Deep Learning Working Group
Statement of Work

August, 2016

Industry Requirement

Deep learning is a rapidly evolving field in modern signal processing, with diverse applications including object detection, gesture tracking, advanced driver assistance, to name a few. Several institutes and companies have released their deep learning toolkits to the public for open use, such as Caffe, TensorFlow, Theano, Torch, Chainer and CNTK – mostly on the server side.

Many modern client devices would also be capable of carrying out such computations, where applications often require deep learning algorithms to work in real-time, processing data obtained from sensors on the fly.

Several hardware vendors have developed accelerated deep learning libraries for their client side products. While most of these companies develop their own APIs, they have to build proprietary data import interfaces to be able to leverage results achieved in the server side frameworks. To complicate the problem, the transfer of data structures from one server side framework to another is also not straightforward.

An open standard that defines common data exchange mechanism would aid the development process from both the viewpoint of hardware vendors and the framework/application developers in creating cross-platform deep learning applications.

Proposed Solution and Impact

Our proposal is to create an open, royalty-free, API independent standard file format for exchanging deep learning data. The standard would specify neural network structure and data exchange formats and commonly used operations and their formal semantics for cross-compatibility between frameworks and to facilitate deployment from frameworks to embedded systems. For example, models trained on a server system could be run on a mobile device, ensuring equivalent computation. The standard would allow hardware vendors to implement optimized and competitive deep learning inference engines. Hardware vendors that comply to the standard and provide import functionality of the file format would not need to work on compatibility with frameworks, but could achieve compatibility with all frameworks through the standard. At the same time, framework developers that comply to the standard and provide export to the file format would ensure that their result is compatible with various embedded devices. Executing trained neural networks on different hardware platforms would become easier.

Design Philosophy

The Deep Learning working group may consider using and refining the following design goals and principles for the proposed standard:

- Keep the standard independent of frameworks and implementing APIs, ensure that the data format is portable across multiple implementing inference engines
- Keep the primary focus on enabling neural network inference on mobile and embedded devices

- Ensure that efficient execution of the neural networks described by the file format is possible on a wide variety of computing devices, including CPUs, GPUs, DSPs, FPGAs and dedicated accelerators
- Ensure that it is possible for one or more Khronos groups (such as OpenVX and OpenCL) to provide efficient conformant runtime implementations
- Ensure that efficient importing of the data format is possible on Khronos APIs
- Aim to design the standard to be future proof in the sense that it would possibly accommodate newly emerging neural network architectures with minimal extension
- Keep the scope of the first version focused on operations and algorithms required for state-of-the-art deep learning with convolutional neural networks, at the same time being open ended to support a wide range of neural architectures that are built from the same components (such as MLPs, RNNs, LSTMs, RBMs, Auto-Encoders, Encoder-Decoder architectures)
- Focus on the main use case of exporting neural networks trained in public deep learning frameworks and importing them to hardware vendor APIs, at the same time be open ended to let the standard be useful for data interchange among frameworks
- Ensure that the file format is prepared for varying precision requirements from the hardware vendors' side, and is prepared for vendor specific extensions

The Deep Learning working group may consider refining and extending the following aspects of an API independent Deep Learning standard

- The use of a computational graph approach to characterize neural network structure
- Enabling the expression of operations at multiple granularities to facilitate the possibility of optimized execution on various hardware
- The formal specification of a set of standardized basic operations required for the expression of various deep learning algorithms
- A textual format that facilitates the description and interchange of computational graphs (neural network structure), using the above mentioned operations
- A set of standardized compound operations that are built from the above mentioned basic operations using the above mentioned textual description format, to facilitate the expression of computations at multiple granularities
- A multidimensional array data structure that takes varying precision requirements into account for interoperable data exchange
- A binary data format (file) using the above mentioned multidimensional array format for interchanging neural network data (weights)

Deliverables

The Deep Learning working group will produce the following deliverables:

- Specification of the API independent data file format and the formal specification of the operations contained within
- Conformance test for the standard
    - File format validation
    - Functional conformance that ensures implementability and output that matches the formal specification up to (possibly multiple) precision level(s) by multiple APIs

Industry Support

It is expected that hardware vendors adopt the standard to deliver efficient deep learning acceleration, and that companies use it for creating high-level algorithms, libraries and applications. Several companies have already expressed their interest including AdasWorks, AMD, MediaTek, Cadence, Axis, Qualcomm, Intel, NXP and Movidius. We would expect strong interest among companies, active in deep learning specific hardware design, to join Khronos to help define this specification.

At the same time, support from framework vendors is also required. It is known to us that the utility of a standard to exchange neural network data among frameworks have already been considered by the deep learning community. It must be ensured that no work is duplicated in this area, so it is important to reach out to framework vendors to ask about their interest and for their input.

A Preliminary Call for Participation summarizing the approach and design principles in this document was circulated among deep learning framework developers to ask them for feedback. All major frameworks responded supporting the idea of creating a standardized exchange format for neural networks. The most important advice we received is that the format has to be general enough to support future network models and computations to avoid becoming obsolete and that it has to support the description of operations on multiple levels of granularity in order to facilitate optimizations on various hardware.

It is expected that framework developers implement export functionalities to the standard exchange format, although their cooperation is not strictly required, since third party conversion tools may also be developed.

Khronos Infrastructure

This is expected to be a typically-organized, medium-sized working group, between 10-20 members. Efficient communication and collaboration with the OpenVX Khronos group must be established in order to avoid duplication of work among the two groups. The OpenVX group would work on an API of an inference engine that can utilize the file formats defined by the Deep Learning group, and compatibility should be ensured.

Milestones

It is suggested that the Deep Learning working group refine the following milestone plan as discussions develop (starting after the SoW is accepted, expected to be finished by CVPR June 2017):

1. 6 months – Agree on the description methods used by the exchange format
    a. Format to describe network structure
    b. Format to describe network weights
2. 9 months – Definition of the API independent standard
    a. Specification of data formats, operations, file formats
    b. Definition of its relation to and compatibility with OpenVX
    c. Specification of conformance tests
3. 9 months – Implementation of the conformance tests
    a. Validation tests
    b. Functional conformance tests