Abstractive Podcast Summarization using BART with Longformer attention

Hannes Karlbom
Uppsala University
hannes.karlbom@gmail.com

Ann Clifton
Spotify
aclifton@spotify.com

Abstract

In this paper, we present our model submitted to the TREC (Text REtrieval Conference) summarization part of the Podcasts track 2020 edition. The goal of this task is to summarize podcast episodes using 100k open-domain podcast transcripts. The challenge lies in the long length of the transcript documents, diverse structures of the podcast format and that neither the creator descriptions nor the transcripts are a perfect gold truth of an episode. We propose a combined model that tackles the length challenge, by using a drop in replacement of the Longformer attention mechanism in a pre-trained BART model and fine-tuning the model on the podcasts dataset.

I. Introduction

Podcasts have become an increasingly popular form of audio media, and they represent a rich potential source of data. Thus, the TREC podcasts dataset and shared task was created with the purpose of better understanding podcast content. The shared task was split into two objectives, summarization and segment retrieval (Clifton et al., 2020). The overview paper provided by the organizers of this years podcasts track gives a more in depth look at the tasks and the dataset (Jones et al., n.d.).

This work focuses on the summarization task. In particular, we train deep neural abstractive summarization models on the large-scale set of podcast transcripts. Neural summarization models are typically trained on relatively short documents, such as news articles; podcast transcripts, however, tend to be significantly longer (often several thousand tokens long). Thus, popular approaches to neural abstractive summarization do not tend to scale to the podcast transcripts in this dataset. Our approach uses a combination of well established a seq2seq models to overcome this length challenge and generate summaries that can meaningfully take into account the entire input transcript document.

In recent years neural models in Natural Language Processing have greatly improved state of the art by pre-training on language modeling tasks and using the attention mechanism to better generate tokens with respect to the input context. To create a model for summarizing the podcast transcripts we use BART (Lewis et al., 2019), a model which has been shown to have state of the art performance on other summarization tasks such as the CNN/DailyMail dataset (Hermann et al., 2015), and we fine-tune it on the podcast transcript dataset. To tackle the length challenge of the input documents, we extend the model by replacing the attention layers with attention mechanism used in the Longformer (Beltagy, Peters, and Cohan, 2020), which uses a combination of global and local windowed attention that scales linearly with the input sequence length.

II. Background

i. Previous Research

There are numerous models for summarization (Lewis et al., 2019; Yan et al., 2020; Zhang et al., 2019; Raffel et al., 2019) and summarization of long texts, however none of them have been a applied to spoken language transcripts. The methods for dealing with longer inputs range from using extractive methods first and the summarizing the result (Liu and Lapata, 2019) whilst others use a divide and conquer approach (Gidiotis and Tsoumakas, 2020).

Recent work aimed at dealing with problem of long input texts for transformers has seen multiple approaches that have been shown to improve the computational complexity of the attention mechanism and at the same time keeping not losing predictive performance (Zaheer et al., 2020; Beltagy, Peters, and Cohan, 2020; Kitaev, Kaiser, and Levskaya, 2020; Tay et al., 2020).

In this work, we combine some of the these improvements in attention computation with a state of the art model for abstractive summarization, starting from a pre-trained model and directly finetuning on the podcast dataset.

ii. BART & Longformer

The basis for the model we used is a large BART model trained for summarization on the CNN/DailyMail dataset. In this way, the model can leverage the significantly larger CNN/DailyMail dataset to learn the summarization task before adapting to the spoken language podcast transcript domain. BART is a denoising autoencoder for training sequence-to-sequence models and has been shown to be effective for finetuning on text generation generation tasks such as summariazation (Lewis et al., 2019).

Longformer, or the "The long document transformer", is a model which aims to address the problem of quadratic time complexity in the attention mechanism which many transformer models, including BART, suffer from. This attention scaling limits the how long input sequences models can handle and therefore could possibly include in a summary.

The attention mechanism in the Longformer model offers a drop-in replacement for the standard self-attention by using a local windowed attention with the option to introduce global attention to important tokens as well. The windowed attention scales linearly with the input sequence length allowing for much longer input lengths (Beltagy, Peters, and Cohan, 2020).

III. Combined Model

Our combined Longformer-BART model is initialized from an already trained checkpoint of BART and replaces the attention layers in the BART model with the Longformer attention layers. In the original Longformer paper, the model is created starting from a RoBERTa checkpoint's pretrained weights. In order to capture much longer sequences, instead of randomly initializing new position embeddings in the Longformer, it reuses RoBERTa's learned absolute position embeddings by copying the position embeddings multiple times.

