# CS 182: project proposal

Aviral Mishra, Nelson Lojo, Derrick Han Sun, Kesava Viswanadha

November 2023

## 1 Overview

We hope to improve inference-time performance (i.e. number of operations) of in-context learning for some of the simple functions proposed in "What Can Transformers Learn In-Context? A Case Study of Simple Function Classes" [1]. We propose applying known transformer inference speedups that approximate the attention mechanism and observing effects on the accuracy of in-context learning. In terms of function classes, we hope to examine performance on linear functions, shallow decision trees, and 2-layer Neural Networks.

We outline exactly what speed-ups we hope to apply:

1. *Combiners*[2] allow for full attention capability while maintaining sub-quadratic complexity via conditional expectation. Ren et al. have shown, within density estimation tasks, their capability to maintain accuracy while boosting speed.

2. Wortsman et al. have shown that for vision transformers, significant accuracy can be maintained when replacing the softmax activation function with ReLU after normalizing by the sequence length [3]. This results in a clear speed-up as ReLU does not require a gather step and is faster than computing an exponential (as in softplus, also explored in [3]).

3. Quantization-aware pretraining as in [5] has been shown to reduce both model size and speed up inference by replacing floating-point operations with Int8 operations.

4. Nyström approximations [4] seek to approximate attention by sampling rows and columns from full query and key matrices and performing smaller matrix multiplications to estimate the full attention matrix.

If time permits, we would like to "limit test" the in-context learning potential of transformers by attempting to express more complicated function types (random forest, 3-layer NN, state transitions of MuJoCo environments). Then, we would apply inference speed-ups to these function types as well, hopefully building upon what we learned from speeding up the transformers of simpler function classes. Thus, our aim would be to improve accuracy while keeping the added inference time at a minimum.

## 2 Sourcing Data, Compute, and Starter Code

We will replicate [1]'s process in sampling x values and functions from predefined $D_x$ and $D_f$. Rather than examine all error regimes as in [1], we will reduce our scope to merely examine out of distribution performance by varying distance from origin, scaling of input, and scaling of output.

We expect free compute to suffice for this project:

- Kaggle offers 20 hours per week per user of P100 GPU time and 12 continuous execution hours of CPU time, unbounded per user per week.
- Github Codespaces provides 150 core hours per user per month in the form of development containers
- Google Colab notebooks provide 12 hours per week per user of V100 GPU time
- We also have approximately $50 GCP cloud credits we can use if the above does not suffice.

Part of our codebase will build off of the provided code in the following ICL paper: https://github.com/dtsip/in-context-learning [1]. This starter code is what we will use to generate our training and test data. We will use this code to sample functions from our function classes and points within the domain of these classes, as well as evaluate performance of the transformers. Most of the speed-up code will have to be written from scratch, as our referenced papers for these optimizations didn't provide us access with a codebase.

## References


[1] Garg et al. "What Can Transformers Learn In-Context? A Case Study of Simple Function Classes" (https://arxiv.org/abs/2208.01066)

[2] Ren at al. "Combiner: Full Attention Transformer with Sparse Computation Cost" (https://arxiv.org/pdf/2107.05768.pdf)

[3] Wortsman et al. "Replacing softmax with ReLU in Vision Transformers" (https://arxiv.org/pdf/2309.08586.pdf)

[4] Xiong et al. "Nyströmformer: A Nyström-Based Algorithm for Approximating Self-Attention" (https://arxiv.org/abs/2102.03902)

[5] Zafrir et al. "Q8BERT: Quantized 8Bit BERT" (https://arxiv.org/pdf/1910.06188.pdf)