Learning Bag-of-Multi-Scale-Activations from Deep Convolutional Nets

Anonymous CVPR submission

Paper ID ****

Abstract

While Deep Convolutional Neural Network (CNN) encodes multi-scale features with their spatial locations, Bag-of-feature representations are more invariant to spatial translations. In this paper, we propose a CNN architecture transformation schema that extracts multi-scale CNN representations with spatial invariance. By directly chaining multiple average-pooled rectified linear units (ReLU) to the final layer of scoring function, we explicitly learn an augmented model from multi-scale activations. We term this transformation Bag-of-Multi-Scale-Activation (BoMSA) Augmentation. The BoMSA augmented model can be trained from scratch or built upon existing pre-trained models, which makes the augmentation more flexible. BoMSA augmentation is also model-independent and can be easily adapted to existing models. Experimental results show consistent improvements achieved by incorporating the BoMSA representation on various models in the literature.

1. Introduction

Deep CNN has shown its versatility in the tasks such object detection and recognition [12, 19, 20, 22, 14] in the field of computer vision. Trained with large number of instances, (such as ImageNet [3]), CNN is also an excel candidate for off-the-shelf feature extractions, results in outstanding performance in various recognition tasks [17]. In the meantime, a considerable amount of effort is spent on how to further improve the performance of CNN. On one hand, many are focus on techniques to efficiently and effective train a CNN. A good initialization need to be carefully selected [2] in the beginning. Data augmentation [12] is recommended to improve the model performance as well. Drop-out [10] and momentum [21] are also necessary to prevent overfitting and obtain superior models. On the other hand, different model components and architectures are proposed. Rectified linear units (ReLU) [12] add non-linearity and enrich the model complexity. Different pooling method, such as Distance Transform Pooling [6], is adopted in [7], allowing local deformations, as in the widely-used deformable part-based model (DPM) [5]. [20] adopts a small 3×3 receptive field to deepen the model, while maintaining less parameters. A network in network [14] is proposed to enhance model discriminability for local patches within the receptive field. The award winning GoogLeNet [22] uses an *Inception* model that is based on the Hebbian principle [9], *i.e.*, neurons that fire together, wire together, the theoretical proof of which is provided by [1] under constraints.

One question to ponder: is there any model-independent potential that is yet to be discovered. Although allowing local deformation, CNN encodes the the spatial information of multi-scale features. Conversely, by ignoring the spatial location of features, bag-of-feature like techniques [13] achieves transformation invariance to some extend. As pointed out in [8], combining both type of features results in a better representation for recognition tasks with large variations, such as Scene classification [23, 16]. Traditionally, when CNN is used for feature extraction, the activation of the first fully-connected (FC) layer is usually considered as the feature [19, 8]. Using an off-the-shelf deep CNN model, [8] explicitly extracts image patches from three scales and computes the feed-forward CNN activations for each patch for feature extraction. This algorithm is cumbersome due to the computation of multiple patches for one image. Besides, it can only encode the multi-scale features to a certain extend limited by the burden of the post processing, *i.e.*, K-means and Principle Component Analysis (PCA). As a matter of fact, CNN feature is meant to encode multi-scale feature at each convolutional layer. Therefore, there is no need to extract multi-scale patches and only one feed-forward computation is sufficient to capture the activations of multi-scale features for one input image.

In this paper, we first conducted an empirical analysis on the feature discriminativeness of every unit of CNN model. By average-pooling at each layer, the output ignores global spatial information and results in a bag-of-feature representation. The bag-of-feature from ReLU is then found to carry the most performance gain at each layer. We have also observed a synergy of multi-scale activations results in a better discriminative representation. These observa-

