Morpho Announces Commercialization of "SoftNeuro™", One of the World's Fastest Deep Learning Inference Engines, Accelerating the Practical Use of Deep Learning

Tokyo, Japan –December 5, 2017– Morpho, Inc. (hereinafter, "Morpho"), a global leader in image processing solutions, announced today the commercialization of "SoftNeuro™", one of the world's fastest*1 deep learning*2 inference engines. Morpho not only licenses the engine alone, but also brings about a large increase in processing speed of their existing engines such as "Morpho Deep Recognizer ™", which is Morpho's image recognition engine, by embedding "SoftNeuro™" into them.

Recently, the development of the deep learning technology has been promoted for artificial intelligence. It has been successfully incorporated into services and products in a variety of fields and its practical use is increasing. However, opportunities to execute inference processing are increasing in the edge environment, and it is evident that there is an issue with the constraints of operating environments and inference speed. Therefore, Morpho developed "SoftNeuro™", a fast inference engine that operates in many environments by utilizing learning results that have been obtained through a variety of deep learning frameworks.
"SoftNeuro™" is a general-purpose inference engine and is available for voice recognition and text analysis as well as image recognition. Therefore, not limited to image recognition. Developer licensing is planned for a variety of inference engines that use deep learning.

"SoftNeuro™" is scheduled to be displayed and demonstrated at Morpho's booth at INTERNATIONAL TECHNICAL EXHIBITION ON IMAGE TECHNOLOGY AND EQUIPMENT 2017 to be held on December 6 to 8, 2017.

【Features of "SoftNeuro™", a deep learning inference engine】

1.One of the fastest in the world (According to Morpho's research as of December 5, 2017) *1
In some applications of inference processing with deep learning, lengthy processing time has become an issue.
"SoftNeuro™" is faster than or equally fast as the mainstream inference engines - when run on CPUs - while providing a set of key features as described later. This high-speed performance has been achieved through a variety of optimizations (neural network, memory usage and others) for each platform.
Please refer to *1 for details.

2.Supports multiple frameworks
There are many frameworks that perform deep learning, including Caffe, Keras and TensorFlow™, which are open source software.
"SoftNeuro™" achieves fast processing by utilizing the learning results of these major frameworks (Fig. 1). It is possible to achieve fast inference processing (the first characteristic) and multi-platform compatibility (the third characteristic) without wasting the learning assets that users have built up so far. Further, the compatibility of "SoftNeuro™" with frameworks and layers will be expanded sequentially.

3. Multi-platform compatibility

Inference processing using deep learning is widely being applied to a variety of places, including smartphones, vehicles and FA equipment, not limited to cloud servers. In these operating environments, different platforms are used, including CPUs and OSs, therefore porting and optimization are necessary. “SoftNeuro™” is scheduled to be applied to a variety of platforms, and appropriate optimization (CPU speed-up instructions, use of GPU and DSP and others) for each platform will be performed.
Multi-platform compatibility enables a flexible response to changes in operating platforms as well as expanded learning results for a wide range of operating platforms.

4.Compatible with secure file formats

Inference using deep learning is widely being applied to a variety of places, including smartphones, vehicles and FA equipment, not limited to cloud servers. To achieve their operation, a network is copied to many places after learning. This increases the risk of leaking learning know-how or results (original network structure, weight parameters and others).
“SoftNeuro™” is capable of encrypting the trained networks, minimizing the risk of leaking the machine learning know-how and the results of learning.

