The Current Landscape of Multimodal Summarization

Atharva Kumbhar
kumbhar.atharva@outlook.com

Harsh Kulkarni
harshkulkarni1105@gmail.com

Atmaja Mali
atmajamali07@gmail.com

Prathamesh Mulay
prathumulay@gmail.com

Sheetal Sonawane
sssonawane@pict.edu

Abstract

In recent years, the rise of multimedia content on the internet has inundated users with a vast and diverse array of information, including images, videos, and textual data. Handling this flood of multimedia data necessitates advanced techniques capable of distilling this wealth of information into concise, meaningful summaries. Multimodal summarization, which involves generating summaries from multiple modalities such as text, images, and videos, has become a pivotal area of research in natural language processing, computer vision, and multimedia analysis. This survey paper offers an overview of the state-of-the-art techniques, methodologies, and challenges in the domain of multimodal summarization. We highlight the interdisciplinary advancements made in this field specifically on the lines of two main frontiers:1) Multimodal Abstractive Summarization, and 2) Pre-training Language Models in Multimodal Summarization. By synthesizing insights from existing research, we aim to provide a holistic understanding of multimodal summarization techniques.

1 Introduction

In today's information age, where vast amounts of data are generated and shared in diverse formats such as text, images, and videos, the ability to distill meaningful insights from this multimedia landscape is paramount. Multimodal summarization, at the intersection of natural language processing, computer vision, and multimedia analysis, stands as a pivotal solution to this burgeoning challenge. Unlike traditional text-only summarization methods, the essence of multimodal summarization lies in its capacity to synthesize information from multiple sources, creating condensed yet comprehensive summaries that capture the richness of textual, visual, and auditory data. This synthesis not only caters to the evolving needs of users seeking efficient content consumption but also finds applications in diverse domains such as multimedia retrieval, content recommendation, and intelligent data analysis. By delving deep into the methodologies, algorithms, and challenges faced in this interdisciplinary field, this survey aims to provide a panoramic view of the existing state-of-the-art techniques.

We primarily concentrate our efforts on two research frontiers:

1. Multimodal abstractive summarization and
2. Pre-training for multimodal summarization.

The former is chosen as it lays a foundation for multimodal summarization research. The prominent datasets used for the same such as How2(Sanabria et al., 2018) are ideal (both from the length and quality point of view) for experimentation and can be used as benchmarks. A lot of research and experimentation models have thus been proposed on these datasets. The latter is a new paradigm that very recently started proliferating in multimodal summarization research. To take multimodal summarization to practical applications, however, the above is not enough.

- Practical applications will have long videos and transcripts which project the problem into long document summarization, standard encoder-decoder models fail to capture long-range dependencies.
- It is quite difficult to find large datasets that are close to the problem statement at hand, thus the need for a pre-trained model that showcases strong generalization capabilities, and can be fine-tuned with a small number of samples.

The pre-training and fine-tuning paradigm has been widely successful in NLP and Computer Vision tasks. It significantly reduces the training

costs on downstream tasks and aids generalizability. Hence the recent work has been shifted toward building a pre-trained language model specifically for multimodal understanding and summarization. However, certain challenges render the conventional pre-training approaches ineffective. We discuss the innovations and research efforts that have been put in to develop multimodal GPLMs(Generatively Pre-trained language models). Based on these developments we also outline a general recipe and pattern in these works and provide a sample architecture/pipeline of these models. We also outline the challenges and future scope of research in this field. Moreover, we provide details of other supplementary resources such as Datasets and Applications for Multimodal Summarization. In short, we provide an exhaustive survey of the past, the present, and the future of Multimodal Summarization research.

The rest of the paper is structured as follows- Comparative analysis of various available datasets (section 2), followed by advances in Multimodal abstractive summarization are discussed in (section 3), discussion on Pre-training in (section 4), Results (sections 5) and Challenges and future work(section 6).

2 Datasets

Datasets play a crucial role in the advancements of a particular field. As there are many different input-output modalities and the job is customizable, there isn't a single standard dataset that can be used as a baseline for evaluation across all techniques for the MMS task. However, we've gathered details regarding the datasets utilized in earlier studies, also found in Table 1 is a thorough analysis of available public datasets.

