Grafting Pre-trained Models for Multimodal Headline Generation

Lingfeng Qiao[†], **Chen Wu**[†], **Ye Liu**[†], **Haoyuan Peng**[†], **Di Yin**[†], **Bo Ren**[§]
[†]Tencent Youtu Lab, Shanghai, China
[§]Tencent Youtu Lab, Hefei, China
{leafqiao,rafelliu,haoyuanpeng,endymecyyin,timren}@tencent.com, overwindows@icloud.com

Abstract

Multimodal headline utilizes both video frames and transcripts to generate the natural language title of the videos. Due to a lack of large-scale, manually annotated data, the task of annotating grounded headlines for video is labor intensive and impractical. Previous researches on pre-trained language models and video-language models have achieved significant progress in related downstream tasks. However, none of them can be directly applied to multimodal headline architecture where we need both multimodal encoder and sentence decoder. A major challenge in simply gluing language model and video-language model is the modality balance, which is aimed at combining visual-language complementary abilities. In this paper, we propose a novel approach to graft the video encoder from the pre-trained video-language model on the generative pre-trained language model. We also present a consensus fusion mechanism for the integration of different components, via inter/intra modality relation. Empirically, experiments show that the grafted model achieves strong results on a brand-new dataset collected from real-world applications.

1 Introduction

In the age of information explosion, generating headlines of videos has been steadily gaining prominence on the short video platform. As the headlines can summarize the videos for people quickly acquiring their essential information. Good headlines are also beneficial for various scenarios, such as video retrieval, recommendation, and understanding (Zhu et al., 2022; Liu et al., 2022). Specially, video headline generation can be regarded as a textual generation task with multimodal inputs (Li et al., 2021), which is called multimodal generation as shown in Figure 1. Given a video with related transcript, algorithm aims to generate a short, concise and readable textual attraction title.

However, to build an effective model, collecting large-scale training data is the main challenge.

Figure 1: An example of multimodal generation.

Especially, in the multimodal headline generation task, the form of triplet data (video, source transcript, target summary) further increases the difficulty of collecting data, which limits the application of video headline generation.

To alleviate the issue in multimodal headline generation, the natural idea is to leverage pre-trained model (PTM). With large-scale corpus, such as GPT (Radford et al., 2018), BART (Lewis et al., 2019), PALM (Bi et al., 2020), etc., have shown great ability to generate readable and informative text. Since the multimodal headline generation combine both video and textual information, we propose that the model can be grafted by PTMs of language generation and video-text matching. The former provides the ability of headline generation, and the latter bridges the semantic gap between the multiple modalities (Radford et al., 2021). In addition, these two tasks have no concern of scarce data. For language generation, the existing pre-training model can be directly adopted. For video-text matching, the model can be trained with sufficient data from the Internet without manual annotations. By this means, multimodal headline generation model can be constructed by fine-tuning the grafted model with limited data collection.

In order to take advantage of the existing PTMs and improve reusability, we propose a **gra**fting mechanism for obtaining the **m**ultimodal summarization pre-training **mo**del (GraMMo). First, language generation and video-text matching tasks are

Proceedings of EMNLP 2022 Industry Track, pages 254–263
December 9–11, 2020.

introduced to pre-train the encoders and decoder, respectively. Then we construct a unified architecture with a video encoder, a text encoder and a text decoder which grafted from different PTMs to initialize the multimodal headline generation model. In addition, a joint-modality layer acted as a modality-balance gate is designed to fuse the video and text features. Unlike previous works which focus on retaining modality-shared feature (Libovickỳ et al., 2018; Yu et al., 2021), this layer uses a two-way attention strategy to capture the commonality and specialty of the modalities. In detail, the video modality can highlight the most relevant and important text tokens by video-text cross attention. The resulted feature is called video-enhanced text feature, which reflect the commonality. On the other hand, since video-enhanced text feature neglects the video specialty which is less related with text modality, we recombine the video embeddings according to video-text attention and exploit the video-specific feature for complementing the fusion representation. Furthermore, dynamic frame sampling (DFS) and masked word prediction are designed in the encoder parts to reinforce the multimodal representation. We summarize the main contributions as follows:

- We propose a grafting video-text pre-training framework for multimodal headline generation. By grafting PTMs of language generation and video-text matching, GraMMo can be efficiently trained without big data collection. It is beneficial for fast deployment of real-world applications.

