Seminar Topics: Large Language Models

Faeze Ghorbanpour

faeze.ghorbanpour@lmu.de
Outline

1. Importance of Data Understanding
2. Data Pruning for LLMs Training
3. Neural Network Pruning for LLMs Training
Importance of Data Understanding

- Large text corpora are the **backbone of language models**
- What do we know about the content of these corpora? such as their **statistics, quality, social factors, and contamination**
- How can large-scale corpora be analyzed? **Mechanisms and tools**
- What are the **effect of the data on model behavior**? What would a model have learned from the given data?
What’s In My Big Data?

Yanai Elazar\textsuperscript{1,2}, Akshita Bhagia\textsuperscript{1}, Ian Magnusson\textsuperscript{1}, Abhilasha Ravichander\textsuperscript{1}, Dustin Schwenk\textsuperscript{1}, Alane Suhr\textsuperscript{3}, Pete Walsh\textsuperscript{1}, Dirk Groeneveld\textsuperscript{1}, Luca Soldaini\textsuperscript{1}, Sameer Singh\textsuperscript{4}, Hanna Hajishirzi\textsuperscript{1,2}, Noah A. Smith\textsuperscript{1,2}, Jesse Dodge\textsuperscript{1}

\textsuperscript{1}Allen Institute for AI
\textsuperscript{2}Paul G. Allen School of Computer Science & Engineering, University of Washington
\textsuperscript{3}University of California, Berkeley
\textsuperscript{4}University of California, Irvine

Does more data lead to better performance?

The data is scraped raw web pages, leading to a substantial portion of the text being noisy and of low quality!

How can we remove the least impactful examples from a pretraining dataset? How to measure the quality of pretraining data?

Data pruning isolates a subset of a larger training dataset so that a model trained on the subset preserves or improves performance.
When Less is More: Investigating Data Pruning for Pretraining LLMs at Scale

Max Marion  
*Cohere for AI*  
maxwell@cohere.com

Ahmet Üstün  
*Cohere for AI*  
ahmet@cohere.com

Luiza Pozzobon  
*Cohere for AI*  
luiza@cohere.com

Alex Wang  
*Cohere*  
alexwang@cohere.com

Marzieh Fadaee  
*Cohere for AI*  
marzieh@cohere.com

Sara Hooker  
*Cohere for AI*  
sarahooker@cohere.com

LLMs model size presents significant challenges in deployment, inference, and training stages.

Neural network compression is independent of the original training data.

How to remove non-critical parts of LLMs while preserving most of their functionality?

Unstructured pruning removes individual weights from the network based on some criteria, resulting in sparse weight matrices that can be stored and processed more efficiently.

Structured pruning eliminates whole components, such as neurons, channels, or blocks, leading to smaller architectures to reduce end-to-end inference latency.
LLM-Pruner: On the Structural Pruning of Large Language Models

Xinyin Ma  Gongfan Fang  Xinchao Wang*
National University of Singapore
maxinyin@u.nus.edu, gongfan@u.nus.edu, xinchao@nus.edu.sg