Self-Attention in Multivariate Time-Series Classification

Aaron Brookhouse
Michigan State University

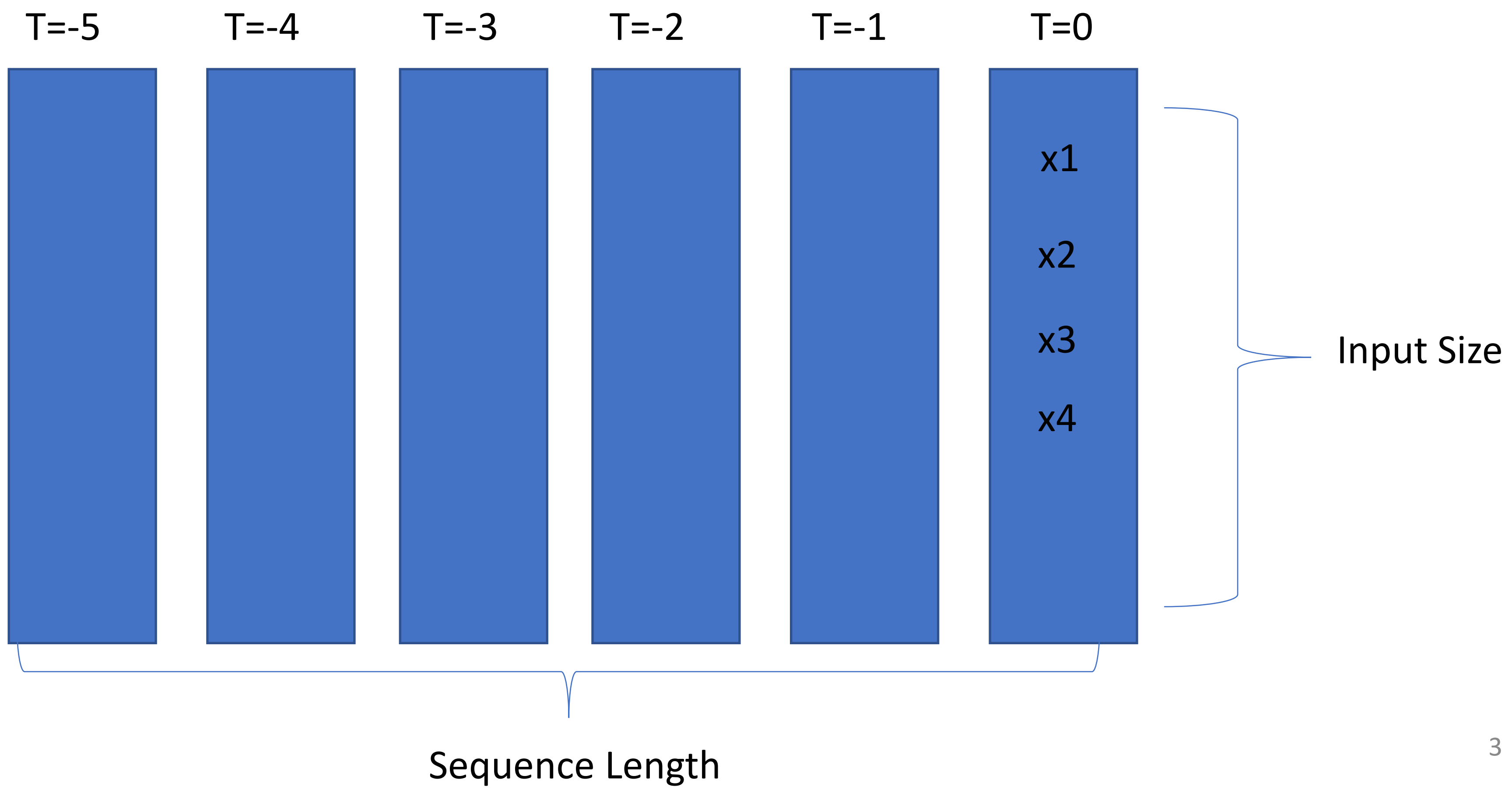
Mentor: Dr. Gebremedhin
Washington State University
Scalable Algorithms for Data Science Lab

Real Time Machine Learning

- Machine learning is often a computationally expensive task
- Want to develop a framework for Human Activity Recognition (HAR) that can be run in real time on a small computer
- HAR could be used to improve assisted living for the elderly

