Thesis Defense

Computer Science Master's Program

"What Makes a Modern Attention Implementation? The Software and Hardware Driving the Transformer Revolution"

By **Brian Slonim**

Abstract:

Since the seminal assertion by Vaswani et al. in 2017 that "Attention Is All You Need," transformer models have risen to ubiquity due to their ability to learn extremely complex patterns from sequence data, culminating in the unprecedented generative capabilities of large language models. These models' strength lies in their scale: hundreds of millions (e.g., BERT-LARGE) to billions or trillions of learned parameters. Running inference with these models, let alone training them, would be intractable without significant innovations in the hardware and software that support them. This need has driven an enormous demand for GPU compute and associated software ecosystems, like that of NVIDIA. With many different accelerator architectures, attention variants, and application constraints, obtaining a holistic view of what goes on behind the scenes of state-of-the-art AI systems can be elusive, even to the well-initiated.

In this work, we seek to provide exactly that: a unified perspective on what makes modern attention implementations fast, efficient, and flexible enough to handle the tremendous computational demands of transformer applications. To this end, we discuss the machine learning operations that motivate this need, together with the algorithmic and systems-level techniques researchers have developed to satisfy it, in terms understandable to readers with a general computer science background. We present five key attributes of modern attention implementations, with detailed explanations of popular frameworks, such as the FlashAttention series, that put them into practice. We demonstrate the tangible performance improvements these yield by evaluating them against attention workloads of various sizes across four generations of NVIDIA GPUs. Our results consolidate the diverse literature surrounding this topic into an accessible, structured framework, and provide researchers a practical reference for navigating and assessing the rapidly evolving field of attention acceleration.

Date: Wednesday, May 27th, 2026
Time: 3:30 PM – 5:30 PM
Location: 14-238b
Zoom: https://calpoly.zoom.us/j/81809114421
Meeting ID: 818 0911 4421
Committee: Dr. Pantoja, Dr. Ventura, Dr. Kurfess