Robust Scene Classification with Cross-level LLC Coding on CNN Features

Zequn Jie[1], Shuicheng Yan[2]

[1] Keio-NUS CUTE Center, National University of Singapore, Singapore
[2] Department of Electrical and Computer Engineering, National University of Singapore, Singapore

Abstract. Convolutional Neural Network (CNN) features have demonstrated outstanding performance as global representations for image classification, but they lack invariance to scale transformation, which makes it difficult to adapt to various complex tasks such as scene classification. To strengthen the scale invariance of CNN features and meanwhile retain their powerful discrimination in scene classification, we propose a framework where cross-level Locality-constrained Linear Coding and cascaded fine-tuned CNN features are combined, which is shorted as *cross-level LLC-CNN*. Specifically, this framework first fine-tunes multi-level CNNs in a cascaded way, then extracts multi-level CNN features to learn a cross-level universal codebook, and finally performs locality-constrained linear coding (LLC) and max-pooling on the patches of all levels to form the final representation. It is experimentally verified that the LLC responses on the universal codebook outperform the CNN features and achieve the state-of-the-art performance on the two currently largest scene classification benchmarks, MIT Indoor Scenes and SUN 397.

1 Introduction

Scene classification is a fundamental problem in computer vision. However, it is not an easy task due to the great diversity of image contents as well as the variations in illumination and scale conditions. Conventional approaches such as Bag-of-Features (BoF) model [1], Bag-of-Parts (BoP) [2], Object Bank [3], and their respective combinations with Spatial Pyramid Matching (SPM) [4], have achieved satisfactory performance in this task. These works [5, 6] utilize hand-crafted features, e.g., SIFT [7] and HOG [8], which require designing lots of tricks and lack image representation power for different complex problems.

Recently, in contrast to hand-crafted features, image features learned from Convolutional Neural Network (CNN) [9] have achieved great success in vision recognition tasks [10–13]. Among these works, one of the greatest breakthroughs is that CNN has achieved an accuracy which is 10% higher than all the hand-crafted feature based methods in ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [14] which contains over 1 million images from 1000 categories. Inspired by the outstanding performance of CNN in large-scale image classification, many works [15–18] consider how to transfer CNN features pre-trained on

ImageNet to small-scale computer vision tasks in which only a limited number of task-specific training samples are available. As generic global image representations, off-the-shelf CNN features pre-trained on ImageNet are successfully applied to various vision tasks, including object detection [18] and image retrieval [19]. Furthermore, to improve the adaptation and representation power of CNN features in specific tasks, the fine-tuned CNN features based on pre-trained ImageNet CNN features are also used and have achieved better performance in these transferred tasks [17, 16, 20].

Despite the great success of CNN in various vision tasks, as global image representations, CNN features retain too much global spatial information and lack invariance to scale transformation since raw pixels are filtered and pooled alternatively within their local neighborhoods in the network. Actually, as shown in [21], feature maps after each layer can be used to reconstruct the original image due to the high spatial order of CNN features. Although the max-pooling layer after each convolution layer provides a certain degree of invariance to local scale transformation, invariance to global scale transformation cannot be guaranteed. Based on the 4096-dimensional global CNN features, their variance to scale transformation will directly lead to the decrease of recognition accuracy when only scale transformed images are available for testing.

To improve the scale invariance of CNN features, a multi-level pooling frameworks has been proposed by [19]. Specifically, CNN features from patches with various sizes in different levels of the framework are extracted as mid-level image representations, followed by an intra-level pooling process over these patches. Within one level, densely distributed patch features cover the whole image and are pooled in an orderless way. By pooling the patch CNN features in each level, the final representation becomes patch-level orderless and scale invariant to a certain degree.

However, when the whole testing image is scaled, all the patches of its finer levels will be scaled by the same scaling ratio accordingly. In this case, CNN features of both the whole image and the patches of all levels will not work well since CNN features of each level are learned in a supervised manner from the training patches in the same level. To demonstrate this, we conduct an experiment on an image from SUN 397 [22] with the model trained on original training samples. Figure 1 shows the prediction of each patch in level 1 and level 2 of both the original image and its scaled version (10/6 ratio). As can be seen, both the whole image (level 1) and patches in level 2 obtain the correct predictions – "tent" by the fine-tuned CNN of their own level. In contrast, the scaled testing image obtains a wrong prediction – "mountain" using the fine-tuned CNN trained on the original non-scaled training images. A similar situation also happens in level 2, where 3 patches of the total 4 obtain wrong predictions. In this case, even if orderless pooling is performed on top of the CNN features of patches, no scale invariance can be guaranteed since the features to be pooled, i.e., CNN features of each patch have changed due to the scaling of the whole testing image.

