Ensemble Modeling For Long Document Summarization

Sai Sahana Bhargavi sbyrapu@gmu.ed
Mrudula Adira madira@gmu.edu
Bantwal Shreyas Mallya bmallya@gmu.edu
Medha Bharadwaj mbharadw@gmu.edu

1 Introduction

1.1 Task / Research Question Description

We want to implement an ensemble model for Long Document summarization and incorporate the findings of (Huang, 2021) along with it to attain better summarization. The paper (Huang, 2021) gives the encoder-decoder attention technique namely – HEPOS (Head wise Positional Strides).we intended to use the HEPOS attention mechanism and ensemble the two best performing summarization models to evaluate if the performance of the ensemble can be better than the individual models

1.2 Motivation and Limitations of existing work

NLP models are used to generate summaries and this task becomes difficult when handling long lengths of data that might lead to (1) loss of context (2) Bad summary quality and (3) Model inefficiency.Attention mechanisms add some context to words in the sentence and uses pairwise words relations between tokens and this can be used to improve the summarization models. Frameworks like Pre-trained seq2seq Transformers take O(n2) time complexity.To reduce such quadratic time in self-attention ,selectively attending to neighboring tokens (or) relevant words helps.Yet, these methods do not apply to encoder-decoder attentions in summarization models since they collaborate and dynamically pinpoint salient content in the source as the summary is decoded. Truncation is commonly used to circumvent the issue. However,training on curtailed content further aggravates “hallucination” in existing abstractive models(Maynez et al., 2020).

We believe using HEPOS reduces computational and memory costs while maintaining the power of emphasizing important tokens and preserving the global context. HEPOS attention successfully doubles the processed input sequence size, when combined with any encoder. Our attempt here is to utilize the efficiency of HEPOS and to create an ensemble model and analyze if this ensemble model utilizing HEPOS attention mechanism provides further improvement in ROUGE scores in comparison with fine-tuned BART and HEPOS. After evaluating the differnt techniques such as Pegasus, Roberta and TextRank our observation is that the Ensemble model of BART and TextRank works well together and can be used for successful summary generation for long documents.

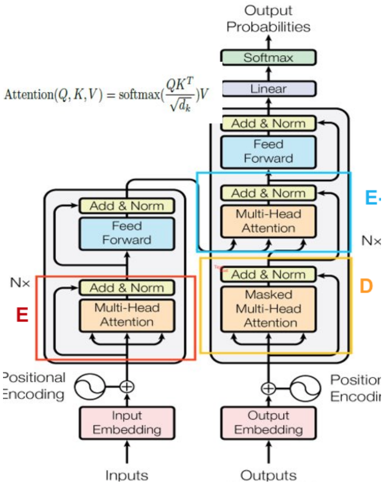


Figure 1:
Transformer

1.3 Proposed Approach

We propose to investigate the performance of abstractive and extractive summarization techniques and propose an ensemble model that could be comparable to the baseline model.

1.4 Key Takeaways

Our baseline motivates us by suggesting that abstractive techniques like BART along with full-attention mechanism fairly improves the performance of long document summarization. Our intention is to enquire further into this by trying to understand the performace of other abstractive and also extractive techniques to understand their merit and build an ensemble model. The idea here is to evaluate whether combining two models that have comaprable summarization performance improves the overall summarization of long documents.

2 Approach

2.1 BART

As mentioned in the previously, we are using BART, a transformer based model(Lewis et al., 2020). This is a denoising autoencoder for pretraining sequence-to-sequence models with bidirectional encoder (BERT) and a left-to-right decoder (GPT).It is trained by (1) corrupting text with an arbitrary noising function, and (2) learning a model to reconstruct the original text.

