Efficient Transformers for Arabic LLM

Project Description. Transformers were initially designed for natural language processing tasks. They have demonstrated impressive performance on complicated tasks such as neural machine translation and question answering. More recently they have been employed on other sequence modeling tasks such as speech and audio processing and image generation.

Transformers are purely based on attention and dense layers, and despite their success, they are difficult to scale for long sequences. The core mechanism of transformers is self-attention, which enables to highlight relevant features of the input data, and this has a quadratic complexity both in terms of memory and computation with respect to the sequence length, i.e., $\mathcal{O}(n^2)$. The scaled dot-product attention of queries $Q \in \mathbb{R}^{n\times d}$, keys $K \in \mathbb{R}^{n\times d}$, and values $V \in \mathbb{R}^{n\times d_v}$ can be written as

$$\textsc{Attention}(Q,K,V) \quad \leftarrow \quad A_\ell(X_{\ell-1}) = \textsc{Softmax}\Big(\frac{QK^T}{\sqrt{d}}\Big)V,$$

where $X_{\ell-1} \in \mathbb{R}^{n\times d}$ denotes a sequence of n (input) feature vectors, and $A_\ell(X_{\ell-1})$ is the self-attention function of the ℓ transformer block that acts across sequences, whereas the transformation function in the dense layer transforms each feature independently of the others. The softmax function is applied rowwise to QK^T resulting in a computational cost of $\mathcal{O}(n^2)$. The same is true for the memory requirements.

In this project, we are interested in experimenting with attention mechanisms that are substantially faster and more memory efficient. This helps in reducing the time and memory complexity of transformers, and allows to train Arabic LLM efficiently on long sequences where softmax transformers will not fit in a GPU.

Project Type. 60% Engineering, 40% Research with a focus on contributing to a serious publication.

Internship Batch. Batch 1 from May 12 to July 12

Duties/Activities. Some code exists, but there is still some work to be done to make it usable, and to produce results.

Required Skills. Python programming.

Preferred Intern Academic Level. B.Sc. (3rd year or 4th year) or MS student.

Learning Opportunities. Students will be exposed to the exciting research in Large Language Models (LLMs), such as ChatGPT or Fanar for arabic LLM, and will learn the details of how these models work, i.e., the transformer architecture, and how to make the self-attention mechanism efficient in terms of computation and memory.

Expected Team Size. It is preferable to have a team of 2 interns.

Mentors. Dr. Abdelkader Baggag, Dr. Sanjay Chawla