In this paper, we present a simple but effective framework, which we refer to as cross-level LLC coding and cascaded fine-tuned CNN (*cross-level LLC-CNN*),

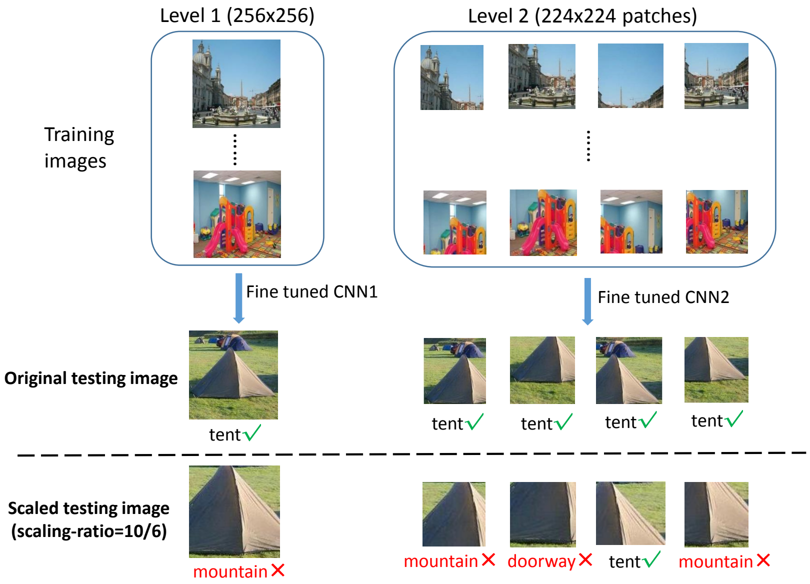

Fig. 1. Predictions of each patch in level 1 and level 2 of both the original image and its scaled version (10/6 ratio) with the CNN trained on original training samples. It is shown that predictions of the original testing image are all correct, while there are many wrong predictions for all levels of the scaled image.

to provide CNN features more robust to scale transformation. The pipeline is illustrated in Figure 2. Details will be presented in Section 3. Our proposed framework first fine-tunes CNNs for each level in a cascaded way, which means the CNN parameters learned in the coarser level are utilized as the initialization of the finer level. Subsequently, CNN features of all the patches in multiple levels are extracted by their own fine-tuned CNNs. Then we learn a universal (cross-level) codebook on all the CNN features of multi-level patches by k-means. Based on this universal codebook, Locality-constrained Linear Coding (LLC) [23] is performed for all the CNN features. The locality-constrained nature of LLC ensures each patch to find its most similar patches among all the patches distributed in multiple levels, even if the image and its patches are scaled. This helps build a more robust representation to scale transformation. Finally, all the LLC features of patches in multiple levels are max-pooled together to build the final image representation.

Extensive experiments on two challenging scene classification datasets, i.e., MIT indoor scenes [24] and SUN 397 [22], verify the superiority of the cross-level LLC coding on the cascaded fine-tuned CNN features over other conventional methods. The rest of the paper is organized as follows. First, we give a survey of typical methods for scene classification in Section 2. Then we elaborate on our framework, cross-level Locality-constrained Linear Coding (LLC) of CNN features in Section 3. After showing experimental results in Section 4, we draw a conclusion in Section 5.

2 Related Work

Scene classification as a fundamental and challenging vision task has attracted much attention and great progress has been achieved in the past decades. Generally, methods which have been proposed to deal with this task can be categorized into two types: Bag-of-Features (BoF) and deep learning.

Conventional methods mostly belong to the Bag-of-Features framework type. Early methods of this type adopted K-means Vector Quantization (VQ) to encode local features [25]. Later, Sparse Coding (SC) [26] was proposed to relax the cardinality constraint of VQ, which requires that only one coefficient of the code words is 1 while the rest are all 0. To add spatial organization information to the orderless Bag-of-Features, Spatial Pyramid Matching (SPM) [4] partitions the entire image into multi-scale patches and performs VQ or SC on each patch. Also, Orientational Pyramid Matching (OPM) [27] was used to partition the image in a more discriminative way, with the consideration of the orientation information. In this type of framework, local scale invariant hand-crafted features are usually relied on, such as SIFT [7] and HOG [8]. The combination between low-level scale invariant features and mid-level orderless pooling builds a more robust representation to scale transformation. The main limitation of this type of framework lies in the designing of hand-crafted features, which needs lots of tricks and is not applicable to some specific complex problems.

The other type of framework, i.e., deep learning, tries to model high-level abstractions of visual data by using architectures containing multiple layers of non-linear transformations. Convolutional Neural Network (CNN), as a typical example of deep learning models, has achieved great success in image classification, including ILSVRC 2012, ILSVRC 2013, tiny image dataset CIFAR-10/100 [28] and hand-written digits recognition [29]. [21] later proved that CNN features do not have invariance to different kinds of geometric transformations, e.g., scale transformation and rotation transformation. To strengthen the representation power of CNN when scale transformation occurs, [19] proposed a multi-scale orderless pooling framework, which includes CNN feature extraction at multiple levels and VLAD [30] pooling over these features. Our approach differs from this work in the different CNN features extracted and the cross-level feature coding and pooling schemes.

