VATMAN : Video-Audio-Text Multimodal Abstractive Summarization with Trimodal Hierarchical Multi-head Attention

Doosan Baek, Jiho Kim, Hongchul Lee*
Department of Industrial & Management Engineering
Korea University
Seoul, South Korea
{97dosan, jihonav, hclee*}@korea.ac.kr

Abstract— **Multimodal Abstractive Summarization is a challenging task that aims to generate concise and informative summaries from diverse modalities, such as video, audio, and text. In this study, we propose VATMAN, a novel approach for multimodal abstractive summarization. To effectively capture the hierarchical relationships and dependencies between modalities, we introduce Trimodal Hierarchical Multi-head Attention (THMA). THMA hierarchically attends to the video, audio, and textual representations, enabling the model to distill salient information and generate cohesive and coherent summaries. VATMAN leverages state-of-the-art generative pretrained language models (GPLMs), specifically Transformer-based models, and applies hierarchical attention at the modality level, which enhances the utilization of contextual information. The proposed VATMAN model on the How2 dataset demonstrates the ability to create more fluent summaries than those generated by human authors, showcasing its potential for utilization in various industrial environments.**

Index Terms—**Trimodal, Abstractive Summarization, Generative Pretrained Language Model, Transformer**

I. Introduction

The task of summarization is classified into two approaches: extractive summarization and abstractive summarization. In extractive summarization, important sentences or paragraphs are extracted from the source document based on statistical or linguistic features to create a summary. On the other hand, abstractive summarization involves a deeper semantic understanding of the entire source document, allowing the generation of new sentences or paragraphs that may not exist in the original text. Abstractive summarization is considered superior to extractive summarization [1] [2].

Research in abstractive summarization includes methods based on recurrent neural networks (RNNs) and Transformer models. Previous studies on RNN-based methods include "Abstractive sentence summarization with attentive recurrent neural networks," "Bottom-up attention," and "MAST." Recently, with the advent of Transformer-based [3] models, research in text generation, including abstractive summarization, has significantly advanced using pretrained language models such as BART, T5, PEGASUS, ProphetNet, and GPT-3.

However, there is a lack of research in the field of multimodal abstractive summarization that leverages the joint usage of video, audio, and text with generative pretrained language models (GPLMs). Thus, this paper proposes a novel multimodal abstractive summarization model called VATMAN, which fuses information from video, audio, and text simultaneously based on pretrained giant language models.

This paper is organized as follows. Section 2 presents the proposed framework. Section 3 shows the results of the experiment. Finally, Section 4 conclude our study.

II. Proposed Method

A. Pre-trained Language Model for Abstractive Summarization

The Transformer-based sequence-to-sequence language model architecture is identical to the one shown in Figure 1, except for the novel video-audio-text fusion block proposed in this study. The input text of this language model undergoes tokenization first and is then transformed into a sequence of token embeddings. To preserve positional information, positional embeddings are added to the token embeddings and used as the input to the encoder.

The encoder consists of L encoder layers, each composed of 1) Encoder Multi-head Self-Attention and 2) Feed-Forward network. Additionally, after passing through each sub-layer, residual connections and layer normalization are applied.

Similar to the encoder, the decoder is also composed of L decoder layers, but with two key differences. Firstly, the multi-head self-attention is masked, preventing it from referencing future words during the next word prediction. Secondly, in the decoder process, a multi-head encoder-decoder sub-layer (Encoder-Decoder Attention) is included to integrate the encoded information by combining decoder embeddings with the output embeddings of the encoder.

In this research, the BART model (Lewis et al., 2019)[4], which introduces a novel pre-training task, is used as the backbone model within the existing Transformer architecture.

