BRNO UNIVERSITY OF TECHNOLOGY FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2023

NXP

Converter between formats of Deep Neural Network models on mobile platforms

Martin Pavella, xpavel39@stud.fit.vutbr.cz

supervisor: Ing. **Radek Kočí**, PhD.

consultant: Ing. **Róbert Kalmár** of NXP

1. The need for conversion

- High popularity and availability of Deep Neural Network (DNN) models in the **ONNX** format
- Superior HW accelerator support on mobile platforms for models in the **TFLite** format
- Existing converters produce sub-optimal models with unnecessary operators due to indirect approach

2. Proposed solution

A **direct** converter from ONNX to TFLite

- Represent an ONNX model using a hierarchy of objects
- Convert it to an equivalent TFLite object model (Fig.1)
- Serialize the model to the output TFLite file

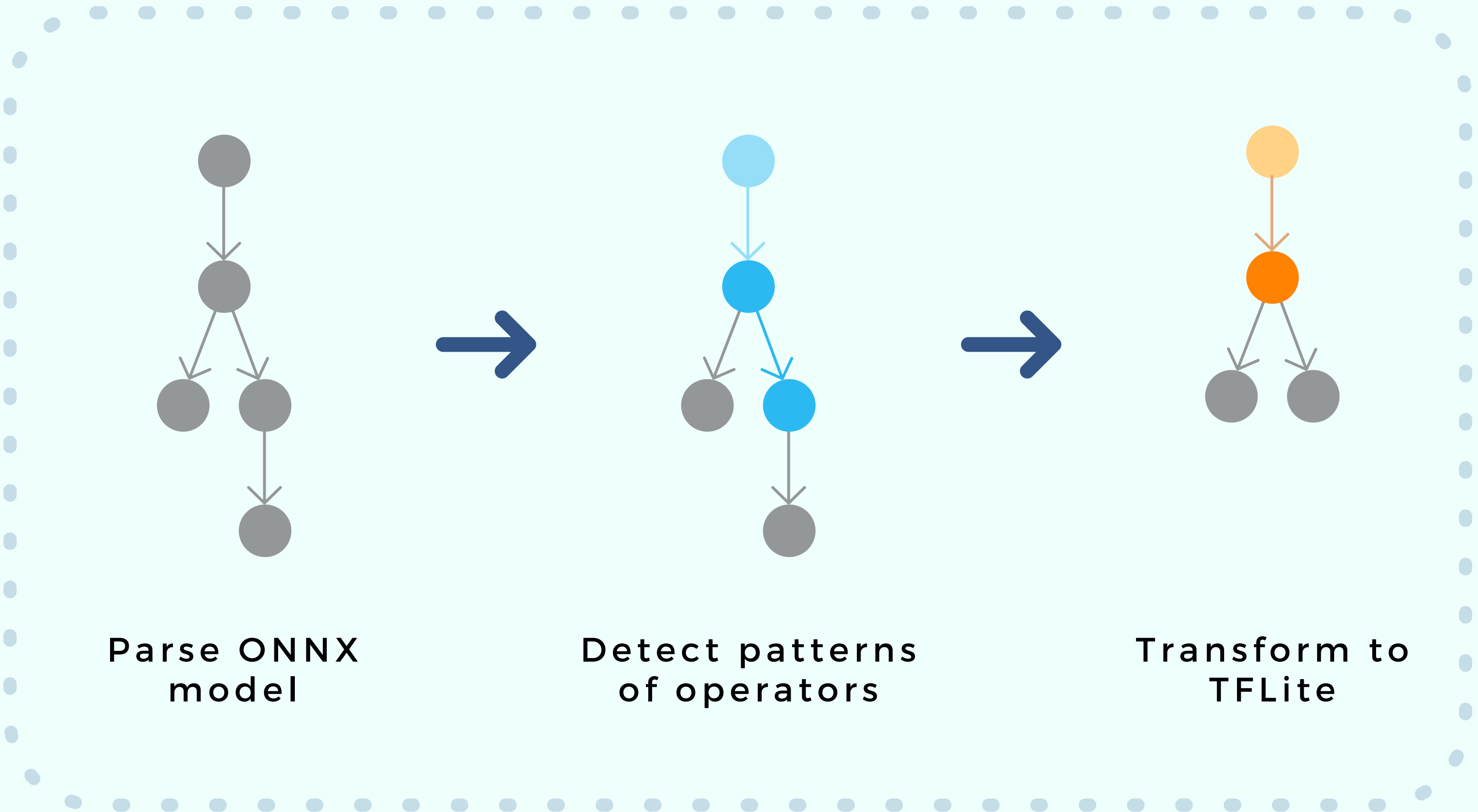
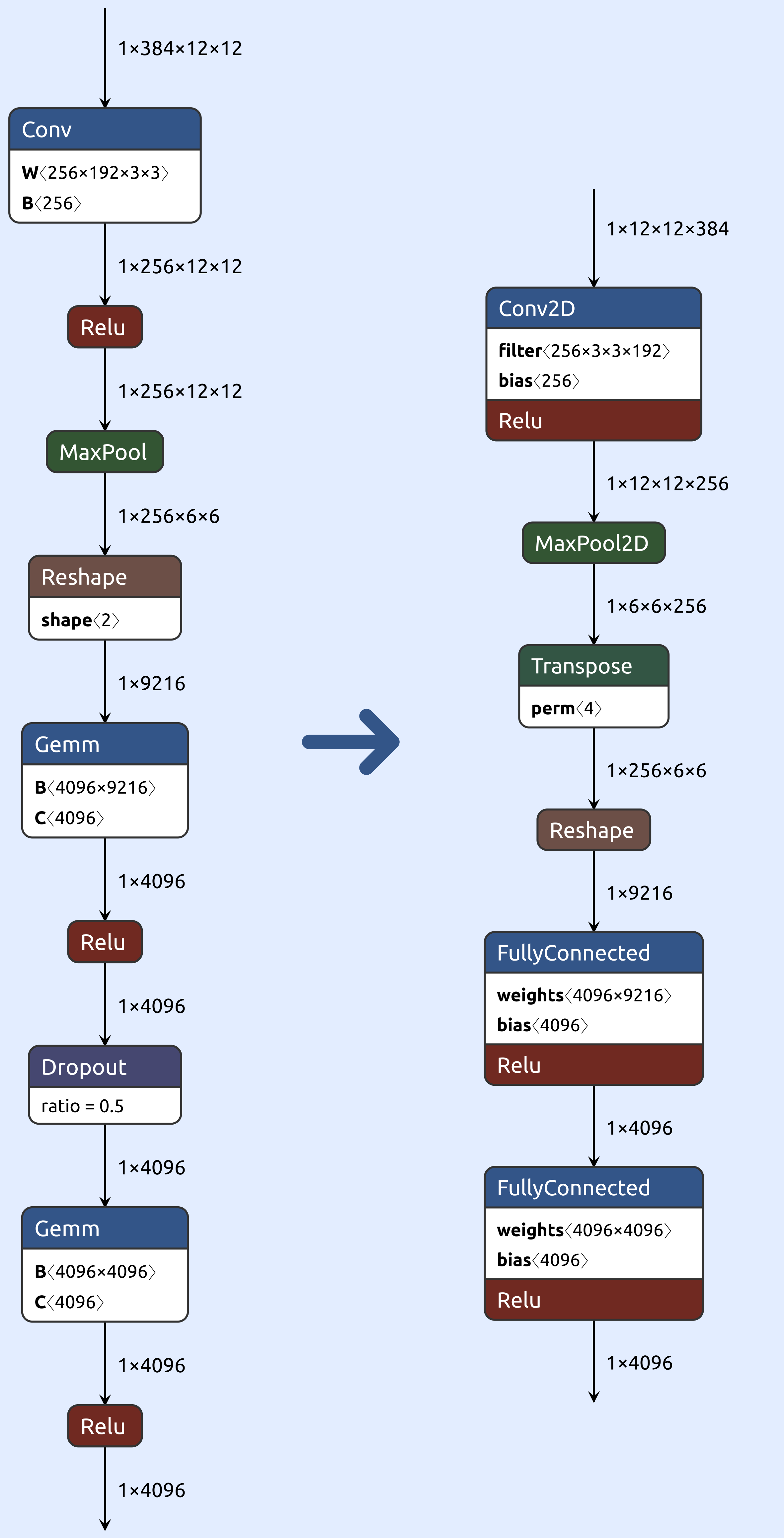


Fig. 1

Process of model conversion

3. Results

- Operator conversion is a complex and evolving problem
- Conversion of all operators is not feasible -> Focus on a subset of commonly used operators
- Successful conversion of models used for classification, object detection, segmentation and analysis of acoustic data
- Model **size reduction** by up to 420kB



