Neurocomputing 456 (2021) 179–189

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

D-MmT: A concise decoder-only multi-modal transformer for abstractive summarization in videos

Nayu Liu [a,b,c,d], Xian Sun [a,b,c,d,*], Hongfeng Yu [a,b], Wenkai Zhang [a,b], Guangluan Xu [a,b]

[a] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China
[b] Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China
[c] University of Chinese Academy of Sciences, Beijing 100190, China
[d] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China

ARTICLE INFO

Article history:
Received 27 August 2020
Revised 19 January 2021
Accepted 20 April 2021
Available online 24 April 2021
Communicated by Zidong Wang

2010 MSC:
00–01
99–00

Keywords:
Abstractive summarization
Multi-modal
Decoder-only
Less model parameters

ABSTRACT

Multi-modal abstractive summarization for videos is an emerging task, aiming to integrate multi-modal and multi-source inputs (video, audio transcript) into a compressed textual summary. Although recent multi-encoder-decoder models on this task have shown promising performance, they did not explicitly model interactions of multi-source inputs. While some strategies like co-attention are utilized for modeling this interaction, considering ultra-long sequences and additional decoder in this task, the coupling of multi-modal data from multi-encoders and decoder needs complicated structure and additional parameters. In this paper, we propose a concise **D**ecoder-only **M**ulti-**m**odal **T**ransformer (D-MmT) based on the above observations. Specifically, we cut the encoder structure, and introduce an in-out shared multi-modal decoder to make the multi-source and target fully interact and couple in the shared feature space, reducing the model parameter redundancy. Also, we design a concise cascaded cross-modal interaction (CXMI) module in the multi-modal decoder that generates joint fusion representations and spontaneously establishes a fine-grained intra- and inter- association between multi-modalities. In addition, to make full use of the ultra-long sequence information, we introduce a joint in-out loss to make the input transcript also participate in backpropagation to enhance the contextual feature representation. The experimental results on the How2 dataset show that the proposed model outperforms the current state-of-the-art approach with fewer model parameters. Further analysis and visualization show the effectiveness of our proposed framework.

1. Introduction

Summarization, which aims to generate a text summary of the original text documents' significant points, is a central problem in natural language processing. Recent years have seen remarkable success in using deep sequence-to-sequence (S2S) neural networks [1] for abstractive text summarization [2–8].

With the increasing multimedia information that emerged on the Internet, multi-modal summarization for videos [9], aiming to integrate the multi-source information (video, audio transcript) of the video into a compressed summary, has gradually attracted interest. An example is shown in Fig. 1. This research uses text descriptions that embody the salient part of the video to replace the previous keyword-based and fuzzy video features, which is of great significance for users to retrieve and recommend videos.

In this field, Sanabria et al. [9] first released the How2 Dataset for multi-modal abstractive summarization for open-domain videos, which also appeared as a track in the ICML 2019 workshop. Afterward, a few related works [10,11] leverage multiple encoders to separately obtain the video and audio transcript features, and a joint decoder to fuse the multi-source encodings to generate a summary, which outperform single-modality models.

Although the existing works have obtained promising results, they treat the video and transcript encoders as two individual modules and did not explicitly model the interactions between multi-source inputs, as shown in Fig. 2(a), while we argue that multi-source data should establish interactions to obtain a thorough multi-modal representation, as shown in Fig. 2(b).

For modeling the interaction of multi-source data, some multi-modal tasks like visual question answering (VQA) [12,13] have pro-

* Corresponding author at: Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China.
E-mail address: sunxian@mail.ie.ac.cn (X. Sun).

https://doi.org/10.1016/j.neucom.2021.04.072
0925-2312/

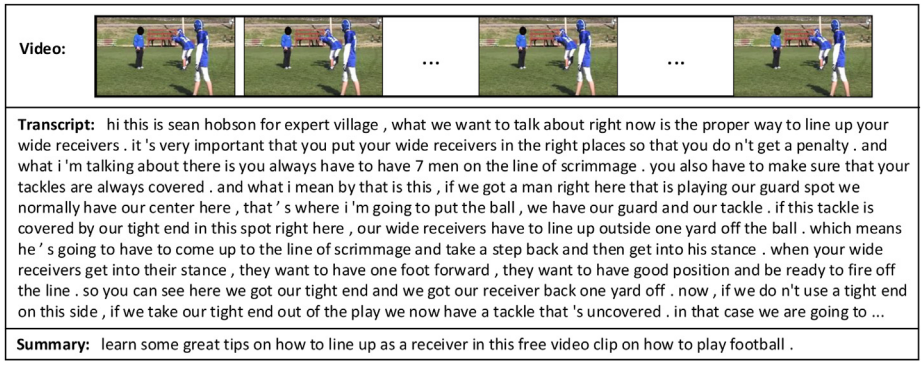

Fig. 1. Given a video and its audio transcription, the goal is to output a abstractive text summary that outlining the video.