979-8-3503-1327-7/23/$31.00 ©2023 IEEE

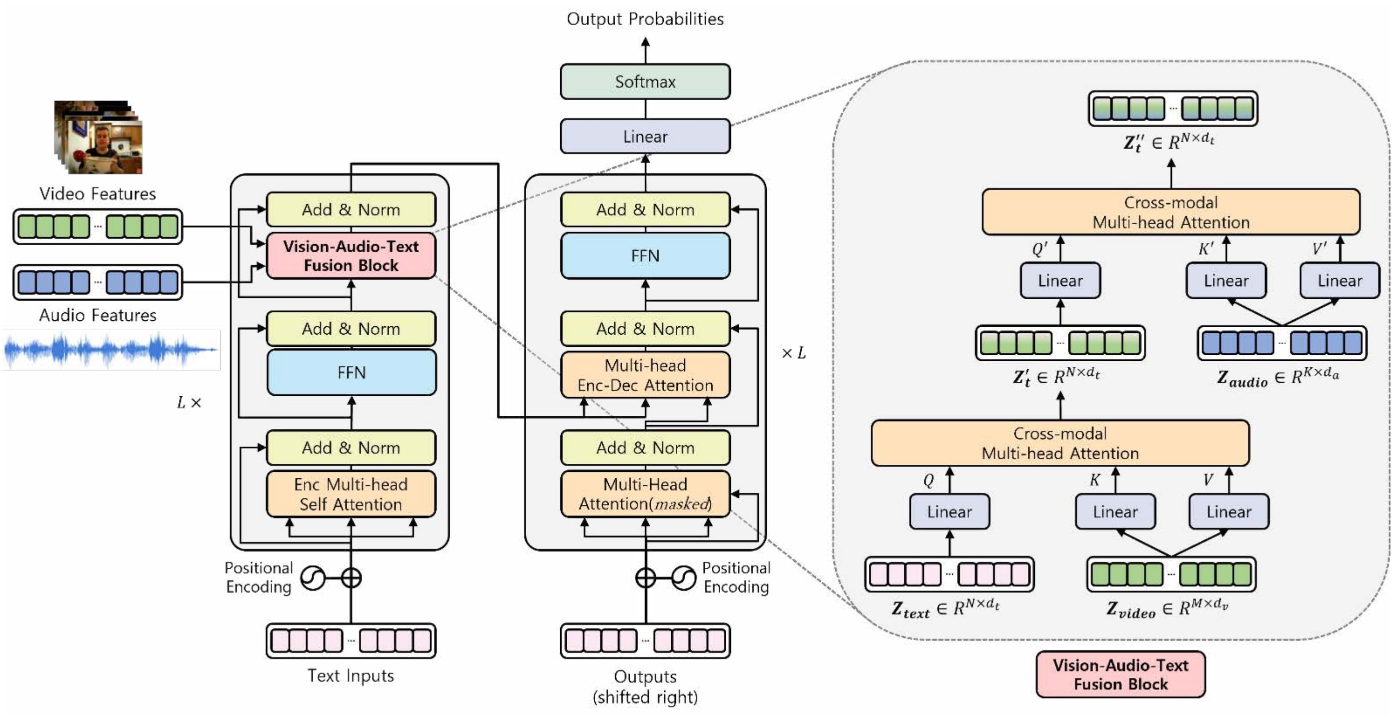
Output Probabilities
Softmax
Linear
Video Features
Audio Features
Add & Norm
Vision-Audio-Text Fusion Block
Add & Norm
FFN
Add & Norm
Enc Multi-head Self Attention
$L \times$
Positional Encoding
Text Inputs
Add & Norm
FFN
Add & Norm
Multi-head Enc-Dec Attention
Add & Norm
Multi-Head Attention(*masked*)
$\times L$
Positional Encoding
Outputs (shifted right)
$Z_t'' \in R^{N \times d_t}$
Cross-modal Multi-head Attention
Q'
K'
V'
Linear
Linear
Linear
$Z_t' \in R^{N \times d_t}$
$Z_{audio} \in R^{K \times d_a}$
Cross-modal Multi-head Attention
Q
K
V
Linear
Linear
Linear
$Z_{text} \in R^{N \times d_t}$
$Z_{video} \in R^{M \times d_v}$
Vision-Audio-Text Fusion Block

