Abstractive Summarization of Long Medical Documents with Transformers

Stanford CS224N Custom Project

Luciano Gonzalez
Computer Science
Stanford University
lucigon@stanford.edu

Sabrina Lu
Mathematical and Computational Science
Stanford University
slu12@stanford.edu

William Buchanan
Computer Science
Stanford University
wbuchan@stanford.edu

Abstract

Using transformer models, we perform the task of long-document summarization through the use of an extractive and abstractive step. Abstractive summarization of long documents is limited by transformer's finite context windows; the extractive step allows us to generate a representative sample of the entire text to feed into our abstractive model. We use this summarization approach to create abstracts for medical papers within the PubMed dataset. This approach has potential for wider applications summarizing long documents. Additionally, work like this may serve as an initial step towards the task of automated understanding of technical language. Our results show that the use of pre-trained transformers lead to improvements in the extractive step and potential qualitative improvements in the abstractive step.

1 Introduction

An effective summarization mechanism has far-reaching benefits. From summarization of legal text to research papers, a well-functioning summarization system has the ability to boost the efficiency and ease with which many users access or parse technical information. Abstractive summarization of long documents has presented itself to be a difficult problem to solve using current Transformer architectures. Most work in the summarization space has been limited to shorter documents due to context-window restrictions for transformer models.

We chose to contribute to the space by implementing a summarization model which combines recent advances in extractive and abstractive to work around this context-window limitation. By building a model that is capable of performing abstractive summarization on longer documents, we widen the space of possible applications for our summarization model.

For this particular project, we focus on the PubMed dataset, where we train our system to generate abstracts for medical papers. Medical papers present a useful dataset to train on since each paper contains its own summary (the abstract), and a model that performs well on this task presents a useful tool for parsing dense medical documents. More importantly, we envision that if the model is able to perform well on the PubMed dataset, then it could be applied to perform summarization on a number of other datasets that aren't limited to research papers.

2 Related Work

The earliest works on summarization focused on extractive techniques, which extracted the most salient words and sentences from the original document. This work includes using neural networks (Kageback et al. 2014; Yin and Pei 2015) and recurrent neural networks (Cheng and Lapata 2016; Nallapti et al. 2017) to map sentences into vectors that would be used to select sentences[1][2]. There was also further work done in combining recurrent neural networks with graph convolutional networks to quantify the salience of each sentence (Yasunaga et al. 2017)[4]. However, extractive summarization limits the summaries to existing phrases in the original document, so recent attention has been placed on abstractive summarization, which allows for greater versatility in the summary.

The first abstractive summarization task was brought up in 2015, in which an attention-based encoder was used to generate a summarization from the input (Rush et al. 2015)[5]. The ensuing years saw further advances from this seminal work through techniques involving a variational auto-encoder (Miao and Blunsom 2016) and neural networks based on the attentional encoder-decoder models (Nallapati et al. 2017)[3][6]. However, though these models achieved high ROUGE scores, they lacked an understanding of what was factual, and only used words within the original vocabulary. See et al (2017) made advancements on this issue by proposing a pointer-generator model which allows from the generation of unseen words in the summary[7].

Most recently, Subramanian et al. (2019) combined the extractive and abstractive steps in order to create a better overall abstractive summarization model [7]. The authors utilize a LSTM model to perform their extractive step before performing abstractive summarization with a transformer that they trained specifically for the task. The authors indicate that the LSTM extraction may be a performance bottleneck in their task. As such, we will improve on their implementation by using a pre-trained transformer model for both the extractive and abstractive step. We believe that this will improve performance on the extractive step, as transformer models have shown vast improvements over traditional RNNs, and on the abstractive step because a pre-trained model will benefit from more information.

3 Approach

In this project, we perform summarization over long technical documents using both an extractive and abstractive step. Transformers have limited-size context windows which limit their ability to perform summarization over long documents; the mixed extractive and abstractive approach attempts to remedy this issue. The models used for extractive and abstractive summarization are trained separately, and then used sequentially to perform our mixed summarization approach.

First, important sentences are extracted from a document through the use of a BERT-based extractive summarizer. The extractive summarizer creates sentence-level embeddings for each sentence in the document, which are then passed into either a classifier or clustering algorithm to identify the best summary sentences.