2.2 HEPOS

We are going to add HEPOS(encoder-decoder attention) on fine-tuned BART(uses full attention). The main reason for using HEPOS is (1)To reduce the redundancy of multi-attention heads, HEPOS Uses separate encoder-decoder heads to cover different subsets of source tokens at fixed intervals.(2)Each head starts at a different position, and all heads collectively attend to the full sequence. The figure below is an example of HEPOS attention, with a stride of 2 and four attention heads. Dark colors indicate that heads 1 and 3 attend to the first and third tokens ("Job" and "home") in the input, heads 2 and 4 look at the second and fourth words ("in" and "care"). Given a stride size of sh, for the h-th head, its attention value between decoder query q(j) (at step j) and encoder key vector k(i) (for the i-th input token) can be formulated as:

$$a_{ji}^{h} = \begin{cases} \text{softmax}(\mathbf{q}_j \mathbf{k}_i), & \text{if } (i-h) \bmod s_h = 0 \\ 0 & \text{otherwise} \end{cases}$$

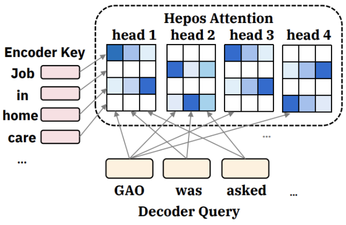


Figure 2: *HEPOS mechanism*

In HEPOS attention, each query token attends to n/sh tokens per head, yielding a memory complexity of O(mn/sh), where m is the output length

2.3 Ensemble Model

In this step we want to investigate how combining BART with other summarization techniques and tools is going to affect the rouge score. We divided this task into 3 steps. (1) Shortlisting the models and techniques, (2) Analyzing the performance individually and (3) Choosing the combination for ensemble.

(1)Shortlisting the models and techniques: Since there are two types of summarization techniques i.e extractive and abstractive summarization techniques respectively, we chose representatives from both these approaches. We chose Pegasus(Zhang et al., 2020) and Roberta(Liu et al., 2019) as the abstractive techniques and TextRank as the extractive technique.

(2)Analyzing the performance individually : Before creating an ensemble model it made sense to understand how each of the chosen techniques behave with our dataset. So the individual performace of the 3 techniques were evaluated on the GovReport dataset.

(3)Choosing the combination for ensemble: After evaluating the individual performances we concluded that TextRank(Mihalcea et al., 2004) is best suited for ensembling with BART. We found that Roberta performed very poorly. Pegasus though had better performance than Roberta, on manual inspection we concluded that the output of Pegasus is a subset of BART output and thus is not favourable for ensemble approach.

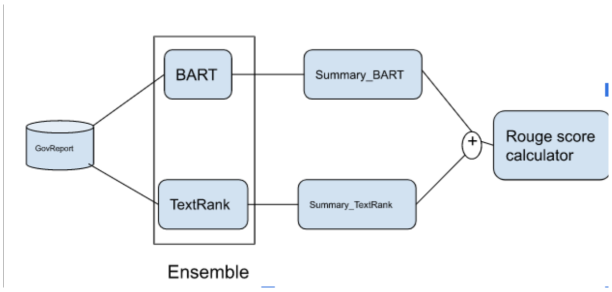


Figure 3: *Our Approach*

3 Experiments

3.1 Datasets

We are using GOVREPORT dataset that contains 19,466 long reports published by U.S. Government Accountability Office (GAO). It contains significantly longer documents (9.4k words) and summaries (553 words). The significance of this dataset is that the salient content is spread throughout the documents, as opposed to cases where summary-worthy words are more heavily concentrated in specific parts of the document. Dataset is split into train, validation and test set by publication date with 17519 training samples, 974 validation documents, and 973 test samples.

The data is in the form of [index, text, summary] thorughout which is very convenient to process.The index indicates the id of the record and the text has the content. Training data : The training data consists of index and the long document and the expected summary for the long documnet. Test data : The test data consists of the index and the document and the expected summary of the document. Val data : The val data consists of a set of data points to be used to validate the training model.

3.2 Baseline

The processed GOVREPORT dataset that contains train,test,validation files of type source and target are taken and converted to byte pair encodings by multiprocessing bpe encoder. Further, these files are converted to dict.source.txt and dict.target.txt.After successful preprocessing, the examples were grouped into 272 iterators and 144 invalid samples. The data is initially being truncated to 1024 tokens. We were also able to obtain the ids of first few examples for iteration1 which are as follows: [5721, 2133, 4499, 468, 14028, 14986, 4658, 5808, 5311, 466]. We are fine-tuning BART on GOVREPORT dataset where the learning rate is set to 0.00001 and learning rate warmup is applied for the first 10,000 steps. Ada-factor optimizer with a gradient clipping of 0.1 is used. Models are trained on 2 X A100 GPU with 400 GB memory. batch size is of 2 per step and accumulated gradient every 32 steps.

