Scientific Paper Abstract Generator

Bryan Coronel, Yifeng Jiang

Github Repo: https://github.com/yf-jiang/abstract_generator

1.- Executive Summary

In this project, we plan to utilize pre-trained text summarization models and fine-tune the parameters on a scientific paper dataset so the model can learn to generate an abstract section given the rest of the paragraphs of a research paper. This model could potentially be used by researchers from different fields to generate abstracts for their papers since writing a perfect abstract section could be difficult and time consuming. Due to the similarity between an abstract and a summarization in this context, the model could also be used to summarize research papers for researchers to skim through and decide if it's necessary to read the entire paper. The dataset we use for this project is about research papers archived on Arxiv. In addition, the fine-tuned model would be deployed as an API or as a standalone application for the intended users.

2.- Data Preprocessing

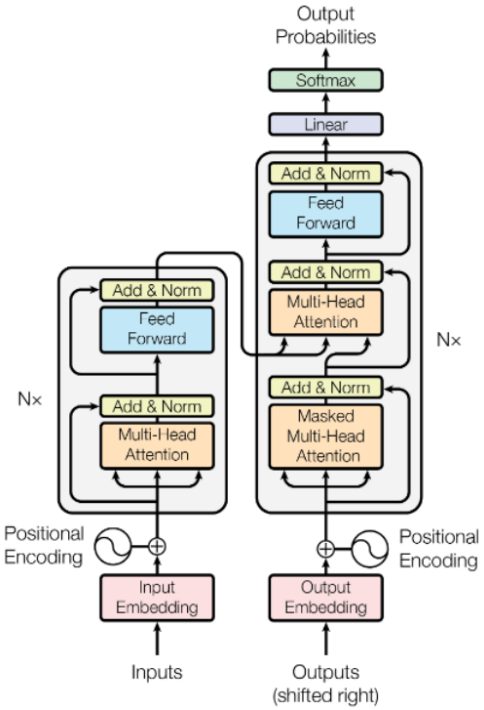
The dataset is stored in JSON format; each line has an abstract, article text, and section names of a scientific paper. The dataset was already divided into three splits which are training, testing, and validation split. There are 215913 instances in total in this dataset and it takes 14.6GB to store the text version of this dataset. Due to the nature of text data, it's unnecessary and almost impossible to visualize the dataset, so a printing output of the dataset is included to demonstrate the structure of this dataset.

	article_text	abstract_text	section_names
0	additive models @xcite provide an important fa...	additive models play an important role in semi...	[introduction, main results on learning rates,...
1	the leptonic decays of a charged pseudoscalar ...	we have studied the leptonic decay @xmath0 , v...	[[sec:introduction]introduction, [sec:detector...
2	the transport properties of nonlinear non - eq...	in 84 , 258 (2000) , mateos conjectured that...	[introduction, regularity and chaos in single-...
3	studies of laser beams propagating through tur...	the effect of a random phase diffuser on fluct...	[introduction, the method of photon distributi...
4	the so - called `` nucleon spin crisis '' rais...	with a special intention of clarifying the und...	[introduction, model lagrangian with pion mass...

Because the dataset was created by other NLP researchers and the data was from published papers on Arxiv, we assume that the text data in this dataset is accurate. Therefore, no data cleaning and outlier detection are conducted in this project. Due to the enormous size of this dataset and the limited computing resources, we have for this project, it's not feasible to fine-tune on the entire dataset. At this stage, we decide to randomly extract 20% of the data for the model training.

3.- Modeling Approach

Initially, the pre-trained model we choose to fine-tune is BART, which was pre-trained by researchers from Facebook for sequence-to-sequence tasks[1]. The model workflow is included below.



However, the result from fine-tuning BART is not ideal since BART was not pre-trained on long text datasets. We found a more advanced model, Longformer Encoder-Decoder, which is designed for this summarizing long text document task. Longformer Encoder-Decoder(LED) uses a different approach to compute the attention which scales linearly to the sequence length instead of quadratically as BART[2]. This attention mechanism allows LED to process longer text compared to BART. The machine learning morphosis of this model is the following.

- Input Space: X = a vector of words
- Output Space: Y = another vector of words
- LM1: Model_Tokenizer(X) + Positional_Embedding(X)
- LM2: ATTN(X_i) = X_i + Q(X_i) * K(X_i)^T * V(X_i)
- LM3: LM1([start token/previous word]) * LM2^16

  Overall: Y = LM1 * LM2^16 * LM3^length of response sequence

After fine-tuning the model, we implement a front-end user interface using Gradio and requests for extracting papers from Arxiv, so the users can use this UI for generating summary/abstracts given the input of papers published on Arxiv or text format of sections of a paper.

4.- Result and Insights

In this project, the evaluation metric we used is rougeLsum[3] which is 2*precision*recall / (precision + recall) of the longest common subsequence between the predicted abstract and the After fine-tuning 20% of the training set and validating on the validation set, the fine-tuned model achieves 0.2475 f1 rougeLsum on the test set which is around 0.2 increase compared to the baseline LED model. We prove that fine-tuning a large language model(LLM) is effective for this sequence-to-sequence generation task. Due to the limited computation resources available to use, we are only able to finetune 20% of the research paper data and that still takes about 15 hours on a single Nvidia Tesla A100 GPU. If we are able to get more computing resources or more quota on a single card, we could potentially fine-tune more data and get a better result.

5.- Conclusions

In conclusion, we have demonstrated the effectiveness of utilizing pre-trained language models for generating abstracts in scientific papers. By fine-tuning the Longformer Encoder-Decoder (LED) model on a dataset of research papers from Arxiv, we were able to improve the performance of the baseline model. Our fine-tuned model achieved a RougeLsum score of 0.2475 on the test set, indicating its ability to generate informative and concise abstracts.

The results suggest that fine-tuning a large language model (LLM) can significantly enhance the sequence-to-sequence generation task, despite the limitations imposed by computational resources. With additional computing power, it is plausible to further improve the model's performance by fine-tuning on a larger portion of the research paper dataset. Moreover, future work could explore incorporating domain-specific knowledge or leveraging domain-specific pre-training to enhance the model's understanding of scientific terminology and context. Additionally, the deployment of the model as an API or standalone application would provide researchers with a valuable tool for automatically generating abstracts and summarizing scientific papers, saving time and effort in the publication process. Overall, our solution and findings would impact researchers from various disciplines in the abstract generation and summarization tasks, paving the way for advancements in scientific communication and dissemination.

6.- Reference:

1 Lewis, Mike, et al. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension." arXiv preprint arXiv:1910.13461 (2019).

2 Beltagy, Iz, Matthew E. Peters, and Arman Cohan. "Longformer: The long-document transformer." arXiv preprint arXiv:2004.05150 (2020).

3 HuggingFace metric: rouge https://github.com/google-research/google-research/tree/master/rouge