VMSMO: Learning to Generate Multimodal Summary for Video-based News Articles

Mingzhe Li[1,2,*], **Xiuying Chen**[1,2,*], **Shen Gao**[2],
Zhangming Chan[1,2], **Dongyan Zhao**[1,2] and **Rui Yan**[1,2,†]
[1] Center for Data Science, AAIS, Peking University,Beijing,China
[2] Wangxuan Institute of Computer Technology, Peking University,Beijing,China
{li_mingzhe,xy-chen,shengao,zhaody,ruiyan}@pku.edu.cn

Abstract

A popular multimedia news format nowadays is providing users with a lively video and a corresponding news article, which is employed by influential news media including CNN, BBC, and social media including Twitter and Weibo. In such a case, automatically choosing a proper cover frame of the video and generating an appropriate textual summary of the article can help editors save time, and readers make the decision more effectively. Hence, in this paper, we propose the task of Video-based Multimodal Summarization with Multimodal Output (VMSMO) to tackle such a problem. The main challenge in this task is to jointly model the temporal dependency of video with semantic meaning of article. To this end, we propose a Dual-Interaction-based Multimodal Summarizer (DIMS), consisting of a dual interaction module and multimodal generator. In the dual interaction module, we propose a conditional self-attention mechanism that captures local semantic information within video and a global-attention mechanism that handles the semantic relationship between news text and video from a high level. Extensive experiments conducted on a large-scale real-world VMSMO dataset[1] show that DIMS achieves the state-of-the-art performance in terms of both automatic metrics and human evaluations.

1 Introduction

Existing experiments (Li et al., 2017) have proven that multimodal news can significantly improve users' sense of satisfaction for informativeness. As one of these multimedia data forms, introducing news events with video and textual descriptions is becoming increasingly popular, and has been employed as the main form of news reporting by news media including BBC, Weibo, CNN, and Daily Mail. An illustration is shown in Figure 1, where the news contains a video with a cover picture and a full news article with a short textual summary. In such a case, automatically generating multimodal summaries, *i.e.,* choosing a proper cover frame of the video and generating an appropriate textual summary of the article can help editors save time and readers make decisions more effectively.

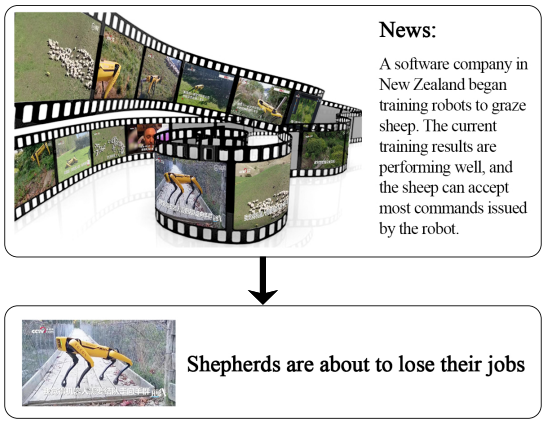

Figure 1: An example of video-based multimodal summarization with multimodal output.

There are several works focusing on multimodal summarization. The most related work to ours is (Zhu et al., 2018), where they propose the task of generating textual summary and picking the most representative picture from 6 input candidates. However, in real-world applications, the input is usually a video consisting of hundreds of frames. Consequently, the temporal dependency in a video cannot be simply modeled by static encoding methods. Hence, in this work, we propose a novel task, Video-based Multimodal Summarization with Multimodal Output (VMSMO), which selects cover frame from news video and generates textual summary of the news article in the meantime.

[*] Equal contribution. Ordering is decided by a coin flip.
[†] Corresponding author.
[1] https://github.com/yingtaomj/VMSMO

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 9360–9369,
November 16–20, 2020.

The cover image of the video should be the salient point of the whole video, while the textual summary should also extract the important information from source articles. Since the video and the article focus on the same event with the same report content, these two information formats complement each other in the summarizing process. However, how to fully explore the relationship between temporal dependency of frames in video and semantic meaning of article still remains a problem, since the video and the article come from two different space.

Hence, in this paper, we propose a model named *Dual-Interaction-based Multimodal Summarizer (DIMS)*, which learns to summarize article and video simultaneously by conducting a dual interaction strategy in the process. Specifically, we first employ Recurrent Neural Networks (RNN) to encode text and video. Note that by the encoding RNN, the spatial and temporal dependencies between images in the video are captured. Next, we design a dual interaction module to let the video and text fully interact with each other. Specifically, we propose a conditional self-attention mechanism which learns local video representation under the guidance of article, and a global-attention mechanism to learn high-level representation of video-aware article and article-aware video. Last, the multimodal generator generates the textual summary and extracts the cover image based on the fusion representation from the last step. To evaluate the performance of our model, we collect the first large-scale news article-summary dataset associated with video-cover from social media websites. Extensive experiments on this dataset show that DIMS significantly outperforms the state-of-the-art baseline methods in commonly-used metrics by a large margin.