tions tie closely to the vanishing gradient [2] issue in the deep CNN literature. During the training of a CNN, the top layer can be easily saturated. The gradients in the back-propagation algorithm will not effective reach the lower levels, resulting in a model that focuses more high-level features. Thus, we proposed a model augmentation schema, BoMSA, that chains the ReLU at each layer to the ultimate loss function. This augmentation approach is not confined by model variations and can be applied to existing pre-trained models or re-training the model from the ground up. By explicitly linking the lower layers to the decision layer, we gear the training towards the bag-of-activations from all layers. Thus, the augmented model is designed to be robust to spatial translations while maintaining its discriminative power in the multi-scale fashion. We apply our augmentation approach on various models (*e.g.*, Caffe [11], Deep19 [20], etc.) in the literature and test them on various scene recognition benchmark datasets, such as SUN397 [23], MIT67 [16], Scene15 [4]. A consistent performance improvement is observed of our multi-scale representation over the best discriminative single-scale feature.

The rest of the paper is organized as follows: we motivate and describe our model augmentation approach in Sec. 2. An in-depth analysis is provided in Sec. 3. Systematic experimental results are included in Sec. 4.

2. Technical Approach

We seek to discover a model-independent feature representation that can be efficiently computed. This representation should conform the property of *multi-scale* and *orderless* which outperform single-scale representation in scene classification tasks.

2.1. Model

In this work, we consider two models in the literature, namely Caffe and Deep19. Caffe [11] is a well-known deep learning framework. It has a pre-trained model which follows the landmark "AlexNet" [12] trained with millions of data in the ImageNet dataset [3]. This model has 6 conv. layers and 2 fully-connected (FC) layers. Deep19 [20] uses very small 3×3 receptive fields for the entire net and the deepest model has 19 layers (16 conv. and 3 FC layers). This model registered the state-of-the-art performance in ILSVRC-2014 classification challenge [18].

We intend to demonstrate the consistency and model-independence of our observations and approach. We have conducted experiments on several CNN models and observed similar patterns. Due to limited space, we only provide the analysis of the aforementioned Caffe and Deep19 models as exemplars in the rest of the paper.

2.2. Multi-scale Orderless CNN Activation

We consider the marginal activation of a rectified linear unit (ReLU) as the *orderless* feature response at its corresponding scale. As described in [15, 12], ReLU models a neuron's output r as a mapping of its input x with the non-saturating nonlinearity $r(x) = \max(0, x)$.

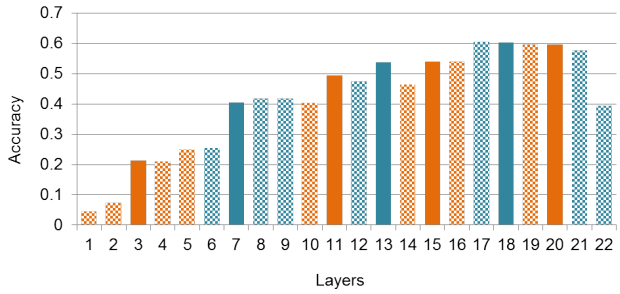
Concretely, to compute the proposed representation, let α_i of dimension $w \times h \times k$ be the output of ReLU at i-th layer ($w = h$ in most CNN models). Each $w \times h$ is the convolutional response map of one of the k filters. It carries the spatial activation information at the current scale. Thus, the *orderless* response, f_i, at each scale can be computed by sum-pooling the first two dimension of α_i, resulting in a k-dimensional vector, *i.e.*,

