“Not so fast, FFT”: Winograd

Deep learning thrives on speed. Faster training enables the construction of larger and more complex networks to tackle new domains such as speech or decision making.

Recently, small convolutional filter sizes have become an important component in convolutional neural networks such as Google’s AlphaGo network or Microsoft’s deep residual networks. While most convolutions are computed with the fast fourier transform (FFT) algorithm, the rising prominence of small 3×3 filter sizes makes the way for a lesser known technique specialized for small filter sizes: Winograd’s minimal filtering algorithms (Lavin and Gray, 2015).

We have implemented the Winograd algorithm on GPUs and benchmarked performance and convergence on state-of-the-art networks. Depending on the network architecture, training with Nervana’s Winograd algorithm yields speed-ups of **2-3x** over NVIDIA’s cuDNN v4 kernels.

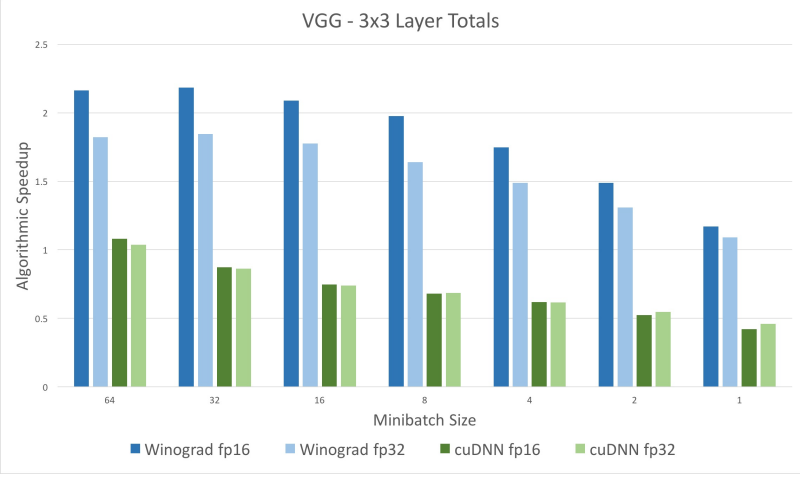
We benchmarked speed on:

- NVIDIA Titan X GPU (fixed at 1 Ghz)
- Intel® Core™ i7 CPU 975 @ 3.33GHz

with several convolutional networks operating on the Imagenet dataset: VGG, GoogleNet, Microsoft’s deep residual networks. We tested different minibatch sizes and compared the computation time of the 3×3 layers between Nervana Winograd and NVIDIA cuDNN v4.

Performance was measured in units of algorithmic speedup. Computation speed (e.g. images/second) was normalized to the maximum theoretical speed of the direct convolutional approach. Values above one indicate how efficiently a faster algorithm is implemented.

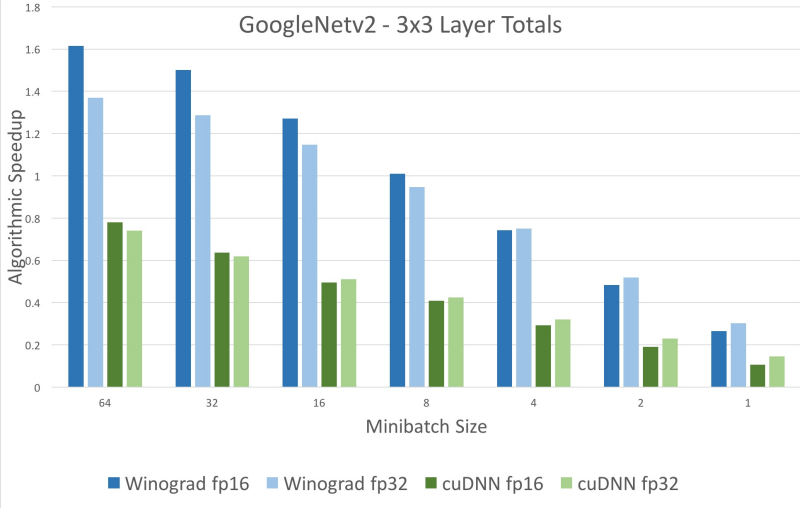
For the 3×3 layers of the VGG model, Nervana Winograd is up to 3 times faster than cuDNN v4 (see Figure below). These speed-ups were also realized when we measured the end-to-end forward propagation and backward propagation times.