To summarize, our contributions are threefold:

- We propose a novel Video-based Multimodal Summarization with Multimodal Output (VMSMO) task which chooses a proper cover frame for the video and generates an appropriate textual summary of the article.
- We propose a Dual-Interaction-based Multimodal Summarizer (DIMS) model, which jointly models the temporal dependency of video with semantic meaning of article, and generates textual summary with video cover simultaneously.
- We construct a large-scale dataset for VMSMO, and experimental results demonstrate that our model outperforms other baselines in terms of both automatic and human evaluations.

2 Related Work

Our research builds on previous works in three fields: text summarization, multimodal summarization, and visual question answering.

Text Summarization. Our proposed task bases on text summarization, the methods of which can be divided into extractive and abstractive methods (Gao et al., 2020b). Extractive models (Zhang et al., 2018; Narayan et al., 2018; Chen et al., 2018; Luo et al., 2019; Xiao and Carenini, 2019) directly pick sentences from article and regard the aggregate of them as the summary. In contrast, abstractive models (Sutskever et al., 2014; See et al., 2017; Wenbo et al., 2019; Gui et al., 2019; Gao et al., 2019a; Chen et al., 2019a; Gao et al., 2019b) generate a summary from scratch and the abstractive summaries are typically less redundant.

Multimodal Summarization. A series of works (Li et al., 2017, 2018; Palaskar et al., 2019; Chan et al., 2019; Chen et al., 2019b; Gao et al., 2020a) focused on generating better textual summaries with the help of multimodal input. Multimodal summarization with multimodal output is relatively less explored. Zhu et al. (2018) proposed to jointly generate textual summary and select the most relevant image from 6 candidates. Following their work, Zhu et al. (2020) added a multimodal objective function to use the loss from the textual summary generation and the image selection. However, in the real-world application, we usually need to choose the cover figure for a continuous video consisting of hundreds of frames. Consequently, the temporal dependency between frames in a video cannot be simply modeled by several static encoding methods.

Visual Question Answering. Visual Question Answering (VQA) task is similar to our task in taking images and a corresponding text as input. Most works consider VQA task as a classification problem and the understanding of image subregions or image recognition becomes particularly important (Goyal et al., 2017; Malinowski et al., 2015; Wu et al., 2016; Xiong et al., 2016). As for the interaction models, one of the state-of-the-art VQA models (Li et al., 2019) proposed a positional self-attention with a co-attention mechanism, which is faster than the recurrent neural network (RNN). Guo et al. (2019) devised an image-

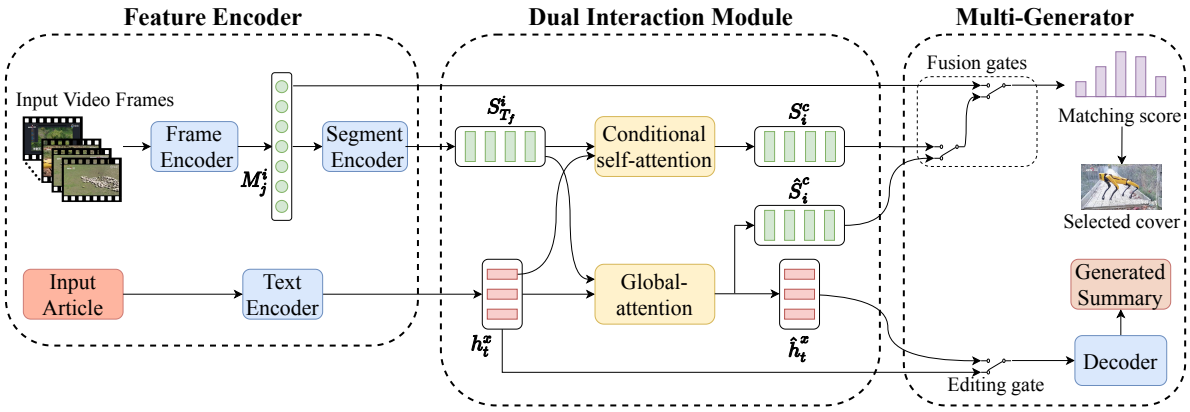

Figure 2: Overview of DIMS. We divide our model into three parts: (1) *Feature Encoder* encodes the input article and video separately; (2) *Dual Interaction Module* learns fused representation of video and article from different level; (3) *Multi-Generator* generates the textual summary and chooses the video cover simultaneously.

question-answer synergistic network, where candidate answers are coarsely scored according to their relevance to the image and question pair and answers with a high probability of being correct are re-ranked by synergizing with image and question.

