CMSC 678
Fall 2024
Homework 7
Due 11:59pm December 17th

(20 points) Describe the differences between traditional RNNs and transformers in terms of their capacity to model sequential data. Be sure to discuss:

- Information retention over long sequences
- Parallelization during training
- Computational complexity

(20 points) Explain why the computational complexity of the self-attention mechanism in transformers is $O(n^2)$ for sequences of length n.

(20 points) Find two recent approaches to reduce the complexity of self-attention (e.g., sparse attention, performer, or reformer). Briefly explain how the work and what tradeoffs are involved in terms of potential reductions in the performance at the sequence prediction task.

(20 points) Prove that attention scores are invariant to uniform scaling of the query and key vectors.

(20 points) In multi-head attention, explain how using multiple heads with separate projections improve the model's expressiveness?