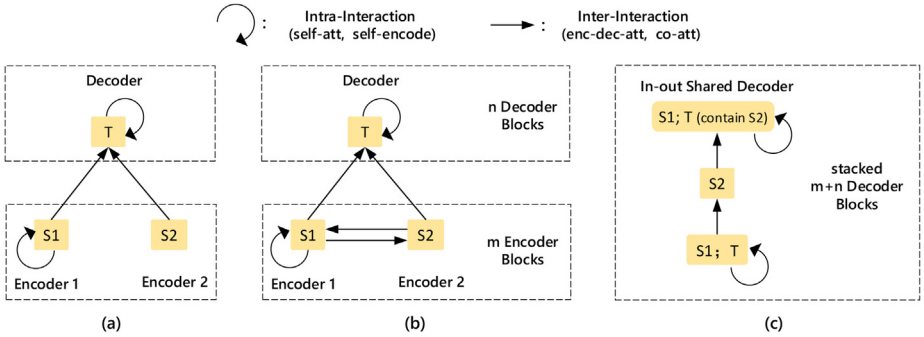

Fig. 2. (a) Previous work uses independent encoders and a joint decoder to generate summaries. (b) Some other tasks propose the collaborative attention of encoders to establish inter-relations. (c) The proposed D-MmT uses an in-out shared decoder to fully interact to learn multi-modal representations in the shared feature space. Owing to the decoder-only architecture, the model can be stacked deeper to gain sufficient feature representations. Note that following the previous work, we also adopt the video features extracted from the pre-trained encoder without assigning its self-encoding parameters. S1: Transcript; S2: Video; T: Summary.

posed co-attention or similar strategies [14–17] to model multi-source interactions between multiple encoders. Considering ultra-long inputs and additional decoder rather than a classifier in this task, although the use of co-attention and target-to-multisource attention establishes a connection between the multi-encoders and the decoder, the feature space of encoders and decoder receiving multi-modality is not fully shared, and the modeling of interactions among multiple encoders and decoder needs complicated structures. This makes the model need to spend more parameters to couple multi-modal data from multiple encoders and decoder to learn adequate multi-modal representations.

Based on above observations, we propose a **D**ecoder-only **M**ulti-**m**odal **T**ransformer (D-MmT) to consider the fine-grained interaction between multi-modalities, utilize the ultra-long input information, and reduce redundant parameters. The core idea is shown in Fig. 2(c). Different from the conventional source input learning in encoder and target output learning in decoder, the D-MmT completely discards the encoders, including encoder self-interaction and multi-encoder-decoder attention. An in-out shared multi-modal decoder is proposed to adequately couple multi-modal data and jointly learn the representation of source and target in the shared feature space, including only decoder masked self-interaction. Concretely, since the source transcript and the target summary have the same language feature distribution, we concatenate the two to a flat sequence to feed into a shared decoder to share parameters. Because the model does not have an encoder, the decoder can also be stacked deeper to allow each modality to learn more adequate feature representation. Since our model cuts the parameters of encoder-decoder and multi-encoder interactions, under the same stacking depth, our model still has fewer parameters.

Also, a cascaded cross-modal interaction (CXMI) module inside the multi-modal decoder is designed to recieve the connective transcript-summary sequence and video sequence in the multi-modal decoder block, capturing the fine-grained intra- and inter-correlations between multi-source input and output and generating multi-modal fusion embeddings. Concretely, a cascaded intra-modal self-attention and inter-modal guided-attention layer are introduced to progressively cover the interactions among the multiple sources and target data, and a simple feature-level fusion layer is applied to control the fusion of multi-modal contextual features at low-level granularity.

As the connective input transcript and output summary in shared decoder possess the same language feature distribution, we introduce a joint in-out loss to make full use of ultra-long input information, making the additional transcript sequence in decoder also participate in backpropagation with target summary. During training, it is like: the output sequence that the model predicts is the connective transcript-summary sequence, instead of the conventional output summary sequence. Additional long sequence information as the supervised information in decoder assists the language feature learning to gain better contextual feature representations and generate a more coherent language summary output.

We conduct experiments on the large scale multi-modal summarization dataset, namely How2 [9]. Experimental results show that our proposed D-MmT outperforms all the compared baselines while the model parameters are reduced. We further analyze the effectiveness of our proposed modules, explore the effect of the CXMI position on the model, and get some interesting findings. Besides, further visualization analysis qualitatively validates that our proposed framework could capture the video keyframes, which

can be used as a supplement to the text summary to improve users' reading satisfaction.

We summarize the main contributions as follows:

- We introduce a in-out shared multi-modal decoder-only framework to make the multi-source input and target output fully interact and couple in the shared feature space, which reduces model parameter redundancy.
- We design a CXMI module in the multi-modal decoder that generates joint fusion representations and spontaneously establish intra- and inter- correlation between multi-modalities.
- We introduce a joint in-out loss to make the additional input transcript participate in backpropagation with target summary to utilize long input information to enhance the contextual feature representations.
- Experimental results show the proposed approach outperforms competitive baselines with fewer parameters under the same model configuration.

2. Related work

Unlike conventional text summarization, multi-modal summarization requires multi-modal data (video, image, audio, text) to generate a summary. Related works are mainly divided into extractive approaches and abstractive approaches, and our framework belongs to the latter. Some works that use decoder-only architecture are also introduced in this section.

2.1. Multi-modal extractive summarization

Multi-modal extractive summarization (MES) directly selects essential information from the multimedia document itself as the output summary, such as crucial sentences, key images. It does not generate the content not contained in the original input. Researchers choose different modalities based on their summarization tasks. Some works [18–21] mainly focus on summarizing texts and images, and some researches [22–24] utilize a collection of video, audio, image, text from a movie, blog, or video platform to generate summaries. Besides, there has also been some work dedicated to generating multi-modal output instead of only textual summaries. For example, Wang et al. [18] and Wang et al. [19] use image-text pairs to generate a pictorial storyline summarization; Evangelopoulos et al. [22] simultaneously extracts visual, audio, and textual saliency as multi-modal outputs as movies.

2.2. Multi-modal abstractive summarization

Unlike extractive approaches selecting summaries from the input content, abstractive methods directly generate a compressed description that does not appear in the original document. With the rise of sequence-to-sequence learning, multi-modal abstract summarization (MAS) has recently attracted interest. The abstractive approaches [25–28] mainly focus on summarizing texts and images until Sanabria et al. [9] first release a large-scale multi-modal dataset for summarizing open-domain videos, namely How2. How2 dataset provides multi-source data, including video, audio, ground-truth transcription, and human-written text summaries that embody salience parts of the video. MAS for open-domain videos, which we study in this paper, first organized in ICML 2019 workshop, is more challenging than text + image modalities because of the complexity of multi-modal information and the obscure nature of video features.

