Report on Bachelor Thesis

Title: Fast Algorithms for Attention Mechanism
Student: Tymofii Reizin
Supervisor: doc. Mgr. Kolman Petr, Ph.D.
Study Programme: Computer Science

Summary of contributions
The thesis deals with transformers, a type of neural network that transforms an input sequence into an output sequence. The focus of the thesis is on the *attention mechanism* in transformers, which is a computational method capturing relationships among words in the input sequence. A principal part of this computation is done by the so-called softmax function applied (element-wise) on a result of a matrix multiplication. A challenging open problem is whether this approach involving the softmax function can be replaced by a different technique that can be implemented faster, without sacrificing the quality of the output.

The thesis is structured as follows. Chapter 1 provides a general description of the structure of transformers. Chapter 2 sketches the known (roughly) quadratic lower bound, with respect to the sequence length, on the time needed to apply the softmax function in the attention mechanism. Chapters 3 and 4 deal, on a theoretical level, with alternatives of the softmax function. Finally, Chapter 5 describes the results of experiments performed by the student in which the softmax function is compared with the alternatives from Chapters 3 and 4. Chapters 2, 3 and 4 are based on recent research papers (the latest is from winter 2024), Chapter 5 is the student‘s contribution.

Evaluation
The student did definitely satisfy the assignment of the thesis: he provided an overview of the general transformer architecture and explored, both theoretically and practically, a couple of alternatives for faster computation of the attention in transformers. By doing so, he demonstrated that he is able to follow the current research trends in AI and theoretical computer science, understand them, apply the results and actively search for improvements on a level that exceeds, in my opinion, an average graduate of our bachelor‘s Computer Science program.
Unfortunately, the thesis was completed in a hurry, which negatively affected its quality. In particular, the experimental setup could have been planned and realized more thoroughly (e.g., sizes of the samples, confidence of the measurements, to name a few). Nevertheless, the thesis satisfies the requirements on a bachelor thesis, and I assume that Tymofii will continue in the research that he started in his thesis.

Overall Assessment
I definitely recommend to accept the thesis.

doc. Petr Kolman, Ph.D.

Prague, June 20 2024