- A joint-modality layer with multimodal fusion module is designed to pay balanced attentions to each modality. It uses a two-way attention strategy to capture the commonality and specialty of multiple modalities, which will reinforce the fusion representation for better headline generation.

- Extensive experiments on a proposed Chinese multimodal headline generation dataset, WB-News, demonstrate that the grafted model can effectively accelerate the downstream fine-tuning procedure and improve generation results. The proposed method has been deployed in an industrial media platform for Chinese video headline generation.

2 Related Work

2.1 Multimodal Generation

Multimodal generation task aims to generate short, concise and readable textual title that can capture the most core information of the input media. The task is closely relevant to text summarization (Zhang et al., 2020; Jiang et al., 2022) while it is much tougher because redundancy and complementary between multiple modalities should be studied (Jangra et al., 2021). Many literatures have been based on pre-extracted unimodal sequential representation and cross attention mechanism to obtain the fused feature (Li et al., 2020d; Khullar and Arora, 2020; Li et al., 2020b; Fu et al., 2020). Besides, some researchers took efforts to sufficiently fusing the multimedia inputs by hierarchical fusion (Liu et al., 2020; Yu et al., 2021; Zhang et al., 2021a). The objective of modality consistency is another tool to guide the learning of multimodal fusion (Zhu et al., 2020; Zhang et al., 2021b).

However, few existing methods studied PTM for multimodal generation (Seo et al., 2022). In this paper, we propose a grafting video-text generation model and a novel joint-modality layer which is designed to capture the commonality and specialty of multiple modalities.

2.2 Video-Text Pre-training

Video-Text pre-training models adopt the "pre-training and then fine-tuning" paradigm, which makes the downstream tasks able to utilize the abundant knowledge included in pre-training data.

One class of work is task-specific pre-training, and contrastive learning is used for zero-shot transfer and video-text retrieval tasks, such as CLIP (Radford et al., 2021) and other related researches (Miech et al., 2019; Patrick et al., 2020; Huang et al., 2021; Xu et al., 2021b). Furthermore, CBT (Sun et al., 2019a), HERO (Li et al., 2020c), VideoAsMT (Korbar et al., 2020) and UniVL (Luo et al., 2020) adopt multi-task learning (MTL) for pre-training on retrieval tasks. The other class of work concentrates on how to interact the multimodal inputs, including VideoBERT (Sun et al., 2019b), Unicoder-VL (Li et al., 2020a), VL-BERT (Su et al., 2019), UNITER (Chen et al., 2020), VLP (Zhou et al., 2018), ActBERT (Zhu and Yang, 2020), VLM (Xu et al., 2021a) and BEiT (Wang et al., 2022).

Currently, few video-text pre-training models focus on multimodal headline generation due to the

