Fisher Vectors Meet Neural Networks: a Hybrid Classification Architecture

Florent Perronnin, Diane Larlus
Computer Vision Group, Xerox Research Centre Europe.

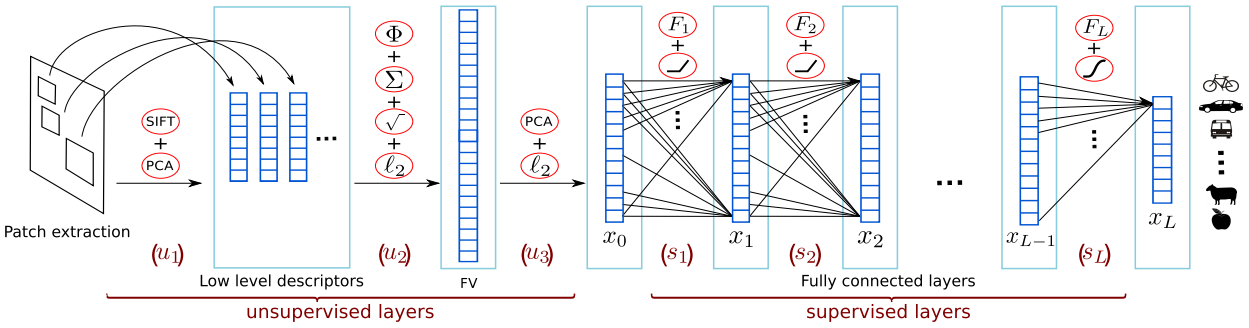

Figure 1: Proposed hybrid architecture. The first three layers – (u_1), (u_2) and (u_3) – are unsupervised. Image patches are described by PCA-projected SIFT or color descriptors (u_1). These descriptors are then embedded using FV-encoding, aggregated at the image level and normalized by square-rooting and ℓ_2-normalization (u_2). The resulting high-dimensional FV is PCA-projected and re-normalized (u_3). The supervised layers (s_1), (s_2), .. (s_{L-1}) involve a linear projection followed by a reLU. The last layer (s_L) involves a linear projection followed by a softmax or a sigmoid and produces the label estimates. Our architecture can be considered deep as it stacks several unsupervised and supervised layers.

Two image classification paradigms have dominated the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [11] in recent years: Fisher Vectors (FV) [12] and Convolutional Neural Networks (CNN) [8]. FVs involve extracting local descriptors, encoding them with high-order statistics, aggregating them, and feeding them to kernel classifiers such as linear SVMs. As for CNNs [9], they are feed-forward architectures which involve multiple computational layers that alternate linear operations such as convolutions and non-linear operations such as max-pooling.

While CNNs have lately shown superior accuracy on large-scale classification tasks, deep architectures come with challenges. This includes the requirement for large amounts of training data, their high computational cost which makes training on GPUs or very large clusters a necessity, or their lack of geometric invariance. This explains why the FV is still very competitive for certain tasks – see the winning system [7] at the Fine-Grained Visual Competition 2013 [2]. Hence, several works [13, 14] have started exploring the combination of FVs and CNNs.

Our **first contribution** is a novel hybrid architecture that combines the best of both worlds. Its first layers are unsupervised and involve the computation and dimensionality reduction of high-dimensional FVs. This is followed by a set of supervised fully connected Neural Network (NN) layers – akin to a Multi-Layer Perceptron (MLP) – trained through back-propagation. See details in the Figure above and in the full paper. We show experimentally on the ILSVRC 2010 dataset that the proposed architecture significantly outperforms previous FV-based pipelines and that it comes close to the accuracy of the "AlexNet" [8]: 17.6% top-5 error rate for our system *vs.* 17.0% for the latter.

Because it is unpractical to collect large amounts of labeled data for each new task, we are also interested in transferring the mid-level features learned by our architecture. Transferring features derived from deep classifiers either to different class sets or even to new tasks (*e.g.* image retrieval or object detection) has been a very active research topic lately [1, 4, 6, 10].

Our **second contribution** is to show that we can derive mid-level features from our hybrid architecture which are competitive with those derived from CNNs. For instance, we conducted experiments where we pre-trained our architecture (with three hidden supervised layers and 4K units per hidden layer) on ILSVRC'12 and, given a new image, we used the output of the penultimate layer as a novel representation. Using such features, we report on PASCAL VOC'07 [5] a competitive 76.2% mean Average Precision (mAP). This is far above the best FV results reported in [3] – 68.0% mAP. Using these mid-level features we also report in the full paper competitive results for the problem of instance-level image retrieval on the INRIA Holidays and UKB datasets.

[1] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky. Neural codes for image retrieval. In *ECCV*, 2014.

[2] A. Berg, R. Farrell, A. Khosla, J. Krause, L. Fei-Fei, J. Li, and S. Maji. Fine-Grained Competition (FGComp). http://sites.google.com/site/fgcomp2013/, 2013.

[3] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: delving deep into convolutional nets. In *BMVC*, 2014.

[4] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. In *ICML*, 2014.

[5] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The Pascal visual object classes (VOC) challenge. *IJCV*, 2010.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[7] P.-H. Gosselin, N. Murray, H. Jégou, and F. Perronnin. Revisiting the Fisher vector for fine-grained classification. *PRL*, 2014.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoff Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012.

[9] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. Handwritten digit recognition with a back-propagation network. *NIPS*, 1989.

[10] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014.

[11] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei. ImageNet large scale visual recognition challenge. arXiv, 2014.

[12] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the Fisher Vector: theory and practice. *IJCV*, 2013.

[13] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep Fisher networks for large-scale image classification. In *NIPS*, 2013.

[14] V. Sydorov, M. Sakurada, and C. Lampert. Deep Fisher kernels – End to end learning of the Fisher kernel GMM parameters. In *CVPR*, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.