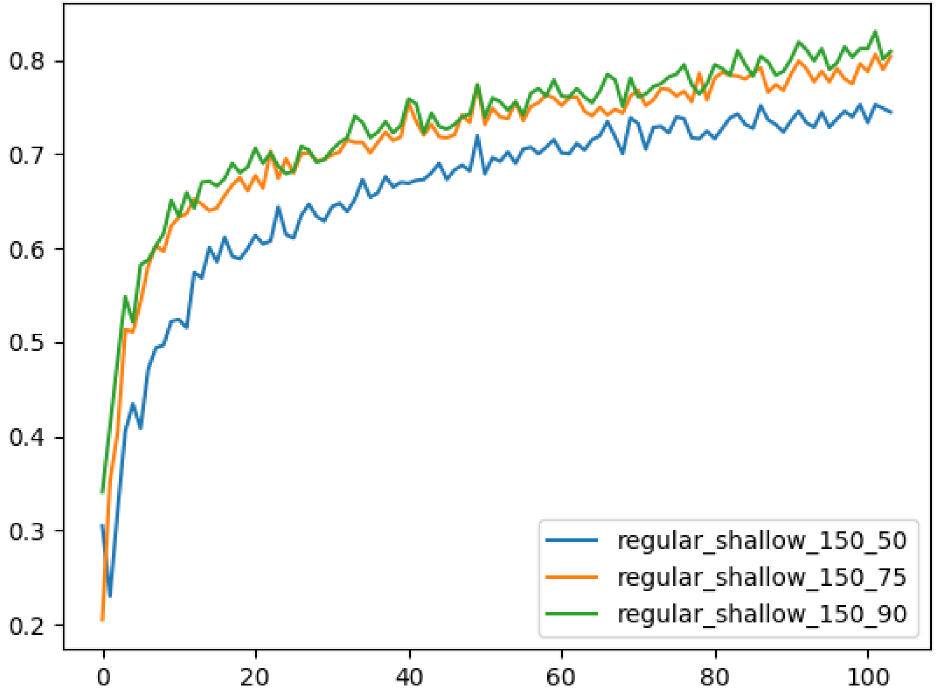
Term Clarification

- Input to time series models is a sequence of vectors representing data at discrete time steps
- Input size is the size of these vectors
- Sequence length is how many of these vectors are passed in