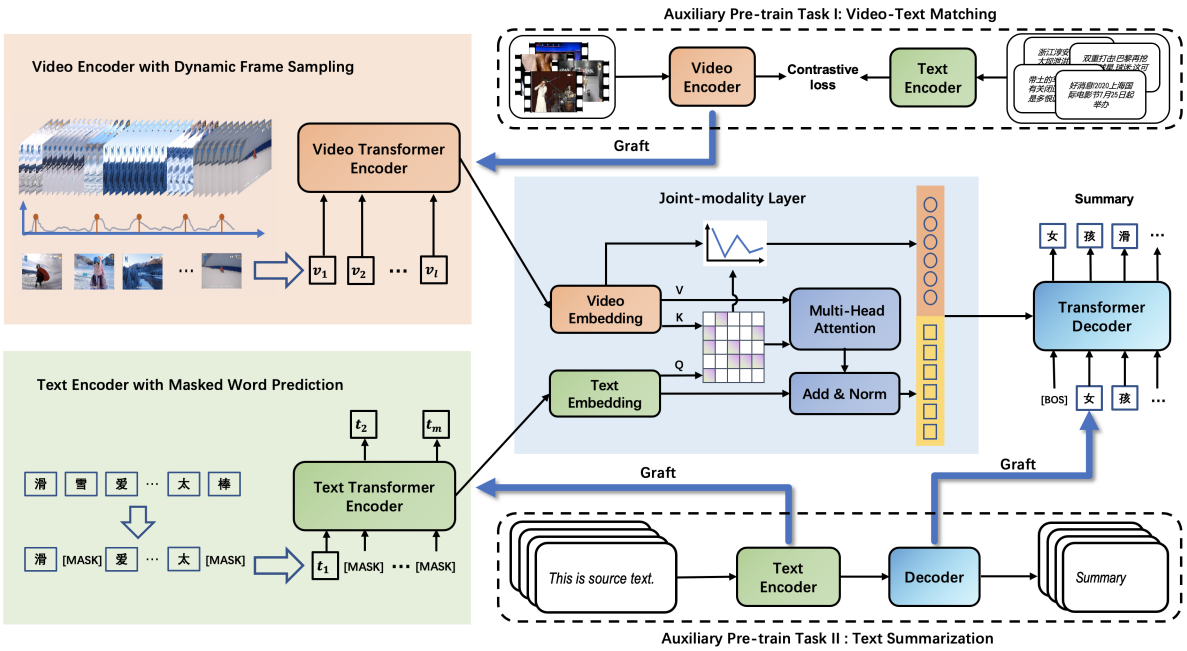

Figure 2: GraMMo framework for multimodal headline generation.

scarce data. GraMMo gives an effective grafting architecture for this task with ready-made PTMs of language generation and video-text matching, which can save a lot of computational costs.

3 Approach

3.1 Grafting Architecture

Given an input video $X_v = \{v_1, v_2, \cdots, v_l\}$ and related source transcript $X_t = \{t_1, t_2, \cdots, t_m\}$, the output is the target textual title $Y = \{y_1, y_2, \cdots, y_n\}$, where l, m, n are the numbers of the corresponding tokens. The goal is to generate a predicted title $Y' = \{y'_1, y'_2, \cdots, y'_n\}$ based on X_v and X_t, which can successfully grasp the main points of the video and transcript.

The proposed architecture is to provide a PTM for multimodal headline generation without large-scale triplet samples $\{X_v, X_t, Y\}$. The concept of grafting architecture GraMMo is illustrated in Figure 2. As a unified structure for multimodal generation, GraMMo consists of a video encoder $E_v(\cdot)$, a text encoder $E_t(\cdot)$, a joint-modality layer $F(\cdot)$ and a text decoder $D(\cdot)$. The encoders are designed for each modality individually and the architecture can be easily extended to various multimodal tasks with different kinds of inputs. Then the embeddings of modalities are fused by joint-modality layer to obtain the multimodal features. The joint-modality layer can provide video-enhanced text feature and video-specific feature, which involve the commonality and speciality of video and text modalities. Finally, text decoder is used to generate headline based on the fused multimodal feature.

To pre-train the video encoder, text encoder and text decoder, we draw support from two auxiliary tasks, i.e. language generation and video-text matching. As Figure 2 shows, the video encoder of video-text matching, the text encoder, and the text decoder of language generation are grafted to obtain the multimodal headline generation model.

3.2 Pre-train Generative Language Model

The language generation model with encoder-decoder structure can be adopted to build text encoder and decoder. In the work, PALM (Bi et al., 2020) which is Transformer-based (Vaswani et al., 2017) architecture is used as the NLG model. The text encoder and text decoder are pre-trained as classic abstractive text summarization task with large-scale unlabeled corpus.

In the pre-training stage, text encoder $E_t(\cdot)$ encodes source transcripts to obtain text embeddings $e_t = E_t(X_t)$, and then the decoder $D(\cdot)$ learns to generate hypothesis summaries. The generation loss and masked word prediction loss are used to guide the learning of text encoder and text decoder.

3.3 Pre-train Video-Text Understanding

We also use Transformer architecture for video encoder. The video features, extracted by I3D net-

work (Carreira and Zisserman, 2017), are first projected to video tokens before being fed into the video Transformer. For a video X_v, it has s frames, which is denoted as $V_f = \{f_1, f_2, \cdots, f_s\}$. Due to the concern of model efficiency, the frames should be sampled and converted to the video tokens with l length. Most conventional methods used pre-extracted features based on uniform sampling from the raw video, e.g. extract one frame every one second duration of the video. However, this sampling method may limit the expression of video. In the framework, to enhance video embeddings, dynamic frame sampling (DFS) is designed. It is a projection layer $DFS(\cdot) : \{1, 2, \cdots, l\} \rightarrow \{1, 2, \cdots, s\}$, which represents the choice from variable frames.

As a result, the video tokens can be obtained by $v_i = f_{DFS(i)}$, where $i = 1, 2, \cdots, l$. Then the I3D features of these tokens are extracted and video embeddings $e_v = E_v(X_v)$ are acquired by a stacked Transformer encoder.

