I. Title of the project:

Lightweight Sequence2Sequence Sentence Predictor

II. Team:

Andrei Cozma, Hunter Price

III. Problem definition:

We propose a Lightweight Sequence2Sequence Sentence Predictor to predict a following collection of tokens given the current or past collection of tokens. Our model will be trained on research papers or essays to help other researchers and authors when they hit writer's block. Our model will attempt to improve similar models by reducing their size and complexity; this drastically limits the widespread use of these types of models.

IV. Motivation:

A. Why is it interesting?

Imagine aiding your creativity via conversing with a deep learning model to help create and develop well-thought-out ideas. We like this area because it shows that we can automate many of our written tasks with a deep learning model. This model could be trained on hundreds of millions of academic recourses to give you the best possible results. It could help current-day research and discovery flourish.

B. Where do you think it's going to be used, i.e., application area?

Our proposed technique could be used on smaller devices such as phones because it is not constrained by the hundreds of millions of weights that current-day models have. This could then be used to adapt to each person gaining almost a personality. This model could help write papers, answer questions, and respond to text messages.

V. Literature review: What reading will you examine to provide context and background? Please put citations of the article/blog posts with full citations.

[1] D. Hutchins, I. Schlag, Y. Wu, E. Dyer, και B. Neyshabur, 'Block-Recurrent Transformers'. arXiv, 2022.

[2] I. Lauriola, A. Lavelli, και F. Aiolli, 'An introduction to Deep Learning in Natural Language Processing: Models, techniques, and tools', Neurocomputing, τ. 470, σσ. 443–456, 2022.

[3] L. Ouyang κ.ά., 'Training language models to follow instructions with human feedback'. arXiv, 2022.

[4] A. Neelakantan et al., "Text and Code Embeddings by Contrastive Pre-Training". arXiv, 2022.

[5] R. Nakano κ.ά., 'WebGPT: Browser-assisted question-answering with human feedback'. arXiv, 2021.

[6] K. Cobbe κ.ά.,'Training Verifiers to Solve Math Word Problems'. arXiv, 2021.

[7] T. B. Brown et al., "Language Models are Few-Shot Learners". arXiv, 2020.

[8] A. Tamkin, M. Brundage, J. Clark, en D. Ganguli, "Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models". arXiv, 2021.

[9] D. Su, P. Xu, en P. Fung, "QA4QG: Using Question Answering to Constrain Multi-Hop Question Generation". arXiv, 2022.

[10] M. Dehghan, D. Kumar, και L. Golab, 'GRS: Combining Generation and Revision in Unsupervised Sentence Simplification'. arXiv, 2022.
[11] M. Chen κ.ά., 'Evaluating Large Language Models Trained on Code'. arXiv, 2021.

VI. Dataset: What data will you use? If you are collecting new data, how will you do it?

We are using text data for this project, and therefore we will require word embeddings. These can be gathered from Stanford's GloVe word embeddings database.

We also need a large collection of text to train our predictive model. For this, we primarily chose the arXiv Bulk Data Access dataset, among a few other specialized datasets. We chose this due to its complete collection of thousands of academic papers. We chose this dataset as we thought it would be interesting to see the results of a predictive sentence model trained solely on research papers and other academic publications. We would like to potentially experiment with other specialized datasets within the fine-tuning phases: The Blog Authorship Corpus, the Cornell Movie Dialog Corpus, and the Elsevier OA CC-BY Corpus.

GloVe: Global Vectors for Word Representation (https://nlp.stanford.edu/projects/glove/)
arXiv Bulk Data Access dataset (https://arxiv.org/help/bulk_data_s3)

VII. Proposed method: What method or algorithm are you proposing? If there are existing implementations, will you use them, and how? How do you plan to improve or modify such implementations? You don't have to have an exact answer at this point, but you should have a general sense of how you will approach the problem you are working on.

We propose a lightweight model for next sentence prediction. Our approach will include our own take on a recurrent transformer model. Current solutions have a significant drawback: their computational cost is extremely large. Models consist of 110-340 million or more learnable parameters that require expensive hardware to use. Our solution attempts to alleviate this issue by creating a model that achieves similar results with far fewer weights and less complexity.

VIII. Evaluation: How will you evaluate your results? Qualitatively, what kind of results do you expect (e.g. plots or figures)? Quantitatively, what kind of analysis will you use to evaluate and/or compare your results (e.g. what performance metrics or statistical tests)?

Our model will use quantitative metrics named BLEU (Bilingual Evaluation Understudy Score) and Rouge (Recall Oriented Understudy for Gisting Evaluation. Using both of these metrics will allow us to fine-tune our model for both precision and recall. Another quantitative measure we will use is Perplexity, which is the probability of producing a sentence by the model trained on a dataset. Additionally, we will attempt to look into and report on the LSA and BLEURT metrics.

We will also quantitatively evaluate our dataset by reading through different results to understand why our model gives us specific results.

BLEU: Bilingual Evaluation Understudy Score

- Measures precision: how much the words (and/or n-grams) in the machine generated summaries appeared in the human reference summaries.
    - numer_of_overlapping_words / total_words_in_reference

Rouge: Recall Oriented Understudy for Gisting Evaluation

- Measures recall: how much the words (and/or n-grams) in the human reference summaries appeared in the machine generated summaries.
    - numer_of_overlapping_words / total_words_in_system

Perplexity

- Probability for a sentence to be produced by the model trained on a dataset.
    - Refers to the power of a probability distribution to predict, or assign probabilities, to a sample.
- The lower the perplexity value, the better the model.
- We ultimately want to check perplexity values on the test set and choose the model with the lowest value for this metric.

LSA: Latent Semantic Analysis

- It doesn't punish word choice variation as much as either Bleu/Rouge
    - i.e. lenient on "good" and "nice", whereas Rouge and Bleu would not be.

BLEURT

- delivers ratings that are robust and reach an unprecedented level of quality, much closer to human annotation