3 Cross-level LLC Coding on Multi-level CNN Features

3.1 Multi-level Cascaded Fine-tuned CNNs

To capture the context information of various sizes of patches, similar to [19], we adopt a multi-level framework to extract fine-tuned CNN features in multiple levels. The patch sizes of level 1 to level 5 are chosen carefully as follows: 256*256, 224*224, 192*192, 160*160, 128*128. Intuitively, transferring the groundtruth label of the whole image to its patches requires the patches not to be too small. The reason is that in scene classification, the groundtruth label is the high-level semantic abstract on the whole image, and too small local patches usually cannot be summarized as the same abstract concept (groundtruth label) as that of the whole image. Actually, we have found that the single patch recognition accuracy of level 5 with patch size 128*128 only achieves 43.6%, while the recognition accuracy of level 1 is 61.46% on the MIT indoor scenes dataset. Fortunately, although in level 5, the single patch recognition accuracy is much lower than that of the whole image, the recognition accuracy using max-pooled features of this level can still obtain a satisfactory result of 64.97%. Thus, we set the smallest patch size as 128*128. The stride of all the 5 levels is 32 pixels, thus we have 1, 4, 9, 16, 25 patches from level 1 to level 5 respectively.

To improve the discrimination and adaptation power of off-the-shelf CNN features on scene classification datasets, we fine-tune the CNN model pre-trained on ImageNet for each level in a cascaded way. We choose the same CNN architecture with [21] for its proven great performance in ILSVRC 2013. It contains five convolutional layers and three fully-connected layers with 60 million parameters. Since the numbers of categories in scene classification datasets differ from that in ImageNet, we change the number of the outputs of the last fully-connected layer, which represents the predicted probability of each target category, from 1000 in ImageNet to 67 and 397 in MIT indoor scenes and SUN 397 datasets respectively. Before fed into this CNN model, all the patches are resized to 256*256. During the stochastic gradient descent training process, the parameters of the first seven layers are initialized by the parameters pre-trained on ImageNet and the parameters of the last fully-connected layer are randomly initialized with a Gaussian distribution. The learning rates of the convolutional layers, the first two fully-connected layers and the last fully-connected layer are initialized as 0.001, 0.002 and 0.01, respectively and reduced to one tenth of the current rates after every 20 epochs (50 epochs in total). By setting the different learning rates for different layers, the parameters in different layers are updated by different rates. The main reasons for this setting are as follows: the first few convolution layers mainly extract low-level invariant features, such as texture and shape, thus the parameters are more consistent from the pre-trained dataset to the target dataset, whose learning rates are set as a relatively low value (i.e., 0.001); for the final few layers, especially the last fully-connected layer which is specifically adapted to the new target dataset, a higher learning rate is required to guarantee its fast convergence to the new optimum.

To strengthen the connections between the fine-tuned CNNs of different levels and reduce the convergence time, we adopt a cascaded fine-tuning strategy. Specifically, we use the model pre-trained on ImageNet as our initialization when training the CNN of level 1. When training on other finer levels, we always use the model trained on the last coarser level as our initialization. For example, the CNN trained on level 1 will be the initialization when training CNN on level 2. In Section 4, we will show the superiority of the cascaded fine-tuned CNN over off-the-shelf CNN and CNN fine-tuned with the pre-trained model on ImageNet in recognition accuracy.

3.2 Cross-level LLC Coding and Pooling on CNN features

Although separate fine-tuning of CNN for each level enhances the discrimination power of CNN features, it is still unstable for scale transformation, as fine-tuned CNNs are trained on the original non-scaled training images and patches, thus naturally characterize the image spatial organization of these non-scaled samples better.

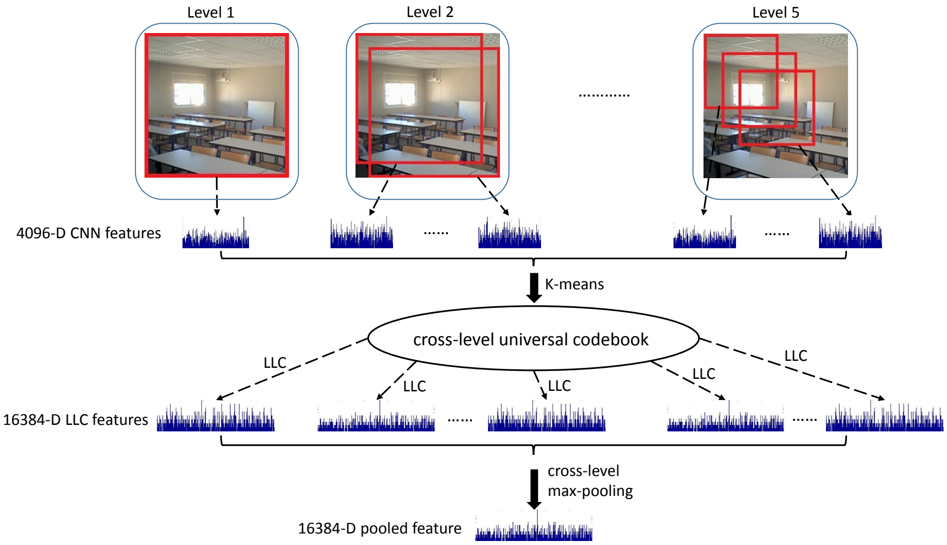

Fig. 2. Pipeline of cross-level LLC coding and max-pooling on CNN features. First, patch CNN features of all the 5 levels are clustered to learn a cross-level codebook. Next, all these CNN features are encoded based on this codebook via LLC coding. Finally, max-pooling is performed on all the encoded features to form a cross-level pooled feature, as the new image representation.

