Numerical Analysis and Scientific Computing Seminar

Galerkin Transformer

Shuhao Cao
Washington University in St. Louis

Abstract: Transformer in "Attention Is All You Need" is now THE ubiquitous architecture in every state-of-the-art model in Natural Language Processing (NLP), such as BERT. At its heart and soul is the "attention mechanism". We apply the attention mechanism the first time to a data-driven operator learning problem related to partial differential equations. Inspired by Fourier Neural Operator which showed a state-of-the-art performance in parametric PDE evaluation, an effort is put together to explain the heuristics of, and to improve the efficacy of the attention mechanism. It is demonstrated that the widely-accepted "indispensable" softmax normalization in the scaled dot-product attention is sufficient but not necessary. Without the softmax normalization, the approximation capacity of a linearized Transformer variant can be proved rigorously for the first time to be on par with a Petrov-Galerkin projection layer-wise. Some simple changes mimicking projections in Hilbert spaces are applied to the attention mechanism, and it helps the final model achieve remarkable accuracy in operator learning tasks with unnormalized data. The newly proposed simple attention-based operator learner, Galerkin Transformer, surpasses the evaluation accuracy of the classical Transformer applied directly by 100 times, and betters all other models in concurrent research. In all experiments including the viscid Burgers' equation, an interface Darcy flow, an inverse interface coefficient identification problem, and Navier-Stokes flow in the turbulent regime, Galerkin Transformer shows significant improvements in both speed and evaluation accuracy over its softmax-normalized counterparts and other linearizing variants such as Random Feature Attention (Deepmind) or FAVOR+ in Performer (Google Brain). In traditional NLP benchmark problems such as IWSLT 14 De-En, the Galerkin projection-inspired tweaks in the attention-based encoder layers help the classic Transformer reach the baseline BLEU score much faster.

Friday, October 15, 2021, 12:30 pm
https://emory.zoom.us/j/94914933211

Mathematics
Emory University