2024 28th International Conference on Information Technology (IT) | 979-8-3503-6961-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/IT61232.2024.10475775

Transformer Based Multimodal Summarization and Highlight Abstraction Approach for Texts and Speech Audios

Turan Goktug Altundogan, Mehmet Karakose and Senem Tanberk

Abstract— **Multimodal summarization is a kind of summarization application in which its inputs and/or outputs can be in different data types like text, video, and audio. In this study, a new approach based on fine tuning of different pre-trained transformers was developed for abstractive and extractive summarization of audio and text data. In the proposed method, abstractive and extractive summaries of text data are provided only as text, while extractive summaries of audio data are presented as both text and audio data. Abstractive summaries of the audio data are presented as text only. Transformers with text2text input-output relationship were used in both extractive and abstractive summarization processes of the proposed method. For the training and inference processes of audio this type of data to be handled in transformers, an ASR step was followed before the summarization step. The experimental results obtained were given in detail and compared with similar approaches in the literature. As a result of the comparison, it was seen that the proposed method achieved better performance than similar prior approaches.**

Keywords - Multimodal Summarization, Audio Summarization, Transformer Fine-Tuning

I. Introduction

Summarization applications have become a very important natural language processing problem with the development of new and high-performance neural architectures such as Transformer and the widespread use of text, video and audio data in digital environments. Summarization is categorized by two different functions. Among these, extractive summarization is an approach based on minimizing or shortening the existing data and consisting of the data contained in the input data as output. In abstractive summarization, summarization is achieved by the proposed method of abstracting the data and producing a completely unique output that is not included in the input data. In this study, a new method focused on extractive and abstract summarization of text data and speech audio data was developed. Before explaining the proposed method in detail and presenting the results, it was deemed useful to examine the existing studies in the literature. In a study made by us, text data was first summarized with a PageRank-based method, and then highlight abstraction was performed from these summarized data with the help of an LSTM Encoder - Decoder [1]. The reason why our approach here included an extractive summarization process was to reduce memory costs during neural training and inference processes. In two of our studies, we developed methods based on fine-tuning pre-trained Transformers to abstractive summarize the medical questions and daily dialogues [2,3]. In some of the studies we reviewed for this study, a method based on automatic speech recognition and Transformer fine tuning was developed to summarize Podcasts [4-6]. In the mentioned approaches, podcast audio is converted into text and then extractive summaries of these texts are obtained from fine-tuned transformer architectures. Although the problem focused on and the solution presented in the studies are quite similar to ours, abstractive summaries of speech audio are also obtained in the method we provide, and the fact that both audio and text data are included in the training data and the results we present shows that our method has a multimodal nature. In another study, a new approach was developed for multimodal summarization by using a combination of audio, text and image data [7]. In the relevant study, audio data was converted to text and a neural perspective architecture was used to express text and image data in a common vector space. One study in the literature focused on extractive summarization of customer service dialogues using a BERT-based method [8]. In another study, a method based on Attention mechanism and Transformer architecture was used for extractive text summarization [9]. In one study, the scoring of sentences for extractive text summarization was carried out with a transformer-based method [10]. Apart from these, there are many text [11-14] and voice summarization approaches [15-17] that use methods similar to the above studies and which we use to compare the performance of our method. On the other hand in this study, we first fine-tune more than one Transformer to perform the extractive summarization process. Then, we used the OpenAI-Whisper architecture to obtain timestamped transcripts of the speech audio data. We performed the audio summarization process by making extractive summaries of the obtained transcripts and time-stamp matching of the

T. G. Altundogan is with the Computer Engineering Department, Manisa Celal Bayar University, Manisa, Türkiye (corresponding author to provide phone: +90-553-253-0175; e-mail: turan.altundogan@cbu.edu.tr).

M. Karakose is with the Computer Engineering Department, Firat University, Elazig, Türkiye. (e-mail: mkarakose@firat.edu.tr).

S. Tanberk is with the Research and Innovation, Huawei Turkey Research and Development Center, Istanbul, Türkiye (e-mail:senem.tanberk@huawei.com).

979-8-3503-6961-8/24/$31.00 ©2024 IEEE

relevant sentences. After this, we fine-tuned multiple transformer architectures to perform highlight abstraction using extractive summarized audio transcript texts and another abstractive text summary dataset.

