
	level2 + level3	8,192	67.24
	level1 + level2 + level3 (MOP-CNN)	12,288	**68.88**
(c)	SPM [14]	5,000	34.40
	Discriminative patches [36]	–	38.10
	Disc. patches+GIST+DPM+SPM [36]	–	49.40
	FV + Bag of parts [37]	221,550	63.18
	Mid-level elements [38]	60,000	64.03

4.4 Image Retrieval Results

As our last experiment, we demonstrate the usefulness of our approach for an *unsupervised* image retrieval scenario on the Holidays dataset. Table 5 reports the mAP results for nearest neighbor retrieval of feature vectors using the Euclidean distance. On this dataset, level1 is the weakest of all three levels because images of the same instance may be related by large rotations, viewpoint changes, etc., and global CNN activations do not have strong enough invariance to handle these transformations. As before, combining all three levels achieves the best performance of 78.82%. Using aggressive dimensionality reduction with PCA and whitening as suggested in [39], we can raise the mAP even further to 80.8% with only a 2048-dimensional feature vector. The state of the art performance on this dataset with a compact descriptor is obtained by Gordo et al. [40] by using FV/VLAD and discriminative dimensionality reduction, while our method still achieves comparable or better performance. Note that it is possible to obtain even higher results on Holidays with methods based on inverted files with very large vocabularies. In particular, Tolias et al. [41] report 88% but their representation would take more than 4 million dimensions per image if expanded into an explicit feature vector, and is not scalable to large datasets. Yet further improvements may be possible by adding techniques such as query expansion and geometric verification, but they are not applicable for generic image representation, which is our main focus. Finally, we show retrieval examples in Figure 7. We can clearly see that MOP-CNN has improved robustness to shifts, scaling, and viewpoint changes over global CNN activations.

Table 4. Classification results on ILSVRC2012/2013. (a) Alternative pooling baselines (see Section 4.1 and Table 1); (b) Different combinations of scale levels; (c) Numbers for state-of-the-art CNN implementations. All the numbers come from the respective papers, except the Caffe numbers, which were obtained by us by directly testing their full network pre-trained on ImageNet. "Global" corresponds to testing on global image features, and "Center+Corner+Flip" corresponds to averaging the prediction scores over ten crops taken from the test image (see Section 3 for details).

	method	feature dimension	accuracy
(c)	Avg (Multi-Scale)	4096	53.34
	Avg (Concatenation)	12,288	56.12
	Max (Multi-Scale)	4096	54.37
	Max (Concatenation)	12,288	55.88
	VLAD (Multi-Scale)	4,096	48.54
(b)	level1	4,096	51.46
	level2	4,096	48.21
	level3	4,096	38.20
	level1 + level2	8,192	56.82
	level1 + level3	8,192	55.91
	level2 + level3	8,192	51.52
	level1 + level2 + level3 (MOP-CNN)	12,288	57.93
(c)	Caffe (Global) [24]	–	54.34
	Caffe (Center+Corner+Flip) [24]	–	56.30
	Krizhevsky et al. [2]	–	59.93
	Zeiler and Fergus (6 CNN models) [22]	–	64.00
	OverFeat (1 CNN model) [12]	–	**64.26**
	OverFeat (7 CNN models) [12]	–	**66.04**



(a) Query (b) level1 (c) MOP-CNN

Fig. 7. Image retrieval examples on the Holiday dataset. Red border indicates a ground truth image (i.e., a positive retrieval result). We only show three retrieved examples per query because each query only has one to two ground truth images.

5 Discussion

This paper has presented a multi-scale orderless pooling scheme that is built on top of deep activation features of local image patches. On four very challenging

Table 5. Image retrieval results on the Holidays dataset. (a) Alternative pooling baselines (see Section 4.1 and Table 1); (b) Different combinations of scale levels; (c) Full MOP-CNN descriptor vector compressed by PCA and followed by whitening [39], for two different output dimensionalities; (c) Published state-of-the-art results with a compact global descriptor (see text for discussion).

	method	feature dimension	mAP
(a)	Avg (Multi-Scale)	4,096	71.32
	Avg (Concatenation)	12,288	75.02
	Max (Multi-Scale)	4,096	76.23
	Max (Concatenation)	12,288	75.07
	VLAD (Multi-Scale)	4,096	78.42
(b)	level1	4,096	70.53
	level2	4,096	74.02
	level3	4,096	75.45
	level1 + level2	8,192	75.86
	level1 + level3	8,192	78.92
	level2 + level3	8,192	77.91
	level1 + level2 + level3 (MOP-CNN)	12,288	78.82
(c)	MOP-CNN + PCA + Whitening	512	78.38
	MOP-CNN + PCA + Whitening	2048	**80.18**
(d)	FV [16]	8,192	62.50
	FV + PCA [16]	256	62.60
	Gordo et al. [40]	512	78.90

datasets, we have achieved a substantial improvement over global CNN activations, in some cases outperforming the state of the art. These results are achieved with the same set of parameters (i.e., patch sizes and sampling, codebook size, PCA dimension, etc.), which clearly shows the good generalization ability of the proposed approach. As a generic low-dimensional image representation, it is not restricted to supervised tasks like image classification, but can also be used for unsupervised tasks such as retrieval.

Our work opens several promising avenues for future research. First, it remains interesting to investigate more sophisticated ways to incorporate orderless information in CNN. One possible way is to change the architecture of current deep networks fundamentally to improve their holistic invariance. Second, the feature extraction stage of our current pipeline is somewhat slow, and it is interesting to exploit the convolutional network structure to speed it up. Fortunately, there is fast ongoing progress in optimizing this step. One example is the multi-scale scheme of Sermanet et al. [12] mentioned earlier, and another is DenseNet [42]. In the future, we would like to reimplement MOP-CNN to benefit from such architectures.