Previous studies [10,11] on this task leverage multi-encoder RNNs to separately encode the video and audio transcript and a joint decoder to attend the two modalities to generate a summary, which neglects the fine-grained interaction and complementarity between multi-source input; while the modeling of interactions among multiple encoders and decoder needs complicated structures and spends more parameters to couple multi-modal data. Based on above observations, we propose our approach.

2.3. Decoder-only architecture

Although the encoder-decoder architecture is now popular to solve the sequence-to-sequence problem, there has been still a few jobs using the decoder-only architecture to achieve gratifying results. A well-known decoder-only architecture is GPT-1, 2, 3 [29–31], which consists of stacked transformer decoder blocks and has achieved state-of-the-art results in multiple NLP tasks. Besides, in the field of text abstractive summarization, Liu et al. [32] uses a decoder-only transformer to handle very long input–output examples and generate more coherent and informative summaries.

While the concept of "decoder-only" has been mentioned a lot in unimodal models (mainly text), few works have been done that apply decoder-only structures on multi-modal tasks. In this paper, we propose a multi-modal decoder-only architecture to summarize the video and audio transcript. To the best of our knowledge, we are the first exploring the decoder-only architecture in the field of multi-modal abstractive summarization, where the ultra-long source text and the same-distributed summary can be learned together as the supervised information with video information in shared multi-modal decoder to gain adequate feature representation.

3. The proposed approach

3.1. Problem formulation

The goal of our multi-modal summarization system is to integrate multi-source information consisting of the video and audio transcript to generate an informative and compressed summary. Formally, let $V = (v_1, \ldots, v_m)$ and $A = (a_1, \ldots, a_n)$ represent the video and audio transcript sequences respectively, and the output textual summary is denoted by $S = (s_1, \ldots, s_o)$. Therefore, the probability of generating the summary S for the given multi-source inputs considering video V and audio transcript A can be computed as:

$$\arg\max_{\theta} P(S|A,V;\theta) = \arg\max_{\theta} \prod_{i=0}^{o} p(s_i|a_1,\ldots,a_n,v_1,\ldots,v_m;s_1,\ldots,s_{i-1};\theta) \tag{1}$$

where θ are learnable parameters.

3.2. Overview

In this work, we consider a decoder-only multi-modal transformer (D-MmT) framework to receive multi-source information of the video and output an abstractive summary. We focus on taking into account modeling multi-source and multi-modal fine-grained interactions for ultra-long sequences with reducing parameter redundancy.

The D-MmT architectures are shown in Fig. 3. The D-MmT model consists of stacked transformer decoder blocks and completely discards the encoder blocks. One of the decoder blocks is the proposed multi-modal transformer decoder that receives multi-source input features simultaneously, in which the cascaded cross-modal interaction (CXMI) module is designed to model the fine-grained interactions between modalities spontaneously and generate joint multi-modal embeddings.

On the base of above definition, we first introduce the proposed multi-modal transformer decoder block and the inside CXMI mod-

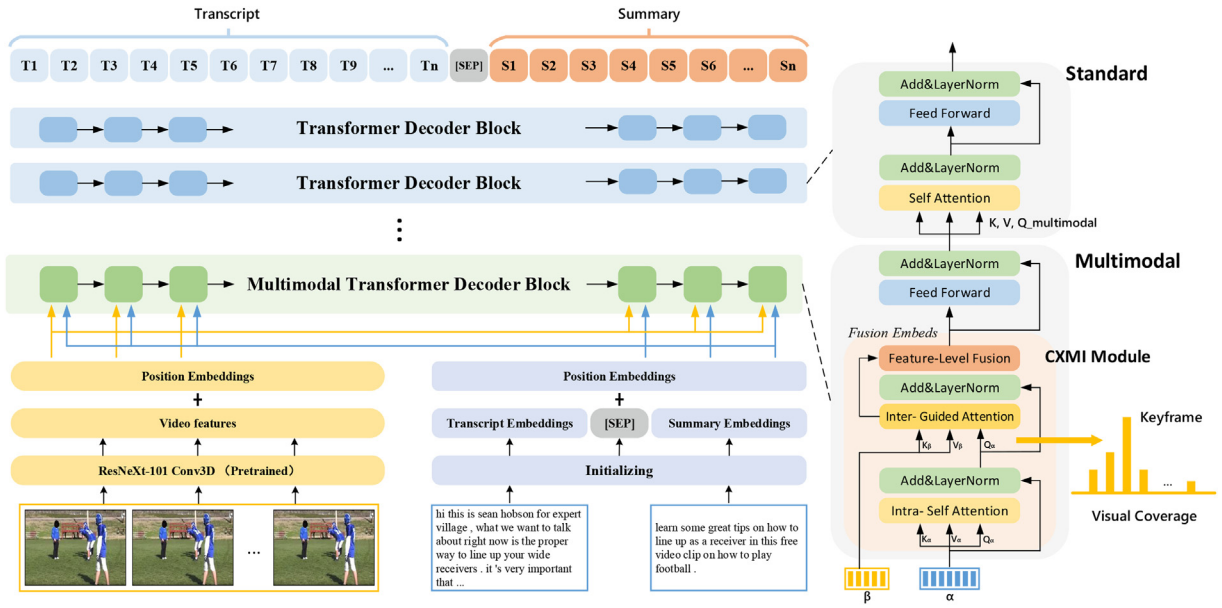

Fig. 3. Our proposed decoder-only multi-modal transformer architecture. We concatenate source input transcript and target output summary as a flat text sequence, which is fed into the multi-modal decoder block via the CXMI module together with the video sequence to generate a joint multi-modal representation. Then, the fusion representation is received by the next standard decoder blocks stacked in-depth for further interaction.

ule in Section 3.3, which is the important part of our model. Then, based on the foreshadowing of Section 3.3, we introduce our complete model in Section 3.4, including the feature representation of multi-source input (Section 3.4.1), how the multi-modal decoder block can be combined with the standard decoder block to form the complete D-MmT framework (Section 3.4.2), the calculation of loss (Section 3.4.3), and the schema of model training and testing (Section 3.4.4). Besides, based on the CXMI, we calculate a visual coverage vector for visualizing the saliency of video (Section 3.4.5).