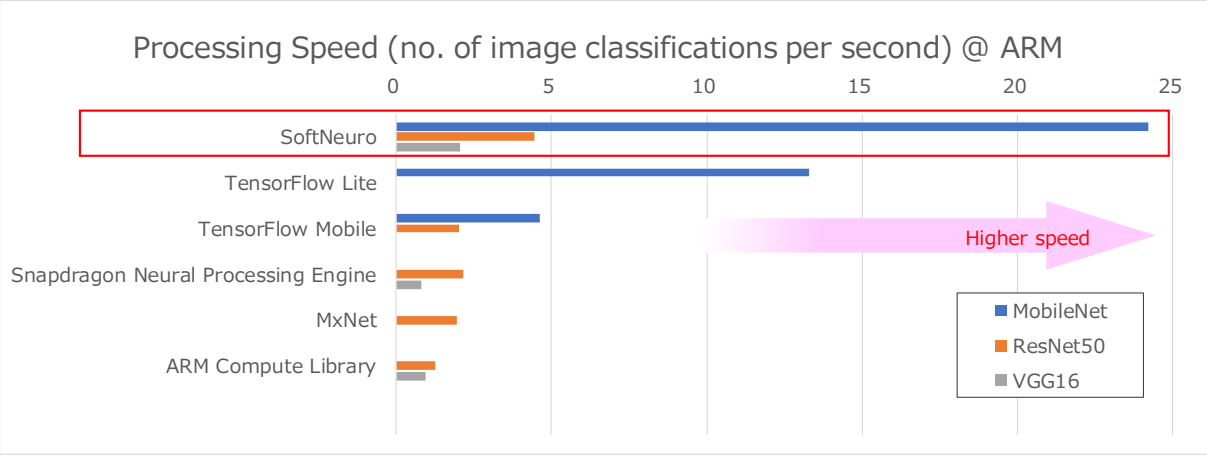
※1 One of the Fastest in the world (According to Morpho’s research in December 2017)

We compared the inference speed of “SoftNeuro™” with that of widely-used inference engines that we could obtain, on ARM and Intel CPU architectures (“SoftNeuro™” does not utilize GPU). Table 1 summarizes the evaluation conditions. Figures 1-1 and 1-2 present the results of the evaluation for these two architectures.

Table 1. Conditions for Processing Speed Comparison

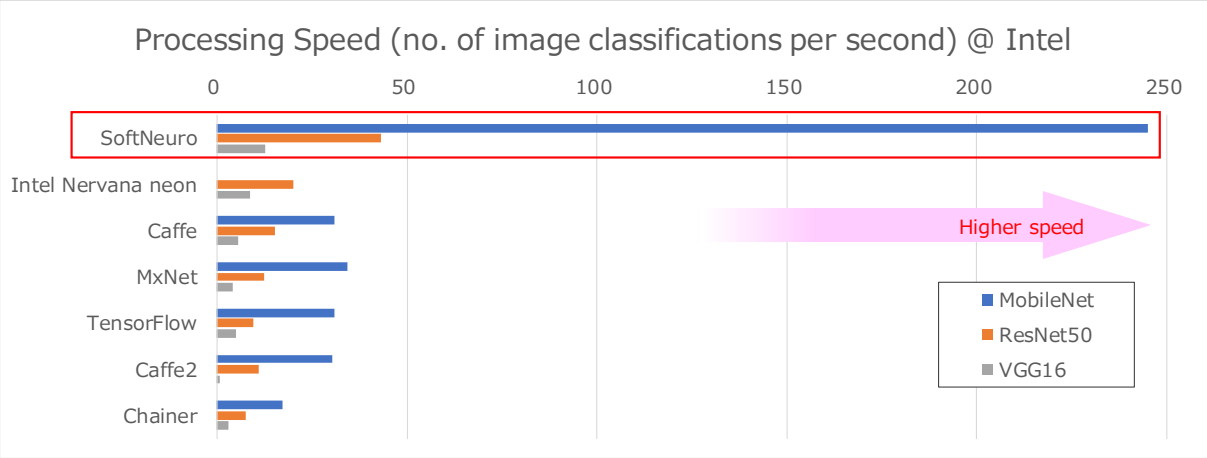
Criterion	Description
Models used for processing speed comparison (Image Classification on 1,000 categories)	➢ MobileNet ➢ ResNet50 ➢ VGG16
Image size used for inference	224x224 pixels

Figure 1-1. Comparison of Inference Speeds (CPU: ARM)
Measured on: Qualcomm Snapdragon 835



*bars were not plotted for models that could not be implemented using a given inference engine, on this CPU.