To pre-train the video encoder $E_v(\cdot)$, video-text matching task is adopted to bridge the semantic gap between video and text modalities. Specifically, we collect large-scale video-text pairs from public video platform on the Internet without manual annotations. Videos and their corresponding descriptions are the natural data for video-text matching. The proposed video encoder and another Transformer-based text encoder are used to encode videos and descriptions respectively. Contrastive loss InfoNCE (Oord et al., 2018) is employed to calculate the correspondence between embeddings and guide the pre-training of encoders.

3.4 Joint-Modality Layer

Joint-modality layer will be used in fine-tuning stage after model grafting. Given text embeddings $e_t \in R^{m \times d}$ and video embeddings $e_v \in R^{l \times d}$, where d is the dimension of the embeddings, joint-modality layer is proposed to fuse them for headline generation. First, video embeddings should be used to highlight the significant elements in text embeddings and make algorithm pay attention to them from redundant source transcripts. Second, video embeddings can supplement key information that is not included in text embeddings to improve the informativeness of headline. To realize these motivations, as Figure 2 shown, the joint-modality layer uses a two-way attention strategy to capture the commonality and speciality of multiple modalities. Denote the query $Q \in R^{m \times d}$ is projected from text embeddings e_t, and the key $K \in R^{l \times d}$ and value $V \in R^{l \times d}$ are projected from video embeddings e_v. With dot-product between Q and K, the video-text attention matrix $M_{vt} \in R^{m \times l}$ is obtained, which represents the relations between text and video tokens. On the one hand, the text features are enhanced by video information based on M_{vt}. Multi-head attention is applied and V is added to the related text tokens to obtain the video-enhanced text feature $g = e_t + M_{vt}V$.

On the other hand, video embeddings e_v can be divided into two aspects, i.e. text-relevant feature and video-specific feature. Text-relevant feature is the part of e_v with large video-text attention score and video-specific feature is the opposite part. The text-relevant feature has already been considered in $M_{vt}V$. On the contrary, the video-specific feature is neglected and should be supplemented. We calculate video-text relevant distribution $p \in R^{1 \times l}$ by summing M_{vt} along the query dimension. The lower value in p means that such a video embedding is less relevant to text, which should be chosen for video-specific feature. As a result, the video-specific feature h is obtained by $h = e_v \odot Norm(1 - p)$, where $Norm(\cdot)$ means the normalized operator and $\odot$ is the element-wise multiplication.

Finally, the fusion embeddings $F(e_t, e_v)$ is obtained by concatenating video-enhanced text feature g and video-specific feature h.

3.5 Fine-tune Multimodal Generation Model

As mentioned above, text encoder $E_t(\cdot)$ and text decoder $D(\cdot)$ are pre-trained by language generation task. Video encoder $E_v(\cdot)$ is pre-trained by video-text matching task. By grafting these modules with joint-modality layer, a multi-modality generation model is established, which can be used for multimodal headline generation task.

For realistic applications, the specific multimodal headline generation triplet data should be collected to fine-tune the model. With GraMMo, we only need to prepare a small amount of data, since most of parameters in model are initialized by grafting, the model can converge rapidly.

4 Experiments

4.1 Datasets and Implementation

We leverage billions of Chinese corpus and millions of videos for pre-training language model and video-text matching models, respectively. For mul-

Methods	R-1	R-2	R-L	B-1	B-2	B-3	B-4	M
Text —> Text								
BART (Lewis et al., 2019)	33.14	19.51	29.41	32.54	25.34	19.60	15.39	29.48
PALM (Bi et al., 2020)	36.73	23.10	33.86	34.01	27.47	21.76	17.37	32.10
Video —> Text								
VLM (Xu et al., 2021a)	5.10	0.61	4.44	4.31	1.54	0.64	0.29	2.89
GraMMo-Video	7.85	1.73	6.90	7.21	3.31	1.76	1.01	5.06
Video+Text —> Text								
VG-GPLM (Yu et al., 2021)	35.35	21.46	32.10	33.81	26.70	20.78	16.38	31.01
MV-GPT (Seo et al., 2022)	37.74	24.04	34.42	36.13	29.27	23.34	18.81	33.73
MMPT (Xu et al., 2021b)	38.21	24.25	35.05	31.65	25.77	20.58	16.65	31.76
GraMMo	**38.87**	**24.85**	**35.38**	**37.80**	**30.65**	**24.43**	**19.65**	**35.30**

Table 1: Headline generation results on WB-News dataset.

timodal headline generation fine-tuning and testing, a new dataset WB-News is built. The details of datasets and the methodology used to obtain the corpus can be found in supplementary materials A. For evaluation on WB-News dataset, three metrics are employed: ROUGE (R-1,R-2,R-L) (ROUGE, 2004), BLEU (B-1,B-2,B-3,B-4) (Papineni et al., 2002), METEOR (M) (Banerjee and Lavie, 2005).

