Supplement to
"Exploiting Linear Structure Within Convolutional Networks
for Efficient Evaluation" NIPS2014

1 Forward propagation time breakdown

Table 1 shows the time breakdown of forward propagation for each layer in the CNN architecture we explored. Close to 90% of the time is spent on convolutional layers, and within these layers the majority of time is spent on the first two.

Layer	Time per batch (sec)	Fraction	Layer	Time per batch (sec)	Fraction
Conv1	2.8317 ± 0.1030	21.97%	Conv1	0.0604 ± 0.0112	5.14%
MaxPool	0.1059 ± 0.0154	0.82%	MaxPool	0.0072 ± 0.0040	0.61%
LRNormal	0.1918 ± 0.0162	1.49%	LRNormal	0.0041 ± 0.0043	0.35%
Conv2	4.2626 ± 0.0740	33.07%	Conv2	0.4663 ± 0.0072	39.68%
MaxPool	0.0705 ± 0.0029	0.55%	MaxPool	0.0032 ± 0.0000	0.27%
LRNormal	0.0772 ± 0.0027	0.60%	LRNormal	0.0015 ± 0.0003	0.13%
Conv3	1.8689 ± 0.0577	14.50%	Conv3	0.2219 ± 0.0014	18.88%
MaxPool	0.0532 ± 0.0018	0.41%	MaxPool	0.0016 ± 0.0000	0.14%
Conv4	1.5261 ± 0.0386	11.84%	Conv4	0.1991 ± 0.0001	16.94%
Conv5	1.4222 ± 0.0416	11.03%	Conv5	0.1958 ± 0.0002	16.66%
MaxPool	0.0102 ± 0.0006	0.08%	MaxPool	0.0005 ± 0.0001	0.04%
FC	0.3777 ± 0.0233	2.93%	FC	0.0077 ± 0.0013	0.66%
FC	0.0709 ± 0.0038	0.55%	FC	0.0017 ± 0.0001	0.14%
FC	0.0168 ± 0.0018	0.13%	FC	0.0007 ± 0.0002	0.06%
Softmax	0.0028 ± 0.0015	0.02%	Softmax	0.0038 ± 0.0098	0.32%
Total	12.8885		Total	1.1752	

Table 1: Evaluation time in seconds per layer on CPU (left) and GPU (right) with batch size of 128. Results are averaged over 8 runs.

2 Theoretical speedups

We can measure the theoretically achievable speedups for a particular approximation in term of the number of floating point operations required to compute the target output. While it is unlikely that any implementation would achieve speedups equal to the theoretically optimal level, the number of necessary floating point operations still provides an informative upper bound on the gains.

Table 2 shows the theoretical speedup of the monochromatic approximation. The majority of the operations result from the convolution part of the computation. In comparison, the number of operations required for the color transformation is negligible. Thus, the theoretically achievable speedup decreases only slightly as the number of color components used is increased.

Figure 1 plots the theoretically achievable speedups against the drop in classification performance for various configurations of the biclustering with outer product decomposition technique. For a given setting of input and output clusters numbers, the performance tends to degrade as the rank is decreased.

3 Combined results

We used the monochromatic approximation with 6 colors for the first layer. Table 3 summarizes the results after fine-tuning for 1 pass through the ImageNet12 training data using a variety of second layer approximations.

Number of colors	Increase in test error			Theoretical speedup
	Original	$\|W\|_{data}$ distance metric	Fine-tuned	
4	24.1%	5.9%	1.9%	2.97×
6	16.1%	2.4%	0.4%	2.95×
8	9.9%	1.4%	0.2%	2.94×
12	3.5%	0.7%	0%	2.91×
16	1.99%	0.8%	-	2.88×
24	1.43%	0.4%	-	2.82×

Table 2: Performance when first layer weights are replaced with monochromatic approximation and the corresponding theoretical speedup. Classification error on ImageNet12 validation images tends to increase as the approximation becomes harsher (i.e. fewer colors are used). Theoretical speedups vary only slightly as the number of colors used increases since the color transformation contributes relatively little to the total number of operations.

Layer 2		Increase in error
Method	Hyperparameters	
Biclustering + outer product decomposition	$G = 48; H = 2; K = 8$	1%
Biclustering + outer product decomposition	$G = 48; H = 2; K = 6$	1.5%
Biclustering + SVD	$G = 2; H = 2; K_1 = 19; K_2 = 64$	1.2%
Biclustering + SVD	$G = 2; H = 2; K_1 = 19; K_2 = 51$	1.4%

Table 3: Cascading approximations.

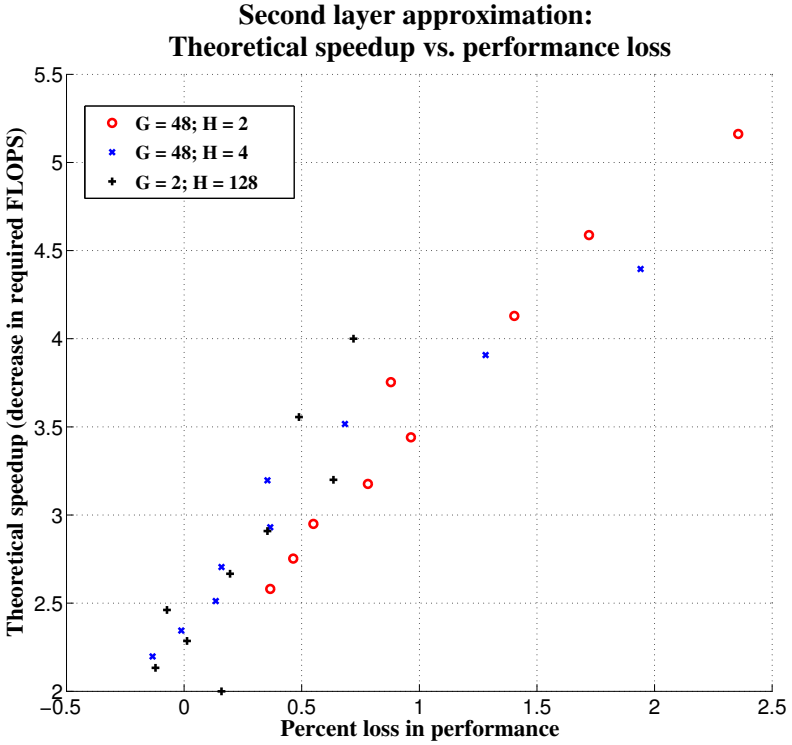

Figure 1: Theoretically achievable speedups vs. classification error for various biclustering approximations.