Evaluating Pre-Trained Language Models on Multi-Document Summarization for Literature Reviews

Benjamin Yu
Georgia Tech
byu92@gatech.edu

Abstract

Systematic literature reviews in the biomedical space are often expensive to conduct. Automation through machine learning and large language models could improve the accuracy and research outcomes from such reviews. In this study, we evaluate a pre-trained LongT5 model on the MSLR22: Multi-Document Summarization for Literature Reviews Shared Task datasets. We weren't able to make any improvements on the dataset benchmark, but we do establish some evidence that current summarization metrics are insufficient in measuring summarization accuracy. A multi-document summarization web tool was also built to demonstrate the viability of summarization models for future investigators: https://ben-yu.github.io/summarizer

1 Introduction

With recent advances in natural language processing and deep learning, large language models are now capable of generating summaries of large volumes of documents that are arguably human readable and logically consistent. With the growing amount of research being published, it has become increasingly difficult to process all the available research and literature in any particular field of study. This has become exceedingly important within the biomedical field as the community has learned with the global COVID-19 pandemic. Speed of research directly impacts patient outcomes and how fast medical practitioners can respond to a constantly changing health landscape. The MSLR22: Multi-Document Summarization for Literature Reviews shared task proposes a challenging research problem that pushes current state of the art multi-document summarization models to generalize over two different datasets: MS^2 Dataset (DeYoung et al., 2021) and Cochrane Dataset (Wallace et al., 2020) We will evaluate in this research study if pre-trained summarization models can successfully solve the proposed task.

2 Related Work

Recent studies in document summarization have mostly focused on Transformer-based models, but applied to the biomedical context either through transfer learning or fine-turning on a specific biomedical dataset (Wang et al., 2021). BioBERTSum is a recent example of using such pre-training methodologies, which used a pre-trained model as an encoder and fine-tuned on a specific task (Du et al., 2020). (Moradi and Samwald, 2019) innovated in this space by applying hierarchical clustering to group contextual embeddings of sentences to select the most informative sentences from a given group to generate summaries. (Sotudeh et al., 2020) also recently proposed a mechanism to leverage domain knowledge and embed it into their SciBERT-based clinical abstractive summarization model.

Scaling such transformer models to longer input sizes has been difficult since the attention layers get exponentially larger and become computationally infeasible to train. Recent advances in model architecture like PEGASUS (Zhang et al., 2019) and Longformer (Beltagy et al., 2020) have introduced different ways around this by introducing sparse attention mechanisms like local attention which replaces the full-attention mechanism with a sparse sliding window. Researchers at Google were able to innovate on these findings further by combining pre-training strategies from PEGASUS along with a new sparse attention mechanism called Transient Global which mimics ETC's local/global attention mechanism and achieve state of the art performance on multiple summarization benchmarks. (Guo et al., 2021)

3 Data Analysis

3.1 MS^2 Dataset

The MS^2 dataset consists of 470k studies mapped to 20k reviews from PubMed (DeYoung et al., 2021). The dataset was further augmented with

Proceedings of the Third Workshop on Scholarly Document Processing, pages 188–192
October 12–17, 2022.

PICO span labels and evidence inference classes. The goal for this dataset is to generate an accurate summary given a set of multiple review abstracts.

To understand the relative difficulty of this summarization task, we measured text similarity between abstracts and their target summaries based on Term Frequency–Inverse Document Frequency (TF-IDF) and Jaccard similarity.

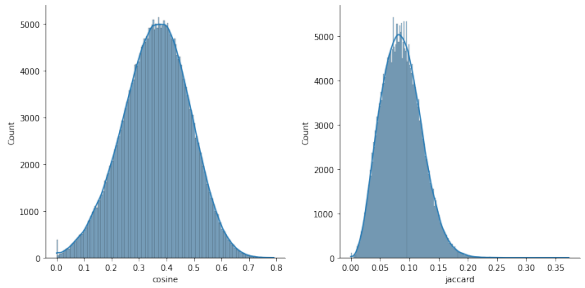

Figure 1: Distribution of MS^2 distances from abstract to target summary

The mean cosine difference was 0.4 and Jacard distance was 0.1. This indicates there was no substantial overlap between the target summaries and their source reviews.

3.2 Cochrane Dataset

This was a smaller dataset of 4.5K reviews collected from Cochrane systematic reviews (Wallace et al., 2020). This dataset was cleaner than the MS^2 dataset, but substantially smaller. The reviews on average included 10 trials each and the average abstract length of included trials was 245 words. We use the authors' conclusions subsection of the systematic review abstract as our target summary (75 words on average).

