
Exploiting Redundancies in Convolutional Networks

Emily Denton DENTON@CS.NYU.EDU
Wojciech Zaremba ZAREMBA@CS.NYU.EDU
Joan Bruna BRUNA@CS.NYU.EDU
Rob Fergus FERGUS@CS.NYU.EDU

Abstract

1. Introduction

Training large neural networks (NN) takes weeks, or even months. This can hinder research, and there have been extensive effort devoted to speed up training procedure. However, resource-wise from perspective of companies executing neural networks on internet-scale data (e.g. annotating images), this is not the main cost. Major cost is in the final stage, where network is evaluated on the target data, which is present in quantities of billions. We focus here on speeding up evaluation of *trained* NN, which directly maps to the cost of executing NN on internet-scale data. Our techniques speed up execution by factor of 2–4 while keeping prediction accuracy within 1% from the original model.

We focus in this work on convolutional neural networks used for computer vision tasks. Most of computation time during evaluation is spend on convolutional layers i.e. $\sim 90\% - 95\%$, while it takes only the small fraction of time $\sim 5\% - 10\%$ to evaluate rest of layers (pooling, local contrast normalization, fully connected). It is worth to note, that most of learnable parameters are kept in fully connected layers $\sim 90\% - 95\%$, and convolutional layers constitutes of very small fraction of parameters $\sim 5\% - 10\%$.

We achieve forward pass speed up by constructing approximations to the convolutional layer kernel. Convolutional kernel is a 4-dimensional tensor, with two spacial dimensions, and two feature maps-to-feature maps dimensions. Kernel of trained network has a lot of redundancies in parameters, which we exploit to

speed up forward pass, while mildly training off prediction accuracy (approximated kernels give prediction within $\sim 1\%$ of the original prediction).

2. Related Work

There have been extensive research devoted speeding up forward pass of neural network. Most efforts are computational architecture specific, i.e. to CPU or GPU architecture.

Vanhoucke et. al. examined how to exploit properties of CPUs to speed up execution. This includes aligning memory, using vectorized operations, and values quantized (this is conceptually most similar to what we do). They were able to achieve $4\times$ speed up by combining this techniques. Methods presented in aforementioned paper are orthogonal to our low-rank approximations, and can be used simultaniously.

Denil et. al (Denil et al., 2013) noticed that

Classical optimizations for convolutions.

Quantization.

Sparse arrays of signatures for online character recognition.

Understanding of ConvNets.

3. Tensor Low Rank Approximation

This section describes a low-rank approximation of a generic convolutional layer.

Let W be a 4-dimensional tensor of dimensions (C, X, Y, F) , and let $I(c, x, y)$ denote an input signal, where $c = 1 \dots C$ and $(x, y) \in \{1, \dots, N\} \times \{1, \dots, M\}$. A generic convolutional layer is defined as

$$I * W(f, x, y) = \sum_{c=1}^C \sum_{x'=-X/2}^{X/2} \sum_{y'=-Y/2}^{Y/2} I(c, x-x', y-y') W(c, x', y', f)$$

3.1. Linear Compression of Convolutional Filter bank

In typical object recognition architectures, the convolutional tensors resulting from the training exhibit strong redundancy and regularity across all its dimensions. This redundancy affects performance since it exposes the architecture to more overfitting, and runtime speed. A particularly simple way to exploit such regularity is to linearly compress the tensors, which amounts to finding low-rank approximations.

Given a 4-tensor W of dimensions (C, X, Y, F) , we

search for decompositions that minimize

$$\|W - \sum_{k \leq K} \alpha_k \otimes \beta_k \otimes \gamma_k \otimes \delta_k\|_F, \quad (1)$$

where α_k , β_k , γ_k and δ_k are rank 1 vectors of dimensions C , X , Y and F respectively, and $\|X\|_F$ denotes the Frobenius norm. Generalization of the SVD.

The rank K approximation (3) can be obtained using a greedy algorithm, which computes for a given tensor X its best rank-1 approximation:

$$\min_{\alpha, \beta, \gamma, \delta} \|X - \alpha \otimes \beta \otimes \gamma \otimes \delta\|_F. \quad (2)$$

This problem is solved by iteratively minimizing one of the monoids while keeping the rest fixed. Each of the step consists in solving a least squares problem. (todo expand).