4.2 Headline Generation Results

On WB-News dataset, according to different types of input, we compare generation methods in three kinds of experimental setups, i.e. *Text —> Text*, *Video —> Text* and *Video+Text —> Text*. **BART** and **PALM** are used as baselines for classic text-only generation. Without text modality, **GraMMo-Video** is designed by pruning text encoder in GraMMo, which is compared with the video caption method **VLM**. For multimodal generation task, latest methods **VG-GPLM**, **MV-GPT** and **MMPT** are compared with the proposed **GraMMo**. As Table 1 shows, GraMMo achieves the best results among related multimodal generation methods and the SOTA text summarization method, PALM, with large margin. It illustrates that the video modality can help text to improve the headline results and GraMMo can better leverage the multimodal information against the related methods.

In the *Video —> Text* scenario, because the factual information, such as the name, cannot be extracted using the video modality, the headline metrics are quite low. Nevertheless, **GraMMo-Video** achieves a better performance and can generate reasonable summaries if neglecting the factual consistency. It also shows that our method has the ability to extract the language semantics from video.

Furthermore, GraMMo has been deployed on an industrial platform for video headline generation, which is shown in the discussion section.

Pre-trained Models				
Text	Video	R-L	B-4	M
		26.98	12.18	24.65
	✓	27.63	12.30	24.69
✓		34.90	18.63	34.06
✓	✓	**35.38**	**19.65**	**35.30**

Table 2: The effectiveness of using grafted model.

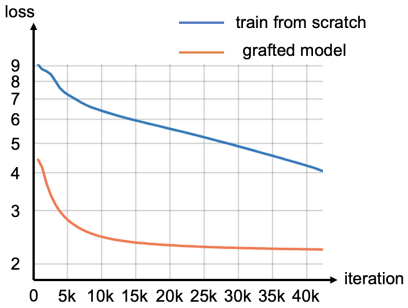

Figure 3: Different initialization methods.

4.3 Grafting for Headline Generation

In Table 2, the generation results are compared by eliminating grafted components. We found that the performance can be significantly improved by grafting. Figure 3 also shows that the learning curve of fine-tuning with grafted model decreases and convergences more rapidly against the case of train-from-scratch.

4.4 Ablation Study

To verify the contributions of each component of GraMMo, we design a series of ablation experiments. The results are shown in Table 3. First, all summary metrics decrease when DFS and video Transformer are removed from the video encoder. The probable reasons are that DFS can improve the generalization of video embedding and video Transformer can model the sequential character of video tokens. Second, different fusion strate-

Ground Truth	GraMMo (Video + Text → Text)	PALM (Text → Text)	GraMMo-Video (Video → Text)
好飒！95后汉服小姐姐雪上绝美驰骋 So cool! A post-1995 girl dressed in Han costume skis on the snow.	95后滑雪爱好者雪中宛如御剑飞行 A post-1995 skier skis like flying on the snow.	95后滑雪爱好者雪中觅食 A post-1995 skier foraged for food in the snow.	东北人在哪里玩雪？一起来解一下冬奥会滑雪 Where do the northeast people play snow? Let's take a look at winter Olympic skiing.
儿子一眼认出奥特曼是爸爸假扮：我不喜欢大肚子奥特曼 Boy recognized at a glance the Ultraman is the impersonation of his father. 'I don't like big belly Ultraman'.	爸爸给儿子制造惊喜奥特曼敲门，儿子：我不喜欢大肚子，我知道是爸爸 Father dressed in Ultraman and knocked at the door to make surprise for his son. Boy:'I don't like big belly. He is my dad.'	爸爸为给儿子庆生制造惊喜，大肚子我知道是爸爸 Father make surprise for his son's birthday. 'Big belly. I know he is my dad.'	可可爱爱！小女孩给消防员送苹果 So lovely! A little girl gave the fireman apples.
幸福时刻！载人航天发射塔架见证航天人婚礼 Happy moment! man-carrying rocket launch tower witness the wedding of aerospace industry staff.	酒泉卫星发射中心为120对航天新人举行婚礼 Jiuquan Satellite Launch Center held a wedding ceremony for 120 newlyweds, which work for aerospace industry.	超甜祝福！神舟飞船启航新人们集体婚礼 Sweet blessing! The Shenzhou spacecraft set off and the newlyweds group wedding.	向全国各族人民致以新春祝福 Happy New Year greetings to the people of the whole country.