3 Problem Formulation

Before presenting our approach for the VMSMO, we first introduce the notations and key concepts. For an input news article $X = \{x_1, x_2, \ldots, x_{T_d}\}$ which has T_d words, we assume there is a ground truth textual summary $Y = \{y_1, y_2, \ldots, y_{T_y}\}$ which has T_y words. Meanwhile, there is a news video V corresponding to the article, and we assume there is a ground truth cover picture C that extracts the most important frame from the video content. For a given article X and the corresponding video V, our model emphasizes salient parts of both inputs by conducting deep interaction. The goal is to generate a textual summary $Y^{'}$ that successfully grasp the main points of the article and choose a frame picture $C^{'}$ that covers the gist of the video.

4 Model

4.1 Overview

In this section, we propose our Dual Interaction-based Multimodal Summarizer (DIMS), which can be divided into three parts in Figure 2:

• *Feature Encoder* is composed of a text encoder and a video encoder which encodes the input article and video separately.

• *Dual Interaction Module* conducts deep interaction, including conditional self-attention and global-attention mechanism between video segment and article to learn different levels of representation of the two inputs.

• *Multi-Generator* generates the textual summary and chooses the video cover by incorporating the fused information.

4.2 Feature Encoder

4.2.1 Text encoder

To model the semantic meaning of the input news text $X = \{x_1, x_2, \ldots, x_{T_d}\}$, we first use a word embedding matrix e to map a one-hot representation of each word x_i into to a high-dimensional vector space. Then, in order to encode contextual information from these embedding representation, we use bi-directional recurrent neural networks (Bi-RNN) (Hochreiter and Schmidhuber, 1997) to model the temporal interactions between words:

$$h_t^x = \text{Bi-RNN}_{\text{X}}(e(x_t), h_{t-1}^x), \quad (1)$$

where h_t^x denotes the hidden state of t-th step in Bi-RNN for X. Following (See et al., 2017; Ma et al., 2018), we choose the long short-term memory (LSTM) as the Bi-RNN cell.

4.2.2 Video Encoder

A news video usually lasts several minutes and consists of hundreds of frames. Intuitively, a video can be divided into several segments, each of which corresponds to different content. Hence, we choose to encode video hierarchically. More specifically, we equally divide frames in the video into several segments and employ a low-level frame encoder and a high-level segment encoder to learn hierarchical representation.

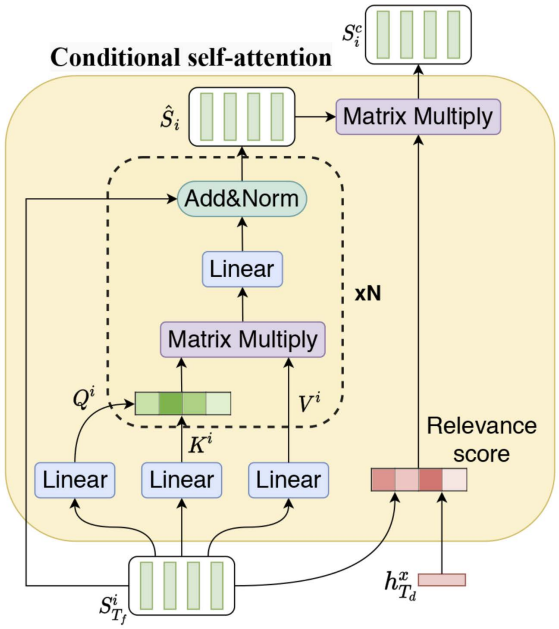

Figure 3: Conditional self-attention module, which captures local semantic information within video segments under the guidance of article representation.

Frame encoder. We utilize the Resnet-v1 model (He et al., 2016) to encode frames to alleviate gradient vanishing (He et al., 2016) and reduce computational costs:

$$O_j^i = \text{Resnet-v1}(m_j^i), \tag{2}$$

$$M_j^i = \text{relu}\left(F_v(O_j^i)\right), \tag{3}$$

where m_j^i is the j-th frame in i-th segment and $F_v(\cdot)$ is a linear transformation function.

Segment encoder. As mentioned before, it is important to model the continuity of images in video, which cannot be captured by a static encoding strategy. We employ RNN network as segment encoder due to its superiority in exploiting the temporal dependency among frames Zhao et al. (2017):

$$S_j^i = \text{Bi-RNN}_S(M_j^i, S_{j-1}^i). \tag{4}$$

S_j^i denotes the hidden state of j-th step in Bi-RNN for segment s_i, and the final hidden state $S_{T_f}^i$ denotes the overall representation of the segment s_i, where T_f is the number of frames in a segment.

4.3 Dual Interaction Module

The cover image of the video should contain the key point of the whole video, while the textural summary should also cover extract the important information from source articles. Hence, these two information formats complement each other in the summarizing process. In this section, we conduct a deep interaction between the video and article to jointly model the temporal dependency of video and semantic meaning of text. The module consists of a conditional self-attention mechanism that captures local semantic information within video segments and a global-attention mechanism that handles the semantic relationship between news text and video from a high level.

