ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Prof. R. Wattenhofer

Deep Learning: To Prune or Not to Prune

Designing (Deep) Neural Networks is difficult: Type, number of layers, size of the layers, etc. need to be chosen by the practitioner. Choosing the model too small will result in bad performance due to underfitting. Choosing large models might overfit, and they use up much more storage space. It has been shown that pruning Neural Networks, i.e., the removing of connections and neurons, can reduce the size of a trained neural network by up to 80% without loss of performance.

We want to gain a better understanding of pruning and try different approaches. Also, we want to understand why a pruned network performs better than a "sparse-from-the-beginning" network, and if there is any way to change that, so that we do not have to go through the pruning-process. We already have many ideas on how to approach these questions, but we would love to hear your ideas as well. If you are interested, do not hesitate contact us.

Requirements: Interest in and willingness to study Machine Learning and Deep Learning. There will be weekly meetings to discuss progress and open questions.

Interested? Please contact us for more details!

Contacts

- Gino Brunner: gino.brunner@tik.ee.ethz.ch, ETZ G63
- Yuyi Wang: yuyi.wang@tik.ee.ethz.ch, ETZ 94