Shriya Chinthak (sc2045) and Sanskar Shah (ss4308)

Professor Stratos

Natural Language Processing

18 April 2023

Milestone Report

Our final project is the recreation of the paper *Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation*. This paper discusses ALiBi (Attention with Linear Biases); a position method that allows for a language model the extrapolate large text inputs after being tested on exclusively short inputs. Within this report, we will discuss our current progress and our next steps prior to submitting our final report.

In terms of the replication process, we have successfully been able to implement ALiBi on our own in order to complete text-to-text processes, such as translation. Primarily, we created a multi-headed attention object and a function to calculate the slope given the number of heads. Using this slope function, we created a function that calculates the ALiBi biases, in matrix form, using the number of heads within the attention layer and the attention mask of the model. In doing so, we were able to reach a benchmark result for translation problems. Through a series of tests, we were able to confirm our ALiBi works for translation.

Our next steps in order to create a successful final report is to finalize the model, test our ALiBi with other models to verify training speeds, as well as cross reference these findings with other peer-reviewed sources. Currently, we have read additional papers regarding the uses of attention for short and long inputs as well as the implications of short inputs on language models. However, we would like to continue researching the importance of multi-headed attention on extrapolation, translation, autocompletion, and many other uses of language modeling.