Brain Inspired Computing Spring 2024

Tej Shah, Priyanshu Shrivastava, Advith Chegu

Project Idea 1 (Tej): Unsupervised Sentiment Neuron in Spiking Legendre Memory Units

What is the computational problem?

Can we train a character-level language model on Amazon Reviews using Legendre Memory Units (LMU) implemented by Spiking Neural Networks (SNN) to display unsupervised sentiment neuron(s)? More generally, can we find interpretable neuron features in certain SNNs?

Why is it interesting and important?

Alec Radford's 2017 discovery of the unsupervised sentiment neuron, which catalyzed the GPT project, highlighted the potential of scaling neural networks for unsupervised learning. Given the heavy computational demands of current foundation models, there's a push towards developing more efficient hardware and software, aligning with Richard Sutton's Bitter Lesson. This drive for efficiency leads to an exploration of whether brain-inspired computing, like neuromorphic hardware, can support capable models. Successfully replicating these preliminary findings in such an environment could trigger a scaling movement in the neuromorphic computing domain.

Why is it hard? What's been done now? Where/Why have previous approaches failed?

Training sequence models with SNNs was difficult until the publication of the LMUs in 2019, which showed better performance than traditional RNNs/LSTMs. Much brain-inspired computing research has focused on hardware with limited focus on software: and, the focus on software has mainly been focused on supervised tasks, not discovering unsupervised behavior and interpretable features. Lastly, interest in more efficient computing paradigms for language models is in the zeitgeist today because of the productivity gains since ChatGPT, LLAMA-2, etc.

What are the key components of your neuro-inspired approach? What model/simulation environments will be used to validate your approach?

We will train a SNN LMU on Amazon Reviews dataset using NengoDL on traditional hardware. We will train a character-level (i.e. ASCII characters) language model using the LMU unit with Spiking Neurons. Afterwards, we will train a linear probe for sentiment analysis. We will investigate which neuron(s) fire in the final layer of the model and see any unsupervised patterns.

If proving the language model proves too difficult with SNNs, we can still investigate the unsupervised sentiment neuron hypothesis by training one SNN layer on top of a traditionally trained model's (without SNNs) embeddings and then investigating the behavior.

We don't need exceptional performance on language modeling, just sufficient is enough. I hypothesize that if unsupervised features are learned, sentiment is an earlier level emergent feature. Hence, we just need a suitable language model that we then probe for sentiment neurons.