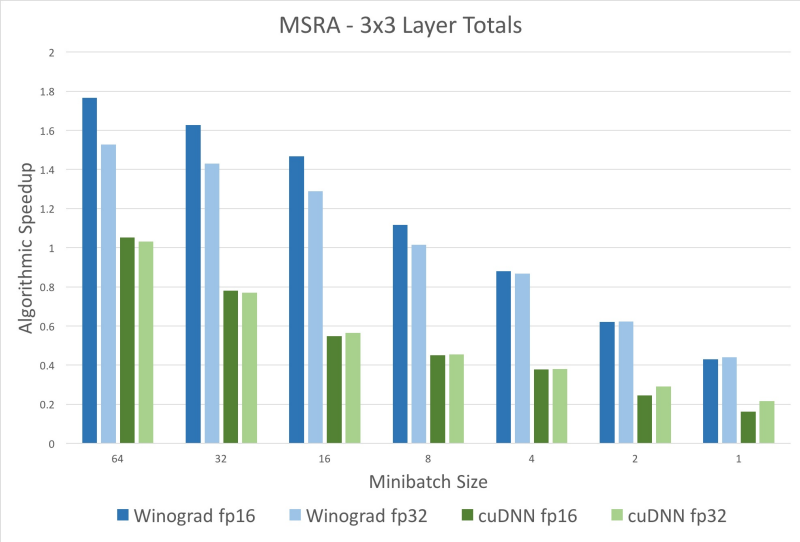


Categories

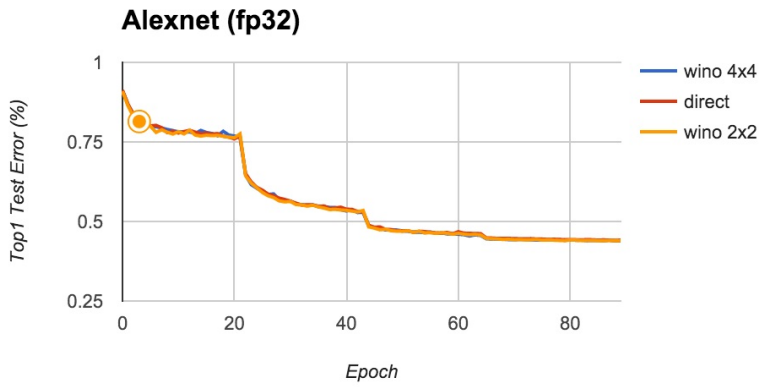
- Developer
- General
- Impact
- Life at Nervana
- Product

We obtained similar results with GoogleNetv2 and MSRA networks.





Not only is Nervana Winograd fast, but also numerically accurate. For example, AlexNet convergence is exactly the same as with direct convolution, as shown below.



These results are also confirmed by third-party benchmarking. Our Winograd implementation is open-source, and can be found in the latest release of our deep learning library, neon. Look forward to a forthcoming part 2 with more technical details on Nervana Winograd. We are actively working on improving Nervana Winograd to allow your networks to train faster and handle more complex problems.

References

Lavin, Andrew and Gray, Scott (2015). Fast Algorithms for Convolutional Neural

Networks. http://arxiv.org/abs/1509.09308

By Scott Gray | Thursday, March 3, 2016 | Developer

Share This Story, Choose Your Platform!

About the Author: Scott Gray

Scott brings over 15 years of experience in software engineering and web infrastructure at a variety of startups and larger companies. His hobbies include computational neuroscience and designing artificial intelligence algorithms for the game of Go. Recently before joining Nervana, he wrote a custom built assembler for the NVIDIA Maxwell architecture to achieve state-of-the-art performance in numerical linear algebra.