Acknowledgments. Lazebnik's research was partially supported by NSF grants 1228082 and 1302438, the DARPA Computer Science Study Group, Xerox UAC, Microsoft Research, and the Sloan Foundation. Gong was supported by the 2013 Google Ph.D. Fellowship in Machine Perception.

References

1. LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., Jackel, L.: Handwritten digit recognition with a back-propagation network. In: NIPS. (1990)
2. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems 25. (2012) 1106–1114
3. Goodfellow, I., Warde-Farley, D., Mirza, M., Courville, A., Bengio, Y.: Maxout networks. In: ICML. (2013)
4. Le, Q., Ranzato, M., Monga, R., Devin, M., Chen, K., Corrado, G., Dean, J., Ng, A.: Building high-level features using large scale unsupervised learning. In: ICML. (2012)
5. Wan, L., Zeiler, M., Zhang, S., Lecun, Y., Fergus, R.: Regularization of neural networks using DropConnect. In: ICML. (2013)
6. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. Arxiv preprint arXiv:1207.0580 (2012)
7. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep fisher networks for large-scale image classification. In: Proceedings Advances in Neural Information Processing Systems (NIPS). (2013)
8. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. arXiv preprint arXiv:1310.1531 (2013)
9. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv preprint arXiv:1311.2524 (2013)
10. Oquab, M., Bottou, L., Laptev, I., Sivic, J., et al.: Learning and transferring mid-level image representations using convolutional neural networks. In: CVPR. (2014)
11. Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: An astounding baseline for recognition. In: CVPR 2014 DeepVision Workshop. (2014)
12. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229 (2013)
13. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR. (2009)
14. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR. (2006)
15. Perronnin, F., Dance, C.R.: Fisher kernels on visual vocabularies for image categorization. In: CVPR. (2007)
16. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: CVPR. (2010) 3304–3311
17. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. CVPR (2010)
18. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: ECCV Workshop on Statistical Learning in Computer Vision. (2004)

19. Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: ICCV. (2003)
20. Grauman, K., Darrell, T.: The pyramid match kernel: Discriminative classification with sets of image features. In: In ICCV. (2005) 1458–1465
21. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **60**(2) (2004) 91–110
22. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional neural networks. arXiv preprint arXiv:1311.2901 (2013)
23. Lee, H., Grosse, R., Ranganath, R., Ng, A.Y.: Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: ICML. (2009) 609–616
24. Jia, Y.: Caffe: An open source convolutional architecture for fast feature embedding. http://caffe.berkeleyvision.org/ (2013)
25. Bergamo, A., Sinha, S.N., Torresani, L.: Leveraging structure from motion to learn discriminative codebooks for scalable landmark classification. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition. CVPR '13 (2013)
26. Perronnin, F., Sanchez, J., Mensink, T.: Improving the Fisher kernel for large-scale image classification. In: ECCV. (2010)
27. Perronnin, F., Liu, Y., Sánchez, J., Poirier, H.: Large-scale image retrieval with compressed Fisher vectors. In: CVPR. (2010)
28. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: SUN database: Large-scale scene recognition from abbey to zoo. In: CVPR. (2010) 3485–3492
29. Pandey, M., Lazebnik, S.: Scene recognition and weakly supervised object localization with deformable part-based models. In: ICCV. (2011) 1307–1314
30. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: CVPR. (2009)
31. Deng, J., Berg, A., Satheesh, S., Su, H., Khosla, A., Fei-Fei, L.: Large scale visual recognition challenge. http://www.image-net.org/challenges/LSVRC/2012/ (2012)
32. Russakovsky, O., Deng, J., Huang, Z., Berg, A., Fei-Fei, L.: Detecting avocados to zucchinis: what have we done, and where are we going? In: ICCV. (2013)
33. Jégou, H., Douze, M., Schmid, C.: Hamming embedding and weak geometric consistency for large-scale image search. In: ECCV. (2008)
34. Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C., et al.: Good practice in large-scale learning for image classification. PAMI (2013)
35. Sanchez, J., Perronnin, F., Mensink, T., Verbeek, J.: Image Classification with the Fisher Vector: Theory and Practice. IJCV **105**(3) (2013) 222–245
36. Singh, S., Gupta, A., Efros, A.A.: Unsupervised discovery of mid-level discriminative patches. In: ECCV. (2012)
37. Juneja, M., Vedaldi, A., Jawahar, C.V., Zisserman, A.: Blocks that shout: Distinctive parts for scene classification. In: CVPR. (2013)
38. Doersch, C., Gupta, A., Efros, A.A.: Mid-level visual element discovery as discriminative mode seeking. In: NIPS. (2013)
39. Jégou, H., Chum, O.: Negative evidences and co-occurrences in image retrieval: the benefit of PCA and whitening. In: ECCV. (2012)
40. Gordo, A., Rodrıguez-Serrano, J.A., Perronnin, F., Valveny, E.: Leveraging category-level labels for instance-level image retrieval. In: CVPR. (2012)
41. Tolias, G., Avrithis, Y., Jégou, H.: To aggregate or not to aggregate: selective match kernels for image search. In: ICCV. (2013)

42. Iandola, F., Moskewicz, M., Karayev, S., Girshick, R., Darrell, T., Keutzer, K.: DenseNet: Implementing efficient convnet descriptor pyramids. arXiv preprint arXiv:1404.1869 (2014)