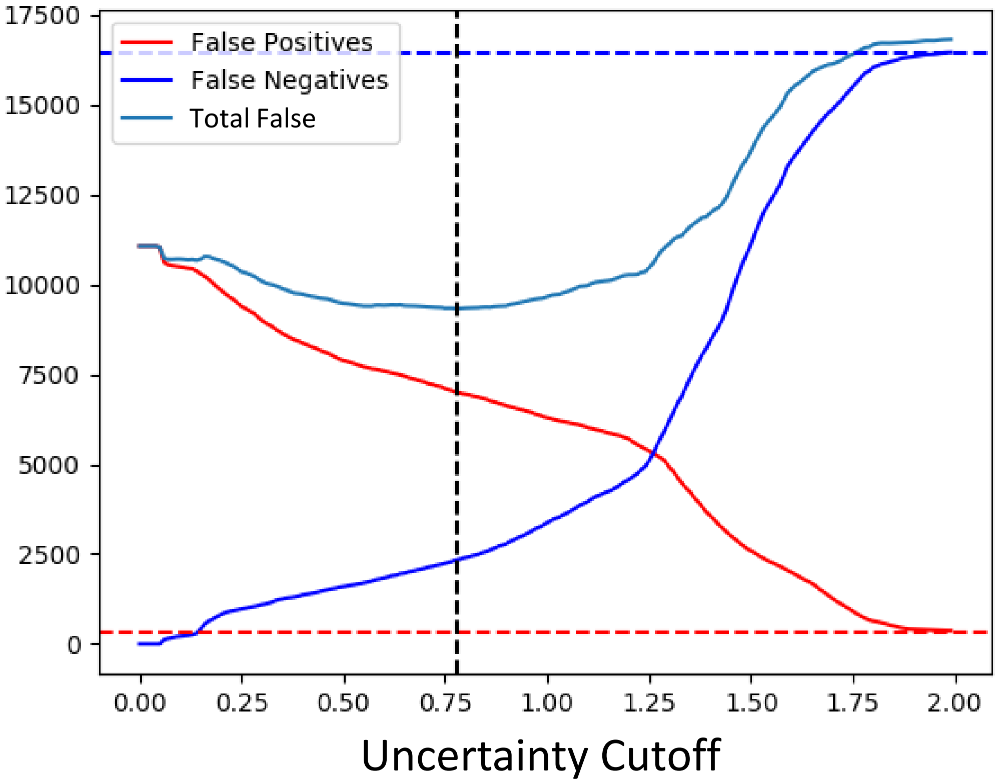
First Approach: LSTMs

- Ran experiments to see which hyperparameters had most impact on accuracy
- Produced some ways to further process model output
- Worked on finding balance between saving time, and maintaining accuracy



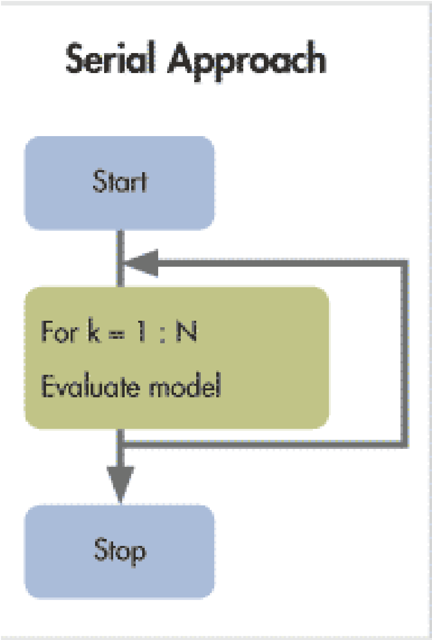
First Approach: LSTMs

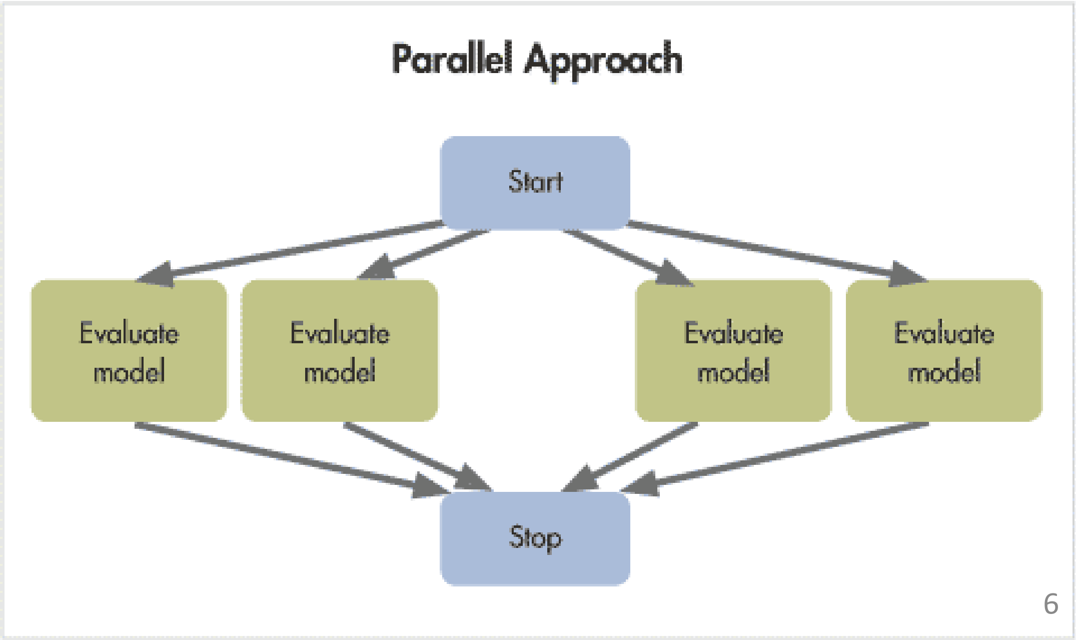
- Noticed that the null class was the most uncertain
- Search for the optimal cutoff point, where if uncertainty > cutoff, the output is null