Figures ? and ?? show low-rank approximations of the first two convolutional layers of the Imagenet architecture.

3.2. Analysis of Complexity

A good low-rank approximation allows a computational speed-up.

Let us assume a fixed stride of Δ in each spatial dimension.

Table ?? shows the number of multiplications required to perform the convolution. In order to optimize the complexity, it is not always a good idea to decompose the full tensor W . Indeed, depending on its dimensions, the approximation cost might be superior than the original. We might consider instead low-rank approximations of W which partition the coordinate space in the most efficient manner.

3.3. Optimizing Cost with Subspace Clustering

We can decompose the 4-tensor W in a collection $W_{k,l}$ of 4-tensors, by considering a partition $G_1, ..G_k, ..., G_N$ of the first coordinate space C and a partition $H_1, ..., H_l, ..., H_M$ of the last coordinate space F . If we assume a uniform partition with N groups of C/N coordinates and M groups of F/M coordinates respectively, and that each tensor $W_{k,l}$ is approximated with K rank-1 tensors, the resulting complexity is

$$K \cdot N \cdot M \cdot \left(\frac{C}{N} + X\Delta^{-1} + Y\Delta^{-2} + \frac{F}{M}\Delta^{-2} \right)$$

How to optimize the groupings on each of the variables? We perform a subspace clustering.

Sharing between blocks:

$$\widetilde{W} = \sum_{k \leq K} \alpha_{i(k)} \otimes \beta_{j(k)} \otimes \gamma_{h(k)} \otimes \delta_{m(k)}, \quad (3)$$

If now each of the separable filters is taken out of a collection smaller than K , we can gain in computation. This can be for instance implemented with a K-Means on the tensor decompositions.

Examples: Monochromatic filtering

Spatially Separable

Memory access constraints

4. Numerical Experiments

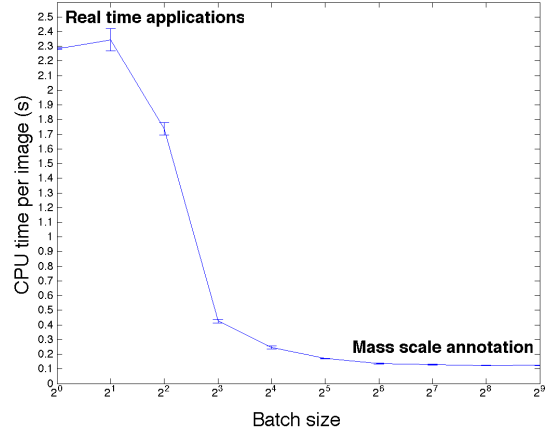


Figure 1. CPU computational time per image for various batch sizes.

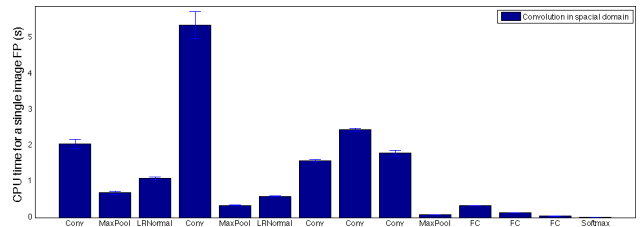


Figure 2. Per layer breakdown of execution time for mini batch of size 128. Such size of mini batch gives optimal per image CPU time.

4.1. Testing time

on GPU: Michael can help.

on CPU

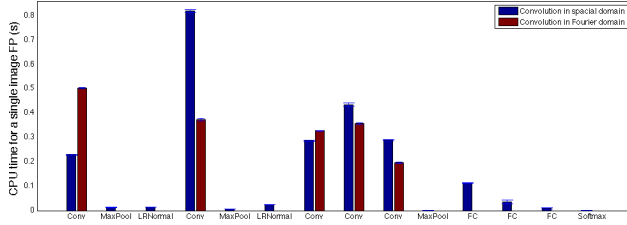


Figure 3. Per layer breakdown of execution for mini batch of size 1. Use for real time applications.

