MM-AVS: A Full-Scale Dataset for Multi-modal Summarization

Xiyan Fu[1], **Jun Wang**[2], **Zhenglu Yang**[1]
[1]Nankai University, China
[2] Ludong University, China
{fuxiyan,junwang}@mail.nankai.edu.cn, yangzl@nankai.edu.cn

Abstract

Multimodal summarization becomes increasingly significant as it is the basis for question answering, Web search, and many other downstream tasks. However, its learning materials have been lacking a holistic organization by integrating resources from various modalities, thereby lagging behind the research progress of this field. In this study, we present a full-scale multimodal dataset comprehensively gathering documents, summaries, images, captions, videos, audios, transcripts, and titles in English from CNN and Daily Mail. To our best knowledge, this is the first collection that spans all modalities and nearly comprises all types of materials available in this community. In addition, we devise a baseline model based on the novel dataset, which employs a newly proposed Jump-Attention mechanism based on transcripts. The experimental results validate the important assistance role of the external information for multimodal summarization.

1 Introduction

Multimodal summarization refines salient information from one or more modalities, including text, image, audio, and video ones (Evangelopoulos et al., 2013; Li et al., 2017). Given the rapid dissemination of multimedia data over the Internet, multimodal summarization has been widely explored in recent years. Meanwhile, some multimodal datasets (Li et al., 2017; Zhu et al., 2018; Sanabria et al., 2018; Li et al., 2020a) have been introduced to advance the development of this research field. However, a majority of them are restricted in scale and too oriented, such as being less than one hundred examples or merely containing Chinese texts. Moreover, the materials from different modalities are rarely collected across the board, especially videos and their accompanying materials that possess abundant external information for multimodal comprehension and fusion.

In this work, we introduce a full-scale Multi-modal Article and Video Summarization (MM-AVS) dataset [1] with documents, summaries, images, captions, videos, audios, transcripts, and titles in English. The significance of MM-AVS for the multimodal summarization community includes but not limited to: 1) MM-AVS is a large-scale multimodal collection compared with existing video containing dataset and its generation codes [1] has been released, which can be readily extended for existing and future multimodal summarization approaches; 2) MM-AVS is collected from CNN[2] and Daily Mail[3], which makes it available to more researchers due to English-based and comparable with the popular text-based CNN/Daily Mail corpus; and 3) MM-AVS firstly collects nearly all types of materials from all modalities, inclusively with videos, audios, transcripts, images, captions, and titles that are rarely assembled.

In addition, we implement a general multimodal summarization baseline based on transcripts for multimodal summarization on MM-AVS. This method employs a Jump-Attention mechanism to align features between text and video. Further, we use the multi-task learning to simultaneously optimize document and video summarizations. Evaluations on MM-AVS illustrate the benefits of external information such as videos and transcripts for multimodal summarization without alignment.

2 Related Work

Multi-modal summarization generates a condensed multimedia summary from multi-modal materials, such as texts, images, and videos. For instance, UzZaman et al. (2011) introduced an idea of illustrating complex sentences as multimodal summaries by combining pictures, structures, and sim-

[1] https://github.com/xiyan524/MM-AVS.

[2] https://www.cnn.com/

[3] https://www.dailymail.co.uk/home/index.html

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5922–5926
June 6–11, 2021. ©2021 Association for Computational Linguistics

Dataset	Doc	Summary		Image		Video			Title
		Abs.Sum	Ext.Label	Image	Caption	Video	Audio*	Transcript*	
MSMO (Zhu et al., 2018)	✓	✓		✓	✓				✓
MMSS (Li et al., 2018)	✓	✓		✓					
E-DailyMail (Chen and Zhuge, 2018)	✓	✓		✓	✓				✓
EC-product (Li et al., 2020a)	✓	✓		✓					✓
MMS (Li et al., 2017)	✓	✓		✓		✓			✓
How2 (Sanabria et al., 2018)		✓				✓		✓	
MM-AVS	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparisons of multimodal corpus. (* means that the audio or transcript is separated from video.)

