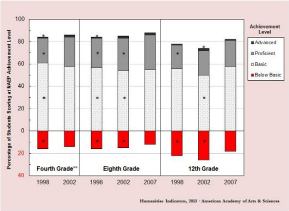
CS 224N: Evaluating Student Writing - Kaggle Competition

Anthony Riley, Sohit Gatiganti, Chris Chankyo Kim

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, rebuttal, 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.



Humanities Indicators, 2013 · American Academy of Arts & Sciences

Fig 1. Writing proficiency of American students over time.

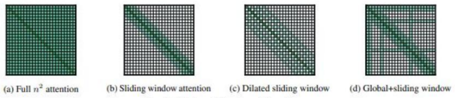
Conclusions

- 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 GiB on our GPU we 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.



(a) Full n^2 attention
(b) Sliding window attention
(c) Dilated sliding window
(d) Global+sliding window

Fig 2. Sliding Window Attention

Gradient Accumulation

- Since our GPUs 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 fixes 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 BigLongBirdformer 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.

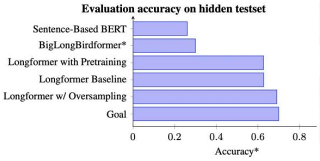
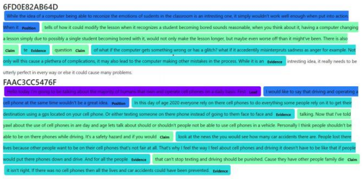


Fig 4. Results of all our models



6FD0E82AB64D

While the idea of a computer being able to reconize the emotions of sudents in the classroom is an intresting one, it simply wouldn't work well enough when put into action. When it Position tells of how it could modify the lesson when it recognizes a student becoming bored sounds reasonable, when you think about it, having a computer changing a lesson simply due to possibly a single student becoming bored with it, would not only make the lesson longer, but maybe even worse off than it might've been. There is also Claim te Evidence question Claim of what if the computer gets something wrong or has a glitch? what if it accedentily misinterpruts sadness as anger for example. Not only will this cause a plethera of complications, it may also lead to the computer making other mistakes in the process. While it is an Evidence intresting idea, it really needs to be utterly perfect in every way or else it could cause many problems.

FAAC3CC5476F

Hello today I'm going to be talking about the majority of humans that own and operate cell phones on a daily basis. First Lead I would like to say that driving and operating a cell phone at the same time wouldn't be a great idea. Position In this day of age 2020 everyone rely on there cell phones to do everything some people rely on it to get their destination using a gps located on your cell phone. Or either texting someone on there phone instead of going to them face to face and Evidence talking. Now that I've told yawl about the use of cell phones in are day and age lets talk about should or shouldn't people not be able to use cell phones in a vehicle. Personally I think people shouldn't be able to be on there phones while driving. It's a safety hazard and if you would Claim look at the news the you would see how many car accidents there are. People lost there lives because other people want to be on their cell phones that's not fair at all. That's why I feel the way I feel about cell phones and driving it doesn't have to be like that if people would put there phones down and drive. And for all the people Evidence that can't stop texting and driving should be punished. Cause they have other people family die Claim it isn't right. If there was no cell phones then all the lives and car accidents could have been prevented. Evidence

Fig 5. Sample labeled output from the model

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.

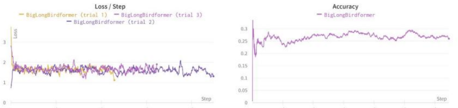


Fig 4. Loss and Accuracy of Longformer + Big Bird

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.

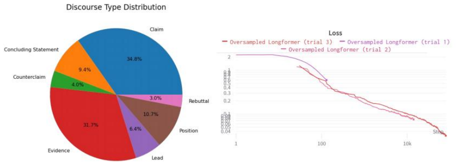


Fig 5.Class Distribution and Loss for Longformer w/ Oversampling

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