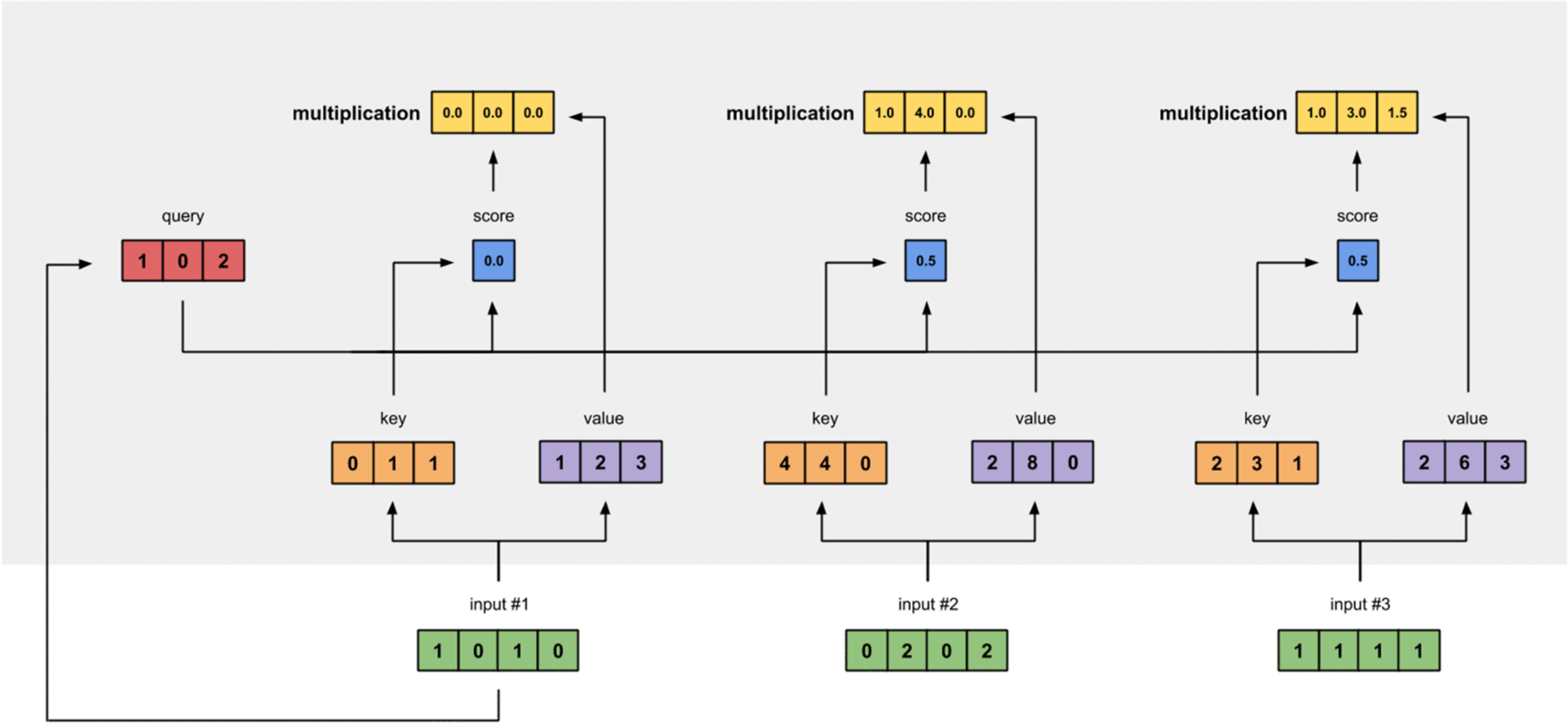


Major Limitation for LSTMs

- LSTMs are required to process time series data sequentially
- Attention based models can process time series data in parallel