II. Proposed Approach

Figure 1 shows the overall block diagram of the proposed method. Since the summarization process in the proposed method is performed on both audio and text data, our approach is multimodal. Obtaining timestamped transcripts from audio data is provided by the Openai-Whisper model. Then, the text data obtained here is subjected to a preprocessing step and split into chunks. The extractive summarization process of each of these resulting chunks is performed by a fine-tuned transformer. Then, voice summarization was carried out by using the timestamps of the sentences in the summary text obtained here. Again, the sentences in the extractive summary text were given as input to the finely tuned transformer model for abstractive summarization, and highlight abstraction of the audio data was carried out in this way. To summarize the text data, the raw text data was subjected to preprocessing and divided into chunks. Then, extractive summarization and abstractive highlighting were carried out with Transformers, just like the audio data. To better understand the proposed method, it would be beneficial to explain some of its main steps in detail.

A. *Automatic Speech Recognition (ASR) with Timestamps*

Automatic Speech Recognition is the general name given to the process of converting speech-containing audio data into text. For this process, neural architectural perspectives are generally used. OpenAI provides the use of models for many solution proposals it has developed in the field of natural language processing through free APIs. OpenAI Whisper is an transformer architecture model that can be used open source by anyone for ASR processing [18]. Figure 2 shows the model diagram of OpenAI Whisper [18]. In the model's working algorithm, first the audio data is divided into 30-second chunks. Then, Log-Mel spectrogram transformation of the audio signal is performed and feature extraction is performed with two one-dimensional Convolution layers. Features are subjected to sinusoidal positional encoding and the audio-to-text conversion process is completed with the help of encoder blocks, cross attention and decoder blocks. Timestamping feature is also implemented with the API provided by OpenAI.

B. *Preprocessing*

The pre-processing process consists of separate steps for text obtained from audio data and raw text data. The transcript obtained from the audio contains Timestamp information for each word. First, the text is split into sentences and the timestamp of each sentence is kept in a dictionary structure. The processes of lower case conversion, elimination of special characters, abbreviations removal and separation into chunks of text and raw text data obtained from audio data are common.

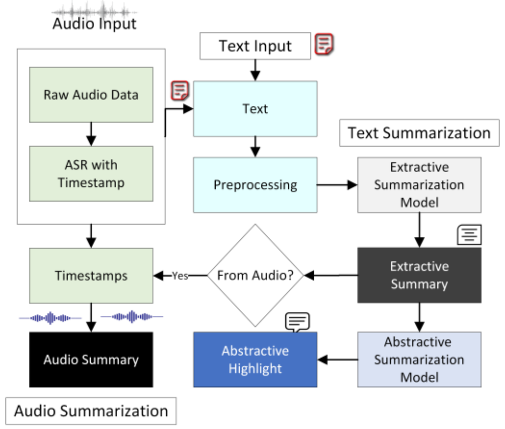

Fig 1. Overall block diagram of proposed approach.

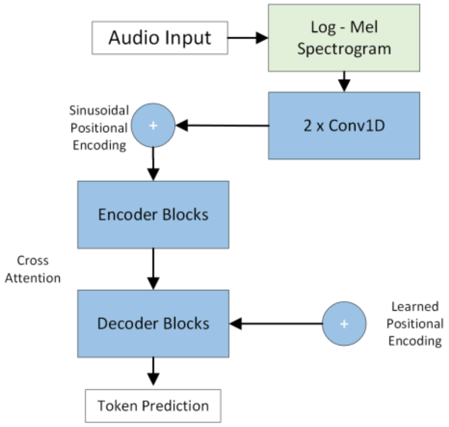

Fig 2. Model diagram of OpenAI-Whisper [18].

C. *Transformer Fine-Tuning*

Transformers are encoder-decoder neural architectures consisting only of embedding, positional encoding and attention layers. Pre-trained transformers are structures that are trained with very large data sets for one or more natural language processing tasks and have the ability to be retrained later to solve another problem. In the proposed method, fine-tuned transformers were used for each of the extractive and abstractive summarization processes. To perform these operations, multiple Transformers were fine-tuned and the performances of different architectures were compared. The data sets used for Transformer fine tuning and the data regarding the training process are given in detail in the experimental results section.