Figure 4: Case study of proposed GraMMo, PALM and GraMMo-Video.

Methods	R-L	B-4	M
GraMMo	35.38	19.65	35.30
Encoder			
w/o DFS	-0.28	-0.13	-0.53
w/o Video Transformer	-0.98	-0.26	-1.03
Fusion			
Naive Concat	-0.89	-0.96	-1.40
Cross Attention	-0.85	-0.66	-1.16

Table 3: Ablation study on model components.

Methods	R@1	R@5	R@10
Train from Scratch			
text-to-video	34.40	53.19	60.28
video-to-text	35.39	52.70	60.21
Grafted Model			
text-to-video	39.65	60.43	67.73
video-to-text	40.35	60.35	67.30

Table 4: Video-Text Retrieval Results

gies are studied by substituting joint-modality layer. Naive concatenate and cross attention are adopted and decrease the summarization performance to great extent. The phenomenon illustrates that the proposed joint-modality layer is more effective in utilizing the multimodal information.

5 Discussion

5.1 Video-Text Matching Helps

One key issue of multimodal headline generation is how to map the video and text into the joint embedding space. Therefore, the alignment of these two modalities is important, which can be reflected by video-text retrieval performance.

We selected 1,400 samples of video and its title pair from WB-News to conduct video-text retrieval experiments. The video embeddings e_v and text embeddings e_t extracted from video encoder $E_v(\cdot)$ and text encoder $E_t(\cdot)$ are computed by dot product to measure the relevant scores. Given a query, the most related items are recalled by sorting the scores. Recall metrics (R@1, R@5, R@10) are used to measure the results. As shown in Table 4, the R@1 achieves about 40% and R@5 about 60%, which means the video and text are well aligned in one common space. Moreover, when using grafted model, the retrieval results are better than the results of train-from-scratch, reflecting the superiority of the grafted model.

5.2 Complementary Relation

Several examples are provided to intuitively understand the effectiveness of modality balance. As shown in Figure 4, the generated hypotheses of GraMMo, PALM and GraMMo-Video are presented. The key information is emphasised by green words, while the red words mean the wrong predictions. Compared with PALM, GraMMo can extract more key information and produce more logical expressions. This effect demonstrates the value of video modality for generating more remarkable headline.

However, the generated hypotheses of GraMMo-Video are totally inconsistent with the ground-truths because the factual information cannot be extracted solely by video modality. In fact, GraMMo-Video can generate the words containing similar topic semantics, which means our method can establish the connection between video and text modalities.

5.3 Human Evaluation

We perform human evaluation from the perspectives of readability and informativeness. For all test samples, the source video, reference headlines, and generated headline are shown to a group of people for evaluation. They need to judge the two aspects of readability and informativeness by giving an integer score in the range of 1-5, with 5 being perfect. Each sample is assessed by 5 people, and the average scores are used as the final score.

Methods	Read.	Info.
Ground Truth	4.35	4.05
PALM	3.65	3.08
MMPT	3.70	3.3
GraMMo	3.71	3.54

Table 5: Human evaluation results on readability (Read.) and informativeness (Info.) of generated headlines.

As shown in Table 5, we find that the GraMMo performs better readability and informativeness scores compared with PALM and MMPT, demonstrating its effectiveness in generating informative headlines. For readability, all the three headline generating methods can generate quite readable language. This is because a large training corpus can make text decoder generate coherent sentences, except for the mistakes of repetition phrases and grammatical errors. For informativeness, the major problems are the fact inconsistency and incomplete key information. They will be investigated in future work to improve the quality of generated headlines.

5.4 Media AI Platform

Our Chinese video headline generation is deployed in an AI platform for industrial media, which is a well-designed video understanding platform with complete video processing services. When generating headline, GraMMo takes the source video and its pre-extracted ASR text as the inputs and then predicts the textual summary as the headline of the video. We give some headline generation examples of real Chinese news videos, as shown in Figure 5.