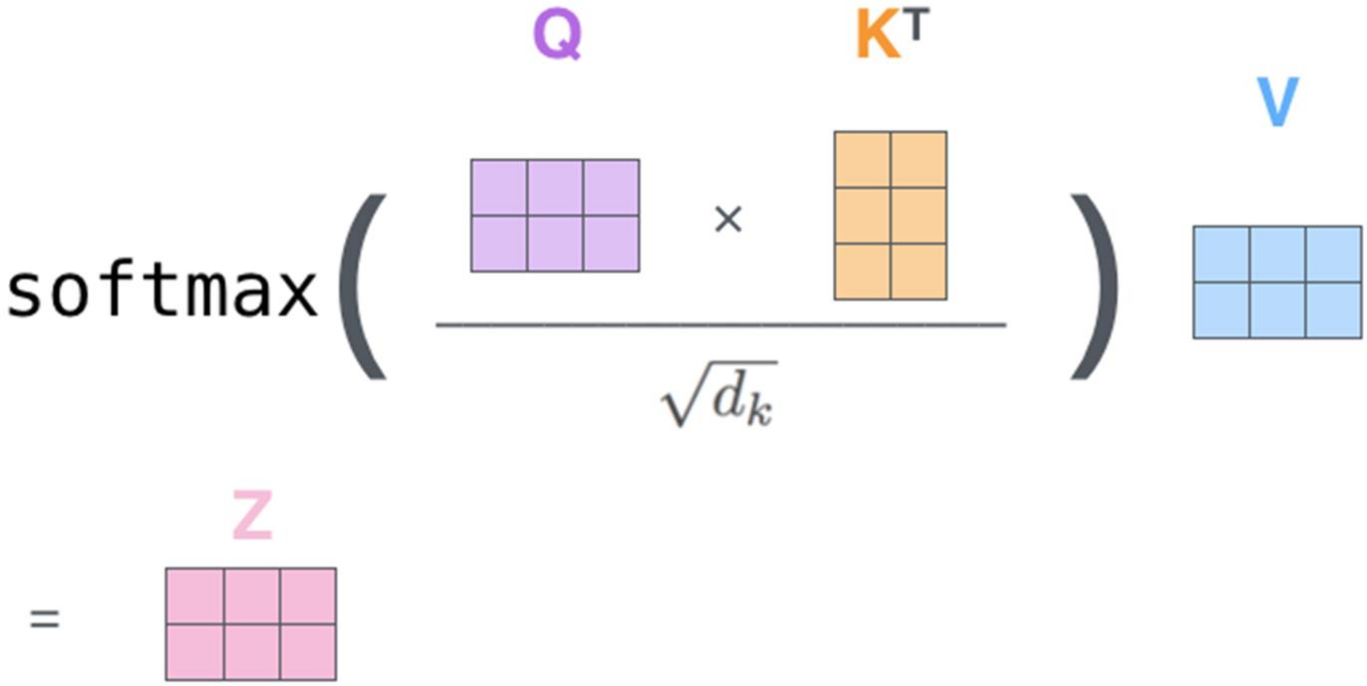




Self Attention Layer

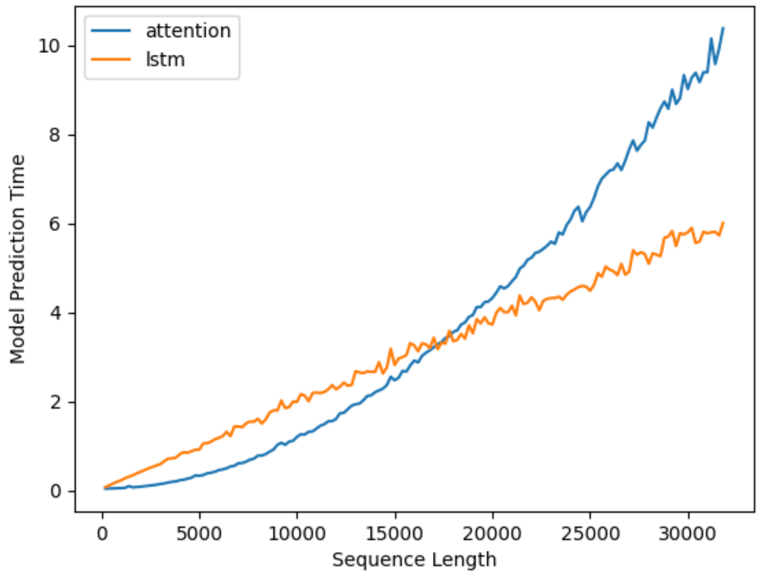
Self Attention Equation

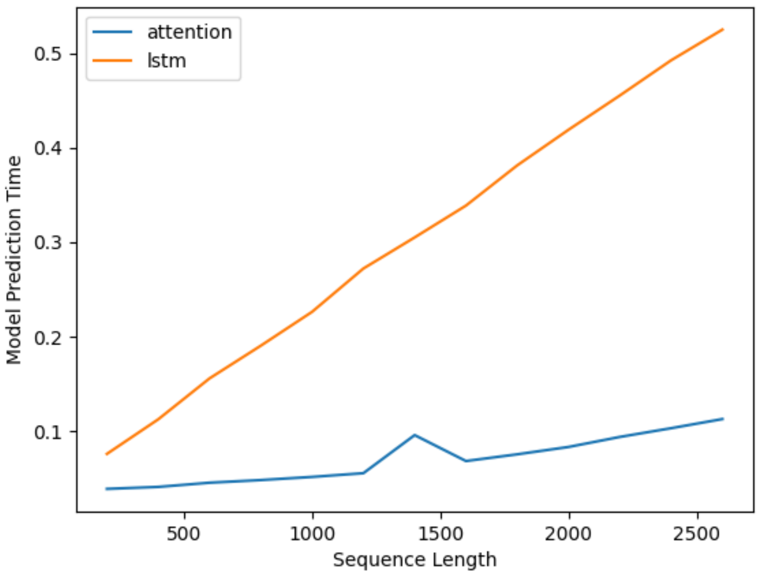
- Derive from input: Q, K, and V
- Output: Z