D. *Inference, Audio Summarization and Evaluation*

In the Inference phase, text chunks are first sent to the fine-tuned Transformer to obtain extractive summaries. The resulting extractive summary text is sent to the fine-tuned transformer for highlight abstraction. If the summarized text is obtained from audio data, timestamps of the sentences in the Extractive summary text are obtained and the summary audio output is obtained by taking the places where these sentences occur in the audio data and combining them. Since Transformers are neural architectures that work in a generative structure, minor differences may be encountered, although very rarely, when producing some original sentences. In this case, sentences that do not have a one-to-one equivalent and sentences in the original text are subject to ROUGE evaluation. The time-stamp of the original sentence with the highest R-Long value is used for the sentence with ambiguity. ROUGE metrics were also used to evaluate the proposed methods performance of extracting summaries and abstracting highlights.

III. Experimental Results

A. Dataset and Text Summarization Results

T5 and Bart transformer neural architectures were Fine Tuned to perform extractive summarization. For this process, 2500 news texts and their extractive summaries taken from the BBC News dataset [19] were used. For the testing process, 200 news texts were used for inference in Transformers and the obtained outputs were evaluated with ROUGE scores. Table 1 includes the training parameters, training - validation results and test results of both models. Figure 3 shows the training and validation loss graphs of both models. To perform Abstractive summarization process, T5 and Bart transformer neural architectures were fine tuned. For this process, extractive summaries of 30000 news texts taken from the CNN News dataset [20], obtained from the Transformers trained in the previous step, extractive text summaries obtained in the same way from 490 audios taken from the Audio Summarization Dataset [21], and their corresponding target abstractive highlights were used. For the testing process, the extractive summary obtained from 500 news texts and 10 audio was used for inference in Transformers. The results obtained were evaluated with ROUGE scores. Table 2 includes the training parameters, results and test evaluations of both models. Figure 4 shows the training and validation loss graphs of both models. When the results of the Transformers with fine-tuning performed for both purposes were examined, it was seen that the T5 architecture for extractive summarization and the BART architecture for abstractive summarization achieved higher performance results. However, when the graphs of the Transformer training performed for both tasks are examined, it is seen that the validation and training loss values of the BART architecture are closer to the loss values of the other model. This actually shows that the BART architecture has not yet fully reached learning satisfaction.

B. Audio Summarization Results

Audio summarization results were evaluated with the text extractive summarization performance of the models. In addition, 10 input audio and abstractive highlight texts in the Audio Summarization dataset were subjected to ROUGE evaluation together with the highlights obtained from inference. Finally, the original and summary times of the summaries obtained from 10 audios were compared. ROUGE evaluation of audio summaries and average original and summary audio durations are given in Table 3. When the results given were examined, it was seen that the abstractive summarization performance of the proposed method was quite high. In addition, it was observed that the original sound durations were shortened by nearly 80% with the proposed method.

C. Comparative Results

In Table 4, the results obtained by the proposed method are compared with the results of similar studies in the literature. When the results were examined, it was observed that the proposed method achieved higher performance results than multi-mode and audio-mode summarization methods. In addition, among the studies discussed, we are the only approach that performs abstractive summarization in multi-modal space.

IV. Conclusions

In this study, a transformer-based method is proposed for multimodal summarization of text and speech audio. In the study, the OpenAI-Whisper model was used to convert audio data into text with timestamp information. Fine tuning of T5 and BART neural architectures, which summarize the obtained texts extractive and abstractive was performed. Audio summarization was carried out using extractive summaries of the texts converted from audio and the previously obtained timestamp information. In addition, the highlight abstraction process was completed by giving extractive summaries obtained from text and audio data as input to the transformer that performs the abstractive summarization process. When the comparative results were examined, it was seen that the proposed method achieved higher performances than the general studies in the literature for all the problems it addressed. In addition, among the existing approaches, we are the only proposed method that offers a multi-modal abstractive summarization approach.