Conditional self-attention mechanism. Traditional self-attention can be used to obtain contextual video representation due to its flexibility in relating two elements in a distance-agnostic manner. However, as illustrated in Xie et al. (2020), the semantic understanding often relies on more complicated dependencies than the pairwise one, especially conditional dependency upon a given premise. Hence, in the VMSMO task, we capture the local semantic information of video conditioned on the input text information.

Our conditional self-attention module shown in Figure 3 is composed of a stack of N identical layers and a conditional layer. The identical layer learns to encode local video segments while the conditional layer learns to assign high weights to the video segments conditioned on their relationship to the article. We first use a fully-connected layer to project each segment representation $S_{T_f}^i$ into the query Q^i, key K^i, and value V^i. Then, the scaled dot-product self-attention is defined as:

$$\alpha_{i,j} = \frac{\exp\left(Q^i K^j\right)}{\sum_{n=1}^{T_s} \exp\left(Q^i K^n\right)}, \tag{5}$$

$$\hat{S}_i = \sum_{j=1}^{T_s} \frac{\alpha_{i,j} V^j}{\sqrt{d}}, \tag{6}$$

where d stands for hidden dimension and T_s is the segment number in a video. $\hat{S}_i$ is then fed into the feed-forward sub-layer including a residual connection (He et al., 2016) and layer normalization (Ba et al., 2016).

Next, we highlight the salient part of the video under the guidance of article. Taking the article information $h_{T_d}^x$ as condition, the attention score on each original segment representation $S_{T_f}^i$ is calculated as:

$$\beta_i = \sigma\left(F_s(S_{T_f}^i h_{T_d}^x)\right). \tag{7}$$

The final conditional segment representation S_i^c is denoted as $\beta_i \hat{S}_i$.

Global-attention mechanism. The global-attention module grounds the article representation

on the video segments and fuses the information of the article into the video, which results in an article-aware video representation and a video-aware article representation. Formally, we utilize a two-way attention mechanism to obtain the co-attention between the encoded text representation h_t^x and the encoded segment representation $S_{T_f}^i$:

$$E_i^t = F_h(h_t^x)\left(F_t(S_{T_f}^i)\right)^T. \quad (8)$$

We use E_i^t to denote the attention weight on the t-th word by the i-th video segment. To learn the alignments between text and segment information, the global representations of video-aware article $\hat{h}_t^x$ and article-aware video $\hat{S}_i^c$ are computed as:

$$\hat{h}_t^x = \sum_{i=1}^{T_d} E_i^t S_{T_f}^i, \quad (9)$$

$$\hat{S}_i^c = \sum_{t=1}^{T_s} \left(E_i^t\right)^T h_t^x. \quad (10)$$

4.4 Multi-Generator

In the VMSMO task, the multi-generator module not only needs to generate the textual summary but also needs to choose the video cover.

Textual summary generation. For the first task, we use the final state of the input text representation $h_{T_d}^x$ as the initial state d_0 of the RNN decoder, and the t-th generation procedure is:

$$d_t = \text{LSTM}_{\text{dec}}(d_{t-1}, [e(y_{t-1}); h_{t-1}^c]), \quad (11)$$

where d_t is the hidden state of the t-th decoding step and h_{t-1}^c is the context vector calculated by the standard attention mechanism (Bahdanau et al., 2014), and is introduced below.

To take advantage of the article representation h_t^x and the video-aware article representation $\hat{h}_t^x$, we apply an "editing gate" γ_e to decide how much information of each side should be focused on:

$$\gamma_e = \sigma\left(F_d(d_t)\right), \quad (12)$$

$$g_i = \gamma_e h_i^x + (1-\gamma_e)\hat{h}_i^x. \quad (13)$$

Then the context vector h_{t-1}^c is calculated as:

$$\delta_{it} = \frac{\exp(F_a(g_i, d_t))}{\sum_j \exp(F_a(g_j, d_t))}. \quad (14)$$

$$h_t^c = \sum_i \delta_{it} g_i, \quad (15)$$

Finally, the context vector h_t^c is concatenated with the decoder state d_t and fed into a linear layer to obtain the generated word distribution P_v:

$$d_t^o = \sigma\left(F_p([d_t; h_t^c])\right), \quad (16)$$

$$P_v = \text{softmax}\left(F_o(d_t^o)\right). \quad (17)$$

Following See et al. (2017), we also equip our model with pointer network to handle the out-of-vocabulary problem. The loss of textual summary generation is the negative log likelihood of the target word y_t:

$$\mathcal{L}_{seq} = -\sum_{t=1}^{T_y} \log P_v(y_t). \quad (18)$$

Cover frame selector. The cover frame is chosen based on hierarchical video representations, *i.e.,* the original frame representation M_j^i and the conditional segment representation S_i^c with the article-aware segment representation $\hat{S}_i^c$:

$$p_j^i = \gamma_f^1 S_i^c + \gamma_f^2 \hat{S}_i^c + (1-\gamma_f^1-\gamma_f^2) M_j^i, \quad (19)$$

$$y_{i,j}^c = \sigma\left(F_c(p_j^i)\right), \quad (20)$$

where $y_{i,j}^c$ is the matching score of the candidate frames. The fusion gates γ_f^1 and γ_f^2 here are determined by the last text encoder hidden state $h_{T_d}^x$:

$$\gamma_f^1 = \sigma\left(F_m(h_{T_d}^x)\right), \quad (21)$$

$$\gamma_f^2 = \sigma\left(F_n(h_{T_d}^x)\right). \quad (22)$$

We use pairwise hinge loss to measure the selection accuracy:

$$\mathcal{L}_{pic} = \sum^{N} \max\left(0, y_{\text{negative}}^c - y_{\text{positive}}^c + \text{margin}\right), \quad (23)$$

where $y_{negative}^c$ and $y_{positive}^c$ corresponds to the matching score of the negative samples and the ground truth frame, respectively. The margin in the $\mathcal{L}_{pic}$ is the rescale margin in hinge loss.

The overall loss for the model is:

$$\mathcal{L} = \mathcal{L}_{seq} + \mathcal{L}_{pic}. \quad (24)$$

5 Experimental Setup

5.1 Dataset

To our best knowledge, there is no existing large-scale dataset for VMSMO task. Hence, we collect the first large-scale dataset for VMSMO task from Weibo, the largest social network website in China. Most of China's mainstream media have Weibo accounts, and they publish the latest news in their

accounts with lively videos and articles. Correspondingly, each sample of our data contains an article with a textual summary and a video with a cover picture. The average video duration is one minute and the frame rate of video is 25 fps. For the text part, the average length of article is 96.84 words and the average length of textual summary is 11.19 words. Overall, there are 184,920 samples in the dataset, which is split into a training set of 180,000 samples, a validation set of 2,460 samples, and a test set of 2,460 samples.

5.2 Comparisons

We compare our proposed method against summarization baselines and VQA baselines.

Traditional Textual Summarization baselines:

Lead: selects the first sentence of article as the textual summary (Nallapati et al., 2017).

TexkRank: a graph-based extractive summarizer which adds sentences as nodes and uses edges to weight similarity (Mihalcea and Tarau, 2004).

PG: a sequence-to-sequence framework combined with attention mechanism and pointer network (See et al., 2017).

Unified: a model which combines the strength of extractive and abstractive summarization (Hsu et al., 2018).

GPG: Shen et al. (2019) proposed to generate textual summary by "editing" pointed tokens instead of hard copying.

Multimodal baselines:

How2: a model proposed to generate textual summary with video information (Palaskar et al., 2019).

Synergistic: a image-question-answer synergistic network to value the role of the answer for precise visual dialog(Guo et al., 2019).

PSAC: a model adding the positional self-attention with co-attention on VQA task (Li et al., 2019).

MSMO: the first model on multi-output task, which paid attention to text and images during generating textual summary and used coverage to help select picture (Zhu et al., 2018).

MOF: the model based on MSMO which added consideration of image accuracy as another loss (Zhu et al., 2020).

5.3 Evaluation Metrics

The quality of generated textual summary is evaluated by standard full-length Rouge F1 (Lin, 2004) following previous works (See et al., 2017; Chen et al., 2018). R-1, R-2, and R-L refer to unigram, bigrams, and the longest common subsequence respectively. The quality of chosen cover frame is evaluated by mean average precision (MAP) (Zhou et al., 2018) and recall at position ($R_n@k$) (Tao et al., 2019). $R_n@k$ measures if the positive sample is ranked in the top k positions of n candidates.

	R-1	R-2	R-L
extractive summarization			
Lead	16.2	5.3	13.9
TextRank	13.7	4.0	12.5
abstractive summarization			
PG (See et al., 2017)	19.4	6.8	17.4
Unified (Hsu et al., 2018)	23.0	6.0	20.9
GPG (Shen et al., 2019)	20.1	4.5	17.3
our models			
DIMS	**25.1**	**9.6**	**23.2**

Table 1: Rouge scores comparison with traditional textual summarization baselines.