3.3. Multi-modal decoder block with the cascaded cross-modal Interaction

The multi-modal decoder block consists of the cascaded cross-modal interaction (CXMI) module and a feed-forward layer. The CXMI is a modular composition of the cascaded intra- and inter-attention layer (Section 3.3.1 and a simple feature-level fusion layer (Section 3.3.2), which is used to generate and fuse the contextual text and video representations at low-level granularity. After passing through the CXMI module, the fusion feature is passed into a feed-forward layer like a conventional transformer [33] to reconstruct their representation.

3.3.1. Cascaded intra- and inter-attention

This part is a modular composition of the cascading of an intra-modal self-attention (SA) layer and an inter-modal guided-attention (GA) layer, which aims to establish the fine-grained intra- and inter- interactions of multi-source data. The intra-modal self-attention layer is composed of a multi-head scaled dot-product attention[33] with masked mechanism and a add&norm layer [34,35]. Taking one group of text input modality features $A = (a_1, \ldots, a_n) \in \mathbb{R}^{n \times d_a}$, the scaled dot-product attention first creates a query and key-value pair for the input feature through parallel computing:

$$[K_a : Q_a : V_a] = AW_a \tag{2}$$

where W_a are learnable parameters. Given the query Q_a, key-value pairs K_a-V_a, the contextual features C_a is obtained by the weighted sum of value, where the weighted sum is determined by the product of query and key, calculated as:

$$C_a = SA(K_a, Q_a, V_a) = softmax\left(\frac{Q_a K_a^T}{\sqrt{d_h}}\right) V_a \tag{3}$$

where d_h is applied to gain more stable gradients. Then the multi-head mechanism [33] is applied to further improve the representation capacity of the contextual features, in which multiple scaled dot-product attention is executed in parallel.

The inter-modal guided-attention layer is designed to establish the cross-modal correlations. It receives two modalities inputs simultaneously, in which the generation of the contextual features of one modality is guided by the other. In particular, the guided-attention is divided into two parts: it first takes the modality A from self-attention layer to creates a query Q_a as a guide, and then takes the modality V and creates the key-value pairs K_v-V_v to receive the guidance of the modality V; in addition to calculating the similarity between the original modality and the guided modality, an adaptive projection layer for the original modality is also designed to reflect its own importance that prevents the loss of its important information. The calculation is defined as:

$$\begin{aligned} Gscore &= softmax\left(\frac{Q_a K_v^T + VW\prime}{\sqrt{d_h}}\right) \\ &= softmax\left(\frac{AW_a\left(VW_v{}^T\right) + VW\prime}{\sqrt{d_h}}\right) \\ &= softmax\left(\frac{A\left(W_a W_v^T\right)V^T + VW\prime}{\sqrt{d_h}}\right) \\ &= softmax\left(\frac{AW_\gamma V^T + VW\prime}{\sqrt{d_h}}\right) \end{aligned} \tag{4}$$

where $W\prime, W_\gamma$ are learnable parameters. W_γ denotes the shared parameters of the two modalities that combine W_a and W_v to simplify calculations. Finally, the contextual representation of modality V guided by modality A is obtained:

$$C_v = GA(K_v, Q_a, V_v) = GscoreV = softmax\left(\frac{AW_\gamma V^T + VW_\prime}{\sqrt{d_h}}\right)V \quad (5)$$

We employed residual connection [35] and layer normalization [34] around the cascaded attention layers.

3.3.2. Feature-level fusion

This module fuses the low-level contextual features of the modality *A* and *V*. After passing through the cascaded intra- and inter- attention layer, the contextual features of the two modalities gain the same sequence length. We apply a simple fusion strategy that concatenates the two modality contextual vectors and feeds them into a linear projection layer. The calculation formulas are below:

$$C_m = ReLU(concat(C_a, C_v)W_m + b_m) \quad (6)$$

where W_m, b_m are trainable parameters. In addition, the alternative solution, merging the two modality sequences with a gate mechanism [36], is also compared in our experiments.

Finally, a multi-modal representation combining two modalities is obtained. More concretely, we generate an A-modal representation carrying aligned V-modal information, which possesses the same length as the A-modality. This could allow us to process only one sequence of fused multi-modalities instead of processing two separate modal sequences.

Following the inner structure of the transformer [33], we also utilize a fully connected feed-forward network at the end of the decoder to reconstruct the representation. It consists of two linear transformations with a GELU activation [37] in between. GELU is the Gaussian Error Linear Unit activation. The calculation is defined as:

$$FFN(x) = GELU(xW_1 + b_1)W_2 + b_2 \quad (7)$$

where W_1, W_2, b_1, b_2 are trainable parameters.

3.4. Decoder-only multi-modal transformer framework

In this section, we describe the overall structure of the proposed decoder-only multi-modal transformer (D-MmT) framework. We first explain the video and text feature representations from the multi-source inputs and output in Section 3.4.1. Then, we introduce the D-MmT in Section 3.4.2, which gives a detail about the stacking of the multi-modal transformer decoder block and standard transformer decoders. Next, the loss calculation and the working schema of the model is described in Section 3.4.3 and Section 3.4.4. In addition, a visual coverage vector is calculated for visualizing the saliency of video in Section 3.4.5.

