AI Research Project Report

Multi-Article Summarization: Beyond Token Limits Using Informed State Search

Ishaan Shrivastava (B20AI013), Jaisidh Singh (B20AI0150)

Both team members contributed equally in the problem formulation, coding and report preparation.

Please find an interactable version of our work below.

Google Colaboratory

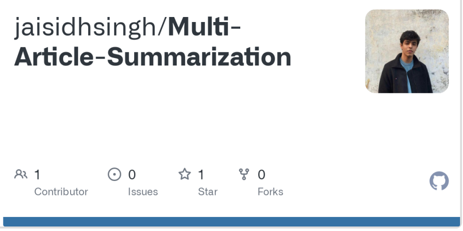
https://colab.research.google.com/drive/1zI626bDeBiPLbrHBK4VgTJ4q94XCV_jY?usp=sharing

A GitHub repository containing our utility code is attached below.

GitHub - jaisidhsingh/Multi-Article-Summarization

transformers (huggingface) sentence-transformers (SentenceBERT) matplotlib nltk (Natural Language Toolkit) The above command loads in an example article set (check

https://github.com/jaisidhsingh/Multi-Article-Summarization



Abstract

Modern summarization approaches in theory can accept infinitely long text sequences, but in practice show poor results after a certain limit. Thus to tackle summarization of multiple articles/large documents, we present a novel clustering-based approach to produce high-quality summaries by solving an informed state search problem. Our method is plug-and-play and allows us to go beyond the token limits of pre-existing summarizers.

1. Brief problem statement

Large articles can be tedious to read for a consumer, thus there exists a need to develop accurate summaries to increase efficiency of work. Modern approaches for the same use sequence-to-sequence modes based on the transformer [3] architecture. However, there exists a limit to the length of their input, called the *token limit*. In theory, a transformer summarization model can consume an infinitely long text sequence as input, however, it is observed that for texts of length beyond the token limit, summarization quality is poor. Thus, the problem statement investigated is: how can we summarize large articles beyond modern token limits accurately ?

2. Background

There have been many works employing SOTA language models for the summarization, the first of which was BART (Bidirectional and Auto-Regressive Transformers) [2]. BART is a Transformer based denoising autoencoder pre-trained on document rotation, sentence permutation, text-infilling, and token masking and deletion objectives. Via this pretraining, BART learns to predict the next words implicitly, making it adaptable to downstream tasks like summarization.

To improve upon the summarization process, several approaches have been adopted. In a work by Google, a transformer architecture called T5 [1] is presented, which utilizes transfer learning to accomplish NLP tasks in a unified text-to-text format. T5 presents high flexibility and can be fine-tuned to a variety of downstream NLP objectives.

Further, Google also develop PEGASUS [4], which follows the same paradigms of self-supervised learning as BERT and T5, while being pre-trained on a more relevant self-supervised learning objective (gap sentence generation) which involves training the model to mimic pseudo-summaries generated by selecting and concatenating key sentences from the documents.

We survey the above models to show background on their plug-and-play usage in our methods and experiments.

3. Methodology

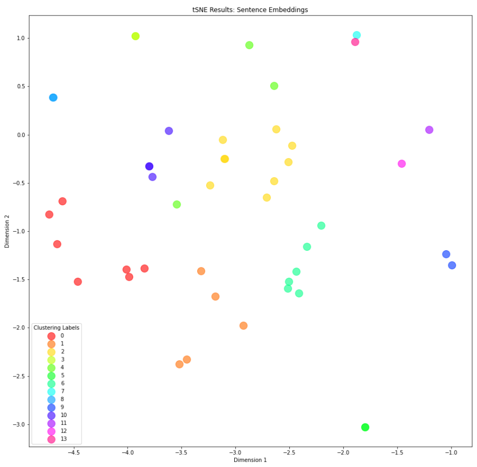
Our methodology presents a novel way to summarize articles larger than the token limits of existing models. Further, this methodology can be used to generate summaries of articles chained together as well. We describe our approach as follows.

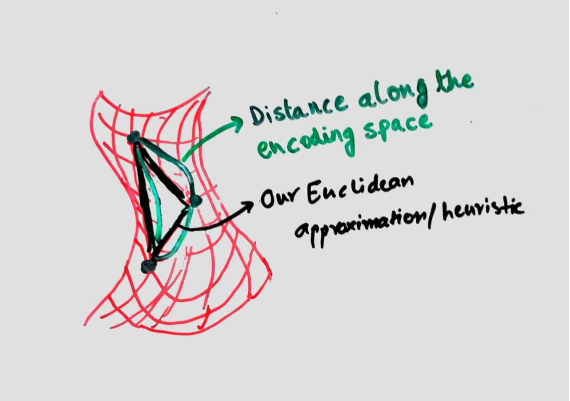
Articles are composed of sentences (sequences of words or tokens). We utilize the information present in the sentences of the article to break it into smaller sub-articles. This is done by first creating an embedding for each sentence. These sentence-embeddings are points in a high dimensional space, which are then clustered using Agglomerative Clustering [5]. Thus, the article shall be broken into clusters of sentences, each of which is summarized independently. These cluster summaries are then presented together as the final summary.

