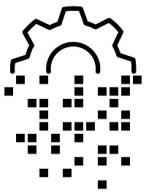
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Prof. R. Wattenhofer

Leveraging Compressed Representations of Text for More Efficient NLP



Large language models tend to be, well, large. However, recent research has shown that in practice, knowledge retrieved by generative LLMs accumulates at a handful of tokens, with the rest playing little to no role in the next token prediction. We aim to leverage this fact by training BERT-like and GPT-like models on compressed representations of text, hoping to achieve comparable results while using significantly lower amount of compute.

To that end, we have already developed a transformer-driven text compression technique for sentences, and we now hope to use it in otherwise compute-hungry downstream tasks.

Candidate Profile. Generally speaking, a good candidate is a master's student, well-versed in Python and with past experience in training deep neural networks. An ideal candidate has a background in natural language processing.

Interested? Please contact us to learn more!

Contact (please send an email with the following as recipients)

- Peter Belcak: belcak@ethz.ch, ETZ G61.3