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). Specifically, 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. 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 grafting mechanism for obtaining the multimodal summarization pre-training model (GraMMo). First, language generation and video-text matching tasks are

![Figure 1: An example of multimodal generation.](image-url)
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 (Libovický 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
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 $X_v = \{v_1, v_2, \ldots, v_l\}$ and related source transcript $X_t = \{t_1, t_2, \ldots, t_m\}$, the output is the target textual title $Y = \{y_1, y_2, \ldots, y_n\}$, where $l, m, n$ are the numbers of the corresponding tokens. The goal is to generate a predicted title $Y' = \{y'_1, y'_2, \ldots, y'_n\}$ 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 $V_f = \{f_1, f_2, \ldots, f_s\}$. 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 of 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 $DFS(\cdot) : \{1, 2, \ldots, l\} \rightarrow \{1, 2, \ldots, s\}$, which represents the choice from variable frames.

As a result, the video tokens can be obtained by $v_i = f_{DFS(i)}$, where $i = 1, 2, \ldots, l$. Then the 3D 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 $h = e_v \odot \text{Norm}(1 - p)$, where $\text{Norm}(\cdot)$ 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 $D_t(\cdot)$ 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-
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 $\rightarrow$ Text, Video $\rightarrow$ Text and Video+Text $\rightarrow$ 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.

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

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

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.

Table 2: The effectiveness of using grafted model.

<table>
<thead>
<tr>
<th>Pre-trained Models</th>
<th>Text</th>
<th>Video</th>
<th>R-L</th>
<th>B-4</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>26.98</td>
<td>12.18</td>
<td>24.65</td>
</tr>
<tr>
<td>✓</td>
<td></td>
<td>27.63</td>
<td>12.30</td>
<td>24.69</td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>34.90</td>
<td>18.03</td>
<td>34.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>35.38</td>
<td>19.65</td>
<td>35.30</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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-
Table 3: Ablation study on model components.

<table>
<thead>
<tr>
<th>Methods</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>GraMMo (Video + Text → Text)</td>
<td>35.38</td>
<td>19.65</td>
<td>35.30</td>
</tr>
<tr>
<td>Encoder</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>w/o DFS</td>
<td>-0.28</td>
<td>-0.13</td>
<td>-0.53</td>
</tr>
<tr>
<td>w/o Video Transformer</td>
<td>-0.98</td>
<td>-0.26</td>
<td>-1.03</td>
</tr>
<tr>
<td>Fusion</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naive Concat</td>
<td>-0.89</td>
<td>-0.96</td>
<td>-1.40</td>
</tr>
<tr>
<td>Cross Attention</td>
<td>-0.85</td>
<td>-0.66</td>
<td>-1.16</td>
</tr>
</tbody>
</table>

Table 4: Video-Text Retrieval Results

<table>
<thead>
<tr>
<th>Methods</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train from Scratch</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>text-to-video</td>
<td>34.40</td>
<td>53.19</td>
<td>60.28</td>
</tr>
<tr>
<td>video-to-text</td>
<td>35.39</td>
<td>52.70</td>
<td>60.21</td>
</tr>
<tr>
<td>Grafted Model</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>text-to-video</td>
<td>39.65</td>
<td>60.43</td>
<td>67.73</td>
</tr>
<tr>
<td>video-to-text</td>
<td>40.35</td>
<td>60.35</td>
<td>67.30</td>
</tr>
</tbody>
</table>

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 headlines 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.
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.

### 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.


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 \( e_v \in \mathbb{R}^{32 \times 512} \) and \( e_t \in \mathbb{R}^{128 \times 512} \). 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