Khronos Group Releases NNEF 1.0 Standard for Neural Network Exchange

BEAVERTON, OR, UNITED STATES, December 20, 2017 /EINPresswire.com/ -- The Khronos™ Group, an open consortium of leading hardware and software companies, an open consortium of leading hardware and software companies creating advanced acceleration standards, announces the release of the Neural Network Exchange Format (NNEF™) 1.0 Provisional Specification for universal exchange of trained neural networks between training frameworks and inference engines. NNEF reduces machine learning deployment fragmentation by enabling a rich mix of neural network training tools and inference engines to be used by applications across a diverse range of devices and platforms. The release of NNEF 1.0 as a provisional specification enables feedback from the industry to be incorporated before the specification is finalized — comments and feedback are welcome on the NNEF GitHub repository.

The goal of NNEF is to enable data scientists and engineers to easily transfer trained networks from their chosen training framework into a wide variety of inference engines. A stable, flexible and extensible standard that equipment manufacturers can rely on is critical for the widespread deployment of neural networks onto edge devices, and so NNEF encapsulates a complete description of the structure, operations and parameters of a trained neural network, independent of the training tools used to produce it and the inference engine used to execute it.

“The field of machine learning benefits from the vitality of the many groups working in the field, but suffers from a lack of common standards, especially as research moves closer to multiple deployed systems,” said Peter McGuinness, NNEF work group chair. “Khronos anticipated this industry need and has been working for over a year on the NNEF universal standard for neural network exchange, which will act as the equivalent of a pdf for neural networks.”

NNEF has been designed to be reliably exported and imported across tools and engines such as Torch, Caffe, TensorFlow, Theano, Chainer, Caffe2, PyTorch, and MXNet. The NNEF 1.0 Provisional specification covers a wide range of use-cases and network types with a rich set of functions and a scalable design that borrows syntactical elements from Python but adds formal elements to aid in correctness. NNEF includes the definition of custom compound operations that offers opportunities for sophisticated network optimizations. Future work will build on this architecture in a predictable way so that NNEF tracks the rapidly moving field of machine learning while providing a stable platform for deployment.

Khronos has initiated a series of open source projects, including a NNEF syntax parser/validator and example exporters from a selection of frameworks such as TensorFlow, and welcomes the participation of the machine learning community to make NNEF useful for their own workflows. In addition, NNEF is working closely with the Khronos OpenVX™ working group to enable ingestion of NNEF files. The OpenVX Neural Network extension enables OpenVX 1.2 to act as a cross-platform inference engine, combining computer vision and deep learning operations in a single graph.

Industry Support Quotes:
“AImotive is proud to be a key player in the development and early deployment of NNEF, not only as instigators of the new standard, but also delegating AI Researcher Viktor Gyenes as specification

editor," said Marton Feher, head of hardware engineering at AImotive. "Having been an early adopter in both our hardware and software technologies for autonomous driving, we fully recognize the importance of having a neutral exchange format for Neural Networks. As the number of development frameworks expands, and the range of execution platforms grows and diversifies, the ability to freely move network topologies and weights from one environment to another is essential for innovation and freedom of supplier choice."

"As we move Deep Learning from the lab to rich customer driven applications, we as an industry needed to drive a Deep Learning interchange solution," said Greg Stoner, CTO Systems Engineering, Radeon Technologies Group, AMD. "We are happy to see Khronos Group's roll out the NNEF specification to support easily moving neural networks between training frameworks and inference engines built on OpenVX."

"Standardizing a format for the interchange of neural network models is a significant step towards improving the portability and optimization of networks and operators between different frameworks, tools, and inference engines," said Robert Elliott, technical director of software, Machine Learning, Arm. "Arm supports NNEF's development, further enabling framework and tool developers to produce models that are run and validated on the wide array of processors and accelerators available in the Arm® ecosystem."

"As neural network processing migrates from the cloud to mobile and edge devices, the need for unified representations of these models increases and helps companies like Qualcomm Technologies focus on delivering the best platform for executing these models on device," said Jeff Gehlhaar, VP of Technology, Qualcomm Technologies, Inc. "As a Khronos Member, Qualcomm Technologies believes consolidation will help growth in this area, and is supportive of standards for representation of neural network models, such as Khronos Neural Network Exchange, which streamlines the migration from cloud to device."

"The NNEF standard and OpenVX integration enables a fast path for deployment of neural networks from a wide collection of training frameworks on our OpenVX enabled GPU, Vision and Neural Net processing IP," said Weijin Dai, chief strategy officer, executive vice president and GM of VeriSilicon's Intellectual Property Division. "By using the common exchange format and our runtime inference, which is optimized for dedicated VeriSilicon IP across all performance tiers, our customers are able to achieve optimal performance on their platform instantaneously, regardless of their chosen training framework."

The new NNEF 1.0 documentation project and specification are available on the Khronos Registry. NNEF open source tools and projects are available on the Khronos NNEF Tools repository. More information on the OpenVX run-time API for vision acceleration and inferencing can be found here. More details on how NNEF provides flexibility for handling flat and compound operations can be found here: https://www.khronos.org/blog/nnef-design-philosophy-network-structure-and-target-use-cases

For more information about The Khronos Group visit Khronos.org.

About The Khronos Group
The Khronos Group is an industry consortium creating open standards to enable the authoring and acceleration of parallel computing, graphics, vision and neural nets on a wide variety of platforms and devices. Khronos standards include Vulkan®, OpenGL®, OpenGL® ES, OpenGL® SC, WebGL™, SPIR-V™, OpenCL™, SYCL™, OpenVX™, NNEF™, COLLADA™, OpenXR™ and glTF™. Khronos members are enabled to contribute to the development of Khronos specifications, are empowered to vote at various stages before public deployment, and are able to accelerate the delivery of their cutting-edge accelerated platforms and applications through early access to specification drafts and

conformance tests.

###

Alexandra Crabb
Caster Communications, Inc.
4013182229
email us here
