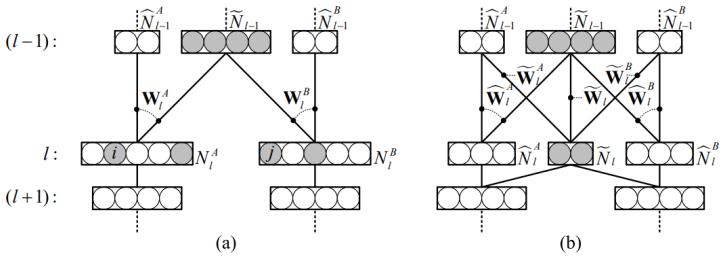
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

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

Prof. Dr. Lothar Thiele

Group/Master Thesis:

Efficient on-device execution of Multi-task Zipping



Although previous work on multi-task merging scheme [1] has provided the possibility to merge multiple pre-trained neural networks into one multi-task network and save memory utilisation, it is yet unclear how to efficiently deploy such merged network in resource constrained devices and allow fast switching between multiple tasks, or even simultaneous execution. Recent advances in on-device multi-task learning [2] provide a potential solution to this problem through memory paging and in-memory execution.

Tasks

The goal of this project is to adopt the weights virtualisation technique introduced by [2] and combine it with Multi-Task Zipping [1] in order to allow fast task-switching or even simultaneous execution.

- Get familiar with codes provided by [1, 2], implement them on test-platform.
- Combine techniques of [1] with [2] and implement fast task-switching.
- Evaluate performance on standard datasets.

Requirements / Skills

- Knowledge in ...
  - neural network enabled machine learning
  - operating system/computer architecture
  - Python programming
  - Tensorflow
- Curiosity, ability to work independently and interest in machine learning and embedded system.

Interested? Please have a look at https://www.tec.ee.ethz.ch/research.html and contact us for more details!

Contacts

- Xiaoxi He: hex@ethz.ch, G77
- Yun Cheng: chengyu@ethz.ch, G77

References

[1] Xiaoxi He, Zimu Zhou, and Lothar Thiele. Multi-task zipping via layer-wise neuron sharing. In *Proceedings of Advances in Neural Information Processing Systems*, pages 6016–6026, 2018.

[2] Lee Seulki and Nirjon Shahriar. Fast and scalable in-memory deep multitask learning via neural weight virtualization. In *Proceedings of ACM Annual International Conference on Mobile Systems, Applications, and Services*, 2020.