ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

TIK
Institut für Technische Informatik
und Kommunikationsnetze
Computer Engineering and
Networks Laboratory

Prof. L. Thiele

Semester Thesis:

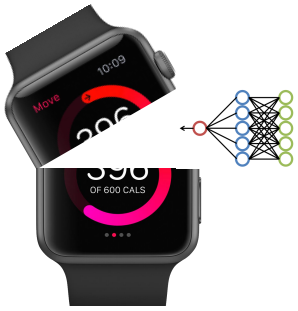
Compressing Deep Neural Networks via Neuron Sharing

Motivation: Despite their increasing popularity, deep neural networks (DNNs) are often too memory and computation intensive to be deployed on mobile and embedded devices. Model compression is an emerging approach to reducing the complexity of DNNs such that deep learning powered applications are able to run locally on-device.

Task: The aim of this thesis is to implement a neuron sharing based model compression technique [1], apply it to popular deep models, and evaluate the performance of the compressed models on mobile platforms.

Specific tasks involve but are not limited to the following.

- Implement the neuron sharing based compression scheme and evaluate it using standard deep models such as AlexNet, VGG-16, etc.
- Deploy the compressed models and evaluate their performance on mobile platforms (Intel UP Squared Grove IoT Development Kit or NVIDIA Jetson TX2).
- Explore the potential of combining the scheme with weight pruning strategies such as layer-wise optimal brain surgeon [2].



Requirements: Proficiency in Python programming; Familiarity with TensorFlow; Knowledge of machine learning and embedded system architectures; Experiences with large-scale datasets and knowledge of deep learning a plus.

References:

[1] He, Xiaoxi, Zimu Zhou, and Lothar Thiele. "Multi-Task Zipping via Layer-wise Neuron Sharing." arXiv preprint arXiv:1805.09791 (2018).

[2] Dong, Xin, Shangyu Chen, and Sinno Pan. "Learning to prune deep neural networks via layer-wise optimal brain surgeon." Advances in Neural Information Processing Systems. 2017.

Contacts

- Xiaoxi He: hex@ethz.ch, ETZ G77
- Zimu Zhou: zzhou@tik.ee.ethz.ch, ETZ G85