5.4 Implementation Details

We implement our experiments in Tensorflow (Abadi et al., 2016) on an NVIDIA GTX 1080 Ti GPU. The code for our model is available online[2]. For all models, we set the word embedding dimension and the hidden dimension to 128. The encoding step is set to 100, while the minimum decoding step is 10 and the maximum step is 30. For video preprocessing, we extract one of every 120 frames to obtain 10 frames as cover candidates. All candidates are resized to 128x64. We regard the frame that has the maximum cosine similarity with the ground truth cover as the positive sample, and others as negative samples. Note that the average cosine similarity of positive samples is 0.90, which is a high score, demonstrating the high quality of the constructed candidates. In the conditional self-attention mechanism, the stacked layer number is set to 2. For hierarchical encoding, each segment contains 5 frames. Experiments are performed with a batch size of 16. All the parameters in our model are initialized by Gaussian distribution. During training, we use Adagrad optimizer as our optimizing algorithm and we also apply gradient clipping with a range of $[-2, 2]$. The vocabulary size is limited to 50k. For testing, we use beam search with beam size 4 and we decode until an end-of-sequence token is reached. We select the 5 best checkpoints based on performance on the validation set and report averaged results on the test set.

[2]https://github.com/yingtaomj/VMSMO

	R-1	R-2	R-L	MAP	R_{10}@1	R_{10}@2	R_{10}@5
video-based summarization							
How2 (Palaskar et al., 2019)	21.7	6.1	19.0	-	-	-	-
Visual Q&A methods							
Synergistic (Guo et al., 2019)	-	-	-	0.588	0.444	0.557	0.759
PSAC (Li et al., 2019)	-	-	-	0.524	0.363	0.481	0.730
multimodal summarization with multimodal output							
MSMO (Zhu et al., 2018)	20.1	4.6	17.3	0.554	0.361	0.551	0.820
MOF (Zhu et al., 2020)	21.3	5.7	17.9	0.615	0.455	0.615	0.817
our models							
DIMS	**25.1**	**9.6**	**23.2**	**0.654**	**0.524**	**0.634**	**0.824**
DIMS-textual summary	22.0	6.3	19.2	-	-	-	-
DIMS-cover frame	-	-	-	0.611	0.449	0.610	0.823
ablation study							
DIMS-G	23.7	7.4	21.7	0.624	0.471	0.619	0.819
DIMS-S	24.4	8.9	22.5	0.404	0.204	0.364	0.634

Table 2: Rouge and Accuracy scores comparison with multimodal baselines.

6 Experimental Result

6.1 Overall Performance

We first examine *whether our DIMS outperforms other baselines* as listed in Table 1 and Table 2. Firstly, abstractive models outperform all extractive methods, demonstrating that our proposed dataset is suitable for abstractive summarization. Secondly, the video-enhanced models outperform traditional textural summarization models, indicating that video information helps generate summary. Finally, our model outperforms `MOF` by 17.8%, 68.4%, 29.6%, in terms of Rouge-1, Rouge-2, Rouge-L, and 6.3%, 15.2% in MAP and R@1 respectively, which proves the superiority of our model. All our Rouge scores have a 95% confidence interval of at most $\pm$0.55 as reported by the official Rouge script.

In addition to automatic evaluation, system performance was also evaluated on the generated textual summary by human judgments on 70 randomly selected cases similar to Liu and Lapata (2019). Our first evaluation study quantified the degree to which summarization models retain key information from the articles following a question-answering (QA) paradigm (Narayan et al., 2018). A set of questions was created based on the gold summary. Then we examined whether participants were able to answer these questions by reading system summaries alone. We created 183 questions in total varying from two to three questions per gold summary. Correct answers were marked with 1 and 0 otherwise. The average of all question scores is set to the system score.

	QA(%)	Rating
How2	46.2	-0.24
MOF	51.3	-0.14
Unified	53.8	0.00
DIMS	**66.7**	**0.38**

Table 3: System scores based on questions answered by human and summary quality rating.

Our second evaluation estimated the overall quality of the textual summaries by asking participants to rank them according to its *Informativeness* (does the summary convey important contents about the topic in question?), *Coherence* (is the summary fluent and grammatical?), and *Succinctness* (does the summary avoid repetition?). Participants were presented with the gold summary and summaries generated from several systems better on autometrics and were asked to decide which was the best and the worst. The rating of each system was calculated as the percentage of times it was chosen as best minus the times it was selected as worst, ranging from -1 (worst) to 1 (best).

Both evaluations were conducted by three highly educated native-speaker annotators. Participants evaluated summaries produced by `Unified`, `How2`, `MOF` and our `DIMS`, all of which achieved high perfromance in automatic evaluations. As shown in Table 3, on both evaluations, participants overwhelmingly prefer our model. All pairwise comparisons among systems are statistically significant using the paired student t-test for significance at α = 0.01.

Figure 4: Visualizations of global-attention matrix between the news article and two frames in the same video.

6.2 Ablation Study

Next, we conduct ablation tests to assess the importance of the conditional self-attention mechanism (-S), as well as the global-attention (-G) in Table 2. All ablation models perform worse than DIMS in terms of all metrics, which demonstrates the preeminence of DIMS. Specifically, the global-attention module contributes mostly to the textual summary generation, while the conditional self-attention module is more important for choosing cover frame.