For the baseline, we are using BART large and finetuning it with GOVREPORT. After getting the best checkpoint model, we are using it with the HEPOS attention mechanism and fine-tuning the BART+HEPOS model on the dataset and then we are evaluating the performance of this fine-tuned BART+HEPOS model via ROUGE.

We used Roberta to summarize the government reports. The implementation of Roberta using HuggingFace transformers was used. We found that Roberta does not work well for very long documents like government reports and only works well for small text such as movie or food reviews. Also, the maximum length that the roberta-base model can process is 512. Due to its limited length, the summaries produced were incorrect and hard to understand.

As can be seen from the above table, R-1 score of 13.77 means that only 13.77 percent of the unigrams in the generated summary is present in the reference summary. R-2 score of 0.02 means that the percentage of bigrams in the generated summary from the reference summary is very less.

We experimented with Pegasus as well for the summarization of government reports. Again, the implementation of HuggingFace Transformers was used. Compared to Roberta, Pegasus worked very well for the government reports. But on manual observation we found that the summary generated by Pegasus was a subset of the summary generated by BART. Though individually this performance is great in terms of comparison with BART we chose not to use Pegasus for ensemble model due to the overlap in the summarisation outputs.

We experimented with TextRank too for summarization of government reports. It is based on Google's PageRank algorithm that finds the most relevant sentences in a text. Here the whole text is split into sentences and the graph is built by the algorithm where the nodes of the graph are the sentences and the links are the overlapped words. Finally, the most important nodes of the network of sentences are identified.We observed here that based on the length of summarization parameter we define the rouge score of the predicted summary using the textrank algorithm becomes more comaparable to that of our SOTA. This inspired us to consider textrank for our ensemble model.

After running individual experiments on BART, Pegasus, Roberta and TextRank we observed that BART has the best rouge score , Roberta has the lowest and Pegasus and TextRank have comaprable rouge scores to that of BART. Looking at this we next considered the idea of creating an ensemble model using BART, Pegasus and TextRank. The approach considered here was to run

BART and Pegasus independantly and then combine the summaries generated by these individual runs as a single summary and pass this into Text Rank Algorithm.

Theoritically this approach seemed to have merit but on closer observation we realised that the individual summary generated by Pegasus is a subset of the summary generated by BART, which does not give any value addition to our goal which is to try and provide a better summary than BART. Hence, we decided to not follow this model.

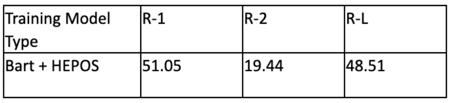
But the idea to ensemble still held merit theoritically and we wanted to see if using just BART and TextRank would give us any better results. Here Text Rank and BART did not have any overlapping sentences and the few samples of overlap that we saw were insignificant as we still had higher non overlapping content than overlapping content. So we continued with this approach and considered experiments within this idea.

The model that we designed here was to first fine-tune long BART using the training and generate the summaries for the test data. Then we ran the TextRank Algorithm on the test data by setting the length of the summarisation ratio. This gave us two set of summaries one extractive from BART and the other based on the TextRank.

3.3 Evaluation Metric

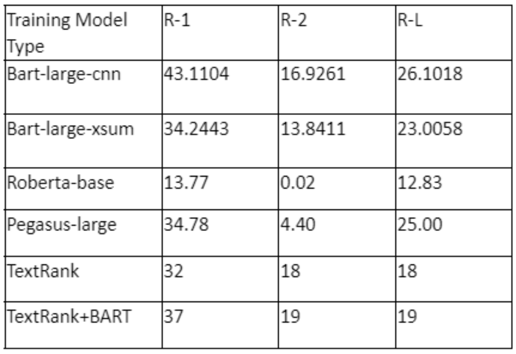
We are using ROUGE score as our evaluation metric.ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It is essentially a set of metrics for evaluating automatic summarization of texts as well as machine translations.It works by comparing an automatically produced summary or translation against a set of reference summaries (typically human-produced). Rouge1: It refers to the overlap of unigrams between the system summary and reference summary. Rouge2: It refers to the overlap of bigrams between the system and reference summaries. RougeL: It measures longest matching sequence of words using longest common subsequence.