3 MultiModal Abstractive Summarization(MAS)

The task of multimodal summarization was first formally defined in (Zhu et al., 2018). The formal definition deduced from the paper is given by: "Given a document containing a set of sentences, and a set of images, the output of automatic multimodal summarization is a pictorial summary, consisting of condensed textual information, and the subsequent set of most relevant images". To extend the pointer generator network (See et al., 2017), which was built for text modality, MSMO (Zhu et al., 2018) introduced a similar pipeline for image modality. Similarly, the attention mechanism was extended to multimodal attention by fusing both the text and image context vectors during the decode step. Finally, the image output was given by an ancillary network that would rank all the images. However, this model was being trained only with text modality in mind, which led to the problem of modality bias, i.e. only the central modality gets trained properly, and the secondary modality's loss is suppressed in the total objective loss in this training mechanism.

To alleviate this lack of multimodal reference, (Zhu et al., 2020) introduces a new multimodal objective function. This objective function consists of the conventional negative log-likelihood over the decoded text and an additional cross-entropy loss of the image selection. This image selection process is guided with the help of the image discriminator module, which gets trained along with the seq-to-seq parameters, thereby mitigating the problem of modality bias.

(Chen and Zhuge, 2018) proposes a model with a hierarchical encoder-decoder model. The proposed model uses 3 multimodal attention mechanisms (text-image caption, text-image and image-image caption) to compute the sentence context in the hierarchical decoder. The model uses RNN(Chung et al., 2014) for text and image captions and CNN (Anvarov et al., 2020) to obtain image vector representation, followed by bi-directional RNN to encode ordered image sets. The attention score of each image is computed with each sentence of the generated summary. Based on the score, images are arranged in descending order and the top K is selected.

Previous abstractive approaches used only image-text as modalities until (Sanabria et al., 2018) introduced the How2 dataset (details in dataset section). Videos contain abundant data that is represented in chronological order, which is crucial for summarization.(Palaskar et al., 2019) proposed baseline framework, in which the videos are encoded with the help of a pre-trained action recognition model: ResNeXt-101 and text is encoded using bi-directional RNN. This is followed by a hierarchical attention mechanism to fuse the modalities and a standard RNN decoder step.

Resnext-101 3D convolutional neural network (Hara et al., 2017) is trained to recognize 400 different human actions 6589 in the Kinetics dataset. This model outputs 2048 dimensional features, ex-

Dataset	Input Modalities	Output Modalities	Data Statistics	Domain
(UzZaman et al., 2011)	Images, Texts	Images, Text	Wikipedia Web crawling content.	Wikipedia entries
How2(Sanabria et al., 2018)	Text, Audio, Video	Abstractive text	2,000 hours of short instructional videos with human-generated transcript and 2-3 sentence summary for each video.	Multiple domain videos such as cooking, sports, indoor/outdoor activities, music, etc
(Li et al., 2017)	Text(Chinese, English),images, audio, video	Extractive text	25 documents in English, 25 documents in Chinese	News (Google News, CCTV.com, Youtube)
(Zhu et al., 2018)	Text, Images	Abstractive text, Images	313k documents, 2.0m images	News
(Jangra et al., 2021)	Text, Images, Audio, Video	Extractive text, Images, Audio, Video	25 topics (contains complementary and supplementary multi-modal references)	News
(Zhu et al., 2018)	Text, Images	Abstractive text, Images	313k documents, 2.0m images	News
(Papalampidi and Lapata, 2023)	Text, Images, Audio, Video	Text, Images, Audio, Video	98 samples with more than 26k episodes consisting of different genres such as drama, comedy, crime etc.	tv shows
(Zhong et al., 2021)	Text, Video, Audio, Images	Text, Video, Audio, Images	232 meeting recordings across multiple domains: Product, academic and committee meetings. 1,808 question-summary pairs in QMSum with an average length of 69.6	Meeting
(Cho et al., 2021b)	Text, Video, Audio, Images	Text, Video,Audio, Images	500 hours of annotated streamed video	livestream
(Fu et al., 2021)	Text, Video, Audio, Images	Abstractive text, Video, Audio, Images	Extends DailyMail and CNN collections to multimodalities with avg. of 2k articles and videos of avg.length of 450.15 hours	News