6.3 Analysis of Multi-task learning

Our model aims to generate textural summary and choose cover frame at the same time, which can be regarded as a multi-task. Hence, in this section, we examine whether these two tasks can complement each other. We separate our model into two single-task architecture, named as DIMS-textual summary and DIMS-cover frame, which generates textural summary and chooses video cover frame, respectively. The result is shown in Table 2. It can be seen that the multi-task DIMS outperforms single-task DIMS-textual summary and DIMS-cover frame, improving the performance of summarization by 20.8% in terms of ROUGE-L score, and increasing the accuracy of cover selection by 7.0% on MAP.

6.4 Visualization of dual interaction module

To study the multimodal interaction module, we visualize the global-attention matrix E_i^t in Equation 8 on one randomly sampled case, as shown in Figure 4. In this case, we show the attention on article words of two representative images in the video. The darker the color is, the higher the attention weight is. It can be seen that for the left figure, the word *hand in hand* has a higher weight than *picture*, while for the right figure, the word *Book Fair* has the highest weight. This corresponds to the fact that the main body of the left frame is two old men, and the right frame is about reading books.

Article: On August 26, in Shanxi Ankang, a 12-year-old junior girl Yu Taoxin goose-stepped like parade during the military training in the new semester, and won thousands of praises. Yu Taoxin said that her father was a veteran, and she worked hard in military training because of the influence of her father. Her father told her that military training should be strict as in the army. 8月26日，陕西安康，12岁的初一女生余陶鑫，在新学期军训期间，她踢出阅兵式般的标准步伐，获千万点赞。余陶鑫说，爸爸是名退伍军人，军训刻苦是因为受到爸爸影响，爸爸曾告诉她，军训时就应和在部队里一样，严格要求自己。
Reference summary: A 12-year-old girl goose-stepped like parade during the military training, "My father is a veteran." 12岁女孩军训走出阅兵式步伐，“爸爸是退伍军人”
QA: *What happened on the 12-year-old girl?* [*She goose-stepped like parade.*] 这个12岁女孩做了什么？[她走出阅兵式步伐。] *Why did she do this?* [*She was influenced by her father*] 她为什么这样做？[她受到爸爸的影响。]
Unfied: 12-year-old gril Yu Taoxin goose-stepped during military training. 12岁女生余陶鑫军训期间阅兵式般的标准步伐
How2: 12-year-old girls were organized military training, and veteran mother parade. 12岁女生组团军训，退伍军人妈妈阅兵式
MOF: A 12-year-old junior citizen [unk]: father gave a kicked like. 1名12岁初一市民[unk]：爸爸踢式点赞
DIMS: A 12-year-old junior girl goose-stepped like parade: My father is a veteran, and military training should be strict as in the army. 12岁初一女生踢出阅兵式：爸爸是名退伍军人，军训时就应和在部队一样
Ground Truth MOF DIMS 

Table 4: Examples of the generated summary by baselines and DIMS.

We show a case study in Table 4, which includes the input article and the generated summary by different models. We also show the question-answering pair in human evaluation and the chosen cover. The result shows that the summary generated by our model is both fluent and accurate, and the cover frame chosen is also similar to the ground truth frame.

7 Conclusion

In this paper, we propose the task of Video-based Multimodal Summarization with Multimodal Output (VMSMO) which chooses a proper video cover and generates an appropriate textual summary for a video-attached article. We propose a model named Dual-Interaction-based Multimodal Summarizer (DIMS) including a local conditional self-attention mechanism and a global-attention mechanism to jointly model and summarize multimodal input. Our model achieves state-of-the-art results in terms of autometrics and outperforms human evaluations by a large margin. In near future, we aim to incorporate the video script information in the multimodal summarization process.

Acknowledgments

We would like to thank the anonymous reviewers for their constructive comments. This work was supported by the National Key Research and Development Program of China (No.2020AAA0105200), and the National Science Foundation of China (NSFC No.61876196, No.61672058). Rui Yan is partially supported as a Young Fellow of Beijing Institute of Artificial Intelligence (BAAI).

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *ICLR*.

Zhangming Chan, Xiuying Chen, Yongliang Wang, Juntao Li, Zhiqiang Zhang, Kun Gai, Dongyan Zhao, and Rui Yan. 2019. Stick to the facts: Learning towards a fidelity-oriented e-commerce product description generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4960–4969.

Xiuying Chen, Zhangming Chan, Shen Gao, Meng-Hsuan Yu, Dongyan Zhao, and Rui Yan. 2019a. Learning towards abstractive timeline summarization. In *IJCAI*, pages 4939–4945.

Xiuying Chen, Shen Gao, Chongyang Tao, Yan Song, Dongyan Zhao, and Rui Yan. 2018. Iterative document representation learning towards summarization with polishing. In *EMNLP*, pages 4088–4097.

