Text Generation Using XLNet

(20BCE0218) ANAND SWAROOP | (20BCE2251) TANMAY MEHROTRA | Prof. RAJESHKANNAN R | SCOPE

Introduction

Natural language processing is at the vanguard of technical innovation in this era of rapid change, revolutionising both our interactions with machines and our understanding of the massive amount of textual data. In this context, the cutting-edge language model XLNet stands out as a source of innovation, providing unmatched text production capabilities. The need for more complex and reliable text creation techniques is growing as the demand for rich, relevant information explodes across multiple disciplines. This work explores the field of text generation by using XLNet's capabilities to raise the bar for linguistic coherence and fluency.

Motivation

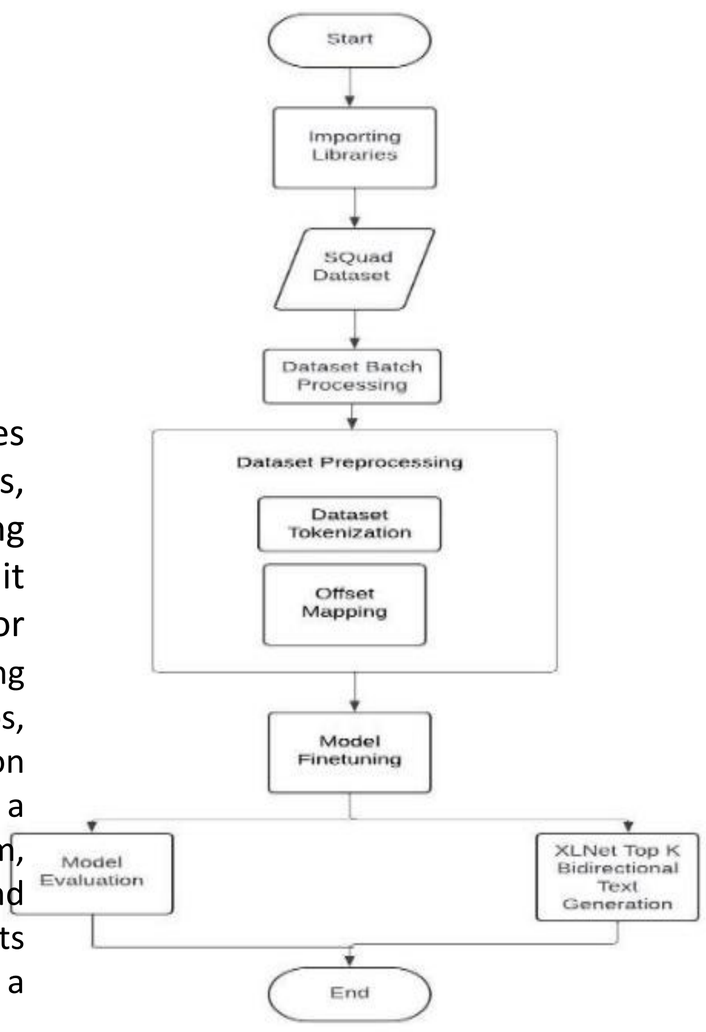
The current landscape of language models, exemplified by GPT-1 and GPT-2, showcases remarkable progress in generating human-like text. However, their unidirectional text generation approach, restricted from left to right, poses inherent limitations on capturing the full context and richness of language. This limitation can hamper coherence and creativity in generated text. By leveraging XLNet's unique pretraining technique, which allows for bidirectional context modelling, we aim to overcome the constraints of unidirectional models

Scope of the project

The goal of this project is to thoroughly investigate the architecture, training goals, and underlying mechanisms of the XLNet text generation model. Acquiring datasets, training models, and fine-tuning them for certain text production tasks will be involved. We shall compare XLNet to other state-of-the-art models in order to show its advantages.

Methodology

The dataset consists of context and a question al with an answer. The st followed by the propo model to generate ans based on the given dataset is



Dataset Cleaning: It involves identifying and correcting errors, inconsistencies, and missing information in a dataset before it can be used for further analysis or modelling. It consists text cleaning for extra spaces, newlines and tabs, spelling correction, Lemmatization which is the process of reducing a word to its base or dictionary form, also known as its lemma and stemming for reducing a word to its base or root form, also known as a stem.