Table 1: Comaprisions of various datasets

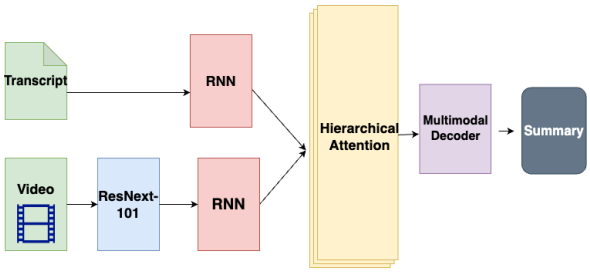

Figure 1: General Architecture of models based on MAS

tracted for every 16 non-overlapping frames in the video. Figure 1 shows the general architecture of How2 based models.

MAST (Khullar and Arora, 2020) which uses the same technique for video encoding, emphasizes the importance of the audio modality along with text and video for generating an abstractive summary. The model considers hierarchical attention between modalities, considering higher priority to the text followed by the video and then by audio. The model separately computes context vectors for audio-text and text-video, at the second level of hierarchical attention, which is combined at the last level. Thus now, the decoder has aggregate information on all the modalities at each timestep.

The model proposed in VMSMO(Li et al., 2020b), proposes a DIMS(Dual Interaction-based Multimodal summarize) which conducts dual interaction between videos and text data after capturing the spatial and temporal dependencies by RNN encoders. The proposed method of a conditional self-attention mechanism which learns locally in a fixed window size, and a global-attention mechanism to learn high-level representations of video-aware articles and article-aware videos. A multi-modal generator enforces Mutli-task learning by extracting a cover frame and textual summary at the same time thus increasing MAP and ROUGE-L score.

Even though video is the most informative and easiest mode of communication, many existing models face the challenge of being based on a single task and fail to capture temporal progression in videos. In (Fu et al., 2020) paper, they have improved by considering feature extraction using hierarchical frameworks and deep learning techniques for both text and video data. This paper introduces two multi-modal attention mechanisms that focus on diverse parts of video frames. The bi-hop attention is utilized to produce a context by simultaneously combining text sentences with a transcript and video frames. Since a transcript is similar to an article text, it uses the BiLSTM to extract its features, addressing the challenge of asynchronism. Feature fusion combines information from text and video, considering both early and late fusion strategies to handle the asynchronous nature of multimodal data. The model is trained jointly for text and video summarization tasks, using sentence classification for text and unsupervised learning with reward-based methods for video. Overall, M2SM aims to extract salient information from articles and videos, improving the summarization process by effectively integrating multi-modal information.

The proposed Multimodal Hierarchical Multimedia Summarization (MHMS) (Qiu et al., 2022) method addresses the task of multimodal multimedia summarization with both visual and language as output, thus being an extension to VMSMO. The model takes as input a multimedia source containing textual documents and videos, along with ground truth textual summaries and cover pictures for the videos. MHMS consists of five modules: Video Temporal Segmentation (VTS) for scene detection, Visual Summarization for extracting keyframes, Textual Segmentation for document segmentation, Textual Summarization using BART for generating abstractive textual summaries, and Cross-Domain Alignment using Optimal Transport to align visual keyframes with textual summaries. VTS uses a binary classification approach on segment boundaries(Bi-LSTM), while Visual Summarization employs an encoder-decoder architecture with attention. Textual Segmentation implements a hierarchical BERT model, and Textual Summarization employs Bidirectional and Auto-Regressive Transformers (BART) (Lewis et al., 2019). Cross-domain alignment uses Optimal Transport for matching visual and textual features, leveraging the cross-domain interaction.