$$f_i = \sum_w \sum_h \alpha_i \tag{1}$$

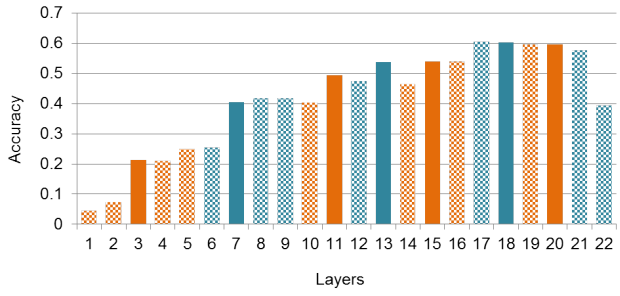
We are motivated to choose the output of the ReLU layer due to the nonlinearity it introduces to the CNN architecture, which boosts the discriminative power of the model. To further provide statistical evidences for the argument, we conduct classification experiments on MIT Indoor Scene (MIT67) dataset [16] using both Caffe and Deep19 models. We consider the orderless response, f_i, from each layer as different features; one-vs-all linear SVMs are then trained and the classification accuracy is reported in Fig. 1. The detailed experimental setups for both Caffe and Deep19 models are similar to the ones described in Sec. 4. As seen in Fig. 1, in general, the discriminativeness of the model increases as the layer becomes higher. More importantly, we observe significant improvement resulting from each ReLU layer. These observations not only inform us that the CNN model owes its performance boost to the nonlinearity of the ReLU layer, but also motivate us to leverage multiple ReLU layer responses to enrich the discriminative ability of the model.

2.3. The Feature Selection from Activation Layer

Since different CNN layer models image patterns at various scales, the first question to ask is that "Does combining activations at all scales help to extract a better feature"? To address this question, we carried out another experiments on the MIT67 data [16]. In the literature when CNN is used for feature extraction, the response of the first fully connected layer is usually selected. Our previous analysis in Fig. 1 also demonstrates the discriminative power of high-level CNN layer. Thus, in this experiment, we start from the f of the last ReLU layer and greedily concatenate the marginal responses from previous ReLU layer. The rest of the experimental setups are the same as in Sec. 2.2 and Sec. 4, *i.e.*, 67-way one-vs-all linear SVM classifiers are trained every time we add one more layer. Since we keep on concatenating the



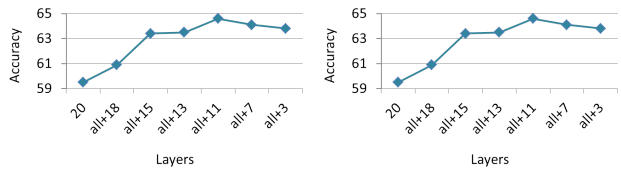
(a) Caffe



(b) Deep19

Figure 1. The classification accuracy on MIT67 [16] using the marginal activation of each layer. The color alternation indicates change of convolutional layers. The solid color fill represents the output of the ReLU layer. There are 7 ReLU layer for the Caffe model and 18 for Deep19. A significant performance improvement is observed at each ReLU layer compared with its previous layer, especially at the early convolutional layers.

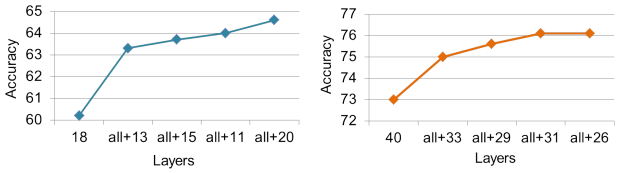
fs to our feature, the feature dimension is monotonically increasing. As the results shown in Fig. 2, the performance increase in the beginning by adding mid-level features to enrich the model discriminativeness. However, including some low level responses to the feature actually hurts the performance, *i.e.*, layer 7 and 3 in Fig. 2(a). We believe it demonstrates the benefit of incorporating more discriminative mid-level and high-level features of CNN (*i.e.*, *parts* and *objects*), but not necessarily the low-level features, such as *corner* or *edge*.



(a) Caffe (b) Deep19

Figure 2. The performance trend when incorporating more f from the ReLU layer. The keyword "all" means previous fs been included as feature. A performance drop is observed when incorporating low level marginal responses.

A more systematic way to select the layers to incorporate as feature is feature forward selection, *i.e.*, greedily add the layer in the ReLU pool which results in the best performance boost in an iterative manner. As seen in Fig. 3(a), the optimal results of this greedy approach is congruent with the previous results in Fig. 2(a), which rejects the low-level features. Similarly for the Deep19 model in Fig. 3(b), although the feature pool is large (18 ReLU layers), the optimal greedy results only selects the mid- or high-level filter responses.