We also did a similar measurement of cosine and Jaccard distances for the Cochrane dataset:

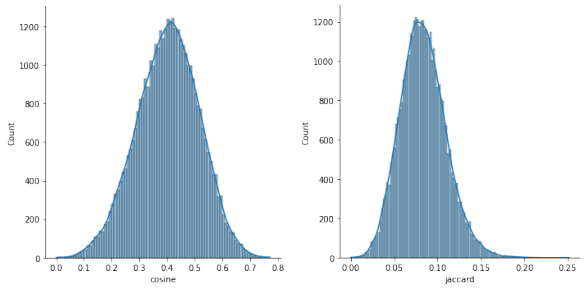

Figure 2: Distribution of Cochrane review distances from abstract to target summary

Similar to the MS^2 dataset, the cosine and Jaccard distances were normally distributed and had roughly the same average difference from their target review. This seemed to indicate that both datasets were similarly difficult and have roughly the same level of sentence overlap.

4 Experiments

The original goal of this study was to experiment with two different approaches to the MSLR22 Shared Task:

1. Fine-tune LongT5 models with both datasets
2. Evaluate existing LongT5 language models on similar datasets like PubMed (Cohan et al., 2018)

We selected the LongT5 model due to its purported state of the art performance numbers and its ability to scale its input size to up to 16384 tokens. We leveraged several cloud providers such as Google Cloud and AWS Sagemaker along with HuggingFace's transformers library for model fine tuning (Wolf et al., 2019). We also experimented with HuggingFace's AutoTrain framework to automatically search for the correct hyperparameters for training. All we had to provide was an initial training and validation datasets, and AutoTrain automated the model training and tuning process. To allow the model to train on multiple documents at once, we pre-processed the training data such that all review abstracts with the same Review ID were appended into a single input string. The single input would then be fed into our model of choice after doing some minimal input validation like checking if the input isn't more than our maximum token length of 16384. We immediately hit several limitations with cloud training including not having sufficient spend to qualify using larger GPU instances for training. HuggingFace's AutoTrain framework also never successfully completed and would often timeout after several days of training. We also attempted to fine tune our models locally, but we only had access to a single RTX 3080 10GB GPU which couldn't even fit the model and dataset even with a batch size of 1. Our conclusion from this experience has demonstrated how the trend towards larger language models might risk increasingly making this type of research inaccessible to hobbyists and practitioners. State-of-the-art model performance will likely only be achieved by researchers with access to compute power and capital unless we prioritize research into reduce model size and resource utilization.

-	Training	Training Target	Test
Characters	1745.81	435.60	1746.66
Words	299.88	68.53	301.42
Sentences	11.2	2.74	11.17

Table 1: MS^2 Dataset Properties

-	Training	Training Target	Test
Characters	1526.79	489.8	1510.42
Words	224.3	72.2	221.14
Sentences	10.2	3.4	10.09

Table 2: Cochrane Dataset Properties

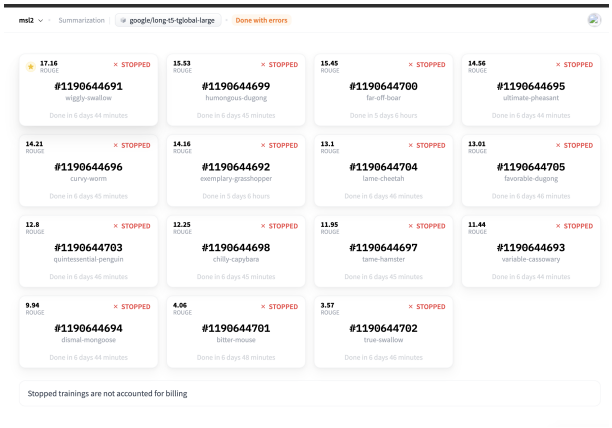

Figure 3: HuggingFace AutoTrain on LongT5

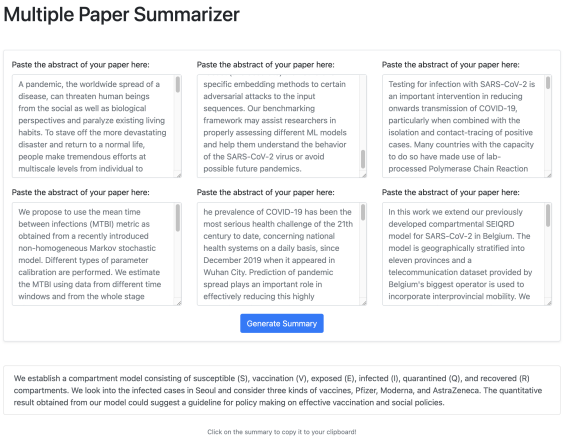

Figure 4: Multi-Document Summarization Tool with HuggingFace Inference API

For our second approach, rather than fine-tuning a base model, we wanted to evaluate if a model that was pre-trained on a similar dataset would still be able to solve this summarization task without any fine-tuning. We found a pre-trained LongT5 model on the PubMed dataset that was trained for around 3k steps (Stancl, 2022). We believed the fine-tuning should be transferable to these datasets as they largely cover the same type of biomedical content and the MS^2 dataset also gets its training data from PubMed. We leveraged HuggingFace's Inference API for model evaluation against the MSLR22 datasets. This also restricted our ability to fine-tune the output size which probably also hindered our performance.