3.4.1. Feature representation

3.4.1.1. Transcript and summary representations. Due to the in-out shared decoder-only architecture, we concatenate the input transcript text and output summary text to a flat sequence to feed the decoder. In particular, a separator is used between transcript and summary, which is denoted by [*SEP*]. It helps the model distinguish between the two and obtain a valid start signal when predicting. We unfold all tokens in transcript-[SEP]-summary into a sequence $(a_1, \ldots, a_n, [SEP], s_1, \ldots, s_o)$, and each token is further transformed into a continuous embedding vector. Following Palaskar et al. [11], we do not use the pre-trained word embeddings on other corpora but build initial word embeddings based on our task dataset. We add learnable position embedding to text features.

3.4.1.2. Video representations. The video representations are denoted by a sequence of video frame features $(v_1, \ldots, v_m)$. Following Palaskar et al. [11], the video features are extracted for every 16 non-overlapping frames from a pre-trained action recognition model: a ResNeXt-101 3D Convolution Neural Network [38] that trained to recognize 400 different human actions in the Kinetics dataset [39]. We add learnable position embeddings to video features.

3.4.2. Stacking of decoder blocks in D-MmT

The D-MmT consists of multiple decoder blocks and one of them is the multi-modal decoder block, which is stacked in-depth to gradually capture the fine-grained intra- and inter- interactions between the multi-source inputs and output. In this section, we take the multi-modal decoder at the bottom block as an example to introduce the D-MmT model. In the experimental analysis, we also analyzed the performance of the multi-modal decoder in different blocks and multiple blocks.

Firstly, taking the initial video features $V = (v_1, \ldots, v_m)$ and the text (transcript-[SEP]-summary) features $A = (a_1, \ldots, a_n, [SEP], s_1, \ldots, s_o)$ as inputs, the bottom multi-modal transformer decoder block with the CXMI module establishes text-to-text and text-to-video association, and outputs the multi-modal fusion features $Z_1 = (z_1, \ldots, [SEP], \ldots, z_{n+o})$, which can be specifically interpreted as a contextual text representation carrying its contextual video representation. Then, the next standard transformer decoder blocks receive the multi-modal fusion embeddings from the previous multi-modal decoder and model intra- and inter- associations for fusion embeddings by its own masked self-attention mechanism.

For the simplicity of the formula notation, the multi-modal transformer decoder is abbreviated as D-MmTrm, and the standard transformer decoder is abbreviated as D-Trm. The entire process can be denoted as:

$$\begin{aligned} Z_1 &= D - MmTrm(A, V) \\ Z_2 &= D - Trm(Z_1) \\ &\cdots \\ Z_n &= D - Trm(Z_{n-1}) \end{aligned} \quad (8)$$

where n is the number of the decoder blocks. For this modeling, source and target data share parameters in the in-out shared decoder, which can fully align the characteristics of the multi-source and the target, and make full use of the data to learn a thorough multi-modal representation in the shared feature space.

3.4.3. Joint in-out loss

Due to the shared decoder structure possesses the same input transcript and output summary feature distribution, we incorporate an additional loss from source transcripts into the original loss function of the target summary to make full use of the ultra-long input information, denoted as:

$$\mathscr{L} = f_{CE}(P_t, Y_t) + f_{CE}(P_s, Y_s) \quad (9)$$

where f_{CE} represents cross-entropy loss function [40]; Y_t and Y_s are the labels for the source transcript and target summary, respectively. The probability distribution of vocabularies is computed by the softmax function. Compared with the conventional target decoding loss, this joint in-out loss utilizes the additional long source sequence as the supervised information in this scenario, which assists the language feature learning to gain better contextual representations.

3.4.4. Training and testing

while the encoder-decoder structure encodes source data in encoder and learns to predict summary text in decoder whether in the training or testing stage, the decoder-only structure performs different patterns during training and testing. During training, the D-MmT model learns to predict the connected transcript-

[SEP]-summary sequence, and the transcript participates in the calculation of joint in-out loss. During testing, the source transcript sequence no longer takes part in prediction; instead, it serves as the current input to predict the remaining summary text sequentially. The pseudocodes of the training&testing details are shown in Algorithm 1 and 2, where A is the transcript feature, V is the video feature, and S is the summary feature.

Algorithm 1. Training

Input: $V = (v_1, \ldots, v_m)$, $A = (a_1, \ldots, a_n)$, and $S = (s_1, \ldots, s_o)$
procedure *Training*(V, A, S)
  concat A and S, get $\widehat{A} = (a_1, \ldots, a_n, [SEP], s_1, \ldots, s_o, [END]) = (\hat{a}_1, \ldots, \hat{a}_{n+o+2})$
  for t in range$(1, n+o+2)$ **do**:
    feed $\left(\widehat{A}[0:t], V\right)$ to *Model*, predict $A[t]$
end procedure

Algorithm 2. Testing

Input: $V = (v_1, \ldots, v_m)$, $A = (a_1, \ldots, a_n)$
Output: $S = (s_1, \ldots, s_o)$
procedure *Testing*(V, A, S)
  get $\widehat{A} = (a_1, \ldots, a_n, [SEP]) = (\hat{a}_1, \ldots, \hat{a}_{n+1})$
  while $\widehat{A}[-1] \neq [END]$ **do**:
    feed $\left(\widehat{A}, V\right)$ to model, predict *token*
    $\widehat{A} \leftarrow concat\left(\widehat{A}, token\right)$
  $S = \widehat{A}[n+2:-1]$
  return S
end procedure