	Daily Mail	CNN
Avg. Num. Articles	1970	203
Avg. Num. Tokens in Article	657.87	951.07
Avg. Num. Tokens in Sentence	33.42	29.84
Avg. Num. Tokens in Summary	59.63	29.73
Avg. Len. Video	81.96	368.19
Avg. Num. Images in Video	91.60	125.50
Avg. Num. Tokens in Transcript	83.74	116.76

Table 2: Corpus statistics of MM-AVS dataset. Each article is paired with a video.

plified compressed texts. Libovickỳ et al. (2018) and Palaskar et al. (2019) studied abstractive text summarization for open-domain videos. Li et al. (2017) constructed MMS dataset and developed an extractive multi-modal summarization method that automatically generated a textual summary based on a topic-related set of documents, images, audios, and videos. Zhu et al. (2018, 2020) combined image selection and output to alleviate the modality bias based on the MSMO dataset. Chen and Zhuge (2018) extended Daily Mail with images and captions to E-Daily Mail dataset and employed a hierarchical encoder-decoder model to align sentences and images. Recently, an aspect-aware model and a large-scale Chinese e-commerce product summarization dataset EC-product were introduced to incorporate visual information for e-commerce product summaries (Li et al., 2020a).

The above mentioned datasets are rarely constructed comprehensively, which ignore the abundant visual information underlying in videos. The only video-containing work is restricted in scale, which hampers its use for deep-learning based methods. In this study, we will build a full-scale multimodal dataset to address these issues.

3 MM-AVS Dataset

To facilitate a straightforward comparison for the multimodal summarization approaches with the text-based ones, MM-AVS extends CNN/DM collections to multimodalities. Each example of MM-AVS contains a document accompanying with multi-sentence summary, title, images, captions, videos, and their corresponding audios and transcripts.

3.1 Dataset comparison

Table 1 compares MM-AVS with the representative multimodal summarization benchmarks. MM-AVS contains documents and abstractive summaries as most of the benchmarks including, while it extends visual information that most existing benchmarks ignore (such as MSMO(Zhu et al., 2018), MMSS(Li et al., 2018), E-DailyMail(Chen and Zhuge, 2018), and EC-product(Li et al., 2020b)). MMS(Li et al., 2017) and How2(Sanabria et al., 2018) also take videos into account; however, MMS only contains 50 examples that are too limited for deep learning and How2 excludes documents, which are the most critical materials for summarization. MM-AVS also keeps image captions for deep descriptions of images as well as document titles for the topic extraction. Further, MM-AVS contains extractive labels for training convenience. In the manner of providing abundant multimodal information, MM-AVS is applicable for existing and future multimodal research in different learning tasks.

3.2 Dataset construction

The concrete statistics of MM-AVS are shown in Table 2[4], incorporating textual and visual modules: **Textual module.** Following (Nallapati et al., 2016), we have crawled all the summary bullets of each story in the original order to obtain a multi-sentence reference, where each bullet is treated as a sentence. Given that the reference is an abstractive summary written by humans, we construct the label of each sentence as (Nallapati et al., 2017) does.

[4] The data scale is determined by its accompanied videos, considering this modality is more space-consuming. The data acquirability code in the project github mentioned above can be used for extension.

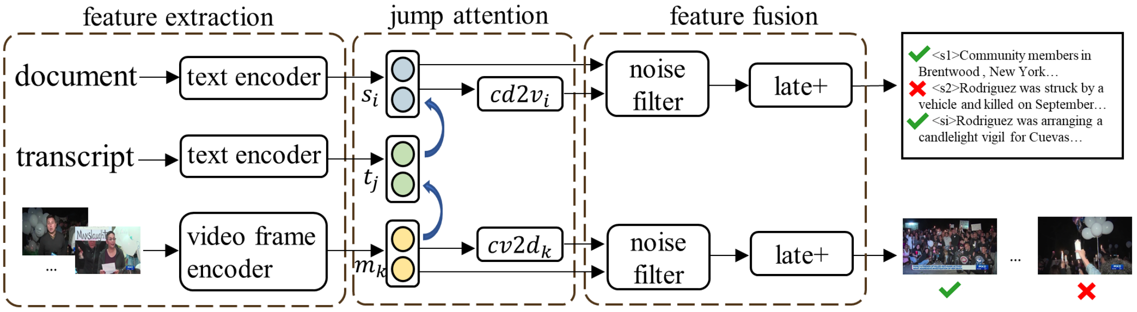

Figure 1: M^{2}SM utilizes multi-modal information for sentence and image extractions in summarization. The probability of the mismatched pair solely calculated based on the single modality is improved with the assistance of multimodalities.

Sentences in the document are selected to maximize the ROUGE (Lin, 2004) score with respect to the gold summary by a greedy approach. As for the document and title, we keep their original formats as shown in the websites.