To solve this problem, we propose to use a cross-level feature coding and pooling scheme on the fine-tuned CNN features extracted from all patches of multiple levels. The pipeline is illustrated in Figure 2. Firstly, a 4096-dimensional feature is extracted in each patch with the fine-tuned CNN of their own level. Subsequently, a cross-level codebook is learned by clustering all these multi-level patch CNN features into 16384 clusters (4 times as the 4096 dimensions) with the k-means algorithm. By doing this, different patch levels of CNN features can be found among the code words of this cross-level codebook such that the codebook gains multi-level representation power. Next, Locality-constrained Linear Coding

(LLC) is performed on the multi-level CNN features based on the learned cross-level codebook. LLC coding enforces the corresponding encoding coefficients to be high if the code words are similar to the feature, and enforces the coefficients of other dissimilar code words to be zero [23]. The underlying hypothesis is that features approximately reside on a lower dimensional manifold in an ambient feature space [31]. Specifically, LLC coding uses the following criteria:

$$\begin{aligned} &\min_{C} \sum_{i=1}^{N} ||x_i - Bc_i||^2 + \lambda||d_i \odot c_i||^2 \\ &\text{s.t.} \;\; \mathbf{1}^T c_i = 1, \forall i. \end{aligned} \tag{1}$$

where N is the number of features to be encoded, x_i represents the ith encoded feature, B is the codebook matrix, c_i is the ith LLC coding result, $\odot$ denotes the element-wise multiplication, and $d_i \in R^M$ is the dissimilarity between the encoded feature and the code words with M denoting the codebook size. Specifically,

$$d_i = \exp\left(\frac{\text{dist}(x_i, B)}{\sigma}\right) \tag{2}$$

where $dist(x_i, B) = [dist(x_i, b_1), ..., dist(x_i, b_M)]^T$, and $dist(x_i, b_j)$ is the Euclidean distance between x_i and b_j. The analytical solution of LLC is as follows:

$$\begin{aligned} &\widetilde{c_i} = (C_i + \lambda \text{diag}(d)) \setminus \mathbf{1} \\ &c_i = \widetilde{c_i} \setminus \mathbf{1}^T \widetilde{c_i} \end{aligned} \tag{3}$$

where $C_i = (B - \mathbf{1}x_i^T)(B - \mathbf{1}x_i^T)^T$ denotes the data covariance matrix. Hence, LLC can be implemented very fast in practice.

By performing LLC on multi-level patch CNN features based on the cross-level codebook, different levels of CNN features extracted from patches of various sizes share a common codebook and can be encoded based on this codebook, regardless of their levels. This naturally enhances the scale invariance of the LLC features since no matter how the whole image and all of its patches are scaled, the CNN features can always find their similar code words in the cross-level codebook, either from the code words of their own levels or from other levels, and use these code words to represent them, leaving the reconstruction coefficients of all the rest dissimilar code words to be zero. As can be seen in Figure 3, CNN features of the original image will probably be represented by the code words of their own level, while CNN features of the scaled image may be similar to the code words from other levels and represented by these code words.

After obtaining LLC features of all the patches from multiple levels, we max-pool these cross-level features together in a mid-level (patch-level) orderless manner to form the final image representation. Finally, a linear SVM is trained based on the cross-level pooled features to obtain the predictions. Experimental results on MIT indoor scenes and SUN 397 datasets shown in Section 4 verify the great discrimination and robustness to scale transformation of the proposed image representation.

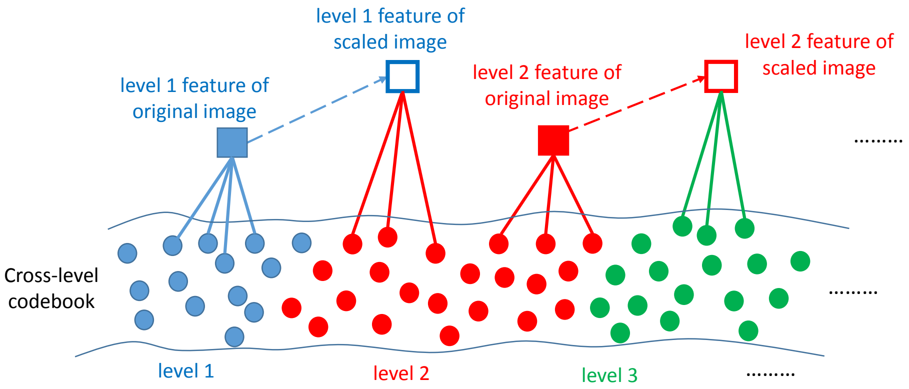
level 1 feature of scaled image
level 1 feature of original image
level 2 feature of original image
level 2 feature of scaled image
....
Cross-level codebook
....
level 1
level 2
level 3
....