To aid in the model development process and also as a validation that these summarization models have a practical use, we created an online tool that allows anyone to invoke the models for any 6 paper abstracts. The tool can be found at: https://ben-yu.github.io/summarizer

5 Discussion

Unsurprisingly the pre-trained models were unable to exceed the dataset benchmarks on the shared task. One key failing came from our inability to configure target generation length using HuggingFace's Accelerated Inference Text2Text Generation API. On the MS^2 Dataset our outputs only had an average sentence length of 1.1 and character count of 87.97, which significantly deviated from our target length of 2.74 sentences and 435.6 characters. This likely due to the out-of-the-box model not properly generalising over the entire PubMed dataset as the model was also only trained for about 3k steps and further training steps would have improved it's performance. The Rouge-L scores were particularly indicative, scoring sometimes up to 50% worse than the benchmarks. Increasing our model output length would have likely dramatically improved our Rouge scores. Our model didn't score that poorly in terms of a delta EI on the MS^2 dataset with only a 0.06 difference from the Long-

Model	R-1	R-2	R-L	EI↓	F1	BERT
BART Benchmark	0.2626	0.0770	0.1950	0.4509	0.4142	0.8636
Longformer Benchmark	0.2637	0.0795	0.1961	0.4621	0.4118	0.8666
LongT5 - Pubmed	0.1200	0.0133	0.0961	0.5280	0.3433	0.8276

Table 3: Model performance on MS^2 Dataset

Model	R-1	R-2	R-L	EI↓	F1	BERT
BART Benchmark	0.2397	0.0671	0.1760	0.2081	0.3348	0.8632
Longformer Benchmark	0.2387	0.0655	0.1755	0.2345	0.3316	0.8641
LongT5 - Pubmed	0.1130	0.0154	0.0903	0.4671	0.2873	0.7863

Table 4: Model performance on Cochrane Dataset

former benchmark. This could be an indicator that delta EI is a flawed metric that doesn't adequately capture the factual correctness of a summary. Recent work by (Otmakhova et al., 2022) evaluated Longformer and BART models along similar metrics and showed that both models failed to pick up and aggregate important details when manually evaluated against with expert human evaluators. Stronger metrics will likely be required in the future if there is to be significant progress in this domain.

We also found that experimenting with language models and training these large language models can be extremely cost prohibitive and potentially inaccessible to hobbyists and novice machine learning practitioners. These models are getting increasingly large and can't be built unless one has access to sufficient GPU-computing or cloud resources. Training these models can take upwards of 48 hours and there is no guarantee that your model is improving or converging at a reasonable rate.

6 Conclusion

We weren't able to improve upon existing benchmarks for either the MS^2 or Cochrane datasets. We did show there is a need for stronger summarization metrics that can capture different linguistic dimensions such as factual correctness and readability. The summaries from our pre-trained model were significantly shorter than the target summaries and often factually incorrect upon manual inspection, but this couldn't directly be inferred from our model scores outside of comparing it to task benchmarks.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. Ms2: Multi-document summarization of medical studies.

Yongping Du, Qingxiao Li, Lulin Wang, and Yanqing He. 2020. Biomedical-domain pre-trained language model for extractive summarization. *Knowledge-Based Systems*, 199:105964.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2021. Longt5: Efficient text-to-text transformer for long sequences.

Milad Moradi and Matthias Samwald. 2019. Clustering of deep contextualized representations for summarization of biomedical texts.

Yulia Otmakhova, Karin Verspoor, Timothy Baldwin, and Jey Han Lau. 2022. The patient is more dead than alive: exploring the current state of the multi-document summarisation of the biomedical literature. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5098–5111, Dublin, Ireland. Association for Computational Linguistics.

Sajad Sotudeh, Nazli Goharian, and Ross W. Filice. 2020. Attend to medical ontologies: Content selection for clinical abstractive summarization.

Daniel Stancl. 2022. Longt5 large 16384 pubmed 3k step checkpoint.

Byron C. Wallace, Sayantan Saha, Frank Soboczenski, and Iain J. Marshall. 2020. Generating (factual?) narrative summaries of rcts: Experiments with neural multi-document summarization.

Benyou Wang, Qianqian Xie, Jiahuan Pei, Prayag Tiwari, Zhao Li, and Jie fu. 2021. Pre-trained language models in biomedical domain: A systematic survey.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.