(Liu et al., 2020) utilizes co-attention between the modalities, namely text-to-video fusion generator (T2VFG) that extracts visual information related to the text, and video-to-text fusion generator(V2TFG) that extracts textual information related to the video. Moreover, they claim that the flow of noise information during cross fusion such as redundant information, image, or text, can greatly inhibit the cross-fusion representation capabilities, and hinder complementarity in multimodal data. To tackle this challenge they introduce a forget gate and memory vector between the infor-

mation flow, this controls the flow of noise and mismatched information. In the decoder, they use a hierarchical attention-inspired decoding technique and factor in three attention maps- text attention, video attention, and Attention over Multimodal attention(AoMA). The AoMA is a high-level context vector computed on top of the two attention maps. This way information of different granularity is captured. The model achieved state-of-the-art results on How2.

One different approach can be seen in (Liu et al., 2020), which comprises of decoder-only multimodal transformer for generating an abstractive summary, which primarily aims to reduce the parameter redundancy with models having multiple encoders and encoder-decoder. The authors of the paper introduce 2 types of decoders, one is the multimodal decoder and the other is the standard decoder. The proposed model uses a sequence of concatenated input transcript and output summary separated by a unique token, which is then fed to a multimodal decoder along with video representation to obtain joint representation, which is then passed on to a standard decoder module. During the training, the model learns to predict the transcript-summary sequence. The paper's authors define a new loss function in which additional loss from the source transcript is added to the original loss function of the targeted summary. During testing, the source transcript is not used for prediction but serves as current input for sequentially predicting the remaining summary text.

4 Pre-Training for Multimodal Summarization

Pre-training and fine-tuning paradigm is widely successful in a variety of natural language processing, and computer vision tasks. Pre-training specific to text summarization is done by using a variation of masked language modelling, in which span corruption is used (Raffel et al., 2020; Lewis et al., 2019; Yang et al., 2021). This substantially improves the downstream summarization performance.

Similarly, research efforts have been put in to develop methodologies for building pre-trained transformer-style encoders that can be used for Vision-Language understanding tasks (Goyal et al., 2017; Tan and Bansal, 2019; Li et al., 2020a). For text-understanding capabilities, most works resorted to standard masked language modelling (Devlin et al., 2019b), and for vision understanding a similar idea was used, just at different levels of granularity. (Huang et al., 2020) considered the pixel level granularity and employed MLM(masked language modelling), Image Text Matching, and Random Pixel Sampling as self-supervised objective functions.

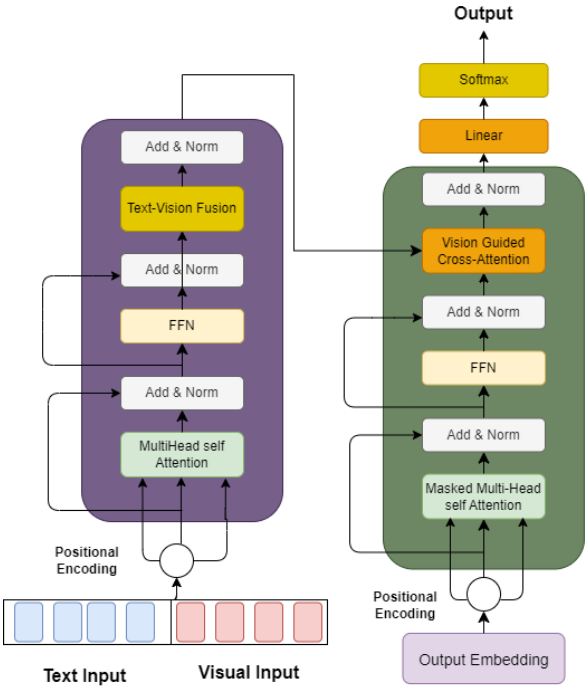

Figure 2: General Architecture for Pre-Trained LM for Multimodal Summarization

(Li et al., 2019) focused on object-level granularity, and introduced masked object detection as their objective function. Furthermore, (Zhang et al., 2021) focused on improving vision representations on similar object detection pre-training lines. (Kim et al., 2021) dealt with patch-level granularity. While these are language-vision understanding models, a similar plethora of research has been conducted on vision and language-guided text generation (Zhu et al., 2020) presented a model for both visual question answering and image captioning (Chen et al., 2015). Similar to pre-trained encoder-decoder T5 (Raffel et al., 2020), to unify all the tasks (Cho et al., 2021a) was introduced.