Next, these extracted sentences are fed as input to the BART transformer model (along with as much of the document's introduction as fits into the remaining space of the context window). The transformer then creates a summary based off these extracted sentences. Performing the extractive step allows the summary to be conditioned on important sentences from throughout the document, despite the limited context window.

3.1 Models

3.1.1 BERT

For our extractive step we used two BERT-based extractive summarizers. BERT, or Bidirectional Encoder Representations from Transformers, is a transformer model designed to create bi-directional embeddings of words via unsupervised pre-training (Devlin et al. 2018) [8]. This is an encoder-only model, which takes text input and encodes it into a high-dimensional vector representation that is useful for a number of downstream tasks. For our initial baseline extractor we followed the lecture-summarizer approach, where BERT sentence embeddings are clustered using K-means (Miller 2019) [9]. We leveraged the `bert-extractive-summarizer` library with an out-of-the-box

pre-trained BERT base as a simple baseline. This creates sentence embeddings by averaging the BERT word embeddings from the second-to-last encoder layer. These embeddings are clustered using the K-means clustering algorithm, making clusters of semantically-similar sentences. Finally, it selects the sentence embeddings closest to the cluster centroids to the extracted sentences. This gives us a summary of K sentences extracted verbatim which are representative of the entire document. Our code uses the `bert-extractive-summarizer` library for encoding and clustering, which we augmented with scaffolding for data processing, testing loops, and calculation of metrics.

For our fine-tuned extractor, we used DistilBERT fine-tuned on Pubmed abstracts as described in the BertSum paper (Liu 2019) [10]. DistilBERT is a version of BERT compressed via knowledge distillation, with 40% fewer parameters than BERT base but comparable results on most tasks (Sanh et. al. 2019) [11]. We used a version of DistilBERT fine-tuned on extracted summarization on the PubMed dataset. The model was fine-tuned to produce sentence embeddings for a binary sigmoid classifier identifying which sentences should be included in a document summary. This was done using the body text of the paper as the input and the ground-truth abstract as the target, with binary classification entropy loss. We downloaded the fine-tuned model from a link in the `TransformerSum` documentation. We use this model to encode each document in several 512-token chunks. Encoding in chunks allows us to get word embeddings within a large context, providing richer embeddings. These are then pooled into sentence embeddings and fed to an inter-sentence transformer layer which extracts the top two sentences from each chunk. Our code uses an implementation of BertSum from the `TransformerSum` package, which we modified to update the tokenizer and use the fine-tuned DistilBERT model. As with our other baseline extractor, we use this package with original code scaffolding for data processing, testing loops, and calculation of metrics.

3.1.2 BART

For the abstractive step of our approach, we use the BART transformer model (Lewis, 2019). BART is a denoising autoencoder designed to be used for pretraining sequence-to-sequence tasks. For pretraining, the model is fed arbitrarily-corrupted text and taught to predict the original text. The model architecture is an encoder-decoder architecture that encodes the corrupted text and then uses the decoder to autoregressively generate its prediction of the original text. BART's encoder-decoder architecture allows it to leverage the benefits of both model types. Encoder models like BERT predict missing tokens independently of each other, and are thus not useful for text generation. Decoder models like GPT predict tokens autoregressively and thus are useful for text generation, but they can only consider dependencies in the context left of the word it is generating (Radford et al., 2018). The encoder-decoder model allows BART to learn bi-directional dependencies in text, but still be useful for text generation.

For our baseline, we used the `sshleifer/distilbart-cnn-12-6` model available on the Huggingface library. We tested this model as both a purely abstractive baseline, and a mixed approach baseline. For the purely abstractive baseline, we fed the model full papers as input (allowing it to truncate the papers to fit in the context window) and compared the generated summaries to the ground truth abstracts. For the mixed approach baseline, we provided sentences extracted using the two extraction models discussed above as input to BART, and compared the generated summaries to the ground truth abstracts.

We later finetuned BART on the pubmed dataset. We trained BART to predict ground-truth abstracts to medical papers, given the sentences extracted by the BERT clustering extractor. The code for finetuning bart was adapted from the following Huggingface script: `https://github.com/huggingface/transformers/blob/master/examples/seq2seq/run_summarization.py`. This code makes use of Huggingface's `Seq2SeqTrainer` class which computes cross entropy loss in order to perform finetuning.

