Grovety Inc.
People and hardware connecting technologies

Translator of NN Algorithms for new AI Accelerator

ORGANIZATION:

The customer develops and manufactures lines of processors and systems on a chip (SoC), based on the NeuroMatrix, Arm, and PowerPC architectures, navigation modules, and neural networking embedded solutions.

INDUSTRY:

semiconductors

CHALLENGE:

The customer has 30+ years of deep expertise in SoC creation and operation, including AI chips.

To expand the customer base, the customer has created an AI chip (NN accelerator) with a new architecture. The software developed for the previous generation AI chips couldn't be applied as-is, while the adaptation required significant investments. Therefore, the customer decided to focus on the NN accelerator design itself and outsource the software infrastructure development.

The Grovety team faced the challenge of selecting a framework and adapting it to the customer's new hardware architecture. The framework must have met the following requirements: support all popular neural network formats and optimize their inference on the new NN accelerator.

SOLUTION:

The solution is to shift from an NMDL compiler developed by the customer to TVM; thus, all popular neural network formats will be supported. In the scope of the project, the Grovety team carried out research on the VTA stack. The investigation revealed that all other TVM stacks process information at lower layers of the IP-core accelerator.

Therefore, we decided to use BYOC (Bring Your Own Codegen), which provides data processing at a high layer. We developed our own backend using BYOC.

Neural network translator was implemented into an executable graph out of a TVM relay Intermediate Representation. During the implementation process, we solved several global tasks: BatchNorm positioning, re-quantizing, such as data preprocessing, and algorithm re-elaboration to provide its work at maximum performance without a drop in precision.

The extended IP-core functionality has been achieved due to TVM framework architecture flexibility and openness. This allows you to support more input formats, makes it easier to add new functionality for processing neural network models, and deploy them on hardware.

RESULTS:

Currently, we created an FPGA prototype with a new IP core and software stack for converting TVM-based neural networks. The design of the SoC which uses this IP core is in progress: first results are expected to be at the end of 2021.

The customer focused on what they are proficient at and got the required result two times faster and at half price than if they had had to do it in-house.

Due to the support of all popular neural network formats, the customer can offer their clients a universal product and, according to preliminary estimations, increase their market share up to 64-70%. As of today, the customer has already succeeded in signing contracts with three new clients for future deliveries.

TOOLS & TECHNOLOGIES:

TVM, TensorFlow, TensorFlow Lite, C++, Python

grovety.com hi@grovety.com