Although prior work has made much progress on VL pre-training, the problem of generating text given text and video input (E.g. the How2 dataset) is not well studied under the VL pre-training setting, except by (Luo et al., 2020), who proposed a dual-stream model for both VL classification and generation with video data. However, compared

to GPLMs in NLP such as BART (Lewis et al., 2019) and T5 (Raffel et al., 2020), their text generation ability is limited as the most prominent dataset How2 (Sanabria et al., 2018) is also comparatively very small, and it contains only short videos and short 2-3 sentence long reference summaries and is thus prone to overfitting. Therefore there are several challenges to replicating a similar success story in the multimodal scenario and any pre-training methodology that initializes a large model from scratch is bound to fail.

The research community found an innovative workaround to this limitation. The research was then manoeuvred to find training objectives on already pre-trained large language models, to render them multimodal, i.e. finding objective functions for injecting visual information in the pre-trained models such as BART. Notably, (Yu et al., 2021), discussed 2 injection strategies: cross-modal dot-product attention and multi-head attention. To further improve the performance of the model, a forget gate (FG)(Liu et al., 2020) is used to filter out redundancy and noisy information in the visual features. The research provides a foundation by answering two main questions: How to apply injection and where to inject visual information.

The most prominent work of pre-trained end-to-end transformer architecture is seen in ((Papalampidi and Lapata, 2022) Adapters for Long Video-to-text Summarization). Firstly, to mitigate the issue of a small dataset they introduce SummScreen, a dataset consisting of 4,575 episodes, their transcriptions, and corresponding reference summaries. As the video and text are very long in this dataset(full-length episodal summary), they take inspiration from a long-document summary regime and explore two methods: making attention lightweight-inspired by (Beltagy et al., 2020) and content selection techniques.

After selecting the utterances, the multimodal embedding vector is appended to the textual data. For utterances <U1, U2, .. Un> we have the corresponding input tokens given by <M1, t1, t2, t3, ...tk, M2, t1, t2, t3,.., Mn, t1, t2, ..tn>, where Mi is multimodal information vector, and ti is the textual input token. To process the multimodal information, they add hierarchical adapters to the encoder and the decoder model of BART, adding only a small fraction of new parameters. These adapters fuse the global-level utterance information with low-level textual information. They end up fine-tuning/training only 3.8% of the total model parameters.

Following the task of MSMO((Zhu et al., 2018)) and (Zhu et al., 2020), wherein a pictorial summary was to be generated with text output and relevant images, (Zhang et al., 2022) proposed a unified framework built upon BART(Lewis et al., 2019) that performs extractive, abstraction summary, as well as image ranking. Firstly, they extend the BART encoder to accept image modality, the images are first divided into patches, flattened, and linearly projected. Similarly, to factor in visual information, they added a visual cross-attention block after every self-attention block, that attended to the visual hidden states. Figure 2 shows the general architecture of Pre-Trained LM for Multimodal Summarization.

To overcome the same limitations of the How2 dataset, as mentioned above,(Atri et al., 2021) developed a new dataset called AVIATE, which is a large-scale dataset for multimodal abstractive summarization with videos of diverse duration and developed from academic presentations. The authors of the paper use DeepSpeech (Hannun et al., 2014) for Automatic Speech Recognition (ASR) to extract transcripts of all the videos. As the dataset contains academic presentations, the useful data from the slides is extracted from Google OCR vision API. Initially, ASR and OCR were both represented using pre-trained model BERT (Devlin et al., 2019a), and then these representations jointly attended to reduce redundant words from both sources. These fused representations of ASR-OCR and video and audio modalities are then fed to Factorised Multimodal Transformer (FMT (Zadeh et al., 2019)) architecture for sequential multimodal learning. The proposed decoder-only network uses a pre-trained transformer (Vaswani et al., 2017) as its basic module and generates summary-like language modelling. Thus, this concept of transformer LM is extended in the multimodal environment using FMT.