4 Experiments

4.1 Data

We used the PubMed dataset of full medical papers. PubMed is a life sciences search engine including 32 million citations from MEDLINE and other medical journals. Our specific dataset included 133k plaintext versions of medical papers. For each paper, the dataset includes metadata, the abstract, and

the various sections of the paper. The lack of tables and figures in the original paper PDFs led to relatively clean text files which required little pre-processing on our part.

4.2 Evaluation method

As our baseline evaluation metric, we use ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores in order to compare our model's performance with those from the previous paper (Lin, 2004). The ROUGE-N metric measures the overlap of n-grams between the generated and reference summaries. ROUGE-L, on the other hand, scores a summary based on the longest common sequence between it and the reference summaries. The metric labeled ROUGE-N Recall refers to the original ROUGE scoring metric, where as ROUGE-N Precision refers to a ROUGE-like scoring metric which focuses on precision rather than recall. These values are similar to BLEU scores, simply lacking the length scoring penalty.

4.3 Experimental details

4.3.1 Baselines

We ran experiments testing our baseline approaches to extractive, abstractive, and mixed summary generation. For our extractive model baseline, we used out-of-the-box pre-trained BERT to perform extraction over the first 500 papers in our test dataset. We used the `bert-extractive-summarizer` model to create sentence embeddings, clustered the sentences with K-means (K=10), and chose the sentences closest to the centroids. For our second BERT baseline extractor (BERT-pub), we encoded each document in several 512-token chunks (a maximum of 6 chunks for computational efficiency) and concatenated the top 2 sentence embeddings from each chunk to make our summary.

For the abstractive model baselines, we generated summaries using the `sshleifer/distilbart-cnn-12-6` model available on the Huggingface library. To maintain consistency with the BERT baseline experiments, we ran this model on the first 500 papers in the test dataset. For a purely abstractive baseline, we provided the entire article as input to the model and allowed the tokenizer to automatically truncate the article to fit the context window. This means that the abstractive BART baseline generated summaries based off the first 512 tokens of each article.

For our mixed approach baseline, we again used the `sshleifer/distilbart-cnn-12-6` model, but this time changed its inputs. For this baseline, we provided BART with extracted sentences as input for which to base summaries off of. We experimented with two mixed baselines, one which used BERT clusting-based extractions, and one which used BERT-pub classification based extractions.

4.3.2 Fine Tuning

For our finetuning experiments, we finetuned BART on the Pubmed dataset. As input, we provided BART with sentences extracted using the BERT extraction method (followed by as much of the rest of the article fit in the context window), and had it train to generate the ground truth abstracts for each article. Due to time and compute constraints, we were limited to training on a subset of the training dataset containing 10,000 examples. We saved the model after 3 epochs of finetuning and again after 5 epochs of finetuning. At test time, we ran both the BERT and BERT-pub extractors on 500 test set examples, and used these as input to the two models.

4.4 Results

Qualitative Results
Below are some examples of our generated abstracts compared the ground truth abstracts.