The authors show that BERT's attention heads have a strong learned bias toward attending to local context and therefore initialization by copying keeps this local structure everywhere except at the partition boundaries (Beltagy, Peters, and Cohan, 2020).

We use the same method for BART, copying the positional embeddings 4 times, with a max position of 1024, thus reaching a max input sequence length of 4096, which we use when finetuning the model on the podcast transcripts.

Parallel and independent to this work the authors of the Longformer paper have released a revision of the paper where they introduce the Longformer-Encoder-Decoder (LED) (Beltagy, Peters, and Cohan, 2020). The LED model uses the same method as we have have in the combined model and manages to achieve state of the art performance on the arXiv summarization dataset (Cohan et al., 2018) with a 16k

token input length.

IV. Methodology

i. Data Preprocessing

The creator descriptions are noisy and are often too short, too long or do not capture what is relevant to the episodes. In addition, many descriptions contain boilerplate texts such as which platform was used to upload the episode. To improve our supervised fine-tuning, we have done the following:

- We removed examples where creator descriptions are either too long or too short with the boundary conditions set to between 10 and 1300 characters.
- We applied a TF-IDF vectorization of the descriptions which were compared to each other using the cosine distance. Any data points with too similar descriptions (threshold 0.95) were filtered out.
- Finally a BERT (Devlin et al., 2018) based sentence classifier was used identify and remove boilerplate sentences. The boilerplate classifier was trained using a small set of 1000 manually labelled episodes (Reddy et al., 2021).

ii. Experimental setup

To create our models, we combined the Huggingface transformers library (Wolf et al., 2019) with the Longformer attention source code [1]. The initial BART model's starting point was "facebook-bart-larg-cnn"[2].

The models where trained using 4 x NVIDIA Tesla P100. We used a validation set of 1000 episodes, randomly sampled after the preposessing step, to do early stopping and picking of the final model.

[1]https://github.com/allenai/longformer
[2]https://huggingface.co/facebook/bart-large-cnn

V. Results

The submissions were ranked according to a four-point quality score and an additional eight yes/no questions to further estimate the quality of the summary. The following descriptions of the quality score and yes/no questions were sent to the evaluators and act as a reference points for the results in the graphs.

Four-point quality scoring:

(3) Excellent: The summary accurately conveys all the most important attributes of the episode, which could include topical content, genre, and participants. In addition to giving an accurate representation of the content, it contains almost no redundant material which is not needed when deciding whether to listen. It is also coherent, comprehensible, and has no grammatical errors.

(2) Good: The summary conveys most of the most important attributes and gives the reader a reasonable sense of what the episode contains with little redundant material which is not needed when deciding whether to listen. Occasional grammatical or coherence errors are acceptable.

(1) Fair: The summary conveys some attributes of the content but gives the reader an imperfect or incomplete sense of what the episode contains. It may contain redundant material which is not needed when deciding whether to listen and may contain repetitions or broken sentences.

(0) Bad: The summary does not convey any of the most important content items of the episode or gives the reader an incorrect or incomprehensible sense of what the episode contains. It may contain a large amount of redundant information that is not needed when deciding whether to listen to the episode.

Yes/No Questions:

Q1: Does the summary include names of the main people (hosts, guests, characters) involved or mentioned in the podcast?

Q2: Does the summary give any additional information about the people mentioned (such as their job titles, biographies, personal background, etc)?

Q3: Does the summary include the main topic(s) of the podcast?

Q4: Does the summary tell you anything about the format of the podcast; e.g. whether it's an interview,

whether it's a chat between friends, a monologue, etc?

Q5: Does the summary give you more context on the title of the podcast?

Q6: Does the summary contain redundant information?

Q7: Is the summary written in good English?

Q8: Are the start and end of the summary good sentence and paragraph start and end points?

We submitted one run referred to as LongformerBart in Figure 1 and Figure 2, which show the model's performance in comparison to the two creator descriptions (filtered and non-filtered). Figure 1 displays a count of episode summaries for each quality rank totaling 179 episodes. Figure 2 shows a percentage of episode summaries which received the answer yes to each respective question.

We see from Figure 1 that the filtered creator descriptions were not always better than the unfiltered, and that our model's summaries were often better than both types of creator descriptions, with fewer bad outputs and more good/excellent outputs.

Table 1 shows the average ROUGE-L scores for recall, precision and f-measure for our model's outputs against the noisy creator summaries, along with values for the 95% confidence interval.