Fig. 3. Illustration of selected representation codewords of CNN features for original image and scaled image. Circles in different colors represent code words from different levels. Blue and red solid squares denote level 1 and level 2 CNN features of the original image, respectively, and blue and red hollow squares represent those of the scaled image respectively. (better viewed in color)

4 Experiments

4.1 Datasets

We evaluate the proposed approach on the two currently largest scene classification datasets: MIT indoor scenes and SUN 397.

MIT indoor scenes is the largest indoor scene dataset, which contains 67 categories and a total of 15620 images. The complex spatial layout of the indoor scene image makes the classification even more difficult than outdoor scene image classification. Therefore, this dataset is chosen as an important benchmark for the evaluation of our approach. The standard training/testing split for the MIT indoor scenes dataset consists of 80 training images and 20 testing images per category.

SUN 397 is the current largest scene classification dataset. It contains 397 scene categories, both indoor and outdoor, with at least 100 images per category. The 10 fixed splits for training and testing images are publicly available. For each category, there are 50 images for training and 50 images for testing. The accuracy is all averaged over all the 10 splits.

4.2 Multi-level Cascaded Fine-tuning

Baselines We compare our cascaded fine-tuned CNN with two baselines: (a) off-the-shelf CNN features extracted by the pre-trained model on ImageNet (here we choose DeCAF_6 [15] as our off-the-shelf CNN feature for its better performance than DeCAF_7); (b) fine-tuned CNN initialized by the pre-trained model on ImageNet.

We conduct the comparison experiments on both the MIT indoor scenes dataset and the SUN 397 dataset. To be fair, we test all these 3 CNN features by simple max-pooling within their own level and training a linear SVM, without cross-level LLC coding and pooling. Please note that for level 1, since the whole image yields only one feature, there is no need to do pooling, and since no coarser level exists, there are no cascaded fine-tuned results. All the fine-tuned CNN features are obtained after 50 epochs of training. L2 normalization is performed on all the CNN features before used to train the SVMs. The SVM parameter (C) is all set as 0.5.

The results on MIT indoor scenes and SUN 397 are shown in Table 1 and Table 2 respectively for comparison. As can be seen, both on the MIT indoor scenes dataset and the SUN 397 dataset, fine-tuned CNN features, including those fine-tuned on ImageNet and cascaded fine-tuned ones on coarser levels of their own datasets, achieve higher accuracy than the off-the-shelf CNN features on all the levels. This is very natural since fine-tuned CNN features gain stronger discrimination power than generic off-the-shelf CNN features after the training on the specific datasets. The comparison between fine-tuned CNN on ImageNet and cascaded fine-tuned CNN shows that cascaded fine-tuned CNN features obtain higher accuracy than CNN fine-tuned on ImageNet by approximately 1% on all levels. This demonstrates that initialization by the trained model on the coarser level of a specific dataset helps the finer level model to converge to a better optimum than initialization by the model pre-trained on ImageNet.

Table 1. Classification accuracy on MIT indoor scenes for off-the-shelf CNN features, fine-tuned CNN features on ImageNet and cascaded fine-tuned CNN features of each level.

	level 1	level 2	level 3	level 4	level 5
off-the-shelf CNN	53.65	57.26	60.75	61.48	61.89
fine-tuned CNN on ImageNet	**61.46**	62.58	63.17	64.03	64.23
cascade fine-tuned CNN	—	**63.77**	**64.27**	**64.39**	**64.97**

Table 2. Classification accuracy on SUN 397 for off-the-shelf CNN features, fine-tuned CNN features on ImageNet and cascaded fine-tuned CNN features of each level.

	level 1	level 2	level 3	level 4	level 5
off-the-shelf CNN	40.53	41.25	41.68	42.07	42.64
fine-tuned CNN on ImageNet	**43.75**	44.88	45.17	45.54	45.81
cascade fine-tuned CNN	—	**45.61**	**46.33**	**46.58**	**46.87**

4.3 Cross-level LLC Coding and Pooling

Baselines We compare our cross-level LLC and pooling approach (*cross-level LLC-CNN*) with multi-level pooled CNN features [19]. We choose multi-level max-pooling as the pooling method since we also perform max-pooling on our *cross-level LLC-CNN* features.