Ground truth abstract for "Bilateral Heel Numbness" paper: we describe the case of a 32-year - old woman who developed bilateral heel numbness after obstetric epidural analgesia . we diagnosed her with bilateral neuropathy of the medial calcaneal nerve , most likely due to longstanding pressure on both heels . risk factors for the development of this neuropathy were prolonged labour with spinal analgesia and a continuation of analgesia during episiotomy . padded footrests decrease pressure and can possibly prevent this neuropathy .	**Generated abstract for "Bilateral Heel Numbness" paper:** introduction : with an incidence rate of 0.92%, maternal puerperal lower extremity nerve injuries are rare. lateral femoral cutaneous neuropathy (meralgia paraesthetica) is the most common, followed by femoral neuropathy. bilateral medial calcaneal nerve neuropathy as a result of external compression is a rare complication of epidural obstetric analgesia.case presentation : a 32-year - old woman presented to our neurology outpatient clinic with tingling and numbness in both heels. she continuously complained about these sensations ever since she had given vaginal birth to her first child 3 months earlier in a hospital.
Ground truth abstract for "Ebola Virus Outbreak" paper: in the wake of the ongoing 2014/2015 ebola virus outbreak , significant questions regarding the appropriate handling of ebola virus - contaminated liquid waste remain , including the persistence of ebola virus in wastewater . to address these uncertainties , we evaluated the persistence of ebola virus spiked in sterilized domestic sewage . the viral titer decreased approximately 99% within the first test day from an initial viral titer of 106 tcid50 ml1 ; however , it could not be determined if this initial rapid decrease was due to aggregation or inactivation of the viral particles . the subsequent viral titer decrease was less rapid , and infectious ebola virus particles persisted for all 8 days of the test . the inactivation constant (k) was determined to be 1.08 (2.1 days for a 90% viral titer decrease) . due to experimental conditions, we believe these results to be an upper bound for ebola virus persistence in wastewater . wastewater composition is inherently heterogeneous ; subsequently , we caution that interpretation of these results should be made within a holistic assessment , including the effects of wastewater composition , dilution , and potential exposure routes within wastewater infrastructure . while it remains unknown if ebola virus may be transmitted via wastewater , these data demonstrate a potential exposure route to infectious ebola virus via wastewater and emphasize the value of a precautionary approach to wastewater handling in an epidemic response .	**Generated abstract for "Ebola Virus Outbreak" paper:** ebola virus is considered a potential bioterrorism agent. we conducted a cell culture assay to determine the initial viral titer and the subsequent response to a recent outbreak of the disease in west africa. the titer decreased rapidly (approximately 99in addition, viral particle aggregation or adsorption to wastewater particles may play a role in the apparent rapid viral decrease and enhanced viral persistence. on the basis of these results , we recommend that wastewater be disinfected prior to disposal of ebola - contaminated liquid waste into sewage systems.
Ground truth abstract for "Malakoplakia" paper: malakoplakia is an uncommon but distinctive type of chronic granulomatous inflammation that occurs most commonly in the genitourinary tract , especially the urinary bladder . most patients have associated conditions characterized by some degree of immunosuppression , as seen in solid - organ transplants , autoimmune diseases requiring steroid use , chemotherapy , chronic systemic diseases , alcohol abuse and poorly controlled diabetes . we report an unusual case of the renal malakoplakia that involved the perirenal space , extending to the descending colon in a 65-year - old korean woman with secondary adrenal insufficiency and diabetes mellitus .	**Generated abstract for "Malakoplakia" paper:** malakoplakia is a rare chronic inflammatory disease that occurs most commonly in the genitinary tract, especially in the urinary bladder. most patients have some degree of immunuppression, as seen in solid - organ transplants, autoimmune diseases requiring steroid use, chemotherapy, chronic systemic diseases, alcohol abuse and poorly controlled diabetes. here , we report an unusual case of renal malaksoplakia involving the perirenal space and extending to the descending colon in a 65-year - old korean woman with secondary adrenal insufficiency and a long history of use of exogenous steroids.

Quantitative Results
Bolded values dictate which model scored highest in a given metric across the models on the same table. Starred (*) values dictate which model scored highest in a given metric across all the tables.

Baseline results:

Model	BERT (Ext)	BERT-pub (Ext)	BART (Abs)	BART/BERT	BART/BERT-pub
ROUGE-1 Precision	26.1	25.3	**51.9**	50.5	50.3
ROUGE-1 Recall	61.5	**63.2***	18.8	18.6	18.3
ROUGE-1 F1	33.8	**34.8**	26.2	25.7	25.4
ROUGE-2 Precision	9.5	9.9	**18.5**	16.2	16.7
ROUGE-2 Recall	22.0	**25.4***	6.6	5.7	6.0
ROUGE-2 F1	12.0	**13.7***	9.3	8.0	8.3
ROUGE-L Precision	13.6	12.8	**33.8***	32.0	32.6
ROUGE-L Recall	31.9	**33.7**	12.2	11.7	11.8
ROUGE-L F1	17.3	**17.9**	17.0	16.1	16.4
Avg Length (words)	536	502	61	62	62

Finetuned results:

Model	BART 5e/BERT	BART 5e/BERT-pub	BART 3e/BERT	BART 3e/BERT-pub
ROUGE-1 Precision	**52.7***	49.9	52.0	49.7
ROUGE-1 Recall	**30.6**	28.8	29.8	28.5
ROUGE-1 F1	**36.7***	34.7	36.0	34.4
ROUGE-2 Precision	**19.8***	18.8	19.2	18.5
ROUGE-2 Recall	**11.3**	10.6	10.9	10.5
ROUGE-2 F1	**13.6**	12.9	13.2	12.7
ROUGE-L Precision	**31.0**	30.0	30.6	30.2
ROUGE-L Recall	**18.1**	17.4	17.7	17.3
ROUGE-L F1	**21.7***	20.9	21.2	20.9
Avg Length (words)	105	105	104	104

In the table above, BART 5e refers to the BERT model finetuned for 5 epochs, and BART 3e to the 3 epoch model.