Time Complexity

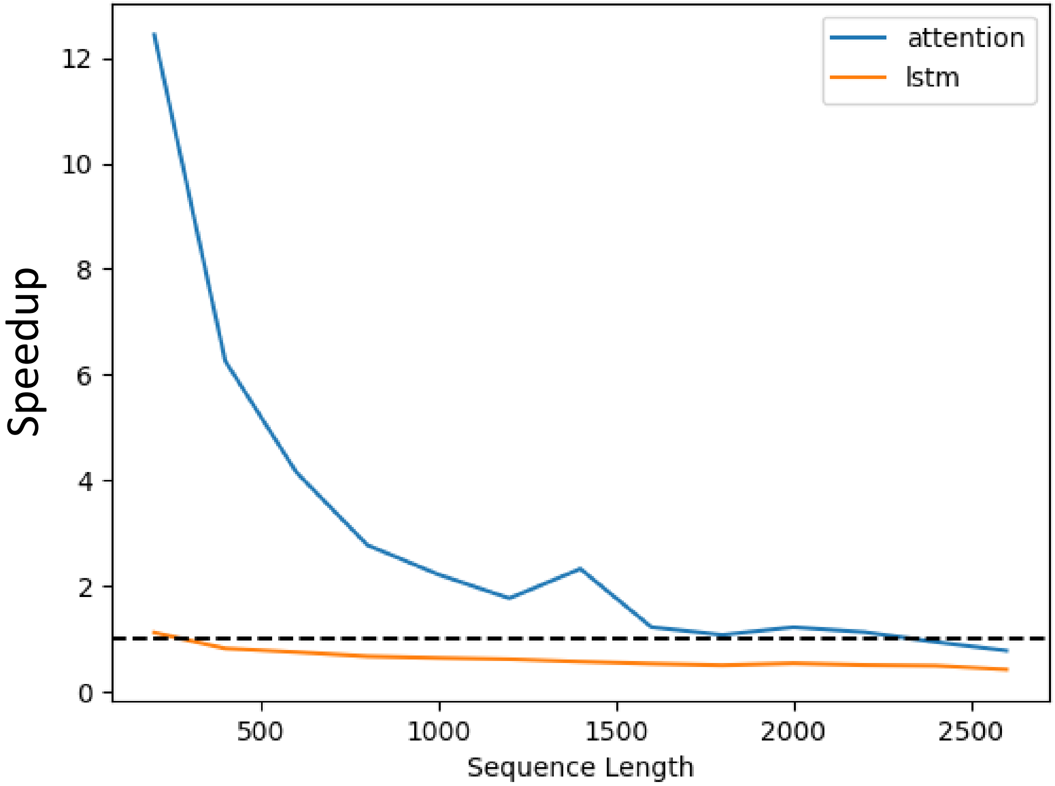
- For sequences shorter than 15,000 attention is faster than LSTM
- 15,000 is a ridiculous length for a machine learning time series problem anyway