Figure 1-2. Comparison of Inference Speeds (CPU: Intel)
Measured on: Intel® Core™ i7-6700K CPU @ 4.00 GHz

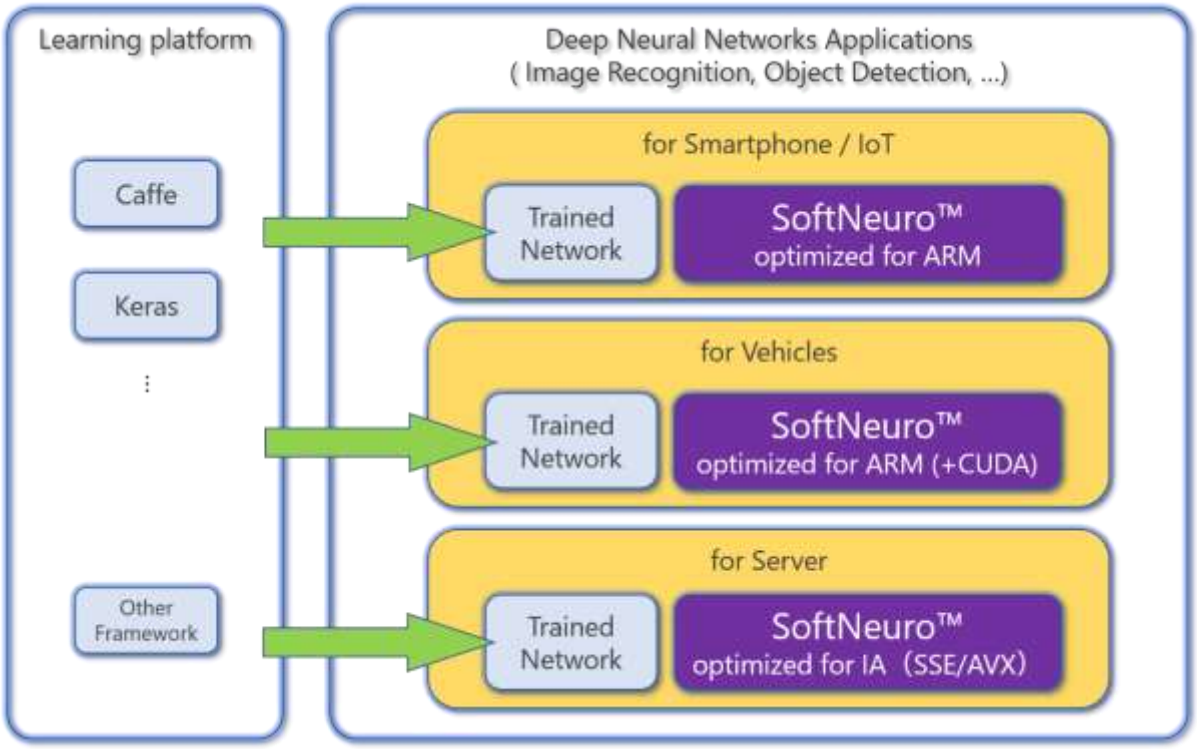


*bars were not plotted for models that could not be implemented using a given inference engine, on this CPU.

***2 About Deep Learning**

Deep learning is a learning method in which artificial intelligence extracts characteristics from learning data, unlike conventional machine learning, which performs recognition based on characteristics defined by a human. Deep learning has mainly been applied to image recognition, voice recognition and language processing. It is popular technology that is expected to be utilized in a variety of fields, including marketing, security, automatic translation and automated driving.

Figure 2. Integrating "SoftNeuro™" into Machine Learning Systems



About Morpho, Inc.
Established in 2004, Morpho, Inc. has built substantial brand recognition in the field of software image processing for mobile devices. Our mission is to provide an environment where a creative group of individuals can develop new imaging technologies, and to introduce innovative technologies in a practical form that contributes to technological development and cultural enrichment. For more information, visit http://www.morphoinc.com/en/ or contact m-info-pr@morphoinc.com.

*Caffe、Keras、TensorFlow™ are software frameworks used for learning and inference of Deep Learning based systems.
*ARM is a CPU architecture, designed by ARM Holdings and widely used in smartphones and embedded systems.
*CUDA is a parallel computing platform and programming model developed by NVIDIA Corporation for general computing on graphical processing units (GPUs).
*IA (Acronym for Intel Architecture) is a generic name for the basic architecture that is used in microprocessors and supporting hardware developed by Intel Corporation.
* SSE/AVX are extensions to the x86 instruction set architecture for microprocessors from Intel and AMD proposed by Intel corporation.