Xiuying Chen, Daorui Xiao, Shen Gao, Guojun Liu, Wei Lin, Bo Zheng, Dongyan Zhao, and Rui Yan. 2019b. Rpm-oriented query rewriting framework for e-commerce keyword-based sponsored search. *arXiv preprint arXiv:1910.12527*.

Shen Gao, Xiuying Chen, Piji Li, Zhangming Chan, Dongyan Zhao, and Rui Yan. 2019a. How to write summaries with patterns? learning towards abstractive summarization through prototype editing. *arXiv preprint arXiv:1909.08837*.

Shen Gao, Xiuying Chen, Piji Li, Zhaochun Ren, Lidong Bing, Dongyan Zhao, and Rui Yan. 2019b. Abstractive text summarization by incorporating reader comments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6399–6406.

Shen Gao, Xiuying Chen, Chang Liu, Li Liu, Dongyan Zhao, and Rui Yan. 2020a. Learning to respond with stickers: A framework of unifying multi-modality in multi-turn dialog. In *Proceedings of The Web Conference 2020*, pages 1138–1148.

Shen Gao, Xiuying Chen, Zhaochun Ren, Dongyan Zhao, and Rui Yan. 2020b. From standard summarization to new tasks and beyond: Summarization with manifold information. In *IJCAI*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. 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.

Min Gui, Junfeng Tian, Rui Wang, and Zhenglu Yang. 2019. Attention optimization for abstractive document summarization. In *EMNLP/IJCNLP*.

Dalu Guo, Chang Xu, and Dacheng Tao. 2019. Image-question-answer synergistic network for visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10434–10443.

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

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735–1780.

Wan-Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun. 2018. A unified model for extractive and abstractive summarization using inconsistency loss. *ACL*.

Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, Chengqing Zong, et al. 2018. Multi-modal sentence summarization with modality attention and image filtering. *IJCAI-PRICAI*.

Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, Chengqing Zong, et al. 2017. Multi-modal summarization for asynchronous collection of text, image, audio and video. *EMNLP/IJCNLP*.

Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. 2019. Beyond rnns: Positional self-attention with co-attention for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8658–8665.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. *ACL*.

Ling Luo, Xiang Ao, Y. S. Song, Feiyang Pan, Min Yang, and Qing He. 2019. Reading like her: Human reading inspired extractive summarization. In *EMNLP/IJCNLP*.

Shuming Ma, Xu Sun, Junyang Lin, and Xuancheng Ren. 2018. A hierarchical end-to-end model for jointly improving text summarization and sentiment classification. *IJCAI-PRICAI*.

Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. 2015. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE international conference on computer vision*, pages 1–9.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *EMNLP*.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. In *NAACL-HLT*.

Shruti Palaskar, Jindrich Libovickỳ, Spandana Gella, and Florian Metze. 2019. Multimodal abstractive summarization for how2 videos. *ACL*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *ACL*.

Xiaoyu Shen, Yang Zhao, Hui Su, and Dietrich Klakow. 2019. Improving latent alignment in text summarization by generalizing the pointer generator. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3753–3764.

I Sutskever, O Vinyals, and QV Le. 2014. Sequence to sequence learning with neural networks. *Advances in NIPS*.

Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019. Multi-representation fusion network for multi-turn response selection in retrieval-based chatbots. *WSDM*, pages 267–275.

Wang Wenbo, Yang Gao, Huang Heyan, and Yuxiang Zhou. 2019. Concept pointer network for abstractive summarization. In *EMNLP/IJCNLP*.

Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2016. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4622–4630.

Wen Xiao and Giuseppe Carenini. 2019. Extractive summarization of long documents by combining global and local context. In *EMNLP/IJCNLP*.

Yujia Xie, Tianyi Zhou, Yi Mao, and Weizhu Chen. 2020. Conditional self-attention for query-based summarization. *arXiv preprint arXiv:2002.07338*.

Caiming Xiong, Stephen Merity, and Richard Socher. 2016. Dynamic memory networks for visual and textual question answering. In *International conference on machine learning*, pages 2397–2406.

Xingxing Zhang, Mirella Lapata, Furu Wei, and Ming Zhou. 2018. Neural latent extractive document summarization. *EMNLP/IJCNLP*.

Bin Zhao, Xuelong Li, and Xiaoqiang Lu. 2017. Hierarchical recurrent neural network for video summarization. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 863–871.

Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-turn response selection for chatbots with deep attention matching network. *ACL*, 1:1118–1127.

Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, Chengqing Zong, et al. 2018. Msmo: multimodal summarization with multimodal output. *EMNLP/IJCNLP*.

Junnan Zhu, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong, and Changliang Li. 2020. Multimodal summarization with guidance of multimodal reference. *Image*, 3:4–64.