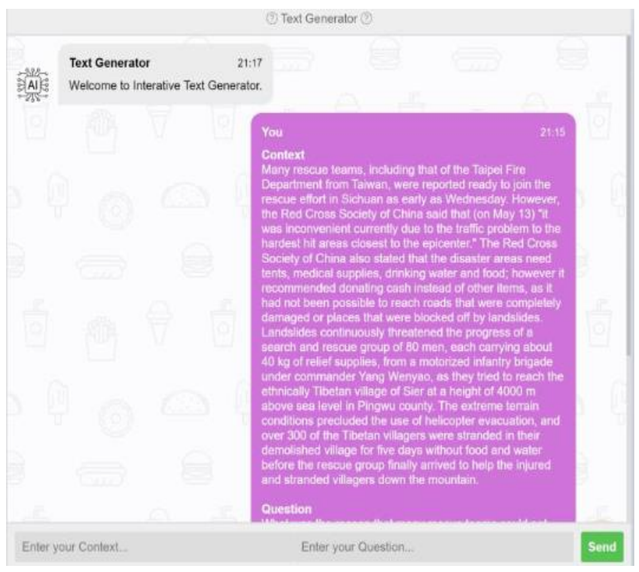
Dataset Preprocessing: Using XLNet tokenizer to convert questions and contexts into sequences of tokens (words or sub words) suitable for the model to train. Offset mapping is applied on tokenized data to obtain a mapping between the original text and the tokenized sequence which will be used later for answer span identification. It involves Tokenization which uses a pre-trained tokenizer and answer span mapping which extract answer text and answer starting positions from the dataset.

Model Finetuning: The pre-processed dataset is passed through the XLnet model. The model will identify the start and end of the answer span within the context. XLNet can leverage its bidirectional processing to understand the relationships between words in the passage and question, improving answer accuracy. It consists of three methods which is Training arguments, Trainer initialization and fine- tuning training.

Top K beam Bidirectional Text Generation: The generated extractive answer is given as input to the Top K Beam Bidirectional Text Generation algorithm for elongated abstractive answer for the question. The steps in the algorithm involves candidate token generation. Based on the directionality provided whether we want to generate text in left or right side or in both directions, using the masked multiheaded attention the next probable token are generated. For as each candidate token it is appended to form a beam of candidate tokens. In this similar way K number of beams are generated by the algorithm and based on the logits value of each beam the beam with the highest probability is chosen and appended to the extracted answer.

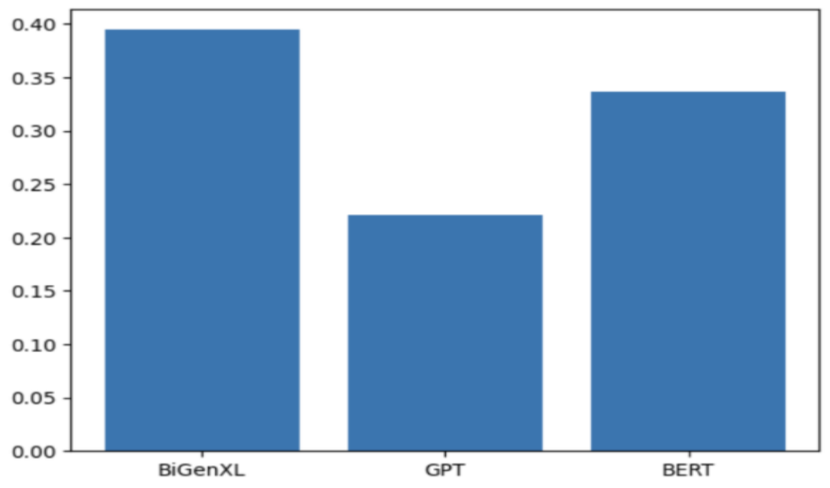
Model Evaluation: The finetuned model is evaluated using F1 Score. The F1 score, also known as the F-measure or F1-measure, is a metric used in machine learning to evaluate the performance of a classification model. It is a harmonic mean that combines two other commonly used metrics: precision and recall. The output obtained by the finetuning the model is compared with the original output given in the dataset and is used to give f1 score.

Result



We successfully implemented our proposed BiGenXL model for generating elongated answers for the given context and Question. The f1 score was used to compare the efficient of Question Answering system. Three models were used namely are BiGenXL, BERT and GPT. The SQuAD dataset was used for evaluating all the efficiency of all three models. The reason for using f1 score was that f1 score is that it provides a good balance between Precision and Recall.

Comparison graph



Conclusion

The successful deployment and testing of the XLNet-based text generation model BiGenXL marks a major step forward in the advancement of NLP. Through careful testing and refinement, we have proven the effectiveness of using bidirectional insight to improve coherence, innovation, and overall output of text generated. Our findings highlight the transformative power of XLNet to overcome the shortcomings of uni-directional models by capturing more complex contextual dependencies.

References

1. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q., V. (2019, June 19). XLNET: Generalized Autoregressive Pretraining for Language Understanding. arXiv.org. https://arxiv.org/abs/1906.08237 .

2. Topal, M. O. (2021, February 16). Exploring Transformers in natural language Generation: GPT, BERT, and XLNET. arXiv.org. https://arxiv.org/abs/2102.08036.

3. XLNET: Generalized Autoregressive Pretraining for Language Understanding (1/3). (2020, May 12). enJOY. https://machinereads.wordpress.com/2020/05/10/xlnet-generalized- autoregressive-pretraining-for-language-understanding-1-3/