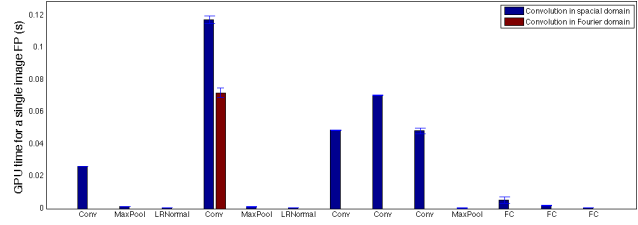


Figure 5. Per layer breakdown of execution time for mini batch of size 128. Such size of mini batch gives optimal per image GPU time.

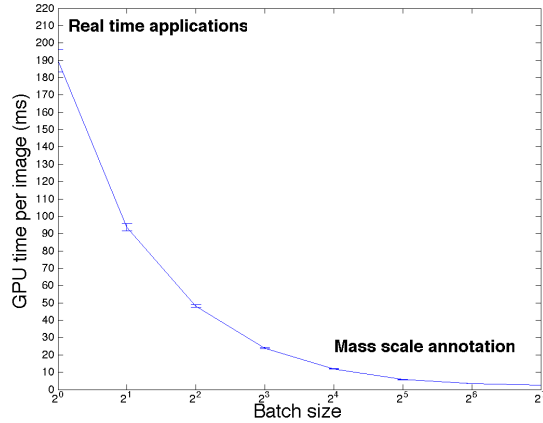


Figure 4. GPU computational time per image for various batch sizes.

4.1.1. MONOCHROMATIC

4.1.2. LINEAR COMBINATION OF FILTERS

4.1.3. SEPARABLE FILTERS

4.2. Denoising (visual inspection, maybe measure)

5. Implications

5.1. Denoising Aspect

we can improve training by simple linear denoting.

5.2. Low-Rank training

Low-rank to avoid over-fitting.

6. Discussion

References

Denil, Misha, Shakibi, Babak, Dinh, Laurent, Ranzato, Marc'Aurelio, and de Freitas, Nando. Predicting parameters in deep learning. *arXiv preprint arXiv:1306.0543*, 2013.

Le, Q. V., Ngiam, J., Chen, Z., Chia, D., Koh, P. W., and Ng, A. Y. Tiled convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2010.

Le, Quoc V, Ranzato, Marc'Aurelio, Monga, Rajat, Devin, Matthieu, Chen, Kai, Corrado, Greg S, Dean, Jeff, and Ng, Andrew Y. Building high-level features using large scale unsupervised learning. *arXiv preprint arXiv:1112.6209*, 2011.

Lowe, David G. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pp. 1150–1157. Ieee, 1999.

Zeiler, Matthew D, Taylor, Graham W, and Fergus, Rob. Adaptive deconvolutional networks for mid and high level feature learning. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 2018–2025. IEEE, 2011.

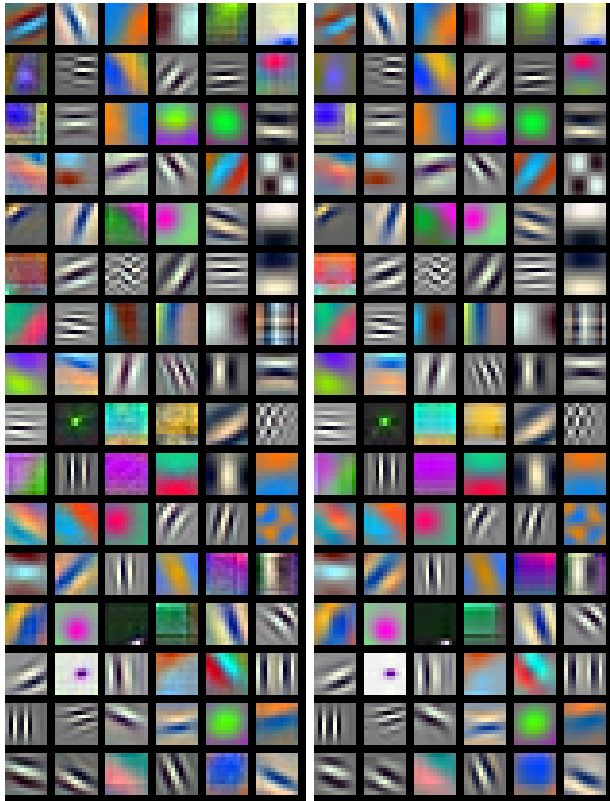


Figure 6. (Left) Original filters, (Right) approximated filters. (this pictures are too big, and should contain white separation).