Classification Accuracy We evaluate our cross-level LLC and pooling approach (*cross-level LLC-CNN*) on the MIT indoor scenes dataset and the SUN 397 dataset. The baseline method, multi-level pooled CNN is also tested for comparison with our *cross-level LLC-CNN*. The comparison results on each level and the combination of all levels are presented in Table 3 and Table 4. Here, cross-level LLC coding and pooling on a single level means that max-pooling is only performed within this level, while a cross-level codebook is still learned over all the levels. For the combination from level 1 to level 5, the final output of multi-level pooled CNN is obtained by concatenating the pooled result of each level together. Before all coding and pooling procedures, all the CNN features are extracted by the cascaded fine-tuned models. All the fine-tuned CNN features are obtained after 50 epochs of training. L2 normalization is performed on all the CNN features after extraction. The SVM parameter (C) is all set as 0.5. For reference, we also include some typical state-of-the-art results to compare with our approach.

In Table 3, from the comparison results between the baseline method and our approach, we can observe that on some finer levels, i.e., level 3, 4 and 5, our *cross-level LLC-CNN* works better than multi-level pooled CNN. The reason may be that more patches are available on these 3 levels (9 patches in level 3, 16 patches in level 4 and 25 patches in level 5), and pooling over more LLC features covers more information compared with original CNN features. Moreover, on the combination of all the 5 levels, our *cross-level LLC-CNN* also achieves higher accuracy than the baseline method, i.e., multi-level pooled CNN, which is 68.96% vs 67.87%, with a lower-dimensional feature. Compared to other state-of-the-arts, *cross-level LLC-CNN* also obtains the highest performance. It is worth mentioning that, to our best knowledge, the former best performance on this dataset is achieved by Multi-scale VLAD pooling on off-the-shelf CNN features, proposed by [19]. Compared to this MOP-CNN framework, our *cross-level LLC-CNN* obtains higher accuracy. Actually, the patches they used, i.e., 25 patches in level 2 and 49 patches in level 3, are much more than ours. The larger number of patches brings higher time cost in codebook learning and VLAD pooling. In contrast, the smaller number of patches utilized in our approach and the fast LLC performing make our *cross-level LLC-CNN* work much faster than their MOP-CNN. Table 4 shows the experimental results on the SUN 397 dataset. Overall, the comparison results are similar with those on MIT indoor scenes dataset. On SUN 397, *cross-level LLC-CNN* outperforms multi-level pooled CNN on some finer levels (level 4 and level 5) and the combination of all the 5 levels. Compared to the state-of-the-arts, our approach achieves the best accuracy (50.87%) on the combination of all the 5 levels with a relatively low feature dimension.

Table 3. Classification results on MIT indoor scenes for (a) baseline: multi-level pooled CNN; (b) *cross-level LLC-CNN*; (c) other state-of-the-arts.

methods		feature dimension	accuracy
(a) multi-level pooled CNN (baseline)	level1	4096	61.46
	level2	4096	63.77
	level3	4096	64.27
	level4	4096	64.39
	level5	4096	64.97
	level1+level2+$\cdots$+level5	20480	67.87
(b) ***cross-level LLC-CNN*** **(Ours)**	level1	16384	60.23
	level2	16384	62.47
	level3	16384	64.66
	level4	16384	65.48
	level5	16384	65.87
	level1+level2+$\cdots$+level5	16384	**68.96**
(c) state-of-the-arts	SPM [4]	5000	34.40
	FV+Bag of Parts [2]	221550	63.18
	Mode Seeking [32]	60000	64.03
	SPM+OPM [27]	—	63.48
	MOP-CNN [19]	12288	68.88

Table 4. Classification results on SUN 397 dataset for (a) baseline: multi-level pooled CNN; (b) *cross-level LLC-CNN*; (c) other state-of-the-arts.

methods		feature dimension	accuracy
(a) multi-level pooled CNN (baseline)	level1	4096	43.75
	level2	4096	45.61
	level3	4096	46.33
	level4	4096	46.58
	level5	4096	46.87
	level1+level2+$\cdots$+level5	20480	49.23
(b) ***cross-level LLC-CNN*** **(Ours)**	level1	16384	40.48
	level2	16384	42.53
	level3	16384	45.89
	level4	16384	47.41
	level5	16384	48.53
	level1+level2+$\cdots$+level5	16384	**50.87**
(c) state-of-the-arts	Xiao et al.[22]	—	38.00
	Decaf [15]	4096	40.94
	Fisher Vector [33]	256000	47.20
	SPM+OPM [27]	—	45.91

Scale Invariance To evaluate the scale invariance of our approach, we randomly select 670 testing images (half of the total) in MIT indoor scenes testing set and scale them by different scaling ratios, i.e., 10/9, 10/8, 10/7, 10/6, 10/5. For SUN 397, we use a random training/testing split (we choose the first split in the experiment) to evaluate the scale invariance. In this split, 1000 testing images randomly selected from the testing set are scaled by the same scaling ratios with those for MIT indoor scenes. Specifically, when scaling by a factor of ρ, we crop the image around the center with $1/\rho$ times of the original size, as illustrated in Figure 4. We compare the recognition accuracy over these scaled testing images of our *cross-level LLC-CNN* and the multi-level pooled CNN. Both methods are trained on non-scaled original training samples and the combination from level 1 to level 5 is adopted. Before all coding and pooling procedures, all the CNN features are extracted by the cascaded fine-tuned models. All the fine-tuned CNN features are obtained after 50 epochs of training. L2 normalization is performed on all the CNN features after extraction. The SVM parameter (C) is all set as 0.5.