(a) Caffe (b) Deep19

Figure 3. The performance trend when using forward selection to incorporate fs from the ReLU layer.

By incorporating marginal responses from multiple layers, we have achieved *multi-scale orderless* feature extraction. We term this feature "MOFeat". We should point out that for the latter analysis and experiments on other datasets, we opt to use the same feature selected by the forward selection algorithm on MIT67 data. An consistent performance boost demonstrates the data-independence of the proposed feature representation.

3. Discussions

3.1. Relationship to the literature

Though inspired by [8], our computation of the *multi-scale orderless* feature is completely different from them. In [8], local patches at three scales are first extracted (roughly 50 patches per image in their setting); each patch is then feed to the entire CNN and the activations of the first fully connected layer (4096-dimension) are used for post processing, *i.e.*, K-means + VLAD pooling.

Our approach has several advantages over [8]:

1. [8] computes multiple feed-forward on a single image due to multiple patches. The computation of convolution is the bottleneck for this procedure. In our implementation, only one feed-forward computation is needed for one image and the activation is available at each scale.

2. [8] uses K-means clustering algorithm in the VLAD pooling, which can result in inconsistent result due to random initialization.

3. [8] conducts two steps of PCA (first PCA to reduce 4096-dimensional activation to 500-dimension, and

second PCA to reduce $50,000$-dimensional VLAD pooled feature to 4096-dimension).

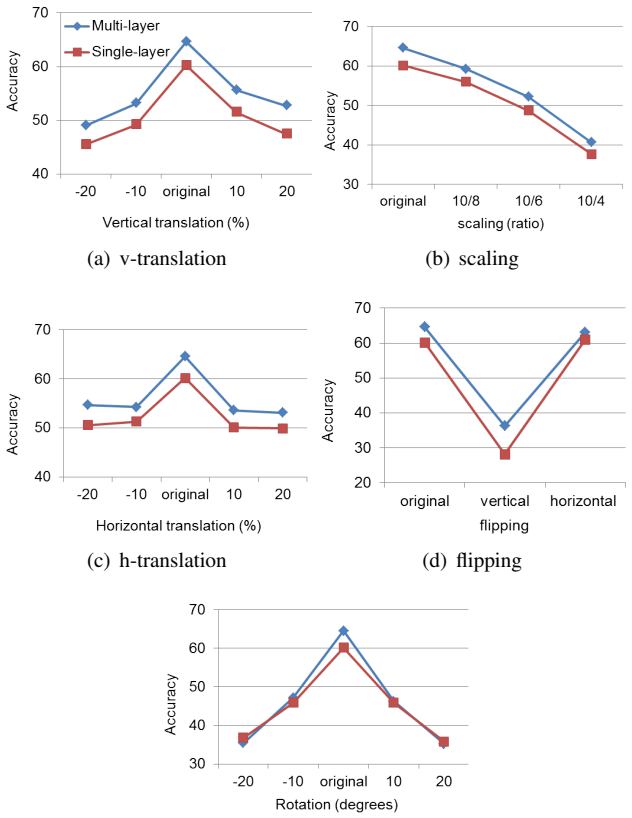
4. the feature extraction procedure in [8] takes more than 20 seconds for one test image while ours only take less than 1 second.

Overall, the implementation of *multi-scale orderless* feature in [8] is not only inefficient but also prone to over-fitting due to a number of hyper-parameter tuning. On the contrary, our implementation is easy to compute and little parameter tuning is needed.