3.4.5. Visual salience

Inspired by Zhu et al. [27] using unsupervised visual attention distribution to visualize image salience for image-text data, we also utilize the attention distribution for visualizing and observing the salience of the video frames. In particular, owing to the decoder-only architecture, the inter-modal guided-attention in the CXMI module can calculate the attention of the other modalities (both transcript and summary) to the video at each decoding time step. We apply a visual coverage mechanism, which sums up the historical guided-attention distribution at all the decoding time steps. At the last time step, the final visual coverage vector *Cover* of the global historical attention distribution is obtained:

$$\begin{aligned} C_{a,j} &= SA(a_1, \ldots a_n, [SEP], s_1 \ldots, sj) \\ Q_{a,j} &= C_{a,j} W \\ Cover &= \sum_j Gscore_j = \sum_j softmax\left(\frac{Q_{a,j} K_{v+VW_I}}{\sqrt{d_h}}\right) \end{aligned} \tag{10}$$

The visual coverage vector *Cover* is the same length as the video input sequence, indicating the salience of the video frames.

4. Experiments

4.1. Dataset

4.1.1. How2 dataset

The How2 dataset [9] is a large-scale open-domain video dataset used for three multi-modal tasks: multi-modal abstractive summarization, multi-modal speech recognition, and multi-modal machine translation, which aims to bring together researchers working on different aspects of multi-modal learning. The How2 video covers more than 22 topics, such as music, cooking, computer, health, etc. The total duration of the video is more than 2000 h and the average length of is 90s. Each video comes with a ground-truth English transcript, a Brazilian Portuguese translation, and a user-generated summary. The average length of the transcript is 291, and the average length of the summaries is 33. The statistics are shown in Table 1.

In this paper, we focus on the multi-modal abstractive summarization task, using the videos, ground-truth transcripts, and summaries to train the model.

4.1.2. How2-300 h dataset

How2-300 h dataset [9] has been built on top of the How2 dataset, which is the subset of the How2 videos. The How2-300 h dataset consists of about 300 h of open-domain instructional videos spanning different topics such as music, cooking, computer, health, indoor/outdoor activities, and more. Like the complete How2 dataset, every video is accompanied with a ground-truth transcript and a 2 to 3 sentence summary. Different from the How2 dataset, How2-300 h additionally provides segmented audio modality information. The training set consists of 13,168 videos totaling 298.2 h. The validation set consists of 150 videos totaling 3.2 h, and the test set consists of 175 videos totaling 3.7 h.

4.2. Baselines

We compare our model with the baseline models of single or multiple modalities: 1) S2S [41], which is a stand encoder-decoder RNN with a global attention mechanism for sequence-to-sequence learning; 2) PG [2], which is a common-used summarization model that combines copying words from source documents and outputs words from a vocabulary dictionary; 3) FT, which is a strong baseline that uses an encoder-decoder transformer model to a flat sequence. 4) BertSumAbs [42], which is a Bert-based summarization model for text modality. 5) VideoRNN [11], which is a baseline of the video-only model implemented on the How2 dataset. 6) MT [43], a transformer-based encoder-decoder architecture that receives sequence features of video for video captions. 7) HA [11] (RNN/Transformer), which is a multi-encoder-decoder model with a hierarchical attention mechanism for the multi-modal summarization task on the How2 dataset. 8) MFN [44], which presents a multi-stage fusion method to fuse multi-source data and designs a forget gate to suppress the flow of multi-modal noise. 9) TrimodalH2 [45], which is first built on video-text hierarchical attention model, and then adds the audio

Table 1
Statistics of How2 dataset.

		Videos	Hours
2000 h	Train	73,993	1766.6
	val	2965	71.3
	test	2156	51.7

modality in the second-level of hierarchical attention. 10) MAST [45], which parallelly combines audio, text, and video information using a three-level hierarchical attention approach. 11) MAST-Binned [45], which is based on the MAST, and groups the features of the audio modality for computational efficiency.

In addition, to comprehensively evaluate our model, we refer to the idea of co-attention in the field of visual question–answer (VQA) and build a multi-encoder-decoder RNN/Transformer with the co-attention module, which establishes the multi-encoder interactions. Concretely, we keep using the configuration of HA, and apple the scaled dot-product co-attention to the multi-source encodings generated by the multiple encoders.

4.3. Experimental settings

We use the 512-dimensional, 8-head, and 8-block transformers (4-block encoder + 4-block decoder or 8-block decoder). For the corpus, We truncate the maximum text sequence length to 1024 and the maximum video sequence length to 1024. During training, our model is optimized using the Adam optimizer [46] with the cross-entropy loss. We train the model for 30 epochs with a batch size of 4–8 an initial learning rate of $1.5e-4$ and use a warmup step of 2000. During decoding for prediction, we use beam search with a beam size of 6 and a length penalty with $\alpha = 1$ [47]. Following Palaskar et al. [11], we take the same 2048-dimensional video features extracted from a ResNeXt-101 3D convolutional neural network [38] as input, and do not use a pre-trained model.

4.4. Results

The models are comprehensively evaluated by multiple metrics: BLEU-1,2,3,4 [48], ROUGE-1,2,L [49], METEOR [50], and CIDEr [51]. The main results on the How2 dataset are shown in Table 2. We can clearly find that 1) compared with single-modality models, all the multi-modal models get better performance; 2) the proposed D-MmT achieves better or comparable performance compared to the state-of-the-art methods in each metric. In particular, compared to the HA, our model improves BLEU-4 by 3.2 points and ROUGE-L by 2.3 points. Compared to the strong baseline HA + Co-Att that we built, our model improves BLEU-4 by 1.5 points and ROUGE-L by 2.0 points. Compared to the MFN, our model still has a slight advantage, and improves ROUGE-L by 0.9 points, which indicates the superiority of our proposed shared multi-modal decoder framework.

