Can Compressing Language Models Improve their Fairness?

Team Project

HWS 2023

Marlene Lutz

Chair for Data Science in the Economic and Social Sciences

Motivation

- Language model based applications have become increasingly popular and integral parts of our lifes

  BUT

- LLM are known to reproduce harmful biases and stereotypes
- Large Language Models (LLMs) have potentially negative environmental consequences from training and deployment due to large model size

- Team Project will focus on the interplay between **fairness** and **compression of language models**

Topic

- Model compression techniques attempt to shrink a model without sacrificing performance
  - E.g. pruning attention heads or groups of network nodes
- Crucial in settings where memory and latency constraints are imposed
- Compression techniques could affect model fairness positively or negatively
  - SOTA research has shown different effects
  - Pruning is underexplored

Goal: How and to what extent do compression techniques impact the fairness of language models?

Logistics

- Language: English
- Duration: 6 months
- Participants: 3 – 4
- Prerequisites:
  - (strong) programming skills in Python
  - previous knowledge in Natural Language Processing, Machine Learning
  - not required but helpful: experience in working with LLMs, HuggingFace library

If you have further questions, contact *marlene.lutz@uni-mannheim.de*