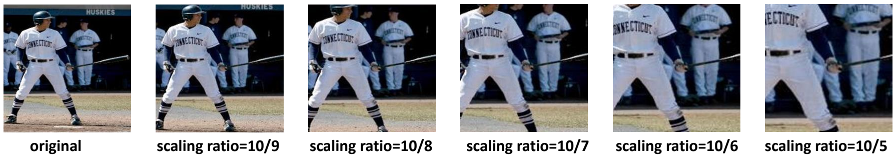

Fig. 4. Illustration of the scaled testing image with different scaling ratios.

The curves of recognition accuracy vs scaling ratio on the MIT indoor scenes dataset and the SUN 397 dataset are shown in Figure 5 and Figure 6, respectively. Both figures reflect the trend that the recognition accuracy decreases with the increase of the scaling ratio, whatever method is used. This shows that CNN features do not have scale invariance, as mentioned by lots of works [19, 21]. However, with our *cross-level LLC-CNN*, the classification accuracy decreases much more slowly than multi-level pooled CNN as the scaling ratio increases. As can be seen, from the original image to the 10/5 ratio scaled image, the difference in accuracy between our approach and the baseline approach is becoming increasingly big as the scaling ratio increases. Specifically, recognition accuracy with our approach when the testing image is scaled by 10/5 can still remain 50.63% and 34.32% for MIT indoor scenes and SUN 397 respectively. In comparison, the accuracy when the scaling ratio reaches 10/5 drops to 35.42% and 24.47% for MIT indoor scenes and SUN 397 respectively. The accuracy differences are all over 10%, showing the great superiority of our approach over the baseline approach. This superiority proves that LLC coding of CNN features on the cross-level codebook produces more robust features to the scale transformation, as LLC coding ensures that scaled CNN features can still be well represented by the cross-level codebook and their discrimination power is retained after scaling.

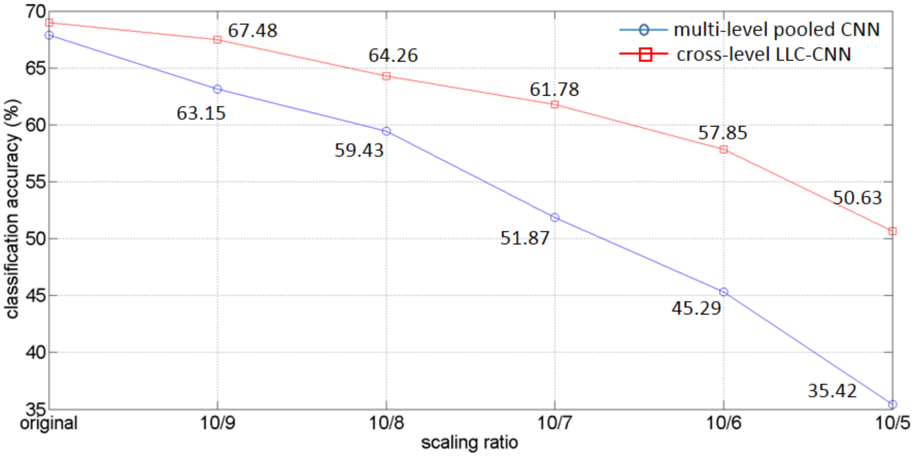

Fig. 5. Classification accuracy comparison between multi-level pooled CNN features and our *cross-level LLC-CNN* for scaled images with different scaling ratios on the MIT indoor scenes dataset.

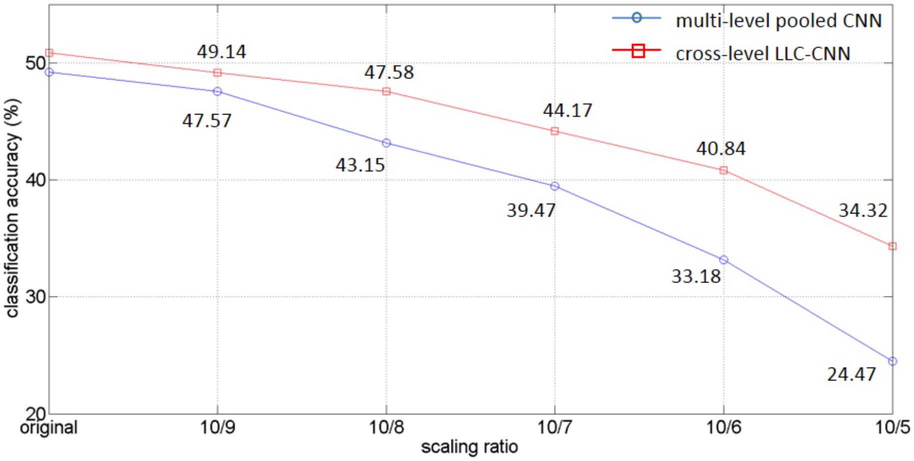

Fig. 6. Classification accuracy comparison between multi-level pooled CNN features and our *cross-level LLC-CNN* for scaled images with different scaling ratios on the SUN 397 dataset.