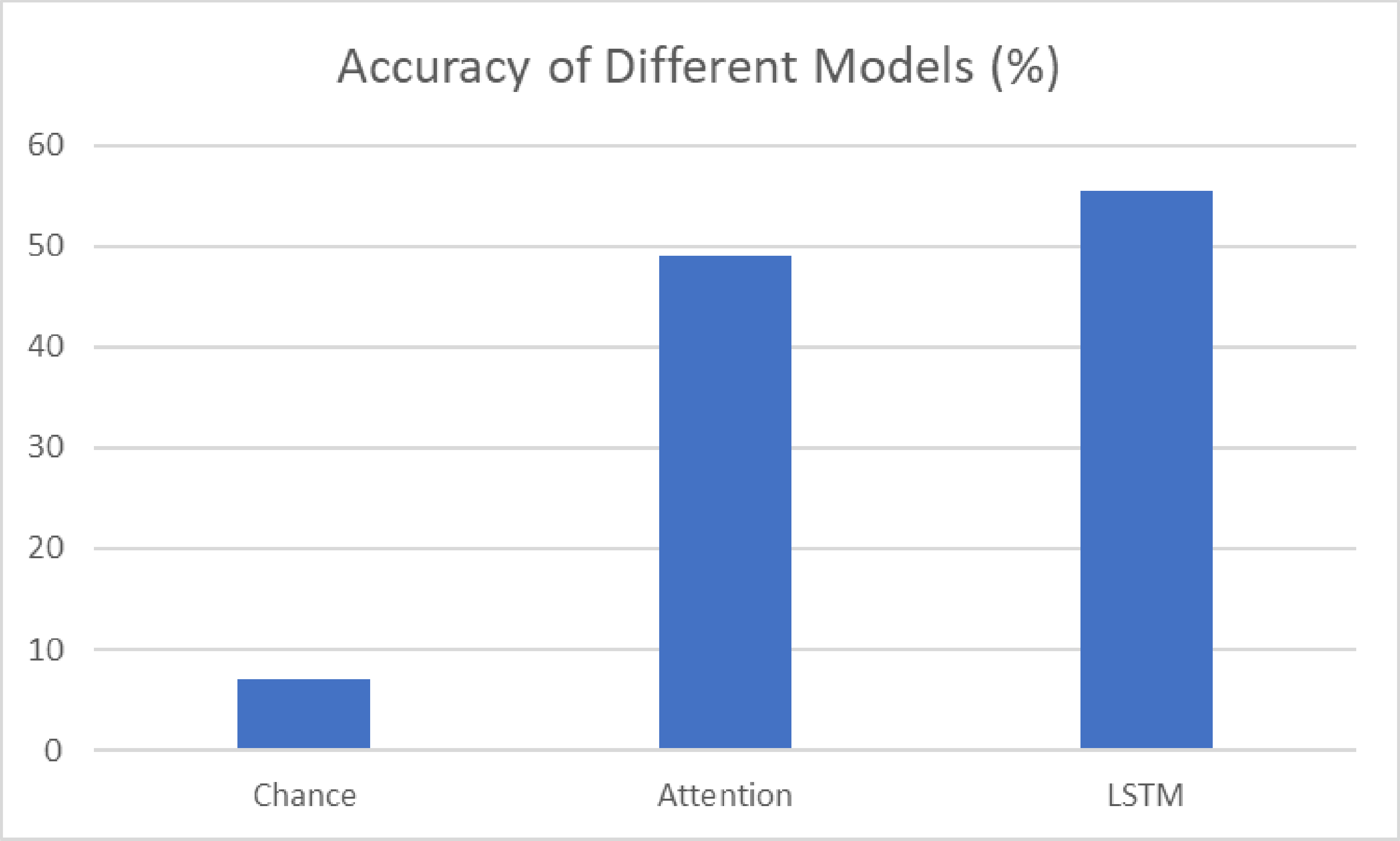
Parallelization Speedup

- Attention model is better able to harness the parallelization benefit of GPU
- 1,500 where this benefit becomes minimal is an extremely long time series in practice



Accuracy

- Trained to 100 epochs, test accuracy was on average 49% for attention, 55.5% for LSTM
- With the amount of classes being predicted, picking randomly is 7%



Conclusions

- Attention based models are significantly faster than LSTMs and can further leverage parallelization
  - However, a single layer is slightly less accurate than a single LSTM layer
- For future work:
  - Attention layer has more parameters than a LSTM layer, I would like to see how these affect accuracy
  - I would like to test the attention model on other datasets to verify these results

Thank You

- Thanks to Dr. Gebremedhin and Skylar Norgaard for working with me this summer
- This material is based upon work supported by the National Science Foundation Research Experiences for Undergraduates Program under Grant No. 1757632.
- Thank you for listening to my presentation

Questions?