Figure 5: The examples of video headline generation results on news videos in the media AI platform. Red text boxes illustrate the generated titles based on the video and ASR text information.

6 Conclusion

In this paper, we propose GraMMo, grafting a pre-trained sequence-to-sequence language model and a video-language understanding model for multimodal headline generation. By fine-tuning the representation components (video-encoder&text-encoder) and generation component (text-decoder) of the model, we alleviate the problem of lacking large-scale dataset in multimodal headline generation. To capture the commonality and specialty of the video and text features, we propose an extra fusion layer to balance modalities and maximally maintain the original architectures. With this approach, we can fully take advantage of the pre-trained models, including well-trained capacity for multimodal understanding and generation. Experiments results show that our method can significantly improve the performance and outperform similar works. Furthermore, the proposed method has also been applied effectively and efficiently in our online system. We will release the WB-News dataset, GraMMo code, and the grafted models.

7 Ethics Considerations

The authors declare that the use of data in our research is permitted. First, the Chinese corpus used in the text summarization auxiliary task is an open-source dataset. Second, for video-text data used in our multimodal headline generation, the data are collected and used in accordance with the privacy policies of short video platforms.

Ethical concerns include the usage of the proposed model for a purpose directly different from the previously mentioned headline generation task, such as hateful memes generation by feeding irrelevant video and text inputs, as well as integration in public opinion manipulation tools.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Bin Bi, Chenliang Li, Chen Wu, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. 2020. Palm: Pre-training an autoencoding&autoregressive language model for context-conditioned generation. *arXiv preprint arXiv:2004.07159*.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Xiyan Fu, Jun Wang, and Zhenglu Yang. 2020. Multimodal summarization for video-containing documents. *arXiv preprint arXiv:2009.08018*.

Po-Yao Huang, Mandela Patrick, Junjie Hu, Graham Neubig, Florian Metze, and Alexander Hauptmann. 2021. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. *arXiv preprint arXiv:2103.08849*.

Anubhav Jangra, Adam Jatowt, Sriparna Saha, and Mohammad Hasanuzzaman. 2021. A survey on multi-modal summarization. *arXiv preprint arXiv:2109.05199*.

Zhuoxuan Jiang, Lingfeng Qiao, Di Yin, Shanshan Feng, and Bo Ren. 2022. Leveraging key information modeling to improve less-data constrained news headline generation via duality fine-tuning. *arXiv preprint arXiv:2210.04473*.

Aman Khullar and Udit Arora. 2020. Mast: Multimodal abstractive summarization with trimodal hierarchical attention. *arXiv preprint arXiv:2010.08021*.

Bruno Korbar, Fabio Petroni, Rohit Girdhar, and Lorenzo Torresani. 2020. Video understanding as machine translation. *arXiv preprint arXiv:2006.07203*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020a. 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.

Haoran Li, Junnan Zhu, Jiajun Zhang, Xiaodong He, and Chengqing Zong. 2020b. Multimodal sentence summarization via multimodal selective encoding. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5655–5667.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. Pretrained language models for text generation: A survey. *arXiv preprint arXiv:2105.10311*.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020c. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*.

Mingzhe Li, Xiuying Chen, Shen Gao, Zhangming Chan, Dongyan Zhao, and Rui Yan. 2020d. Vmsmo: Learning to generate multimodal summary for video-based news articles. *arXiv preprint arXiv:2010.05406*.

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

Nayu Liu, Xian Sun, Hongfeng Yu, Wenkai Zhang, and Guangluan Xu. 2020. 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.

Ye Liu, Lingfeng Qiao, Di Yin, Zhuoxuan Jiang, Xinghua Jiang, Deqiang Jiang, and Bo Ren. 2022. Os-msl: One stage multimodal sequential link framework for scene segmentation and classification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6269–6277.

Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. 2020. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, Joao Henriques, and Andrea Vedaldi. 2020. Support-set bottlenecks for video-text representation learning. *arXiv preprint arXiv:2010.02824*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Lin CY ROUGE. 2004. A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization of ACL, Spain*.

Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. 2022. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17959–17968.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. 2019a. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019b. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2022. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. 2021a. Vlm: Task-agnostic video-language model pretraining for video understanding. *arXiv preprint arXiv:2105.09996*.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021b. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al. 2020. Clue: A chinese language understanding evaluation benchmark. *arXiv preprint arXiv:2004.05986*.

Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. 2021. Vision guided generative pre-trained language models for multimodal abstractive summarization. *arXiv preprint arXiv:2109.02401*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Litian Zhang, Xiaoming Zhang, Junshu Pan, and Feiran Huang. 2021a. Hierarchical cross-modality semantic correlation learning model for multimodal summarization. *arXiv preprint arXiv:2112.12072*.

Zhengkun Zhang, Xiaojun Meng, Yasheng Wang, Xin Jiang, Qun Liu, and Zhenglu Yang. 2021b. Unims: A unified framework for multimodal summarization with knowledge distillation. *arXiv preprint arXiv:2109.05812*.

Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. 2018. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8739–8748.

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*, volume 34, pages 9749–9756.

Linchao Zhu and Yi Yang. 2020. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755.

Peng Zhu, Dawei Cheng, Siqiang Luo, Ruyao Xu, Yuqi Liang, and Yifeng Luo. 2022. Leveraging enterprise knowledge graph to infer web events' influences via self-supervised learning. *Journal of Web Semantics*, page 100722.

A Experimental details

A.1 Dataset

The datasets for pre-training and fine-tuning are listed as follows.

A.1.1 Pre-training for NLG

We leverage 14GB high quality Chinese corpus of CLUE-small[1] (Xu et al., 2020). It contains following genres:

News This sub-corpus is crawled from the We Media (self-media) platform, with a total of 3 billion Chinese words from 2.5 million news articles from roughly 63K sources.

WebText With 4.1 million questions and answers, the WebText sub-corpus is crawled from Chinese Reddit-like websites such as Wukong QA, Zhihu, Sogou Wenwen, etc.

Wikipedia This sub-corpus is gathered from the Chinese content on Wikipedia (Chinese Wikipedia), containing around 1.1 GB of raw texts with 0.4 billion Chinese words on a wide range of topics.

Comments These comments are collected from E-commerce websites including Dianping.com and Amazon.com by SophonPlus[2]. This subset has approximately 2.3 GB of raw texts with 0.8 billion Chinese words.

A.1.2 Pre-training for Video-Text Matching

We collect 3.2 million videos from the Chinese video platform. The topics of video cover news, sports, entertainments, etc. For each video, its headline information is edited by uploader so that the video-text matching task can be conducted without additional manual annotations.

A.1.3 Multimodal Generation Fine-tuning

We establish the WB-News dataset for multimodal generation fine-tuning and evaluation. It contains more than 43,000 samples that are collected from official Weibo accounts of China's main media.

When building WB-News, we first filter the weibo contents which have corresponding videos. Then, the raw data is manually annotated and cleaned to produce the triplet form, i.e. video, source transcript and target summary. The average video duration of these samples is about one minute, the average length of transcript is 120.6 words and the average length of summary is 21.2 words. For testing, 697 samples are selected, which can evaluate the performance of Chinese video headline generation.

A.2 Hyper-parameters

Text Encoder We use PALM as the text pre-training model, in which 6-layer Transformers are used for both text encoder and decoder. The text tokens of samples are padded to 128 lengths.

Video Encoder For the video encoder, DFS extracts 32 video tokens from each video. Then a 2-layer Transformer encoder with 8 attention heads is applied to get video embedding.

Joint-modality Layer After obtaining the outputs encoded by text and video Transformers, linear projection layers are used to project them into the same 512 dimension. As a result, the feature dimension of $e_v \in R^{32\times512}$ and $e_t \in R^{128\times512}$. Then a video-text attention matrix M_{vt} with 8 heads is established to produce the fusion embeddings.

Decoder In the decoding stage, we use beam search with a beam size of 5. The decoding process will not stop until an end-of-sequence (EOS) token is emitted or the length of the generated summary reaches to 64 tokens.

Training Details GraMMo is realized by fairseq toolkit[3]. In training phrase, we use learning rates 3e-4 to pre-train the encoders and decoder, and 1e-5 to fine-tune the model. Batch size is set to 64 and dropout rate is 0.1. Adam optimizer is adopted with 0.01 weight-decay and 0.1 clip-norm. The training procedure runs on 8 NVIDIA V100 GPU cards and costs about 7 days for NLG pre-training, 5 days for video-text matching pre-training and 1 hour for multimodal generation fine-tuning.

[1] https://www.cluebenchmarks.com/

[2] https://github.com/SophonPlus/ChineseNlpCorpus/

[3] https://github.com/pytorch/fairseq