4. Study

The intuition behind our approach is information. Generally, sentences closer together convey the same information (in a context). Thus, these sentences shall be present together in a cluster, leading to meaningful sequential summarization by our process.

The sentence-embedding space is a high-dimensional manifold, where there can exist a distance/metric representing the meanings of two sentence-embeddings, which may not be Euclidean. However, we present an approximation/heuristic for this metric by Euclidean distance. This was developed by taking inspiration from A* search examples. We show the same in the figure below:

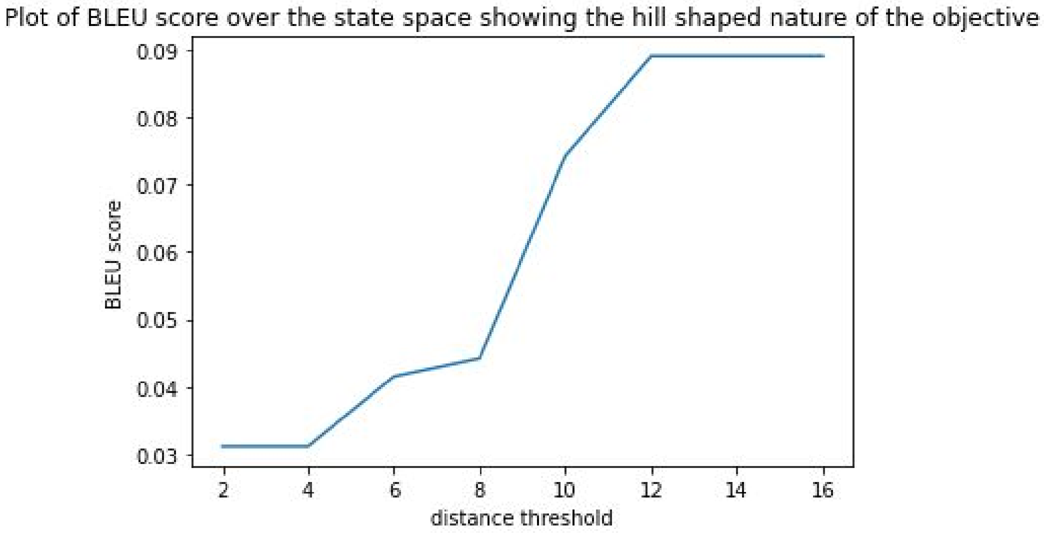



Distance along the encoding space
Our Euclidean approximation/ heuristic

Our approximation (top) shown in a hand-drawn diagram of the high dimensional manifold of the sentence-embedding space, along with sentence-embedding clusters for a test case on the left.

Further, we highlight the clustering hyper-parameter called the *distance threshold.* If two points have distance less than this threshold, they are said to lie in the same cluster. This threshold controls the sizes of the clusters and by extension the summaries. To set the threshold, we consider *informed searches, particularly, hill climbing search* described by the following.

We use text-summary pairs. A predicted summary is generated by our method with an initial value of the threshold. Next, we modify the value of the distance threshold to result in an increase in the BLEU score between the predicted summary and the reference summary. A higher BLEU score implies higher quality summarization. Now, as soon as we detect a drop or a plateau, we stop. This is similar to moving to a state which will place us at a higher ground. Thus, we utilize informed hill climbing search to perform an otherwise very computationally demanding state space search.



5. Results

We provide comparison test reports with existing techniques in the purview of document summarization. We evaluate the summarization quality using our approach versus BART, PEGASUS, and T5. For testing purposes, we report the average BLEU score in the following table:

Document Type	Ours	BART	PEGASUS	mT5
Small	0.07	0.11	0.0	0.0
Large	0.09	0.04	0.0	0.0

On manual inspection, PEGASUS and T5 were found to create summaries that did not successfully encapsulate all of the important information in the inputs provided. This was not the case in BART which performed well on small documents, and produced summaries which were not overly verbose, which was the drawback of our method. However, this drawback become the advantage of our method when using large documents which is beyond the token limits of the other models, showed by the high BLEU score.

6. Conclusions with limitations and future work

Our approach is developed to break the limitation of token limits on summarization tasks. By design, the proposed approach performs well using a plethora of inputs and settings considering the pre-trained modules we use.

It remains to be seen how sentence-embeddings in one cluster can contribute to another. Thus, future work can investigate the effect of composite clusters on the quality of summaries.

7. References

1. Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *J. Mach. Learn. Res.* 21.140 (2020): 1-67.
2. Lewis, Mike, et al. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension." *arXiv preprint arXiv:1910.13461* (2019).
3. Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).
4. Zhang, Jingqing, et al. "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization." *International Conference on Machine Learning*. PMLR, 2020.
5. Müllner, Daniel. "Modern hierarchical, agglomerative clustering algorithms." *arXiv preprint arXiv:1109.2378* (2011).