OpenCL Accelerated Deep Learning for Visual Understanding

Jeremy Bottleson, Sungye Kim, Jeff Andrews, Preeti Bindu, Deepak N. Murthy, Joseph Spisak, Jingyi Jin
Graphics Initiatives Team, VPG, Intel Corporation

BACKGROUND

Visual understanding is a broad term used to describe efforts to extract meaning and knowledge from images computationally. Much effort has been spent by researchers to develop methods to allow for information to be gained from images the way humans do. However, it has proven to be a very difficult problem. Since the 1950s researchers have been attempting to come up with better methods of understanding images. The results however were generally very limited in application and far worse than what an average person was capable of.

In recent years some major advances in the field are seeing vision systems come about that are performing well at general recognition in images for large sets of objects. All of these new systems have been based around the concepts of convolutional neural networks (CNN) and deep learning. Convolutional neural networks which are a more specific form of the general neural network were first proposed in 1980 [1]. They were later improved by LeCun et al. [2] in 1998 to create Lenet-5 which was shown to be trainable to recognize hand written digits. Then in 2012, Krizhevsky et al. [3] made another break through in applying concepts of CNN to image classification, and won the ImageNet competition with the highest hit score. Nowadays, as of 2015, Baidu, Google and Microsoft have announced that they have deep learning systems that outperform humans on the ImageNet challenge.

DEEP LEARNING FRAMEWORK: CAFFE

Caffe [4] is an open source software framework for easily designing and testing CNN networks. It was developed by the Berkley Vision and Learning Center. It is open source and very popular having been forked over 1,000 times during its first year. In addition it has been used for many of the ImageNet competition entries. Caffe currently supports both CPU and GPU for forward/backward passes, as well as options for using several different CPU BLAS libraries and CUDA/cuDNN based GPU implementations.

MOTIVATION & GOAL

Much of the research and software development focus related to CNNs has been focused so far on developing better and faster methods for training networks. Because of this most software frameworks for CNN are targeted only at high end discrete GPUs. That focus however is beginning to expand, as networks are now available that can perform well on many vision tasks, it is now desirable to start creating applications which use CNN based vision systems. This presents new challenges however as the devices which are most likely to be used for classifying and detecting objects in a video stream are mobile power constrained platforms that typically do not have a large discrete GPU available.

We see this as a great opportunity to leverage smaller, more power efficient GPUs which are often found in mobile devices. One of the hurdles we identified to this use case however is that almost all efforts and software frameworks for CNNs thus far have focused solely on discrete GPUs and proprietary APIs for accessing them.

As a first target to explore the advantages of utilizing integrated GPUs to accelerate CNN processing we decided to target a very popular open source framework, Caffe, and create an OpenCL accelerated branch.

	Calories	
🟩	toast	72/slice
🟨	salmon	235/125g
🟧	cake	240/unit
🟥	corn	124/unit
🟪	bread	95/unit
🟥	tomato	16/unit

This is an example of the type of future use case we hope to enable.

REFERENCES

[1] K. Fukushima. "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position". Biological cybernetics, 1980

[2] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition". IEEE, 1998

[3] A. Krizhevsky, I. Sutskever, and G. Hinton. "Imagenet classification with deep convolutional neural networks". NIPS, 2012.

[4] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. "Caffe: Convolutional architecture for fast feature embedding". arXiv preprint arXiv:1408.5093, 2014.