3.2. The Analysis of Invariance

Besides the insight on visualizing deep CNN features, [24] provides transformation invariance analysis of their model on several images. This is done by comparing the feature vector distance between the original and transformed images. Realizing that several examples do not necessarily represents the overall statistics, [8] improve the invariance analysis by experiments on the SUN397 scene database [23]. Unfortunately, only the results on 4 categories are provided in their paper. The performance for the entire dataset is still unknown

To cope with the aforementioned gap and analyze the invariance of our representation, we carried out experiments on the entire MIT Indoor dataset [16] and report the average performance. We intend to compare the tolerance of the proposed multi-scale MOFeat compared with single-scale CNN features. Concretely, the single-scale CNN feature in comparison is the activation of CNN model which achieves best classification performance (not necessarily the output of the first FC layer). We first trained 67-way one-vs-all linear SVM classifiers for all 67 classes using features extracted from the original images. During test phase, four transformations are considered for the analysis, *i.e.*, translation, scaling, flipping, and rotation. Fig. **??** illustrates all the transformations and their corresponding parameters, which are similar to the ones in [8]. After applying the transformations to test images, features are extracted for the proposed MOFeat and single-scale feature. The classification is conducted using the trained SVMs, and the accuracy is shown in Fig. 4. Each data point is the performance of the entire test data. As seen in Fg. 4, the transformations almost always hurt the performance except for horizontal flip. This observation is similar to the one in [8]. We also see that multi-scale feature, MOFeat, consistently out-performs the single-scale feature under every degree of all transformations. As far as the invariance is concerned, we observe the same level of performance decrease, if not more, for MOFeat compared with single-scale. This means that both features have the same level of tolerance to transformations. This immediately suggest that "training jittering", *i.e.*, including transformations in the training data, will improve the performance. Since "training jittering" is conducted during the training of CNN model for classification [20], it is interesting that the observation of our invariance analysis shows that when using CNN for feature extraction, it is beneficial to perform another round of "training jittering" to train the classifiers.



(a) v-translation (b) scaling

(c) h-translation (d) flipping

(e) rotation

Figure 4. The classification accuracy of various transformations on test images in MIT67 data [16]. The legends shown in Fig. 4(a) is the same in the other four figures.

4. Experimental Results

In this section, we carried out experiments on benchmark dataset, SUN397 [23], MIT67 [16], and Scene15 [4] to demonstrate the consistency of our approach. We should stress that our MOFeat representation is learned on MIT67 and applied to all the experiments.

5. Experiment Setup

We follow the standard image pre-processing steps in the literature [12, 11, 20]. Since both Caffe [11] and Deep19 models [20] have a fixed input size of 224×224, we first resize the image such that the smaller side matches 224, and then crop an 224×224 patch from the center. The mean

RGB value of each model learned on ImageNet [3] is then subtracted.

consistent performance improvement on all scene dataset

1. SUN397
2. MIT67
3. Scene15

6. Conclusion

References

[1] S. Arora, A. Bhaskara, R. Ge, and T. Ma. Provable bounds for learning some deep representations. *CoRR*, abs/1310.6343, 2013. 1

[2] Y. Bengio and X. Glorot. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of AISTATS 2010*, 2010. 1, 2

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1, 2, 5

[4] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005. 2, 4

[5] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 2010. 1

[6] P. F. Felzenszwalb and D. P. Huttenlocher. Distance transforms of sampled functions. *Theory of Computing*, 2004. 1

[7] R. B. Girshick, F. N. Iandola, T. Darrell, and J. Malik. Deformable part models are convolutional neural networks. *CoRR*, abs/1409.5403, 2014. 1

[8] Y. Gong, L. Wang, R. Guo, and S. Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *ECCV*, 2014. 1, 3, 4

[9] D. Hebb. *The Organization of behaviour*. New York: Wiley & Sons, 1949. 1

[10] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580, 2012. 1

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 2, 4

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*. 2012. 1, 2, 4

[13] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 1

[14] M. Lin, Q. Chen, and S. Yan. Network in network. *CoRR*, abs/1312.4400, 2013. 1

[15] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010. 2

[16] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *CVPR*, 2009. 1, 2, 3, 4

[17] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. *CoRR*, abs/1403.6382, 2014. 1

[18] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014. 2

[19] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, abs/1312.6229, 2013. 1

[20] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, 2014. 1, 2, 4

[21] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *ICML*, 2013. 1

[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014. 1

[23] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 1, 2, 4

[24] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013. 4