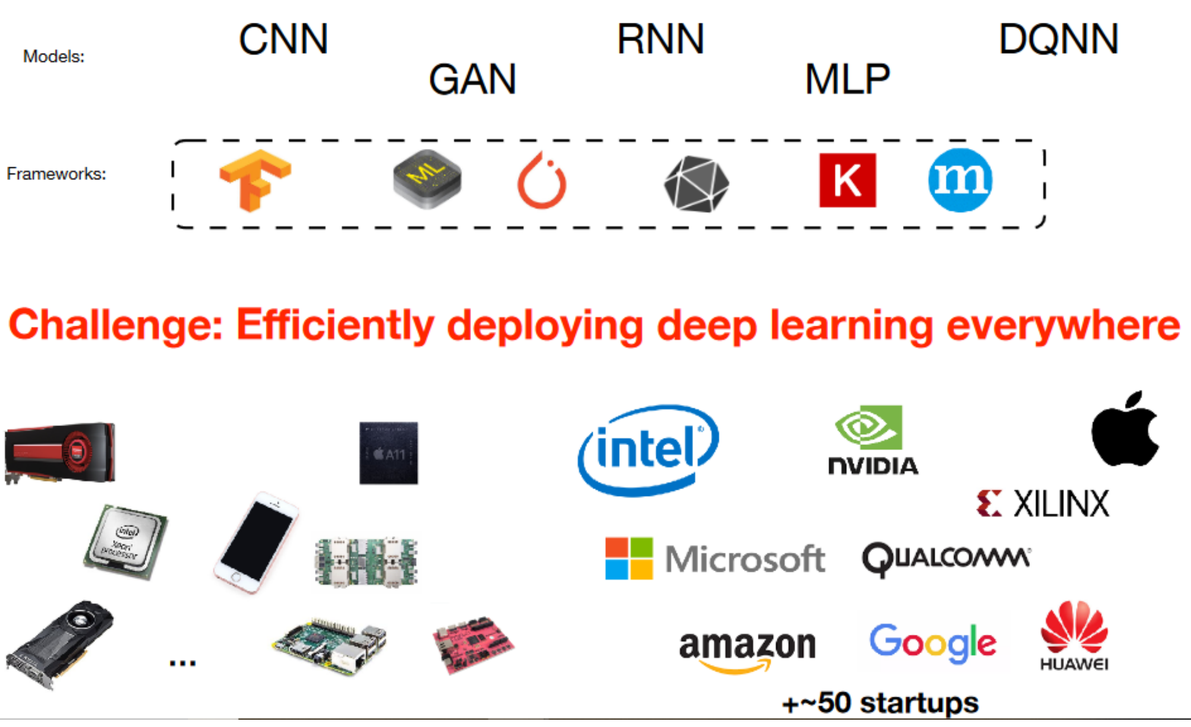
TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

Electronic System Group

Automating efficient deployment of DNN applications by leveraging compiler techniques