5 Results

Most prominently the performance of the models is benchmarked on How2(Palaskar et al., 2019), and the VMSMO(Li et al., 2020b) dataset. However, some research papers employed their datasets such as (Papalampidi and Lapata, 2022), (Fu et al., 2020), etc. We report the widely used evaluation metrics for summarization tasks- Rouge-1, Rouge-

Paper	Dataset-version	R1	R2	RL	FC
MAST (Khullar and Arora, 2020)	300h	48.85	29.91	43.23	35.40
Forget Gate (Liu et al., 2020) with RNN	2000h	62.3	46.1	58.2	*
Forget Gate (Liu et al., 2020) with Transformers	2000h	61.6	45.1	57.4	*
(Palaskar et al., 2019)	2000h	*	*	54.9	48.9
(Liu et al., 2020)	300h	61.43	44.67	58.03	*
(Liu, 2019) BERT SUM	*	48.26	*	44.02	36.4
(Yu et al., 2021) VG-BART+Forget gate+Vision Transformer Encoder(VTE)	*	*	68.0	63.3	69.7

Table 2: Results of Various Papers on How2 dataset (Sanabria et al., 2018)

2, Rouge-L, and Content-F scores in Table2 [1] and Table3 [2].

Paper	R1	R2	RL	FC
(Qiu et al., 2022) MHMS	27.1	9.8	25.4	*
(Li et al., 2020b) VMSMO	25.1	9.6	23.2	*

Table 3: Results of Papers based on VMSMO dataset

6 Challenges and Future Frontiers

Whilst there have been innovative measures to inject visual information into the pre-trained LMs, it is not yet clear how effectively is that information getting encoded and how effectively is it used by the decoder. (Zhang et al., 2022) attempted to visualize the impact of the added vision component. However, they only showcased the relationship between the input text and the images, in the encoder step, and not the relationship between the decoded text and the image coverage being considered. As the LM is largely still pre-trained on the text modality, the downstream tasks might face the problem of modality bias (Zhu et al., 2020). The task of standard abstractive summarization saw exponential progress with the advent of Pegasus(Zhang et al., 2020), BART(Lewis et al., 2019), Z-Code++(He et al., 2023) etc. The primary reason for this jump was that these approaches came up with specific self-supervised objective functions that were specifically helpful for downstream summarization tasks. For a pre-trained model to be successful it has to satisfy- 1) The objective function will help downstream tasks, and 2) The dataset used for pre-training must be large, and generalized. It is hard for multimodal summarization to satisfy either of the criteria. The availability of such datasets is limited, moreover, the pre-training objective functions for multimodal understanding notoriously hinder the generation capabilities (Yu et al., 2021). Most applications of Multimodal summarization deal with long videos, and documents, long document summarization is a challenging field, as it is hard to capture long-range relationships, and is computationally expensive. Recently, (Reed et al., 2022) and (Jaegle et al., 2022) have shown promising results that map from arbitrary length input to arbitrary length output in a transformer-style model, and fixed size parameters(however, they have not been used for summarization yet). Research on such new frontiers could help solve the challenges of multimodal summarization.

[1] *-Not provided in particular paper

[2] *-Not provided in particular paper

7 Conclusion

This survey paper presents a comprehensive exploration of the multifaceted landscape of multimodal summarization. We then offer a detailed overview of the datasets needed to complete the MMS challenge. Additionally, we provide a brief description of the several approaches taken to complete the MMS assignment as well as the assessment metrics applied to determine the quality of the summaries generated. Lastly, we also offer some potential avenues for future MMS study. In essence, this survey serves as a compass guiding researchers, practitioners, and enthusiasts through the dynamic landscape of multimodal summarization.

References

UzZaman, N., Bigham, J. P., and Allen, J. F. Multimodal summarization of complex sentences. In *Proceedings of the 16th International Conference on Intelligent User Interfaces*, IUI '11, page 43–52. Association for Computing Machinery, New York, NY, USA, 2011. ISBN 9781450304191. doi: 10.1145/1943403.1943412.

Chung, J., Gülçehre, Ç., Cho, K., and Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555, 2014.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., and Ng, A. Y. Deep speech: Scaling up end-to-end speech recognition. 2014.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913. 2017.