TABLE I. Fine Tuning Parameter and Results (Extractive)

	T5			BART		
	Epoch	Param Count.		Epoch	Param Count.	
	5	60506624		5	139420416	
Loss	Training	Validation		Training	Validation	
	0.3447	**0.4903**		1.60	1.69	
Test	R-1	R-2	R-L	R-1	R-2	R-L
	0.547	**0.321**	**0.529**	0.193	0.014	0.166

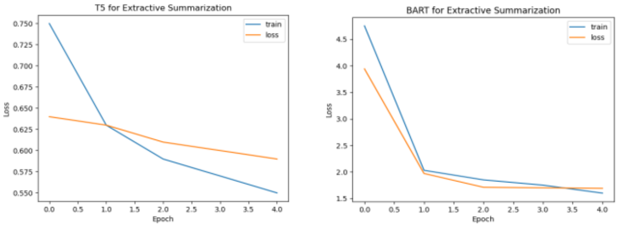

Fig 3. Loss graph of transformers fine-tuned for extractive summarization.

TABLE II. Fine Tuning Parameter and Results (Abstractive)

	T5			BART		
	Epoch	Param Count.		Epoch	Param Count.	
	5	60506624		5	139420416	
Loss	Training	Validation		Training	Validation	
	1.249	1.495		**0.923**	**1.043**	
Test	R-1	R-2	R-L	R-1	R-2	R-L
	0.251	0.174	0.246	**0.343**	**0.258**	**0.324**

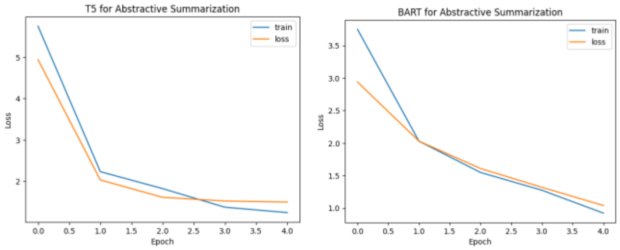

Fig 4. Loss graph of transformers fine-tuned for abstractive summarization.

TABLE III. AUDIO SUMMARIZATION RESULTS

ROUGE-1	**ROUGE-2**	**ROUGE-L**
0.4423	0.2947	0.4201
	Original	**Summary**
Avg. Duration	947sec	204sec

TABLE IV. COMPARATIVE RESULTS

Ref	Modal	Summarization Type	R-1	R-2	R-L
[4]	Audio	Extractive	0.52	**0.39**	0.51
[5]	Audio	Extractive	**0.54**	0.29	0.45
[6]	Audio	Extractive	0.25	0.08	0.21
[7]	Multimodal	Extractive	0.44	0.16	0.20
[8]	Text	Extractive	0.48	0.29	0.46
[9]	Text	Extractive	0.42	0.20	0.39
[13]	Text	Extractive	0.44	0.21	0.41
[14]	Text	Abstractive	0.34	0.25	0.32
[17]	Multimodal	Extractive	0.43	0.17	0.20
[Ours]	Multimodal	Extractive	**0.54**	0.32	**0.52**
[Ours]	Multimodal	Abstractive	0.44	0.29	0.42

ACKNOWLEDGMENT

This study was supported by the TUBITAK (The Scientific and Technological Research Council of Turkey) under Grant No: 5220154.

REFERENCES

[1] T. G. Altundogan and M. Karakose, "LSTM Encoder Decoder Based Text Highlight Abstraction Method Using Summaries Extracted by PageRank," 2023 27th International Conference on Information Technology (IT), Zabljak, Montenegro, 2023, pp. 1-4, doi: 10.1109/IT57431.2023.10078652..

[2] T. G. Altundogan, M. Karakose, S. Tanberk and O. B. Mercan, "Fine Tuning and Comparative Analysis of Pre-Trained Transformers for Dialogue Texts Abstractive Abstractive Summarization," 2023 Internationcal Conference on Advances and Innovations in Engineering (ICAIE), Elazig, Turkey, 2023, pp. 81-89

[3] T. G. Altundogan, M. Karakose and O. Tokel, "BART Fine Tuning based Abstractive Summarization of Patients Medical Questions Texts," 2023 International Conference on Data Analytics for Business and Industry (ICDABI), Sakhir, Bahrain, 2023

[4] A. Vartakavi, A. Garg and Z. Rafii, "Audio Summarization for Podcasts," 2021 29th European Signal Processing Conference (EUSIPCO), Dublin, Ireland, 2021, pp. 431-435, doi: 10.23919/EUSIPCO54536.2021.9615948

[5] N. Vanjari, S. Pinjari, S. Labde, P. Patil, P. Patil and A. Sahitya, "Podcast Summarization System," 2022 5th International Conference on Advances in Science and Technology (ICAST), Mumbai, India, 2022, pp. 379-382, doi: 10.1109/ICAST55766.2022.10039668.

