Problem & Background

- In society, proficiency in writing ability is a key skill for success. Despite this, less than one-third of high school seniors in the United States demonstrate writing proficiency.
- In this project, our group seeks to explore a successful autoregressive learning model that identifies argumentative components of written documents - such as claims, evidence, rebuttals, conclusions, etc.
- Identification of argumentative writing structure is a difficult natural language processing task that requires establishing abstract relationships between local and global textual context.
- This presents a potential solution to help students improve their writing abilities with little to no cost.

Conclusions

- With 3 days remaining in the competition, we placed 188 out of 1997 teams. The model performed with very high accuracy: the best models in the competition performed less than 5% better than our model and we performed less than 4% worse than human annotators.
- However, we had our fair share of issues. Despite having 16 GB of memory, the model frequently ran out of memory when trying to train our models. This was a huge issue and led us to adapt our models with gradient accumulation and using small batch sizes. This also meant we could not use larger models with more parameters.

Methods

**The Longformer**

- The Long-Transformer (Longformer) is a state-of-the-art Transformer model designed for processing long sequences.
- Transformer attention scales quadratically with sequence length. Since our model will be reading and classifying long pieces of text (some over 4000 words) we require a faster attention computation. Therefore, we used a sliding window attention pattern.

**Gradient Accumulation**

- Since our CPUs only have 16GB of memory and we are training very large transformer models, it became nearly impossible to use a batch size greater than 1.
- Gradient accumulation fixed this issue by effectively simulating a higher batch size by computing and storing the gradient for the last N forward passes, averaging them, and then doing a single backpropagation step with this mean gradient. This leads to more accurate weight updates and faster training.

Results

We were quite surprised to find that the Longformer that was pre-trained with additional data performed worse than when loading the original pretrained embeddings. The BigLongformer also performed much worse than expected. The Longformer with Oversampling model surprisingly boosted our accuracy by over 6% a hefty improvement to the baseline Longformer. We attribute this to the additional feed-forward layer and extra training time on data that was better distributed because of the oversampling.

Notable Experiments

**Sentence Based BERT**

- Since argumentative components of essays are sentences, we thought a sentence-based classifier would naturally produce respectable results.
- This was not the case as the accuracy was only 26%.

**Longformer + Big Bird**

- The idea behind the model was that maybe using contextual embeddings from two different transformers would provide better contextual information to the feed-forward layers.
- This did not seem to work as loss and accuracy stagnated very early.

**Longformer with Oversampling**

- After finalising the model architecture to be a single Transformer, we decided to oversample the data as some classes were severely under-represented.

- This model had significantly better performance than our baseline and BigBird Model with the loss consistently dropping across epochs.