Fig. 1: Overview of proposed framework

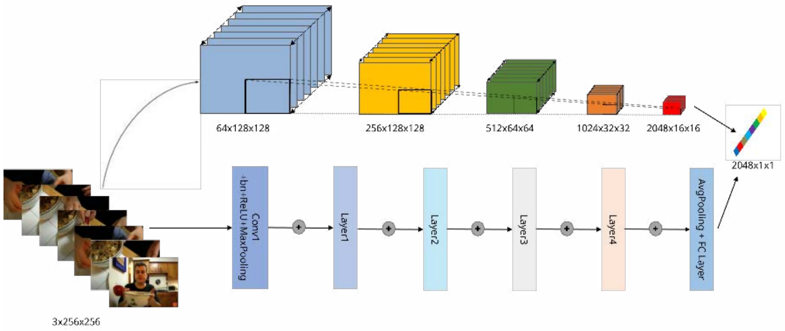

Fig. 2: Video Features through ResNeXt-101 Network

B. *Feature Extraction*

The video data consists of 16 frames per second, as seen in prior research [5][6][7][8], and is processed through the pre-trained 3D ResNeXt-101 network (Hera et al., 2018)[9], which yields 2,048-dimensional features per frame as depicted in Figure 2. As for the audio data, we utilize Kaldi (Povey et al., 2011)[10] to extract 40-dimensional filter bank features and 3-dimensional pitch features, which are then combined to form 43-dimensional features. Additionally, to introduce variability in speaker characteristics for each video, we apply Cepstral Mean and Variance Normalization (CMVN). The resulting 2,048-dimensional video features and 43-dimensional audio features serve as inputs for the Video-Audio-Text fusion methodology.

C. *Multimodal Fusion*

In our approach, we insert a fusion unit block at the end of each encoder unit block, as depicted in Figure 1. This sublayer includes the Video-Audio-Text fusion mechanism, residual connections, and layer normalization. We propose the Video-Audio-Text fusion mechanism, illustrated on the right side of Figure 2. Given the text input, video input, and audio input after passing through their respective modality embeddings, the output of the first fusion (Video-Text fusion) mechanism is a hidden layer of dimension, containing both the attention information from text and images.

Subsequently, the output of this hidden layer ($Z_t^{'}$) serves as the query, while the audio input (Z_{audio}) acts as the key and value inputs for the second fusion (Video-Audio-Text fusion) mechanism. The final output of this process is an hidden layer of dimension $Z_t'' \in R^{N \times d_t}$, which retains the same dimension as the text input (Z_{text}) before passing through the Video-Audio-Text fusion layer. This characteristic ensures that the dimension remains unchanged even when multiple layers are stacked.

TABLE I. EVALUATION RESULTS OF BASELINES AND OURS

Input Modality	Method	Rouge			BLEU
		1	2	L	1
Text	T5	45.0766	21.5857	37.0916	37.0916
	BART	49.0718	26.7150	41.6582	41.2951
Text + Video	T5	47.4345	23.3531	39.9249	42.7832
	BART	50.1433	26.5368	42.0547	43.2625
Text + Video + Audio	RNN	48.85	29.51	43.23	-
	VATMAN(ours)	52.5315	29.4880	44.1799	49.4860

ID : g3pXM5X3-Xw

Reference : proper tool preparation will make your ceramic charger plate easier and cleaner to work with . learn how to prepare your clay tools with tips from a master potter in this free ceramics video .

Video-Text(T5) : the a charger is a clay pottery wheel . learn how to make a clay charger with tips from a ceramics expert in this free ceramics video .

Video-Text(BART) : the ceramic charger is a great tool for making dinnerware . learn how to make ceramic charger plates with tips from a master potter in this free ceramics video .

Video-Audio-Text(ours) : clay chargers are traditionally used to put dinnerware on top of a large plate . learn how to use clean tools to make ceramic chargers from a potter in this free ceramics video .

ID : g0S7FaqIweA

Reference : firefighters often use the coat method to don a scba . learn how to use the coat method to don scba (self-contained breathing apparatus) in this free video on donning firefighter gear .

Video-Text(T5) : learn how to use the coat method in this free video on how to use the scba in this

Video-Text(BART) : learn the coat method for a firefighter with tips from a firefighter in this free video on firefighter training .

Video-Audio-Text(ours) : donning a firefighter scba using the coat method involves reaching down and grabbing the top of the shoulder straps with the thumbs facing each other . learn how to donning firefighters with tips from a certified fire instructor in this free video on firefighting

Fig. 3: Examples of summary generation from various models:

III. Experiment and Result

A. *Data*

The How2 dataset consists of a total of 79,114 video/transcript/summary data pairs, divided into two sets: a 2,000-hour set and a 300-hour set, each containing 13,445 audio/video/transcript/summary data pairs. This dataset comprises short instructional videos covering various domains, such as cooking, music, indoor/outdoor activities, and sports. Alongside the videos, transcripts are provided, which are text data obtained by converting the audio spoken by the speakers. These transcripts encompass the overall content of the videos.

