Domain-adaptive deep network compression

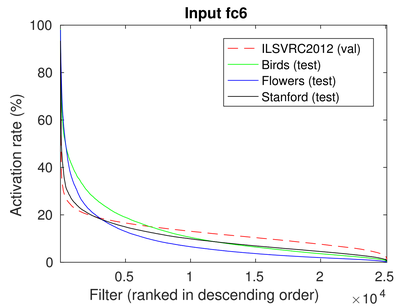
Marc Masana, Joost van de Weijer, Luis Herranz, Andrew D Bagdanov, Jose M Alvarez
Computer Vision Center, Universitat Autònoma de Barcelona, Spain

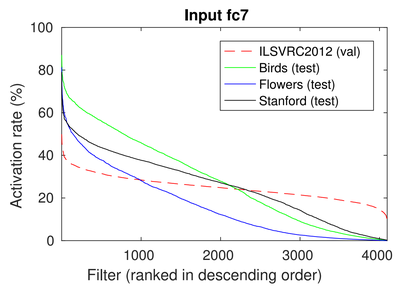
Problem Statement

Present situation: Networks trained on large datasets can be easily transferred to new domains with less examples using fine-tuning [1].
Problem: Networks from the source task are often too large for the target task.
Our approach: We use the statistics of network activations to improve the existing compression algorithms based on low-rank matrix decomposition.

Motivation

Activation rate: fraction of images in which a neuron has non-zero response.
Question: Is there a shift of the activation distributions in deep networks when changing domain?





Compression by Matrix Decomposition

Consider a single fully-connected layer being:

$$y = Wx + b,$$

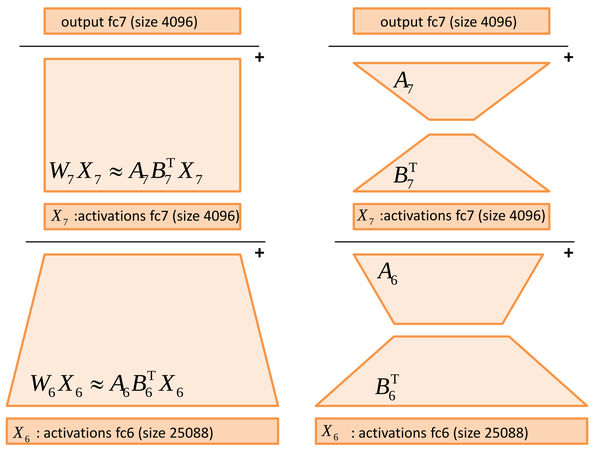
or considering a set of p inputs to the layer:

$$Y = WX + b1_p^T.$$

The truncated SVD method [2, 3] applies SVD decomposition so that the layer weights W can be approximated by keeping the k most significant singular vectors,

$$W = USV^T \quad \longrightarrow \quad \hat{W} = \hat{U}\hat{S}\hat{V}^T.$$

Compression is obtained by replacing the original layer by two new ones: the first with weights $\hat{S}\hat{V}^T$ and the second with weights $\hat{U}$.



Compressing two layers of the VGG-16 network (fc6, fc7). The original weight matrix is approximated by two matrices. As main novelty, we consider the input unit activations X of each layer when compressing the corresponding weight matrix W.

Domain Adaptive Low Rank Decomposition

The incorporation of input X is done by minimizing $||Y - \hat{Y}||_F$. We want to decompose the layer with weights W into two layers according to:

$$W \approx \hat{W} = AB^T. \tag{1}$$

We want the decomposition which minimizes:

$$\min_{A,B} ||Y - \hat{Y}||_F = \min_{A,B} ||WX - AB^TX||_F, \tag{2}$$

where we have set $\hat{b} = b$ and subsequently removed it from the equation. Eq. 2 is a rank constrained regression problem which can be written as:

$$\begin{aligned} &\arg\min_C ||Z - CX||_F^2 + \lambda||C||_F^2 \\ &\text{s.t. } \operatorname{rank}(C) \leq k, \end{aligned} \tag{3}$$

where $C = AB^T$ and $Z = WX$, and we have added a ridge penalty which ensures that C is well-behaved even when X is highly co-linear.

A closed form solution exists for such minimization problems [4], based on the SVD of Z. Applying SVD we obtain $Z = USV^T$. Then the matrices A and B in Eq. 1 which minimize Eq. 3 are:

$$\begin{aligned} A &= \hat{U} \\ B &= \hat{U}^T ZX^T \left(XX^T + \lambda I\right)^{-1} \end{aligned} \tag{4}$$

where $\hat{U}$ consists of the first k columns of U.

