Raising an LLM
Language Model Pre-Training With Scalable Transformers

Project supervisor: Kevin Glocker

Large Language Models (LLMs) have rapidly gained popularity in industry and academia through their fluent language generation, general-purpose task-solving, and reasoning capabilities. However, achieving these results requires a large number of model parameters and internet-scale datasets, leading to high costs and substantial environmental impact.

Today, a range of training methods and architectures are employed to minimize the number of active parameters during training. For instance, recently introduced transformer scaling methods, such as HyperCloning [1] and the Tokenformer architecture [2], gradually increase the number of parameters throughout training. The results show that these methods can achieve comparable or better benchmark scores with 2–4 times faster training compared to standard LLM pretraining setups [1], [2].

Further research efforts target the data efficiency of language models. Motivated by the fact that children are exposed to less than 100 million words by the age of 13, while LLMs are trained on trillions of tokens, the BabyLM challenge series [3] has been introduced. A primary goal of the challenges is to maximize data efficiency under such more developmentally plausible conditions.

The goal of this project is to apply recent advances in scalable transformers in a more developmentally plausible, data-constrained setting to investigate data-efficient language models that gradually grow in size and complexity.

Your tasks will include:

- Using heuristics or machine learning methods to estimate pre-training data complexity in English or Swedish
- Training scalable language models by gradually growing the number of parameters with increasing data complexity using the Nanotron transformer pretraining library[1]
- Evaluating and comparing different model growth and data selection strategies on established benchmarks

Depending on your progress and interests, the project can be further extended with:

- Implementing the Tokenformer scalable transformer architecture [2] in Nanotron
- Training and additional evaluation of Tokenformer models
- Including code or other formal language data in the pre-training schedule and evaluating its impact

References

[1] M. Samragh, I. Mirzadeh, K. A. Vahid, *et al.*, “Scaling smart: Accelerating large language model pre-training with small model initialization,” in *ENLSP*, 2024.

[2] H. Wang, Y. Fan, M. F. Naeem, *et al.*, “Tokenformer: Rethinking transformer scaling with tokenized model parameters,” in *The Thirteenth International Conference on Learning Representations*, 2025.

[3] M. Y. Hu, A. Mueller, C. Ross, *et al.*, “Findings of the second BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora,” in *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, M. Y. Hu, A. Mueller, C. Ross, *et al.*, Eds., Miami, FL, USA: Association for Computational Linguistics, Nov. 2024, pp. 1–21.

[1] `https://github.com/huggingface/nanotron`