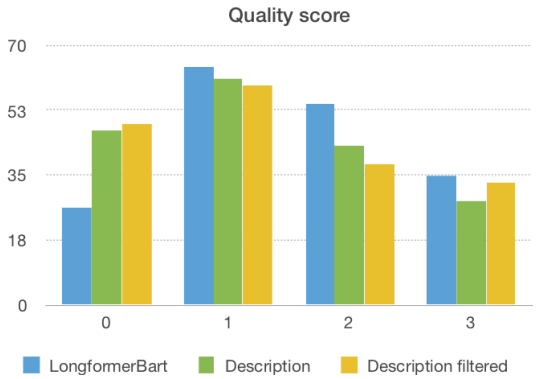

Figure 1: *Quality scores reported for the run in comparison to creator descriptions*

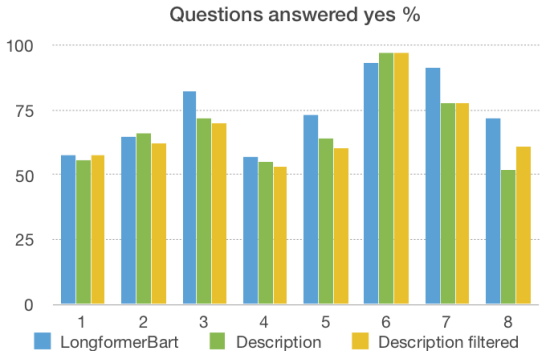

Figure 2: *Percentage of yes answers per questions with creator descriptions for comparison*

VI. Conclusion and Discussion

This paper describes our submission to the summarization task of the TREC 2020 podcast track in which we produced summaries for podcast episodes using the transcribed texts as input. Our method uses the summarization state of the art seq2seq model BART and extends it with a sparse attention mechanism from the Longformer to be able to take into account much long input sequences. Our approach copies the learned embeddings in the BART model to the Longerformer attention layers and finetunes the model on the podcast dataset. This method is effective at capturing the entirety of the long input sequences and thus outperforms the descriptions made by the podcast creators themselves when evaluating over both the question set and the quality score.

References

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

Clifton, Ann et al. (2020). "100,000 Podcasts: A Spoken English Document Corpus". In: *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*.

ROUGE-L Average scores		
Recall	Precision	F-measure
0.18983	0.26462	0.19234
Recall (95%-conf.int)	Precision (95%-conf.int)	F-measure (95%-conf.int)
0.25362 - 0.27536	0.25362 - 0.27536	0.18083 - 0.19886

Table 1: *Rouge scores average report for 179 episodes evaluated*

Cohan, Arman et al. (2018). "A discourse-aware attention model for abstractive summarization of long documents". In: *arXiv preprint arXiv:1804.05685*.

Devlin, Jacob et al. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

Gidiotis, Alexios and Grigorios Tsoumakas (2020). "A Divide-and-Conquer Approach to the Summarization of Academic Articles". In: *arXiv preprint arXiv:2004.06190*.

Hermann, Karl Moritz et al. (2015). "Teaching machines to read and comprehend". In: *Advances in neural information processing systems*, pp. 1693–1701.

Jones, Rosie et al. (n.d.). "Overview of the TREC 2020 Podcasts Track". In:

Kitaev, Nikita, Łukasz Kaiser, and Anselm Levskaya (2020). "Reformer: The efficient transformer". In: *arXiv preprint arXiv:2001.04451*.

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

Liu, Yang and Mirella Lapata (2019). "Text summarization with pretrained encoders". In: *arXiv preprint arXiv:1908.08345*.

Raffel, Colin et al. (2019). "Exploring the limits of transfer learning with a unified text-to-text transformer". In: *arXiv preprint arXiv:1910.10683*.

Reddy, Sravana et al. (2021). "Detecting Extraneous Content in Podcasts". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*.

Tay, Yi et al. (2020). "Sparse Sinkhorn Attention". In: *arXiv preprint arXiv:2002.11296*.

Wolf, Thomas et al. (2019). "HuggingFace's Transformers: State-of-the-art Natural Language Processing". In: *ArXiv* abs/1910.03771.

Yan, Yu et al. (2020). "Prophetnet: Predicting future n-gram for sequence-to-sequence pretraining". In: *arXiv preprint arXiv:2001.04063*.

Zaheer, Manzil et al. (2020). "Big bird: Transformers for longer sequences". In: *arXiv preprint arXiv:2007.14062*.

Zhang, Jingqing et al. (2019). "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization". In: *arXiv preprint arXiv:1912.08777*.