Master thesis:

Training Sparse Neural Networks with Knowledge Distillation

Context:

Deep Neural Networks (DNN) are powerful machine learning models for a large number of applications. However, they may have an enormous number of parameters and require large amounts of memory and computational resources and thus incur a high energy consumption, which makes their use for edge computing difficult.
Several approaches have been proposed to alleviate this problem [5-7], e.g. pruning, quantisation or architectural optimisations such as Neural Architecture Search. Although more and more efficient solutions exist also on the practical side (TensorFlow Lite, PyTorch quantization, NVIDIA Tensor RT etc.), more fundamental research is still required to reduce the complexity of these models and reduce the inherent redundancy in the parameters.
On a more general level, a major concern in reducing the energy consumption related to AI in the cloud as well as on the edge is to make these models more efficient and more accessible to a larger public.

Objectives:

The aim of this Master thesis is to conceive a new algorithm for effectively training sparse neural networks based on knowledge distillation [1]. The principal idea is based on a teacher-student approach, where a smaller network (i.e. the student) is trained to mimic the output of the original (large) model (i.e. the teacher). Many extensions and variants exist [2]. For example, this algorithm has been further extended to mimic *intermediate* layer activations [3]. In this Master thesis, the current state of the art in knowledge distillation will be studied, and a new approach that reduces the number of neurons layer-by-layer will be developped. An appropriate loss function (e.g. MSE) and regularisation term (e.g. based on the l1 norm) needs to be proposed that favors the sparsity within each layer as well as an efficient training strategy ensuring convergence and minimal loss in performance while avoiding overfitting. The study will primarily focus on post-training compression but a compression-aware training strategy [4] could also be envisaged.
The approach should be validated experimentally on Multi-Layer Perceptron (MLP) models and potentially extended to Convolutional Neural Networks (CNN) using standard machine learning benchmarks.

Prerequisites:

- Good knowledge in machine learning, statistical data analysis and neural networks
- Good python programming skills
- Curiosity and scientific methodology
- Autonomy

Environment:

The 6 month internship will take place at the LIRIS laboratory in Lyon (https://liris.cnrs.fr) (Doua Campus, Villeurbanne) under the supervision of Stefan Duffner, Imagine team.

Contact:

stefan.duffner@liris.cnrs.fr

[1] Hinton, G., Vinyals, O. & Dean, J. "Distilling the knowledge in a neural network", NIPS workshop, 2015
[2] Jianping Gou, Baosheng Yu, Stephen John Maybank, Dacheng Tao. "Knowledge Distillation: A Survey", International Journal of Computer Vision, 2021
[3] Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., & Bengio, Y. "Fitnets: Hints for thin deep nets. ICLR, 2015
[4] Jose M. Alvarez, Mathieu Salzmann. Compression-aware Training of Deep Networks, NIPS 2017
[5] Renato Cintra, Stefan Duffner, Christophe Garcia & André Leite. "Low-complexity Approximate Convolutional Neural Networks". IEEE Transactions on Neural Networks and Learning Systems, 2018
[6] Anthony Berthelier, Yongzhe Yan, Thierry Chateau, Christophe Blanc, Stefan Duffner & Christophe Garcia (2021). « Learning Sparse Filters In Deep Convolutional Neural Networks With A l 1 /l 2 Pseudo-Norm ». CADL 2020 : Workshop on Computational Aspects of Deep Learning - ICPR 2020
[7] Anthony Berthelier, Thierry Chateau, Stefan Duffner, Christophe Garcia & Christophe Blanc (2020). « Deep Model Compression and Architecture Optimization for Embedded Systems: A Survey ». Journal of Signal Processing Systems