B. *Experiment settings*

In this study, we need to verify the performance of utilizing three modalities simultaneously by injecting auditory context (audio). To achieve this, we sample 300 hours of text/video/audio data based on the IDs of the 300-hour audio dataset. The dataset comprises a total of 13,445 data samples, with 12,798 samples for training, 520 samples for validation, and 127 samples for testing.

we employ the PyTorch (Paszke et al., 2019)[11] deep learning framework and implement the code using PyTorch-Lightning for distributed training. Furthermore, we conduct the experiments using four Nvidia GeForce RTX 3090 GPUs.

Hyperparameters are variables that adjust and control machine learning models or deep learning algorithms. the common dimension is set to 256, the number of epochs is 150, the learning rate is 3×10^{-5}, and the optimizer used is Adam.

C. *Experiment results*

Table 1 presents a comparison of the summarization performance of various models, including both prior work [6] and the proposed multimodal summarization model. The evaluation is conducted on the same test dataset, using two key metrics: Recall-Oriented Understudy for Gisting Evaluation (ROUGE) and Bilingual Evaluation Understudy (BLEU).

ROUGE-N measures the overlap between the N-grams of the model-generated summaries and the reference summaries. Specifically, ROUGE-L represents the longest common subsequence between the model-generated and reference summaries. Similarly, BLEU-N is another N-gram count-based metric commonly used for evaluating natural language processing performance.

In Figure 3, we compare the summarization results of various models. The first summary pertains to a ceramics video. When using only text data, the summary includes a directive to learn about tool usage. However, with the addition of video and audio information, the summary provides further details on the technique of using the tool. The second generated summary pertains to firefighters donning their gear. In the bimodal models, the generated summaries lack specific details about the gear donning process. However, when incorporating video and audio information, the summaries provide more concrete instructions on the proper gear donning technique, including descriptive explanations from the video.

IV. Conclusion

In this paper, we propose the VATMAN (Video-Audio-Text Multimodal Attention Network) model, which leverages the pre-trained language model BART (Bidirectional and Auto-Regressive Transformers) and incorporates a Trimodal Hierarchical Multi-head Attention mechanism. In previous research on multimodal summarization tasks, there were transformer-based bimodal models that focused mainly on using text and video to generate text-based summaries, but they did not fully utilize additional modalities such as audio, which limited their performance.

To overcome this limitation and take advantage of the untapped information from additional modalities, we introduce a multimodal structure that incorporates audio information alongside text and video. By preserving the overall bimodal architecture while adding an additional attention layer, we can effectively enhance the performance of summary generation by exploiting the synergy of all three modalities.

With the proposed Trimodal Hierarchical Multi-head Attention mechanism, our VATMAN model achieves a more comprehensive understanding of the input data, resulting in improved summarization performance compared to traditional bimodal models. The integration of audio information allows the model to capture richer and more diverse contextual cues, ultimately leading to more accurate and informative summaries.

Acknowledgment

This research was supported by the BK21 FOUR funded by the Ministry of Education of Korea and National Research Foundation of Korea.

References

[1] Chopra, Sumit, Michael Auli, and Alexander M. Rush. "Abstractive sentence summarization with attentive recurrent neural networks." Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies. 2016.

[2] Nallapati, Ramesh, et al. "Abstractive text summarization using sequence-to-sequence rnns and beyond." arXiv preprint arXiv:1602.06023 (2016).

[3] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[4] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL).

[5] Palaskar, Shruti, et al. "Multimodal abstractive summarization for how2 videos." arXiv preprint arXiv:1906.07901 (2019).

[6] Khullar, Aman, and Udit Arora. "MAST: Multimodal abstractive summarization with trimodal hierarchical attention." arXiv preprint arXiv:2010.08021 (2020).

[7] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: a large-scale dataset for multimodal language understanding. arXiv preprint arXiv:1811.00347.

[8] Yu, Tiezheng, et al. "Vision guided generative pre-trained language models for multimodal abstractive summarization." arXiv preprint arXiv:2109.02401 (2021).

[9] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[10] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G., and Vesel, K. (2011), The Kaldi speech recognition toolkit, IEEE 2011 Workshop on Automatic Speech Recognition and Understanding .

[11] Paszke, Adam, et al. "Pytorch: An imperative style, high-performance deep learning library." Advances in neural information processing systems 32 (2019).