IMPLEMENTATION OF EFFICIENT, LOW POWER DEEP NEURAL NETWORKS ON NEXT-GENERATION INTEL CLIENT PLATFORMS

Michael Deisher, Andrzej Polonski

Intel Corporation

ABSTRACT

In recent years many signal processing applications involving classification, detection, and inference have enjoyed substantial accuracy improvements due to advances in deep learning. At the same time, the "Internet of Things" has become an important class of devices. Although the paradigm of local sensing and remote inference has been very successful (e.g., Apple Siri, Google Now, Microsoft Cortana, Amazon Alexa, and others) there exist many valuable applications where sensing duration is very long, the cost of communication is high, and scaling to millions or billions of devices is not practical. In such cases, local inference "at the edge" is attractive provided it can be done without compromising accuracy and within the thermal envelope and expected battery life of the edge device.

This demonstration presents a very low power neural network co-processor capable of continuous inference on local battery powered devices. A pre-production reference platform will be demonstrated. Continuous acoustic model likelihood scoring will be shown using models trained with the open source Kaldi framework. Acoustic likelihoods will be presented in a real-time spectrogram-like display (likelihood-o-gram). First, a speed test will be shown pitting the application processor (e.g., CPU, DSP, or GPU) against the neural network co-processor. Figure 1 shows the elements of the demo display. The application processor and accelerator block will be used to score acoustic feature vector inputs as fast as possible while the scrolling output is displayed. A stopwatch timer will be shown for each as a way of comparing performance. A resource monitor showing application processor utilization will also be shown. Second, real-time scrolling likelihood-o-gram displays will be shown as live audio input from the microphone is processed by feature extraction and neural network acoustic model likelihood scoring. Participants will visualize the acoustic likelihood scores corresponding to their speech as they speak into the microphone. Lightweight fully connected networks (i.e., DNN) as well as larger more complex networks (deep CNN, RNN, LSTM) will be selectable. A brief presentation of the programming model and best known methods will be given. Examples using the Intel® Deep Learning SDK will be shown.

The value of this demonstration to participants is several-fold. It introduces researchers and practitioners to early technology showing that commodity hardware for remote, highly accurate, low power, continuous inference is around the corner. It allows them to familiarize themselves with the technology, tools, and infrastructure in advance. Also, real-time likelihood-o-gram is introduced as a way to visualize neural network performance on live audio. Techniques for rapid prototyping of efficient neural network based applications will be discussed giving plenty of opportunity for brainstorming and sharing of ideas. Participants will gain knowledge that can be applied across future Intel embedded, mobile client, and PC/workstation platforms.

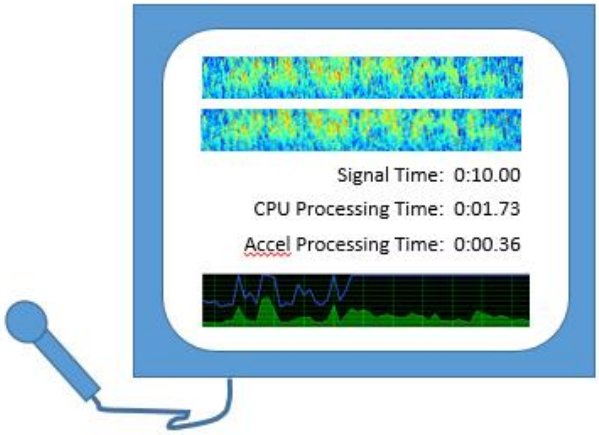

Figure 1. Representation of demo screen.

Index Terms— neural network, hardware, visualization

1. REQUIREMENTS

This demonstration requires a table at least 3 feet by 6 feet and a power strip with 5-6 power outlets. Space to hang a standard ICASSP poster is also requested.