APPLIED AND NUMERICAL ANALYSIS SEMINAR

**Thursday Sep 16th Period 9**

The zoom link
is https://ufl.zoom.us/j/94160055012?pwd=OUdFTlU5TGRSeldaQkxIQTFnQVhMZz09

**Speaker:** Shuhao Cao, Washington University in St. Louis

**Title:** Galerkin Transformer

**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 parametric partial differential equations. Inspired by Fourier Neural Operator which showed a state-of-the-art performance in parametric PDE evaluation, we put together an effort to explain the heuristics of, and improve the efficacy of the self-attention. We have 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 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, surpassing the evaluation accuracy of the classical Transformer applied directly by 100 times. Meanwhile in all experiments including the viscid Burgers' equation, an interface Darcy flow, and an inverse interface coefficient identification problem, the newly proposed simple attention-based operator learner, Galerkin Transformer, shows significant improvements in both speed and evaluation accuracy over its softmax-normalized counterparts, as well as other linearizing variants such as Random Feature Attention or FAVOR+ in Performer by Google and Deepmind's researchers.