Original Paper Results:

Model	Lead-10 (Ext)	Sent-CLF (Ext)	Sent$_P TR(Ext)$	TLM-I (Abs)	TLM-I+E(G,M) (Mix)
ROUGE-1 Recall	37.45	**45.01**	43.30	37.06	42.13
ROUGE-2 Recall	14.19	19.91	**17.92**	11.69	16.27
ROUGE-L Recall	34.07	**41.16***	39.47	34.27	39.21

5 Analysis

Extractive Model
In their paper, Subramanian et. al. introduce two novel approaches to long document summarization - splitting the process into an extractive and abstractive step, and employing transformers for the abstractive step. We expand on this methodology by adopting their two-fold approach, while also testing the effectiveness of using a transformer model for the extractive step. Our hypothesis was that utilizing a transformer for the extraction would further improve summarization results. Our results using the ROUGE metrics show that both of our extractive transformers were able to achieve higher ROUGE scores than the three extractors that the authors tested, supporting our hypothesis.

This was expected, since in general transformers have demonstrated the ability to outperform a wide range of legacy models on various tasks. More specifically, because transformers are trained on large corpuses, they develop a better understanding of the natural language, and are therefore able

to generate expressive sentence embeddings. In the case of the clustering-based extractor, these embeddings led to meaningful clustering, and thus high quality extractions. For the classification based extractor, BERT-pub, expressive sentence embeddings and the ability of tranformers to capture long-range dependencies led to high-quality extractions.

Mixed Model
Our mixed model came short of performing at the same level as the model in the original paper. This is likely due to a number of limitations. First, due to our compute power and time limitations, we were forced to finetune on our model on only 10k medical papers, as opposed to the whole dataset of over 133k papers. Additionally, we only finetuned for 5 epochs. Our results indicated a noticeable performance boost from from training for 5 epochs as opposed to 3, so it is likely that more training time would have led to additional performance gains. Additionally, due to shortage in time, we were only able to finetune our mixed model using the original, clustering-based extraction method. Had we more time, we could have finetuned another model using the the BERT-pub extractions as input. Though this model wouldn't be guaranteed to achieve better performance, it did hold the potential to.

Generated Abstracts
In taking a look at our generated abstracts in comparison to the ground truth abstracts, there are a few disparities that stand out. The first is that our generated abstracts appear to be more generalized than the ground truth abstracts, which often jump directly into the case being examined. To get a better understanding of why, we take a look at the "Malakoplakia" paper. In this case, our generated abstract is very similar to the ground truth abstract. A closer examination at the original report shows that in this particular paper, much of the language in the ground truth abstract is present in the introduction. The reason why our model predicted this abstract so accurately may be that the model has learned to pay close attention to the introductions of papers. Sentences in the introduction typically provide generalized context for the paper, and this would explain why our generated abstracts tend to be so generalized.

Another interesting difference is that our abstracts appear to be written in a more vernacular language that the ground truth abstracts, which are often filled with complex medical terminology. This is likely because our transformers are trained on a large corpus of data filled with everyday language. As such, it is more likely to choose the words and phrases that it has seen before. This could be seen as a benefit, because it increases the readability of these technical papers, making them more accessible. While the papers themselves are technical, abstracts are meant to be able to reach a broader audience, and thus our generated abstracts serve this goal.

Returning to the "Malakoplakia" paper example, we also noticed an interesting transfer of information throughout the paper into our abstract. In particular, the sentence "a 65-year-old korean woman with secondary adrenal insufficiency and a long history of use of exogenous steroids" caught our attention. In the original paper, there is no explicit link between the 65 year old Korean women with a history of use of "exogenous steroids". Instead, the paper mentions a few drugs that the Korean women has taken, and later on classifies these drugs as exogenous steroids. What's particularly impressive is that the sentences with these two links are also quite far apart within the paper. Our model is therefore able to take information that is far apart and make connections between them. Though we are unable to know for certain, we hypothesize that because the extractive step allows us to capture sentences from all over the paper; this, combined with the long-term dependencies afforded by our transformers, allows for such transfers of knowledge.