5 Conclusion

In this paper, we proposed a cross-level Locality-constrained Linear Coding and pooling framework (*cross-level LLC-CNN*) on multi-level CNN features to enhance the discrimination and scale invariance of the image representation for scene classification problems. Based on the cascaded fine-tuning scheme, the CNN features gain stronger discrimination in scene classification. In addition, with cross-level Locality-constrained Linear Coding and pooling on these multi-level fine-tuned CNN features, robustness to scale transformation is improved. We evaluated our approach on the MIT indoor scenes dataset and the SUN 397 dataset. Experimental results demonstrated that significant improvements in classification accuracy are achieved for both original and scaled testing images. In the future, we will explore how to improve the discrimination power and scale invariance of CNN in other vision tasks.

Acknowledgement.This research is supported by the National Research Foundation, Prime Ministers Office, Singapore under its International Research Centre @ Singapore Funding Initiative and administered by the Interactive & Digital Media Programme Office.

References

1. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on statistical learning in computer vision, ECCV. (2004)
2. Juneja, M., Vedaldi, A., Jawahar, C., Zisserman, A.: Blocks that shout: Distinctive parts for scene classification. In: IEEE CVPR. (2013) 923–930
3. Li, L.J., Su, H., Fei-Fei, L., Xing, E.P.: Object bank: A high-level image representation for scene classification & semantic feature sparsification. In: NIPS. (2010) 1378–1386
4. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE CVPR. (2006) 2169–2178
5. Dong, J., Xia, W., Chen, Q., Feng, J., Huang, Z., Yan, S.: Subcategory-aware object classification. In: IEEE CVPR. (2013) 827–834
6. Dong, J., Chen, Q., Yan, S., Yuille, A.: Towards unified object detection and segmentation. In: ECCV. (2014)
7. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **60** (2004) 91–110
8. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE CVPR. (2005) 886–893
9. LeCun, Y., Bengio, Y.: Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks **3361** (1995)
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. (2012) 1097–1105
11. Jarrett, K., Kavukcuoglu, K., Ranzato, M., LeCun, Y.: What is the best multi-stage architecture for object recognition? In: IEEE ICCV. (2009) 2146–2153

12. Ouyang, W., Wang, X.: Joint deep learning for pedestrian detection. In: IEEE ICCV. (2013) 2056–2063
13. Sun, Y., Wang, X., Tang, X.: Hybrid deep learning for face verification. In: IEEE ICCV. (2013) 1489–1496
14. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE CVPR. (2009) 248–255
15. 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)
16. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv preprint arXiv:1311.2524 (2013)
17. Oquab, M., Bottou, L., Laptev, I., Sivic, J., et al.: Learning and transferring mid-level image representations using convolutional neural networks. arXiv preprint (2013)
18. 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)
19. Gong, Y., Wang, L., Guo, R., Lazebnik, S.: Multi-scale orderless pooling of deep convolutional activation features. arXiv preprint arXiv:1403.1840 (2014)
20. Wei, Y., Xia, W., Huang, J., Ni, B., Dong, J., Zhao, Y., Yan, S.: Cnn: Single-label to multi-label. arXiv preprint arXiv:1406.5726 (2014)
21. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional neural networks. arXiv preprint arXiv:1311.2901 (2013)
22. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: IEEE CVPR. (2010) 3485–3492
23. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: IEEE CVPR. (2010) 3360–3367
24. Quattoni, A., Torralba, A.: Recognizing indoor scenes. IEEE CVPR (2009)
25. Bosch, A., Zisserman, A., Muñoz, X.: Scene classification via plsa. In: ECCV. (2006) 517–530
26. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: IEEE CVPR. (2009) 1794–1801
27. Xie, L., Wang, J., Guo, B., Zhang, B., Tian, Q.: Orientational pyramid matching for recognizing indoor scenes. In: IEEE CVPR. (2014)
28. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Computer Science Department, University of Toronto, Tech. Rep (2009)
29. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86** (1998) 2278–2324
30. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: IEEE CVPR. (2010) 3304–3311
31. Shabou, A., LeBorgne, H.: Locality-constrained and spatially regularized coding for scene categorization. In: IEEE CVPR. (2012) 3618–3625
32. Doersch, C., Gupta, A., Efros, A.A.: Mid-level visual element discovery as discriminative mode seeking. In: NIPS. (2013) 494–502
33. Sánchez, J., Perronnin, F., Mensink, T., Verbeek, J.: Image classification with the fisher vector: Theory and practice. IJCV **105** (2013) 222–245