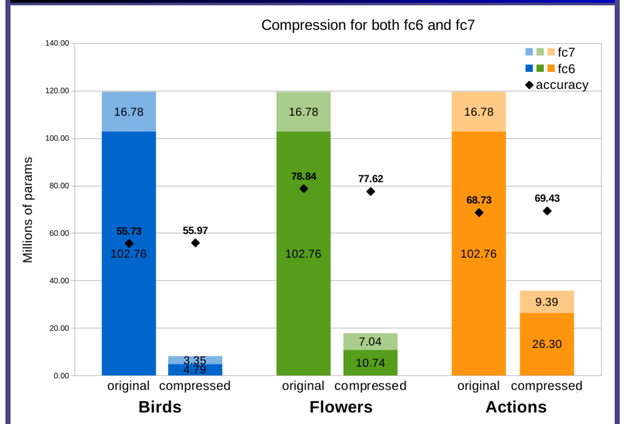
Network compression is obtained by replacing the layer weights W by two layers with weights B and A, just as in the truncated SVD approach. The first layer has no biases and the original biases b are added to the second layer.

Code Available

GitHub

https://github.com/mmasana/DALR

DALR Iterative Compression

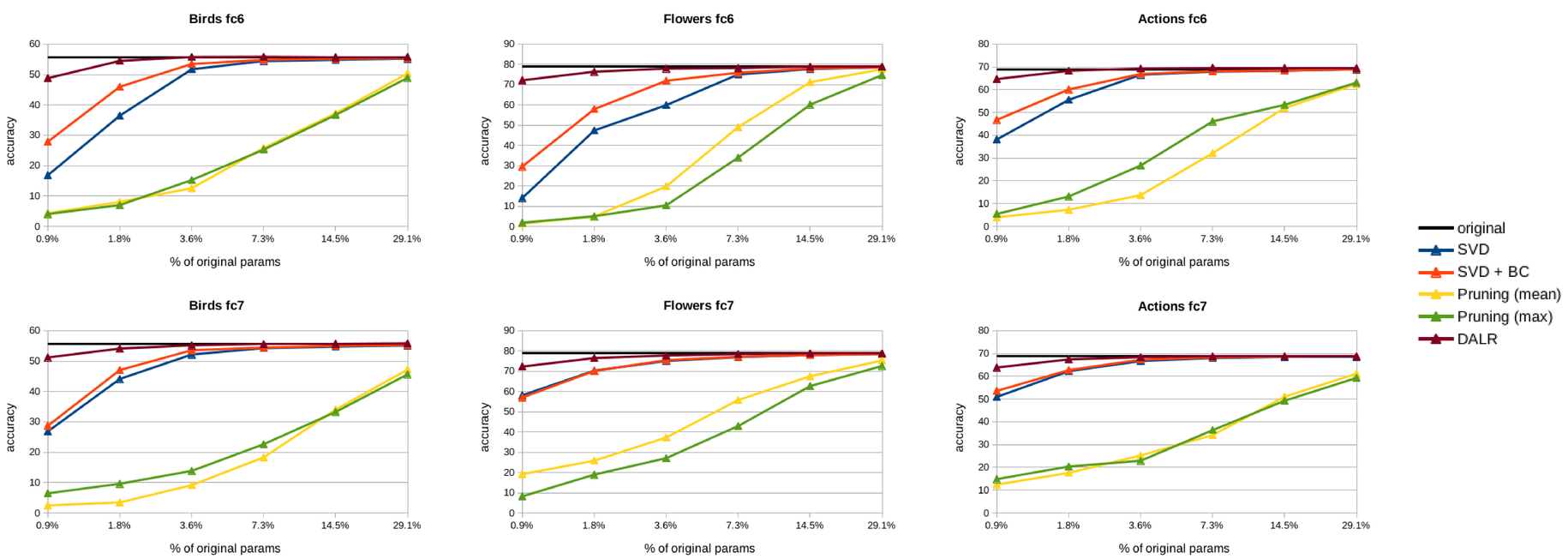


We implemented an iterative solution to find compression pairs for both fc6-fc7. At each step we slightly increase the compression on both layers. Then, both options are evaluated on the validation set, and the compression rate with better performance is applied to the network and passed onto the next iteration. Once the stop condition is met (here 1% accuracy drop), we evaluate on the test set.

Conclusions

- DALR on image recognition, the fc6 and fc7 layers of VGG-19 can be compressed 4x-8x more than the truncated SVD alone for CUB-200 Birds, Oxford-102 Flowers and Stanford-40 Actions.
- DALR provides a slight boost in performance when compressing to 25% of the original parameters for the above datasets.
- DALR on object recognition with fast RCNN on PASCAL 2007 does not impact performance if reduced to 25% of the parameters, while SVD has a drop of 0.3% mAP. If fc6 is reduced further to 22%, DALR reduces the error (0.2% mAP) obtained with truncated SVD (0.4% mAP).
- We found that higher compression rates can be applied in target domains further away from the source domain.

Experimental Results



References

[1] Maxime Oquab, Leon Bottou, Ivan Laptev and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2014.

[2] Yann LeCun, John S. Denker and Sara A. Solla. Optimal brain damage. In *Neural Information Processing Systems (NIPS)*, 1989.

[3] Jian Xue, Jinyu Li and Yifan Gong. Restructuring of deep neural network acoustic models with singular value decomposition. In *Interspeech* pages 2365–2369, 2013.

[4] Ashin Mukherjee. Topics on Reduced Rank Methods for Multivariate Regression. In *PhD Thesis* at The University of Michigan, 2013.