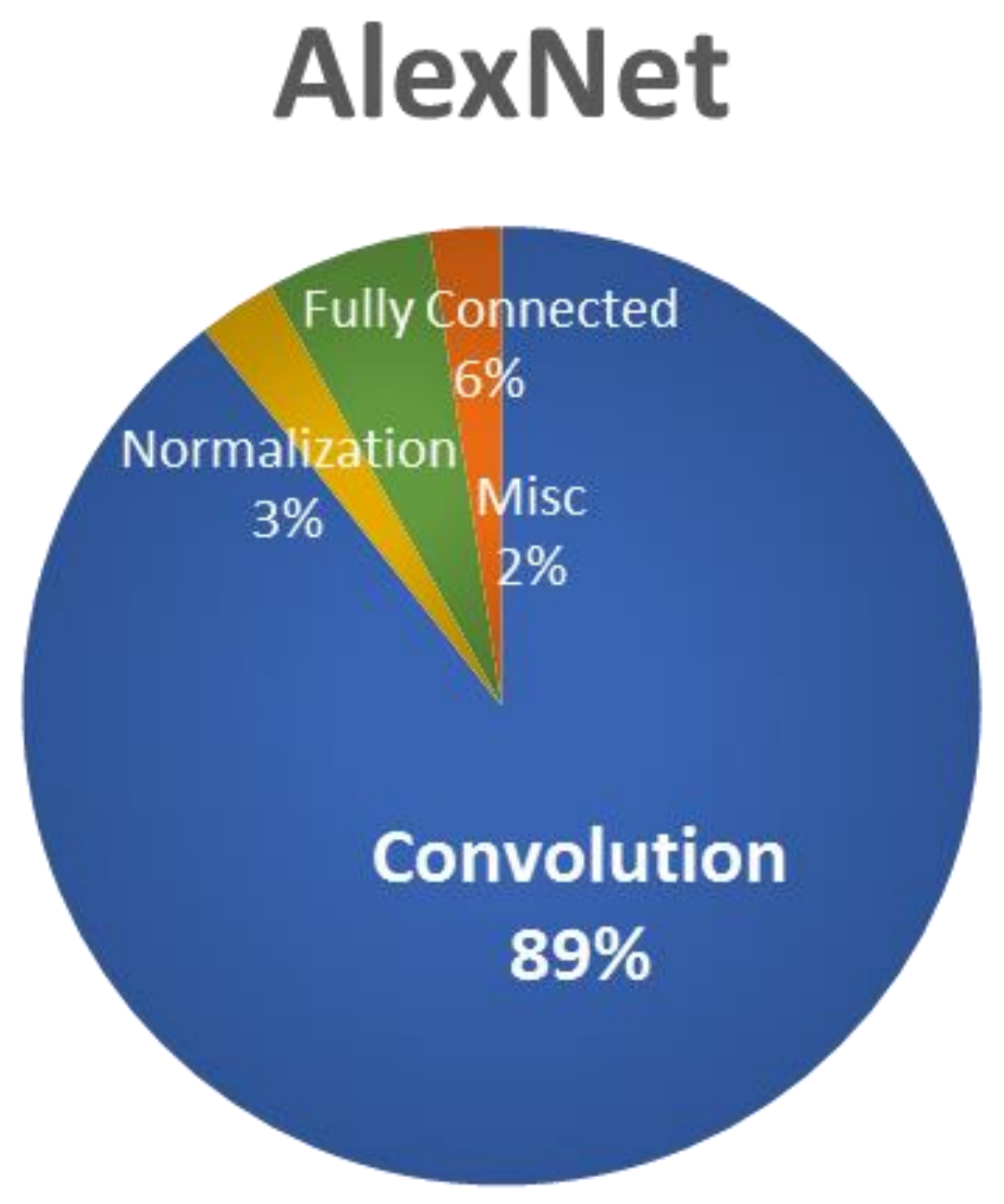
OPENCL ACCELERATED CNN

By targeting our OpenCL support to Caffe we enable researchers to start testing and designing on a much larger range of hardware through the support of an open standard. In particular we see compelling advantages to utilizing the GPU in power constrained mobile devices to enable exciting new use cases. Below shows a brief list what we have done in our OpenCL accelerated Caffe framework.