Hara, K., Kataoka, H., and Satoh, Y. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? *CoRR*, abs/1711.09577, 2017.

Li, H., Zhu, J., Ma, C., Zhang, J., and Zong, C. Multimodal summarization for asynchronous collection of text, image, audio and video. In M. Palmer, R. Hwa, and S. Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1092–1102. Association for Computational Linguistics, Copenhagen, Denmark, 2017. doi:10.18653/v1/D17-1114.

See, A., Liu, P. J., and Manning, C. D. Get to the point: Summarization with pointer-generator networks. *CoRR*, abs/1704.04368, 2017.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Chen, J. and Zhuge, H. 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. Association for Computational Linguistics, Brussels, Belgium, 2018. doi: 10.18653/v1/D18-1438.

Sanabria, R., Caglayan, O., Palaskar, S., Elliott, D., Barrault, L., Specia, L., and Metze, F. How2: a large-scale dataset for multimodal language understanding. *arXiv preprint arXiv:1811.00347*, 2018.

Zhu, J., Li, H., Liu, T., Zhou, Y., Zhang, J., and Zong, C. MSMO: Multimodal summarization with multimodal output. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4154–4164. Association for Computational Linguistics, Brussels, Belgium, 2018. doi:10.18653/v1/D18-1448.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota, 2019b. doi:10.18653/v1/N19-1423.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. 2019a.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

Li, G., Duan, N., Fang, Y., Gong, M., Jiang, D., and Zhou, M. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. 2019.

Liu, Y. Fine-tune bert for extractive summarization. 2019.

Palaskar, S., Libovický, J., Gella, S., and Metze, F. Multimodal abstractive summarization for how2 videos. In A. Korhonen, D. Traum, and L. Màrquez, editors,

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 6587–6596. Association for Computational Linguistics, Florence, Italy, 2019. doi:10.18653/v1/P19-1659.

Tan, H. and Bansal, M. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

Zadeh, A., Mao, C., Shi, K., Zhang, Y., Liang, P. P., Poria, S., and Morency, L.-P. Factorized multimodal transformer for multimodal sequential learning. 2019.

Anvarov, F., Kim, D. H., and Song, B. C. Action recognition using deep 3d cnns with sequential feature aggregation and attention. *Electronics*, 9(1), 2020. ISSN 2079-9292. doi:10.3390/electronics9010147.

Beltagy, I., Peters, M. E., and Cohan, A. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Fu, X., Wang, J., and Yang, Z. Multi-modal summarization for video-containing documents. 2020.

Huang, Z., Zeng, Z., Liu, B., Fu, D., and Fu, J. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020.

Khullar, A. and Arora, U. Mast: Multimodal abstractive summarization with trimodal hierarchical attention. *arXiv preprint arXiv:2010.08021*, 2020.

Li, G., Duan, N., Fang, Y., Gong, M., and Jiang, D. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11336–11344. 2020a.

Li, M., Chen, X., Gao, S., Chan, Z., Zhao, D., and Yan, R. VMSMO: Learning to generate multimodal summary for video-based news articles. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9360–9369. Association for Computational Linguistics, Online, 2020b. doi:10.18653/v1/2020.emnlp-main.752.

Liu, N., Sun, X., Yu, H., Zhang, W., and Xu, G. Multistage fusion with forget gate for multimodal summarization in open-domain videos. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1834–1845. Association for Computational Linguistics, Online, 2020. doi:10.18653/v1/2020.emnlp-main.144.

Luo, H., Ji, L., Shi, B., Huang, H., Duan, N., Li, T., Li, J., Bharti, T., and Zhou, M. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

Zhang, J., Zhao, Y., Saleh, M., and Liu, P. J. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. 2020.

Zhu, J., Zhou, Y., Zhang, J., Li, H., Zong, C., and Li, C. Multimodal summarization with guidance of multimodal reference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9749–9756, 2020. doi:10.1609/aaai.v34i05.6525.