Visual module. To enrich visual information for multimodal summarization, we collect images and videos for each example. Image caption is preserved to assist further explorations such as feature extraction and alignment to documents. Given long videos, we separate the audios and extract the transcripts[5] to alleviate the pre-process pressure for large-scale or online learning.

4 Summarization Method: M^{2}SM

4.1 Feature Extraction

We utilize the hierarchical bi-directional long short term memory (BiLSTM) (Nallapati et al., 2017) based on word and sentence levels to read tokens and induce a representation for each sentence denoted as s_i. Each sentence in a transcript is denoted as t_j. In terms of videos, we employ ResNet (He et al., 2016) for feature extraction and BiLSTM to model the sequential pattern in video frames. Each image is represented as m_k.

4.2 Feature Alignment-Jump Attention

Given that the transcript extracted from a video shares the same modality with a document and accurately aligns with a video, we take it as a bridge to deepen the relationship between two modalities. We apply the jump attention based on transcripts to assist modality alignment, which focuses on transcripts to video images and then on documents to transcript attention context. The video-aware context $cd2v_i$ is denoted as

$$cd2v_i = \sum_{j=1}^{NT} \sum_{k=1}^{NM} b_i^j d_j^k m_k, \qquad (1)$$

where NT and NM are the lengths of transcripts and image frames. b_i^j and d_j^k are the attention weights and can be calculated as follows (taking d_j^k for illustration):

$$d_j^k = \sigma(V^T(q_j \odot r_k + q_j + r_k)), \qquad (2)$$

where V is the training parameter, q_j and r_k are the feature mappings of each modality that are calculated as $q_j = tanh(W_m m_k + b_m)$ and $r_k = tanh(W_t t_j + b_t)$. The jump attention can be reversed to obtain an article context vector for video summarization.

4.3 Feature Fusion

Given that modalities may be not accurately aligned, we employ **late+ fusion** by fusing unimodal decisions. Inspired by (Liu et al., 2018), we induce noise filters to eliminate noises as $F(W_s f(s_i), W_c g(cd2v_i))$, where the filters W_s and W_c are calculated as follows:

$$W_s = [1 - g(cd2v_i)]^\beta, W_c = [1 - f(s_i)]^\beta, \qquad (3)$$

where β is a smoothed coefficient for penalty intensity, and $f(\cdot), g(\cdot)$, and $F(\cdot)$ are feedforward networks.

4.4 Multi-task Training

We employ the multi-task training to enhance summarization. The loss function is the weight mix of

[5] We use IBM Watson Speech for the text service https://www.ibm.com/watson/services/speech-to-text/.

	R-1	R-2	R-L
text-only	39.11	16.42	28.56
+video frames	40.86	17.48	30.23
+transcripts	41.26	17.95	30.98

Table 3: Summarizations based on the materials of documents, videos, and transcripts.

		R-1	R-2	R-L
document	CNN	25.29	5.70	11.11
	Daily Mail	8.78	2.15	5.45
reference	CNN	9.67	1.93	6.23
	Daily Mail	6.02	0.84	4.18

Table 4: Word overlap statistics of video transcripts with documents and references.

each task loss as follows:

$$\mathcal{L} = \alpha_{ts}\mathcal{L}_{ts} + \alpha_{vs}(R_{\text{div}} + R_{\text{rep}}),$$
$$\mathcal{L}_{ts} = -\frac{1}{NS}\sum_{n=1}^{NS}\left[y_n \log \hat{y}_n + (1 - y_n)\log(1 - \hat{y}_n)\right], \tag{4}$$

where $\mathcal{L}_{ts}$ is the training loss for extractive summarization, y_n and $\hat{y}_n$ represent the true and predicted labels, and α_{ts} and α_{vs} are balance parameters. Following (Zhou et al., 2018), we use unsupervised learning by reinforcement learning methods for video summarization whose loss can be separated into the diversity reward R_{div} (measuring frames dissimilarity) and the representativeness reward R_{rep} (measuring similarity between summary and video) as follows:

$$R_{\text{div}} = \frac{1}{|\mathcal{M}|(|\mathcal{M}|-1)}\sum_{j\in\mathcal{M}}\sum_{j'\in\mathcal{M},\, j'\neq j} d\left(m_j, m_{j'}\right),$$
$$R_{\text{rep}} = \exp\left(-\frac{1}{NM}\sum_{j=1}^{NM}\min_{j'\in\mathcal{M}}\left\|m_j - m_{j'}\right\|_2\right), \tag{5}$$

where $\mathcal{M}$ is the set of the selected video frames and $d(\cdot)$ is the dissimilarity function.