1. Enabled OpenCL GPU support in Caffe
2. Passed all existing unit tests in Caffe
3. Support multiple convolution approaches: GEMM, Spatial, FFT-based
4. Employed OpenCL based math libraries (clBLAS, clFFT, Random123)

*(*We appreciate open source effort for OpenCL based math libraries.)*

To enable this functionality we utilize the clBLAS library as well as clFFT. In order to increase performance even further we are implementing a number of optimization efforts focused primarily on the convolution layer as our profiling identified it as the largest bottleneck. The profiling chart below shows the amount of time spent in each layer for AlexNet in our OpenCL implementation.

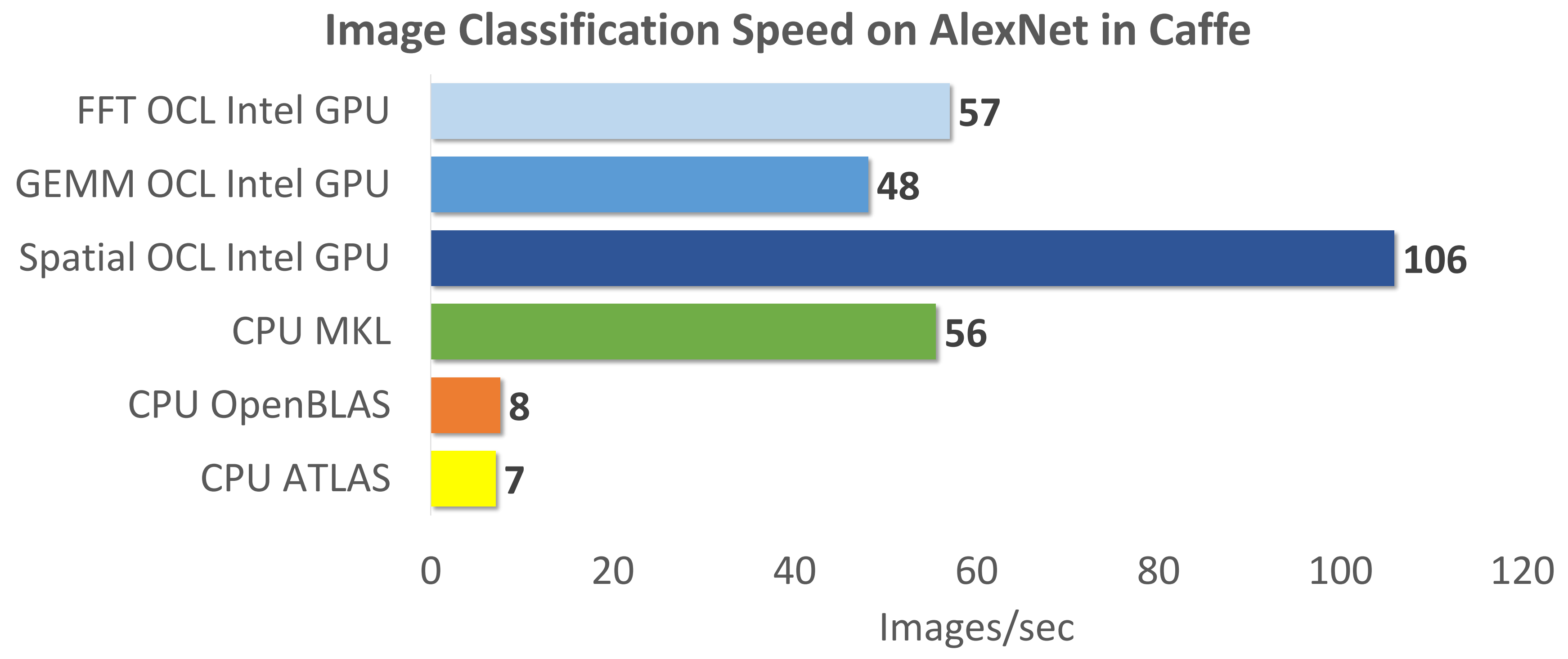


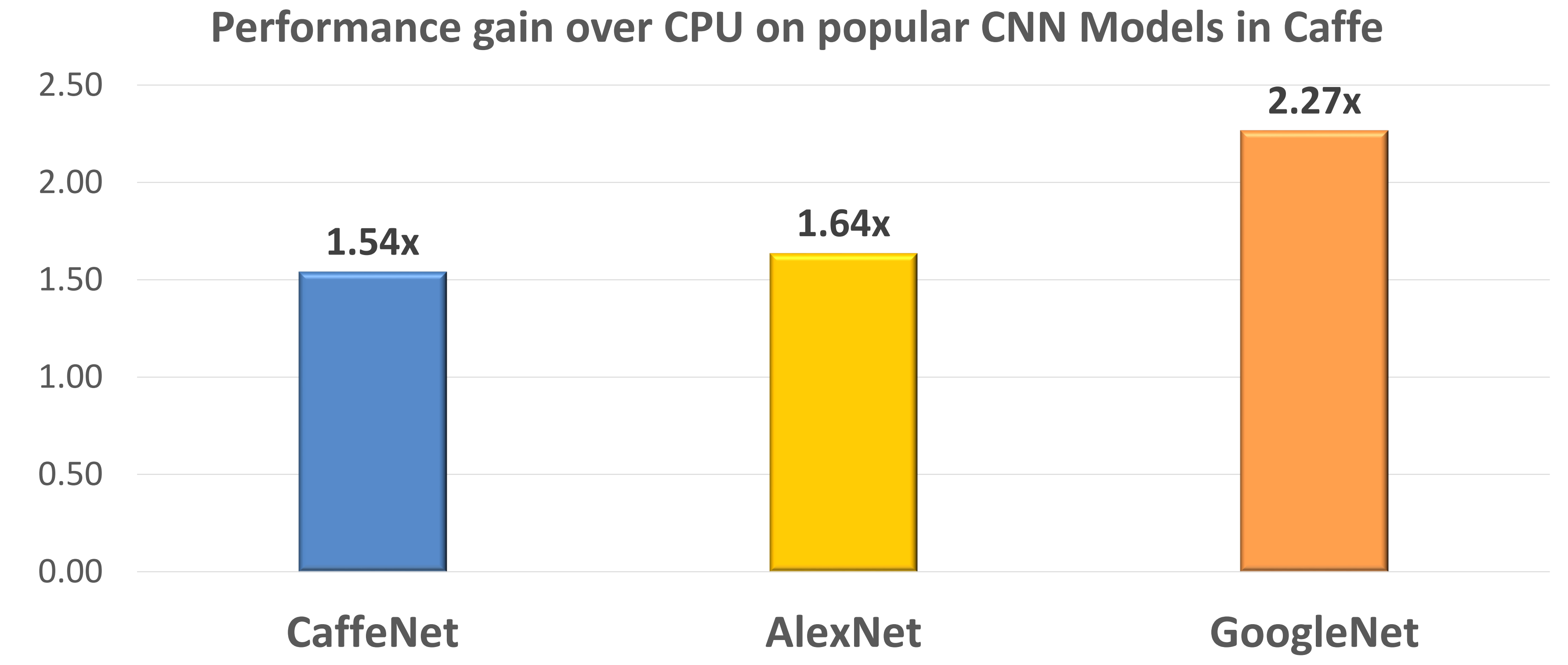
Compute profile of AlexNet in Caffe

Layer	Mathematical backend
Convolution	GEMM (Matrix multiplication)
Fully connected	Inner product

EARLY PERFORMANCE RESULTS

While still in the early stages we are already at a point in which we are seeing very promising results from testing on our Intel Integrated GPU equipped systems. Our current prototype is already passing all unit tests and fully functional. In performance comparisons we are seeing a large performance increase in classification speed when compared to the CPU in the same system. We expect to see even larger increases as the OpenCL approach is modified to leverage more advanced features and OpenCL libraries become more mature.





*All tests performed on i7-4960HQ processor.
*Caffe CPU mode with Intel Math Kernel Library v.11.2 (1.x)

*Code drop is available by request, please contact: jingyi.jin@intel.com

IWOCL 2015
Stanford University, USA | 12-13 May