[6] G. Vico and J. Niehues, "TED Talk Teaser Generation with Pre-Trained Models," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 8067-8071, doi: 10.1109/ICASSP43922.2022.9746700.

[7] H. Li, J. Zhu, C. Ma, J. Zhang and C. Zong, "Read, Watch, Listen, and Summarize: Multi-Modal Summarization for Asynchronous Text, Image, Audio and Video," in IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 5, pp. 996-1009, 1 May 2019, doi: 10.1109/TKDE.2018.2848260.

[8] B. Ma, H. Sun, J. Wang, Q. Qi and J. Liao, "Extractive Dialogue Summarization Without Annotation Based on Distantly Supervised Machine Reading Comprehension in Customer Service," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 87-97, 2022, doi: 10.1109/TASLP.2021.3133206.

[9] K. Shi, X. Cai, L. Yang, J. Zhao and S. Pan, "StarSum: A Star Architecture Based Model for Extractive Summarization," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 3020-3031, 2022, doi: 10.1109/TASLP.2022.3207688.

[10] Q. Zhou, N. Yang, F. Wei, S. Huang, M. Zhou and T. Zhao, "A Joint Sentence Scoring and Selection Framework for Neural Extractive Document Summarization," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 671-681, 2020, doi: 10.1109/TASLP.2020.2964427.

[11] A. Mishra, A. Sahay, M. a. Pandey and S. S. Routaray, "News text Analysis using Text Summarization and Sentiment Analysis based on NLP," 2023 3rd International Conference on Smart Data Intelligence (ICSMDI), Trichy, India, 2023, pp. 28-31, doi: 10.1109/ICSMDI57622.2023.00014.

[12] P. M. Hanunggul and S. Suyanto, "The Impact of Local Attention in LSTM for Abstractive Text Summarization," 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 2019, pp. 54-57, doi: 10.1109/ISRITI48646.2019.9034616.

[13] X. Zhang, Q. Wei, Q. Song and P. Zhang, "An Extractive Text Summarization Model Based on Rhetorical Structure Theory," 2023 26th ACIS International Winter Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD-Winter), Taiyuan, China, 2023, pp. 74-78, doi: 10.1109/SNPD-Winter57765.2023.10223980.

[14] Y. Chen and Q. Song, "News Text Summarization Method based on BART-TextRank Model," 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, China, 2021, pp. 2005-2010, doi: 10.1109/IAEAC50856.2021.9390683.

[15] S. R. Chauhan, S. Ambesange and S. G. Koolagudi, "Speech Summarization Using Prosodic Features and 1-D Convolutional Neural Network," 2022 IEEE 7th International Conference on Recent Advances and Innovations in Engineering (ICRAIE), MANGALORE, India, 2022, pp. 14-19, doi: 10.1109/ICRAIE56454.2022.10054315.

[16] K. Yamamoto, H. Banno, H. Sakurai, T. Adachi and S. Nakagawa, "A Study of Speech Recognition, Speech Translation, and Speech Summarization of TED English Lectures," 2023 IEEE 12th Global Conference on Consumer Electronics (GCCE), Nara, Japan, 2023, pp. 451-452, doi: 10.1109/GCCE59613.2023.10315471.

[17] T. Hayashi, T. Yoshimura, M. Inuzuka, I. Kuroyanagi and O. Segawa, "Spontaneous Speech Summarization: Transformers All The Way Through," 2021 29th European Signal Processing Conference (EUSIPCO), Dublin, Ireland, 2021, pp. 456-460, doi: 10.23919/EUSIPCO54536.2021.9615996.

[18] OpenAI Whisper, Access Link: https://openai.com/research/whisper, Access Date: 1/11/2024

[19] BBC News Summary Dataset, Access Link: https://www.kaggle.com/datasets/pariza/bbc-news-summary, Access Date: 1/11/2024

[20] CNN Daily Mail Dataset, Access Link: https://huggingface.co/datasets/cnn_dailymail, Access Date: 1/11/2024

[21] Audio Summarization Dataset, Access Link: https://www.kaggle.com/datasets/nfedorov/audio-summarization, Access Date:1/11/2024