5 Experiments

We conduct experiments on the MM-AVS dataset and evaluate the performance by ROUGE (Lin, 2004). R-1, R-2, and R-L respectively represent ROUGE-1, ROUGE-2, and ROUGE-L F1-scores, which are widely used to calculate the n-grams overlapping between decoded summaries and references.

	Inform	Satis
document	3.65	3.76
video	2.73	2.78
document+video	3.87	4.30

Table 5: Manual summary quality evaluation.

5.1 Assistance of External Information

Videos, audios, or transcripts are less concerned than documents and images, as revealed in Table 1. Accordingly, the multimodal corpus assembling all of them has been absent so far, till MM-AVS is built in this study. To verify the importance of these materials for multimodal summarization, we test a text-only baseline and its two extensions. As for the baseline, we construct a hierarchical framework that concentrates on word and sentence levels with a feedforward classification layer. Its two extensions respectively take videos and transcripts for additional considerations.

As shown in Table 3, both videos and transcripts can contribute to improving multimodal summarizaions by fusing documents. This validates that the external information complementary for texts can facilitate capturing the core ideas of documents and inducing high-quality summaries.

5.2 Analysis of Transcript

To further investigate the nature of transcripts, we compare them with documents and references. As shown in Table 4, the video transcripts in MM-AVS are distinct from documents with low overlaps, indicating that they are not repeating documents but provide useful assistant information. While Table 4 also illustrates that the transcripts are lowly correlated with references, suggesting that transcripts can assist summary generation but are not enough for the final excellent summaries.

5.3 Manual Evaluation

The *document*, *video*, and *document with video* summarization results on 200 groups of MM-AVS examples are scored by five computer science graduates in terms of their informativeness (Inform) and satisfaction (Satis). Each summary is scored from 1 to 5, where a higher score denotes more informative or satisfied, and we record the average scores in Table 5. It shows that the summaries induced via *documents and videos* are more close to human comprehensions, which is in accord with the observations in Section 5.1, verifying the importance of external information such as videos for excellent summaries.

5.4 Conclusions

In this work, we contribute a full-scale dataset for multimodal summarization, which extensively assembles documents, summaries, images, captions, videos, audios, transcripts, and titles. A novel multimodal summarization framework is proposed based on this dataset to be taken as a baseline for the future research in this community.

Acknowledgement

We sincerely thank Wei Liu for his constructive collaboration during development of this paper and Nayu Liu for the helpful discussions.

References

Jingqiang Chen and Hai Zhuge. 2018. Abstractive text-image summarization using multi-modal attentional hierarchical RNN. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4046–4056.

Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas, and Yannis Avrithis. 2013. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, 15(7):1553–1568.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Haoran Li, Peng Yuan, Song Xu, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020a. Aspect-aware multimodal summarization for chinese e-commerce products. In *Proceedings of the AAAI Conference on Artificial Intelligence*, page in press.

Haoran Li, Peng Yuan, Song Xu, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020b. Aspect-aware multimodal summarization for chinese e-commerce products. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8188–8195.

Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, and Chengqing Zong. 2018. Multi-modal sentence summarization with modality attention and image filtering. In *International Joint Conference on Artificial Intelligence*, pages 4152–4158.

Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2017. Multi-modal summarization for asynchronous collection of text, image, audio and video. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1092–1102.

Jindrich Libovickỳ, Shruti Palaskar, Spandana Gella, and Florian Metze. 2018. Multimodal abstractive summarization of opendomain videos. In *Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL). NIPS.*

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 8.

Kuan Liu, Yanen Li, Ning Xu, and Prem Natarajan. 2018. Learn to combine modalities in multimodal deep learning. *arXiv preprint arXiv:1805.11730.*

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3075–3081.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290.

Shruti Palaskar, Jindřich Libovický, Spandana Gella, and Florian Metze. 2019. Multimodal abstractive summarization for how2 videos. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6587–6596.

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. In *Workshop on Visually Grounded Interaction and Language (ViGIL), NIPS.*

Naushad UzZaman, Jeffrey P. Bigham, and James F. Allen. 2011. Multimodal summarization of complex sentences. In *ACM International Conference on Intelligent User Interfaces*, pages 43–52.

Kaiyang Zhou, Yu Qiao, and Tao Xiang. 2018. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7582–7589.

Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2018. MSMO: Multimodal summarization with multimodal output. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4154–4164.

Junnan Zhu, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong, and Changliang Li. 2020. Multimodal summarization with guidance of multimodal reference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, page in press.