3.4 Results



Training Model Type	R-1	R-2	R-L
Bart + HEPOS	51.05	19.44	48.51

Figure 4: *Baseline*



Training Model Type	R-1	R-2	R-L
Bart-large-cnn	43.1104	16.9261	26.1018
Bart-large-xsum	34.2443	13.8411	23.0058
Roberta-base	13.77	0.02	12.83
Pegasus-large	34.78	4.40	25.00
TextRank	32	18	18
TextRank+BART	37	19	19

Figure 5: *Our Approach*

From the tables we can observe that the baseline paper exhibits an R1 score of 51.05. And this is the score we have tried to emulate in our experiments. Our individual implementations of BART, Pegasus, Roberta and TextRank Scores on GovReport are comparable with each other with BART performing the best followed by TextRank, Pegasus and Roberta. Using these score we have picked the 2 best models to ensemble which is BART and TextRank.

Our experiments show that Roberta has the lowest rouge score and on analysis this is because the pre-training model used in Roberta-Base considers words more than sentence context and hence does not have the capacity to generate substantial summaries for long documents.

Pegasus has comparable results but on manual observation we found that the summary of Pegasus is a subset of BART and hence was not considerd for the Ensemble model.

TextRank is an extractive summary technique that has good rouge score and also complements the abstractive nature of BART summarization and hence has been chosen for the ensemble model.

4 Related Work

Transformer based models are known to use full attention mechanism to perform the NLP tasks which leads to increased time and memory complexities. In order to perform the tasks in reduced running time and memory, solutions have been proposed to achieve it by sparse attention mechanism by selectively attending to the tokens(Beltagy et al., 2020; Child et al., 2019) or relevant words(Kitaev et al.,2020; Tay et al., 2020a). However, these were only applicable for encoder side attentions but not for encoder-decoder attention. When working with long documents for summarization tasks, it is necessary for handle both full and encoder-decoder attentions (Huang et al., 2021). We are going to use the idea proposed by (Huang et al., 2021) called HEPOS and apply encoder-decoder attentions on long documents to get better summarization performance.

5 Conclusions and Future Work

In conclusion we observed that the performance of Pegasus and TextRank were comparable to the baseline performance of BART. Summarization using Pegasus gives an output that is a subset of the summary generated using the baseline model. TextRank is an extractive summarizaion technique that works well with the long document summarization. It also complements the summarizaiton generated by the baseline(BART). Due to the complementing nature of TextRank we propose an ensemble model(BART+TextRank) that has comparable results to that of the baseline(BART).

Though we have achieved comparable results using our ensemble model it has not reached the performance of the BART+HEPOS model or surpassed it. By our analysis this can be achieved by further finetuning our model. Also HEPOS is a techniques we observed works well only with BART and does not have a successful implication on other summarization techniques like Pegasus and TextRank. The future of this research problem would be to understand the nature of Pegasus and TextRank further and device a mechanism that can emulate the performance amplification that HEPOS provides for BART.

6 References

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. *Big bird: Transformers for longer sequences. arxiv e-prints, art. arXiv preprint arXiv:2007.14062.*

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. *BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880, Online. Association for Computational Linguistics.*

Rada Mihalcea and Paul Tarau. 2004. *TextRank: Bringing Order into Texts. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pages 404-411, Association for Computational Linguistics.*

Luyang Huang, Shuyang Cao, Nikolaus Parulian and Heng Ji, Lu Wang1. 2021. *Efficient Attentions for Long Document Summarization. arxiv e-prints, art. arXiv preprint arXiv:2104.02112.*

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. *Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.*

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. *Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In International Conference on Machine Learning, pages 11328–11339. PMLR.*