Internet Technologies and Systems
Prof. Dr. Christoph Meinel
Multimedia and Deep Learning Research Group
Master Project Summer Term 2021

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Accelerating Training and Inference of Large-scale Pre-trained Language Models

Background

The recently emerged large-scale pre-trained language models based on the Transformer model, such as GPT-3, have brought about a series of breakthroughs in many Natural Language Processing tasks, such as translation, semantic understanding, and sentiment analysis.

Problem

The training of these large-scale models is computationally extremely expensive. Training a single Transformer model with neural architecture search emits as much CO_2 as 5 complete life cycles of a US car[1]. Moreover, these models generally have millions of parameters, making it challenging to conduct inference on resource-limited devices, such as smartphones.

Thus, this project belongs to the **clean-IT initiative** to improve the sustainability of future AI models and applications. More specifically, this project is committed to solve the current large-scale language model's excessive consumption of energy and thus reduce the carbon emissions needed to train and run such models.

Goal

This project will dive into the training process and inference of large-scale language models for a variety of Natural Language Understanding and Natural Language Generation tasks. The goals of this project are to:

- Understand how these models work in detail
- Study and implement different approaches to decrease their space and time complexity during training and/or inference
- Evaluate the models on different benchmark datasets
- Estimate or measure the accuracy, energy consumption, and run-time of the implemented methods

Advisors

Prof. Dr. Christoph Meinel
Joseph Bethge, Ting Hu, Dr. Haojin Yang

1 - Emma Strubell, Ananya Ganesh, and Andrew McCallum. "Energy and Policy Considerations for Deep Learning in NLP."