Atri, Y. K., Pramanick, S., Goyal, V., and Chakraborty, T. See, hear, read: Leveraging multimodality with guided attention for abstractive text summarization. *Knowledge-Based Systems*, 227:107152, 2021. ISSN 0950-7051. doi:https://doi.org/10.1016/j.knosys.2021.107152.

Cho, J., Lei, J., Tan, H., and Bansal, M. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021a.

Cho, S., Dernoncourt, F., Ganter, T., Bui, T., Lipka, N., Chang, W., Jin, H., Brandt, J., Foroosh, H., and Liu, F. StreamHover: Livestream transcript summarization and annotation. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6457–6474. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021b. doi:10.18653/v1/2021.emnlp-main.520.

Fu, X., Wang, J., and Yang, Z. MM-AVS: A full-scale dataset for multi-modal summarization. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5922–5926. Association for Computational Linguistics, Online, 2021. doi:10.18653/v1/2021.naacl-main.473.

Jangra, A., Saha, S., Jatowt, A., and Hasanuzzaman, M. Multi-modal supplementary-complementary summarization using multi-objective optimization. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 818–828. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450380379. doi:10.1145/3404835.3462877.

Kim, W., Son, B., and Kim, I. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

Yang, J., Ma, S., Huang, H., Zhang, D., Dong, L., Huang, S., Muzio, A., Singhal, S., Awadalla, H. H., Song, X., et al. Multilingual machine translation systems from microsoft for wmt21 shared task. *arXiv preprint arXiv:2111.02086*, 2021.

Yu, T., Dai, W., Liu, Z., and Fung, P. Vision guided generative pre-trained language models for multimodal abstractive summarization. *arXiv preprint arXiv:2109.02401*, 2021.

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., and Gao, J. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588. 2021.

Zhong, M., Yin, D., Yu, T., Zaidi, A., Mutuma, M., Jha, R., Awadallah, A. H., Celikyilmaz, A., Liu, Y., Qiu, X., and Radev, D. QMSum: A new benchmark for query-based multi-domain meeting summarization. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921. Association for Computational Linguistics, Online, 2021. doi:10.18653/v1/2021.naacl-main.472.

Jaegle, A., Borgeaud, S., Alayrac, J.-B., Doersch, C., Ionescu, C., Ding, D., Koppula, S., Zoran, D., Brock, A., Shelhamer, E., Hénaff, O., Botvinick, M. M., Zisserman, A., Vinyals, O., and Carreira, J. Perceiver io: A general architecture for structured inputs outputs. 2022.

Papalampidi, P. and Lapata, M. Hierarchical3d adapters for long video-to-text summarization. *arXiv preprint arXiv:2210.04829*, 2022.

Qiu, J., Zhu, J., Xu, M., Dernoncourt, F., Bui, T., Wang, Z., Li, B., Zhao, D., and Jin, H. Mhms: Multimodal hierarchical multimedia summarization. 2022.

Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S. G., Novikov, A., Barth-Maron, G., Gimenez, M., Sulsky, Y., Kay, J., Springenberg, J. T., Eccles, T., Bruce, J., Razavi, A., Edwards, A., Heess, N., Chen, Y., Hadsell, R., Vinyals, O., Bordbar, M., and de Freitas, N. A generalist agent. 2022.

Zhang, Z., Meng, X., Wang, Y., Jiang, X., Liu, Q., and Yang, Z. Unims: A unified framework for multimodal summarization with knowledge distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11757–11764, 2022. doi: 10.1609/aaai.v36i10.21431.

He, P., Peng, B., Wang, S., Liu, Y., Xu, R., Hassan, H., Shi, Y., Zhu, C., Xiong, W., Zeng, M., Gao, J., and Huang, X. Z-code++: A pre-trained language model optimized for abstractive summarization. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5095–5112. Association for Computational Linguistics, Toronto, Canada, 2023. doi: 10.18653/v1/2023.acl-long.279.

Papalampidi, P. and Lapata, M. Hierarchical3D adapters for long video-to-text summarization. In A. Vlachos and I. Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1297–1320. Association for Computational Linguistics, Dubrovnik, Croatia, 2023. doi: 10.18653/v1/2023.findings-eacl.96.