Limitations
The summarization task is a particularly hard one, in large part because of the difficulty in evaluating its performance. Though ROUGE scores are useful because they offer an automated metric, they are ultimately not good measures of readability, coherence, or truth.

As seen in our results section, there are several rouge metrics where the extractive summarizers outperform the mixed approach summarizers. Since ROUGE scores simply measure N-gram overlap, it is expected that ROUGE scores would be higher when creating summaries by extracting text written by the same authors. ROUGE scores have no way of capturing the abstractiveness of a summary, which is another important aspect of the summaries generated by our model.

There is also no automated way to test the truth-value of the generated summaries. In the examples we evaluated by hand, we found examples of generated abstracts which contradicted or misrepresented the content of the original paper. At the moment there seems to be no evaluation method for this

issue other than human evaluation, and thus training a summarizer which preserves truth remains a challenge.

6 Conclusion

Our project demonstrates the effectiveness of using transformers in the task of abstractive summarization. In particular, we show that it quantitatively increases performance in the extractive step and qualitatively provides more context and readability to the abstract. We were able to achieve higher ROUGE scores on our BERT extractor than those from the authors of the original paper we based our project on.

However, our abstractive transformer models were unable to achieve the same scores that Subramanian et. al. attained, largely due to computing limitations. We chose to use the `sshleifer/distilbart-cnn-12-6` model because it was able to train within the limits of our Azure virtual machine. If given greater capacity, we would opt for a larger model such as Pegasus or `bart-large`. In addition, due to our credit limit we were unable to train our `sshleifer/distilbart-cnn-12-6` model for too long, and therefore limited the training to 5 epochs. Again with a larger capacity, we would be able to spend more time fine tuning our abstractive model. We believe that a combination of a larger model and a greater epoch would increase the performance of our abstractive models, and subsequently the mixed models.

For future work, as mentioned above we would like to train larger models to create better summarizations. We would also like to expand the use case of our summarization task to more general long document summarization.

References

[1] Mikael Kageb ack, Olof Mogren, Nina Tahmasebi, and Devdatt Dubhashi. 2014. Extractive summarization using continuous vector space models. *In Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC),* pages 31–39.

[2] Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. I*n Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* volume 1, pages 484–494.

[3] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. *In Proceddings of the 2017 Association for the Advancement of Artificial Intelligence,* pages 3075–3081.

[4] Michihiro Yasunaga, Rui Zhang, Kshitijh Meelu, Ayush Pareek, Krishnan Srinivasan, and Dragomir Radev. 2017. Graph-based neural multi-document summarization. *In Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 452–462.

[5] Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing,* pages 379–389.

[6] Yishu Miao and Phil Blunsom. 2016. Language as a latent variable: Discrete generative models for sentence compression. *In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing,* pages 319–328.

[7] Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointergenerator networks. *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* volume 1, pages 1073–1083.

[7] Subramanian, S., Li, R., Pilault, J., & Pal, C. (2019). On extractive and abstractive neural document summarization with transformer language models. *arXiv preprint arXiv:1909.03186*.

[8] Miller, D., 2019. Leveraging BERT for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*.

[9] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[10] Liu, Y., 2019. Fine-tune BERT for extractive summarization. *arXiv preprint arXiv:1903.10318*.

[11] Sanh, V., Debut, L., Chaumond, J. and Wolf, T., 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

A Appendix

Here is the abstract for this paper, generated by the model we developed:

abstractive summarization of long documents has presented itself to be a difficult problem to solve using current Transformer architectures. most work in the summarization space has been limited to shorter documents due to context-window restrictions for transformer models. we chose to contribute to the space by implementing a summarization model which combines recent advances in extractive and abstractive to work around this context- window limitation. our model outperforms current transformer architectures in both the extractive step and in the abstractive step. in particular, we show that our model quantitatively improves performance in both extractive steps, and qualitatively provides more context and readability to the abstract. importantly, we imagine that the