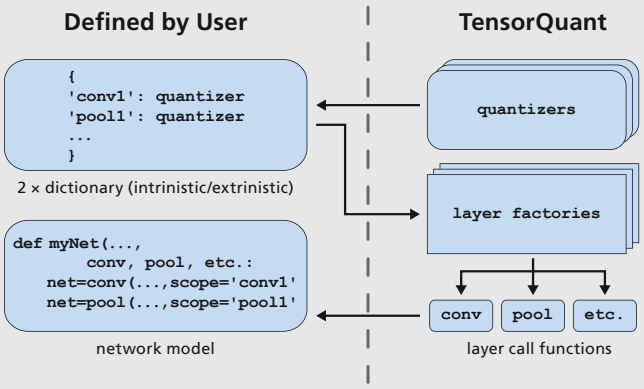
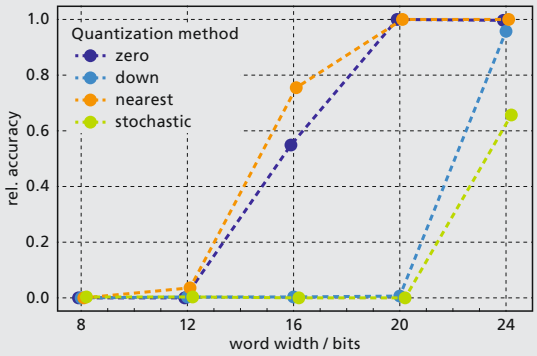


1



2

TENSORQUANT BRINGS DEEP LEARNING TO MOBILE APPLICATIONS

Machine learning methods are being used more and more in the industrial and service sector. Especially artificial neural networks or Deep Learning (DL) have a high impact on the development of intelligent systems. Research continuously provides new Deep Learning methods, which open a wide range of possibilities for these algorithms in many different practical application scenarios.

However, a significant technological hurdle on the way to such applications in production is the enormous computational effort required to calculate and evaluate the DL models.

This explains why the development of specialized DL hardware has recently come into focus. In the future, new chip and memory architectures will enable the use of high performance hardware components that save energy and, at the same time, expand the use of DL, for example to autonomous vehicles, mobile phones, or integrated production controls.

Learning does not require high precision in numerical processing

We exploit a mostly mathematical feature of DL: Learning and evaluating models can be reduced to a numerical computation of a small number of operations using tensor-algebra (for example, matrix-multiplications). In addition, tensor calculation works well with much less precision in terms of numerical processing than it is typically the case with physical simulations. In comparison to general computational units such as CPUs and GPUs, these features enable a highly efficient hardware implementation.

TensorQuant allows the emulation of machine learning hardware

In the development of DL applications on specialized hardware, difficulties are encountered as the minimum requirements for computational precision vary significantly between the individual models. As a result, the simultaneous optimization of DL models and hardware is difficult in terms of computational performance, power consumption, and predictive accuracy. Our TensorQuant (TQ) software lets developers identify critical tensor operations and emulate DL models with numerical processing and computing accuracy, which in effect accelerates development. TQ has already been used in collaborative research projects with the automobile industry.

1 *TensorQuant allows the automatic emulation of given TensorFlow models with any number representations of individual tensor operations.*

2 *The evalulation of the well-known ResNet-50 model shows that the concrete choice of the number representation has a considerable influence on the performance of DL applications, which is difficult to estimate in advance without the simulation in TensorQuant.*