Speeding up Convolutional Neural Networks with Low Rank Expansions

Max Jaderberg
max@robots.ox.ac.uk
Andrea Vedaldi
vedaldi@robots.ox.ac.uk
Andrew Zisserman
az@robots.ox.ac.uk

Visual Geometry Group
Department of Engineering Science
University of Oxford
Oxford, UK

The focus of this paper is speeding up the application of convolutional neural networks (CNNs). While delivering impressive results across a range of computer vision and machine learning tasks, these networks are computationally demanding, limiting their deployability. Convolutional layers generally consume the bulk of the processing time, and so in this work we present two simple schemes for drastically speeding up these layers. This is achieved by exploiting cross-channel or filter redundancy to construct a low rank basis of filters that are rank-1 in the spatial domain. Our methods are architecture agnostic, and can be easily applied to existing CPU and GPU convolutional frameworks for tuneable speedup performance. We demonstrate this with a real world network designed for scene text character recognition [1], showing a possible 2.5× speedup with no loss in accuracy, and 4.5× speedup with less than 1% drop in accuracy, still achieving state-of-the-art on standard benchmarks.

Approximation Schemes. We provide the frameworks for two methods to approximate the N 3D filters W_n of a convolutional layer acting on the input z with C channels z_n^c such that $W_n * z = \sum_{c=1}^{C} W_n^c * z^c$. Both methods exploit the redundancy that exists between different feature channels and filters of convolutional layers.

Scheme 1 builds upon the work of Rigamonti *et al.* [2] and approximates the original set of full-rank filters as a linear combination of a smaller set of separable (rank-1) filters. The separability of these filters allows convolutions to be computed much more efficiently than the full-rank filters by splitting the full convolution in to horizontal convolution followed by vertical convolution. For a layer of N convolutional filters W_n^c, $n \in [1 \dots N]$, where each filter acts on a single channel c of the 3D input, $c \in [1 \dots C]$, we learn a basis of M separable filters s_m^c, $m \in [1 \dots M]$, where $M < N$, as well as the coefficients a_n^{cm} to linearly combine them, such that the original filter $W_n^c \simeq \sum_{m=1}^{M} a_n^{cm} s_m^c$, offering a speedup due to the separable convolution and smaller basis of filters required for convolution.

Scheme 2 also employs the idea of separable convolutions, but uses a separate basis of vertical filters and horizontal filters. The original convolutional layer is approximated by a vertical convolution layer with K vertical filters $\{v_k : k \in [1 \dots K]\}$ followed by a horizontal convolutional layer with horizontal filters $\{h_n : n \in [1 \dots N]\}$. This results in the original filters being approximated by the sequence of these two layers, *i.e.* $W_n^c \simeq \sum_{k=1}^{K} h_n^k * v_k^c$. The advantage of this method over Scheme 1 is that it can be plugged straight in to any CNN toolbox that supports rectangular filters.

For both schemes, the speedup can be tuned by varying the number of filters for each basis.

Optimization. The separable approximations can be optimized using two methods – *filter reconstruction optimization* and *data reconstruction optimization*. Filter reconstruction optimization aims to minimise the reconstruction error of the original filters by the approximation, whereas data reconstruction optimization aims to reconstruct the output of the original layer by the approximated layer for the training data inputs, minimising the reconstruction error using traditional back-propagation of errors.

Results. We provide the results of our approximation schemes on scene text character recognition using the state-of-the-art classifier of [1], under different scenarios and settings. A 4.5× speedup can be obtained with virtually no loss in classifier accuracy (Fig. 2), with data reconstruction optimization improving the accuracy of both schemes.

[1] M. Jaderberg, A Vedaldi, and A. Zisserman. Deep features for text spotting. In *European Conference on Computer Vision*, 2014.

[2] R. Rigamonti, A. Sironi, V. Lepetit, and P. Fua. Learning separable filters. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2754–2761. IEEE, 2013.

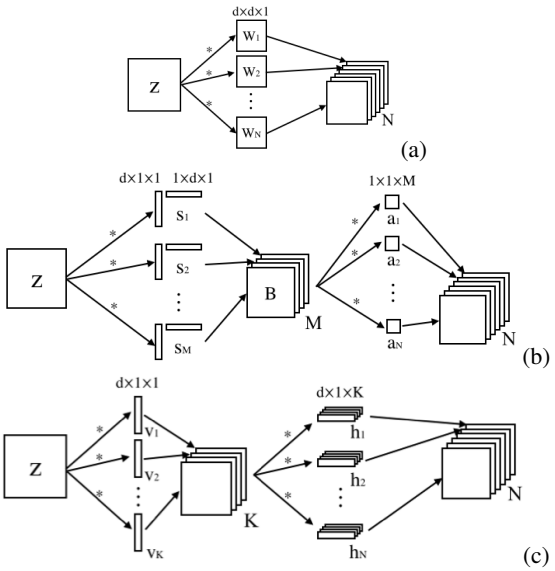
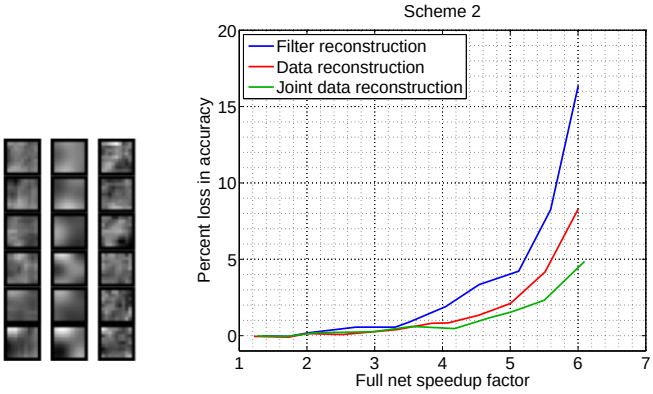


Figure 1: **Approximation Frameworks** (a) The original convolutional layer acting on a single-channel input *i.e.* C=1. (b) The approximation to that layer using the method of Scheme 1. (c) The approximation to that layer using the method of Scheme 2.



(a) (b)

Figure 2: **Approximation Results** (a) A selection of Conv2 filters from the original CNN (left), and the reconstructed versions under Scheme 1 (centre) and Scheme 2 (right), where both schemes have the same model capacity corresponding to 10x theoretical speedup. (b) The percent loss in performance as a result of the speedups attained with Scheme 2 (c). Joint data reconstruction optimizes the solution of multiple layers' approximations jointly, rather than optimizing each layer in isolation.

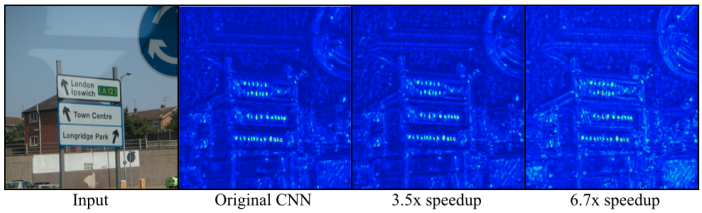


Figure 3: **Qualitative Result** Text spotting using the CNN character classifiers run in sliding window mode. The maximum response map over the character classes of the CNN output with Scheme 2 indicates the scene text positions. The approximations have sufficient quality to locate the text, even at 6.7× speedup.