ONNX model

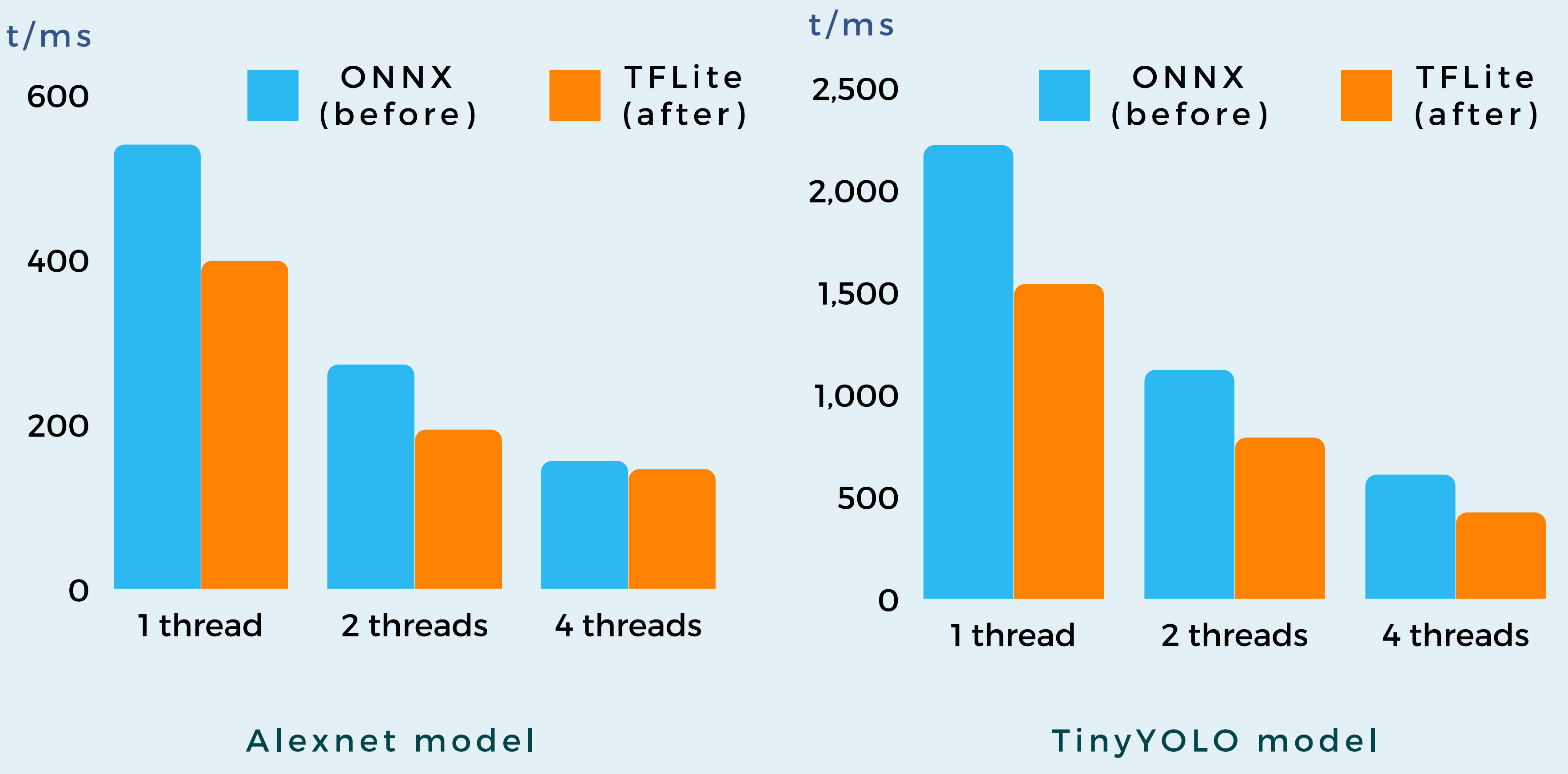
Equivalent **TFLite** model

Fig. 2

Example of a section of a convolutional Alexnet model before and after conversion

4. Impact on inference

- Converted models produce identical outputs as the original ones
- Experiments in collaboration with the NXP company show a significant **improvement of inference speed** on target platforms



Tab. 1

The time duration of DNN model inference on target platforms before and after conversion

5. Limitations

- Limited subset of supported operators
- Conversion is not always possible
- Some models can be converted, but are not efficient