Table 3 lists the parameter size of the main modules in our framework and the state-of-the-art methods. We can find that our decoder-only framework is more concise, which is attributed to the decoder only retaining self-attention parameters like an encoder, cutting the encoder-decoder and multi-encoder attention parameters. Compared with the previous multi-modal decoder with a hierarchical structure, the multi-modal decoder with the CXMI module also reduces the parameters. This is because the previous multi-modal decoder needs the summary-to-transcript attention, summary-to-video attention, and a high-level attention over the two attention context vectors, while our multi-modal decoder only needs the unidirectional attention from the connective transcript-summary to video.

Table 3
The parameter size of the main modules. Our decoder-only model is more concise that cuts the encoder-decoder and multi-encoder attention parameters. Compared to the previous multi-modal decoder with hierarchical attention, our multi-modal decoder with the CXMI module also reduces parameters.

Parameter size	D-MmT	HA	HA + Co-Att
Encoder	-	3.2 M×4	3.2 M×4
Co-Att	-	-	10.3 M
Decoder	3.2 M×7	4.2 M×3	4.2 M×3
Multi-modal Decoder	5.5 M×1	6.2 M×1	6.2 M×1
Vocabulary Embedding	40.0 M	40.0 M	40.0 M
Position Embedding (video, text)	2.7 M	2.7 M	2.7 M

Table 4 lists the comparison of total parameter size and performance between our framework and the state-of-the-art methods on the How2 dataset. We gradually reduce the number of transformer blocks until we reach half the depth of the comparison model. The results show that our model gains competitive results with fewer parameters. In detail, our approach (No. 4) obtains comparable performance of the HA (No. 1) while the parameters are reduced by 23.4%. With the deeper model stacking (No. 5), our approach achieves better performance than the HA (No. 1) and HA + Co-Att (No. 2). When our stacking depth of the model is the same as the previous (No. 6), our model significantly outperforms the HA (No. 1) and HA + Co-Att (No. 2) in each evaluation metric with $p < 0.05$ under t-test, and the parameters are still reduced by 6.6% compared to HA and 17.9% compared to HA + Co-Att. Compared to the MFN, our achieves better or comparable performance, while the parameters are significantly reduced by 26.0%.

The experimental results on the How2-300 h dataset are shown in Table 5. We can observe from the table that our proposed D-MmT outperforms the existing methods in each metric. In particular, our model, which only utilizes video and text modality, not only outperforms all the baselines in bimodal methods, but also outperforms the current state-of-the-art trimodal (video, audio, and text) model - MAST by 1.2 ROUGE-1 points, 0.8 ROUGE-2 points, and 1.3 ROUGE-L points. This demonstrates the effectiveness of our proposed shared multi-modal decoder framework.

Table 2
Results of different models on the How2 test set. "P" refers to the position of the multi-modal decoder block in our model, "P = 0" means that the multi-modal decoder is applied to the bottom decoder block. B: BLEU; R: ROUGE.

Modality	Method	B-1	B-2	B-3	B-4	R-1	R-2	R-L	METEOR	CIDEr
Ground-truth transcript	S2S	55.25	45.63	39.92	35.84	58.62	40.61	53.82	27.64	2.349
	PG	55.35	45.62	39.82	35.76	57.28	39.51	52.82	26.81	2.134
	FT	56.63	46.71	40.82	36.67	59.05	41.06	54.38	27.72	2.296
Video	VideoRNN	44.15	32.92	26.93	22.74	46.58	26.26	41.53	19.92	1.149
	MT	49.61	38.47	32.91	27.45	51.92	32.03	46.81	22.92	1.461
Ground-truth transcript + Video	HA (RNN)	57.24	47.71	41.83	37.52	60.37	42.51	55.70	28.84	2.476
	HA (Trm)	58.61	48.35	43.34	38.12	60.28	43.16	55.94	28.95	2.512
	HA + Co-Att (RNN)	57.19	48.16	42.73	38.71	60.15	43.56	56.04	28.20	2.426
	HA + Co-Att (Trm)	59.21	49.72	44.02	39.82	60.60	43.74	56.27	29.05	2.591
	MFN	60.00	**50.90**	45.30	**41.30**	61.60	45.10	57.40	29.90	2.671
	D-MmT(P = 0)	60.12	50.47	45.16	41.02	61.43	44.67	58.03	30.03	2.650
	D-MmT(P = 2)	**60.32**	50.75	**45.43**	41.28	**61.60**	45.12	**58.26**	30.12	2.670
	D-MmT(P = 4)	60.19	50.84	45.17	41.09	61.52	**45.23**	58.19	**30.28**	**2.689**

Table 4
Comparison of performance and parameters between our model and the state-of-the-art approaches on the How2 dataset.

No.	Method	Parameters	B-1	B-2	B-3	B-4	R-1	R-2	R-L	METEOR	CIDEr
1	HA	75.2 M (4enc + 4dec)	58.61	48.35	43.34	38.12	60.28	43.16	55.94	28.95	2.512
2	HA + Co-ATT	85.5 M (4enc + 4dec)	59.21	49.72	44.02	39.82	60.60	43.74	56.27	29.05	2.591
3	MFN	94.9 M (4enc + 4dec)	60.00	**50.90**	45.30	**41.30**	61.60	45.10	57.40	29.90	**2.671**
4	D-MmT	57.6 M (4dec)	58.72	48.13	43.24	38.09	60.15	43.21	55.71	29.01	2.463
5		63.9 M (6dec)	59.60	50.40	44.62	40.37	60.63	43.75	57.81	29.89	2.589
6		70.2 M (8dec)	**60.32**	50.75	**45.43**	**41.28**	**61.60**	**45.12**	**58.26**	**30.12**	2.670

Table 5
Results of different models on the How2-300 h dataset. Our model achieves state-of-the-art performance.