Wei Sun, Sander Stuijk, Andrew Nelson, Henk Corporaal

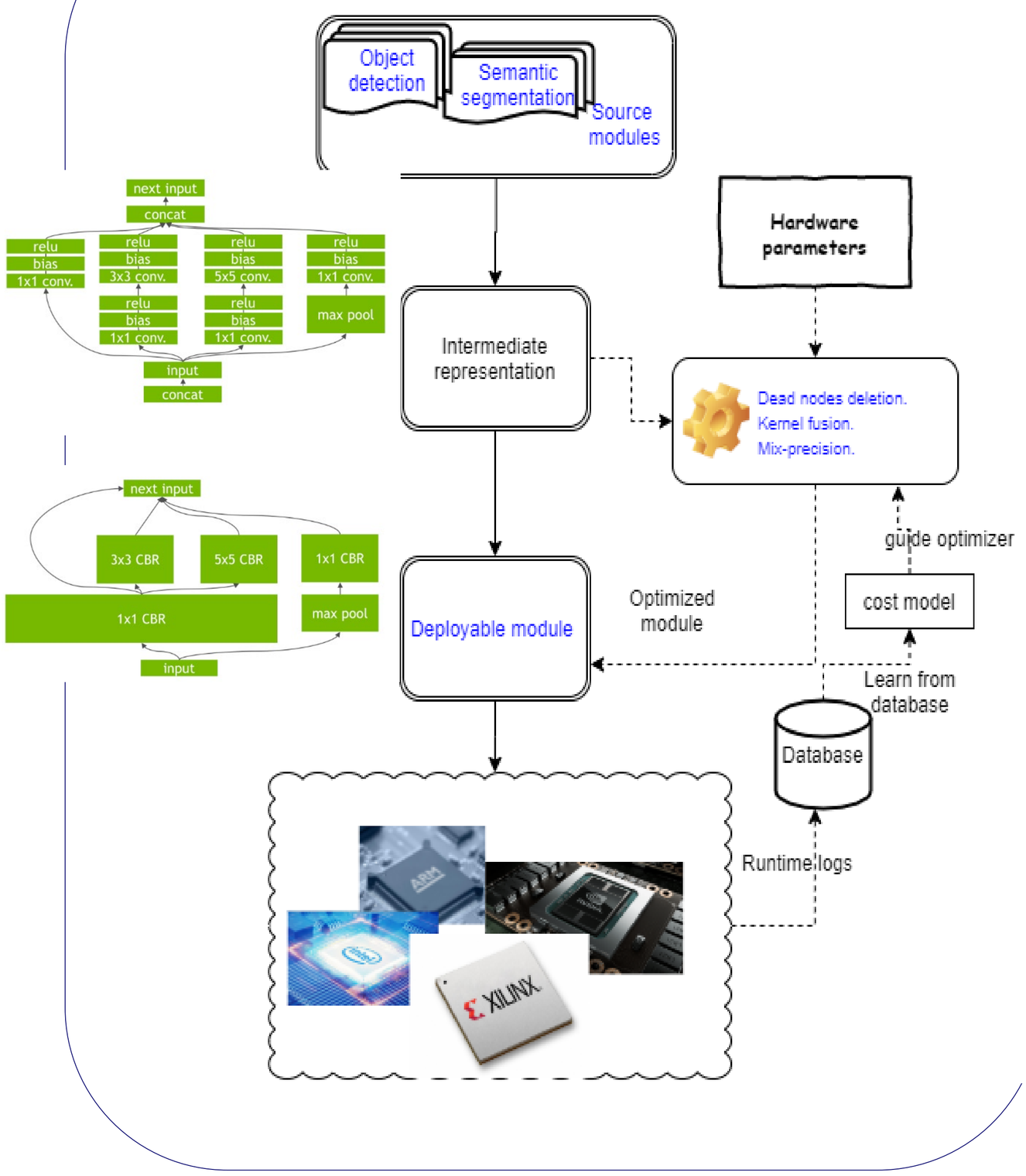


Research Problem: Developing a framework which enables end-to-end DNN applications deployment on resource-constrained hardware platforms.

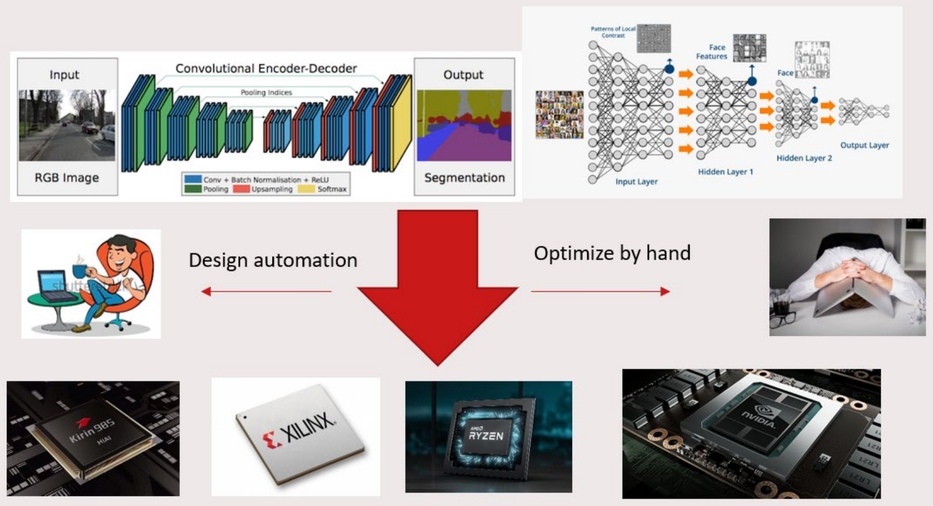
Research goals:

- For users: maximum automation
  - Leverage autotuning to teach the framework to search the optimal software configurations on target hardware.
  - End-to-end deployment, input trained DNN model, output executable program.
- For Designers: Reusability
  - Decouple the framework into individual reusable components
  - Reuse by modifying front-end and back-end
- For Community: Open-sourced whenever possible
  - Leverage open-sourced project, contribute to open-sourced community

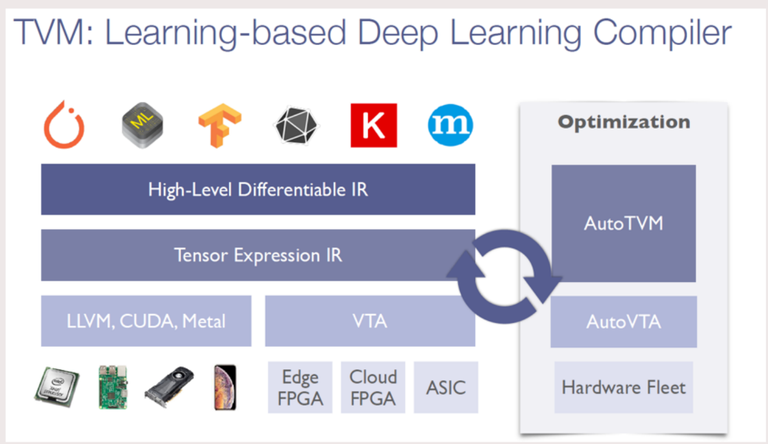
Deep learning compiler workflow



Problem statement



State-of-the-art solution



- Automatically explore the design parameters based on the hardware targets

References:

- Nvidia Tensor RT
- Halide programming language
- TVM open-source deep learning compiler

Wei Sun
w.sun@tue.nl