Modality	Model name	ROUGE-1	ROUGE-2	ROUGE-L
Unimodal	Text Only S2S	46.01	25.16	39.98
	BertSumAbs	29.68	11.74	22.58
	Video Only S2S	39.23	19.82	34.17
	Audio Only S2S	29.16	12.36	28.86
Bimodal	Audio-Text HA	34.56	15.22	31.63
	Video-Text HA	48.40	27.97	42.23
	D-MmT(ours)	**49.58**	**30.30**	**44.56**
Trimodal	TrimodalH2	47.85	28.46	42.17
	MAST-Binned	46.22	25.94	40.34
	MAST	48.85	29.51	43.23

4.5. Ablation

We conduct ablation experiments to better understand the influence of the internal components.

First, we cancel the feature-level fusion layer and remove the transcript input, and retain the text sequence [SEP]-summary instead of the previous transcript-[SEP]-summary to receive only video input. The result is shown in Table 6. Interestingly, our video-only model degenerates into an encoder-decoder architecture, and obtains a similar performance compared to the MT which is an encoder-decoder transformer that receiving video features and output text descriptions for video caption task.

Then, we ablate the inter-modal guided-attention in CXMI and remove the video input to compare our proposed decoder architecture and the conventional encoder-decoder structure under the text-only model. The ablated text-only model is similar in structure to GPT-2 model [30]. As shown in Table 6, the text-only decoder model gains better performance in each metric, indicating the effectiveness of the in-out shared decoder to learn a deeper feature representation.

Next, we retain the full model and remove the loss computation of the source transcript part. The experimental results in Table 6 show that the model performance gain notably with keeping the joint in-out loss, which indicates the source transcript involving in loss calculation plays an important role in enhancing the feature representations.

Further, we apply the gate mechanism [36] on the feature-level fusion layer to replace the simple concatenate&linear operation in the CXMI to discuss the effect of the more complex multi-modal fusion strategy. The experimental results in Table 6 show that the model performance has not been improved obviously. For a concise implementation, we cancel the gate mechanism.

4.6. More analysis

In this study, we further analyze our proposed D-MmT framework by implementing the multi-modal decoder block at different positions of the model. We trained the 4-block, 6-block, and 8-block models, and applied multi-modal decoder block to one or more other blocks. As shown in Table 7, we obtain two interesting experimental findings: 1) When the multi-modal decoder block is applied in the middle of the model, the performance is better than when applied to the bottom and top blocks; 2) Applying the multi-modal decoder block to all positions does not improve the model performance.

For the first findings, we believe that the reason is: when multi-modal fusion occurs at the bottom, the single-modality context representation before fusion is insufficient; when multi-modal fusion occurs at the top, the fusion embedding is unstable; when multi-modal fusion occurs in the middle, the single-modality first obtain a sufficient representation through multiple standard blocks, and after the fusion, the next multiple blocks allow the multi-modal fusion embedding to be fully interactive and achieve the best performance. For the second findings, we believe that the flow of lots of noise in the ultra-long sequences during multi-modal fusion makes the complex fusion schema conversely reduce the performance.

4.7. Visualization

We take a case study for visualization, as shown in Fig. 4. We use D-MmT to generate an abstractive summary for it, and visualize the inter-modal guided-attention in the CXMI module that contacts video and text modality. Further, we display the significant video keyframes obtained from the visual coverage vector.

Table 6
Ablation results on the How2 test set. The ablation experiments are based on the 8-decoder block D-MmT framework. V: video; T: transcript; FF: feature-level fusion; GA: inter-modal guided-attention.

Modality	D-MmT	B-1	B-4	R-1	R-L	Comparing	B-1	B-4	R-1	R-L
V only	-FF	50.21	27.46	52.05	47.14	MT	49.61	27.45	51.92	46.81
T only	-GA	58.28	38.04	59.98	55.94	FT	56.63	36.67	59.05	54.38
V + T	-loss	58.82	38.18	60.04	55.73	Full	60.12	41.02	61.43	58.03
	+Gate	59.98	40.98	61.49	58.05					

Table 7
We experiment using the multi-modal decoder block with the CXMI module in different locations of the D-MmT. As we can see, the model gains better performance when the multi-modal decoder block is in the middle of the model; using multi-modal decoder blocks in multiple positions does not improve performance.

Stacking	Position	B-1	B-2	B-3	B-4	R-1	R-2	R-L	METEOR	CIDEr
4 Blocks	0	58.72	**48.13**	**43.24**	**38.09**	60.15	43.21	55.71	29.01	2.463
	0,1	**58.83**	48.07	43.22	38.01	**60.22**	**43.28**	**55.80**	**29.05**	**2.472**
	0,1,2	58.05	47.91	42.87	37.70	60.02	42.96	55.25	28.85	2.421
	0,1,2,3	57.61	47.20	42.23	37.11	59.67	42.21	55.03	28.29	2.377
6 Blocks	0	59.60	50.40	44.62	40.37	60.63	43.75	57.81	29.89	2.589
	2	**59.84**	**50.84**	**45.17**	**40.99**	**60.92**	**44.03**	**57.98**	**30.16**	**2.689**
	4	59.37	50.01	44.19	39.96	60.60	43.74	57.61	29.05	2.601
8 Blocks	0	60.12	50.47	45.16	41.02	61.43	44.67	58.03	30.03	2.650
	2	**60.32**	50.75	**45.43**	**41.28**	**61.60**	45.12	**58.26**	30.12	2.670
	4	60.19	**50.84**	45.17	41.09	61.52	**45.23**	58.19	**30.28**	**2.689**
	7	58.48	48.88	42.96	38.68	60.60	43.74	56.28	29.77	2.533

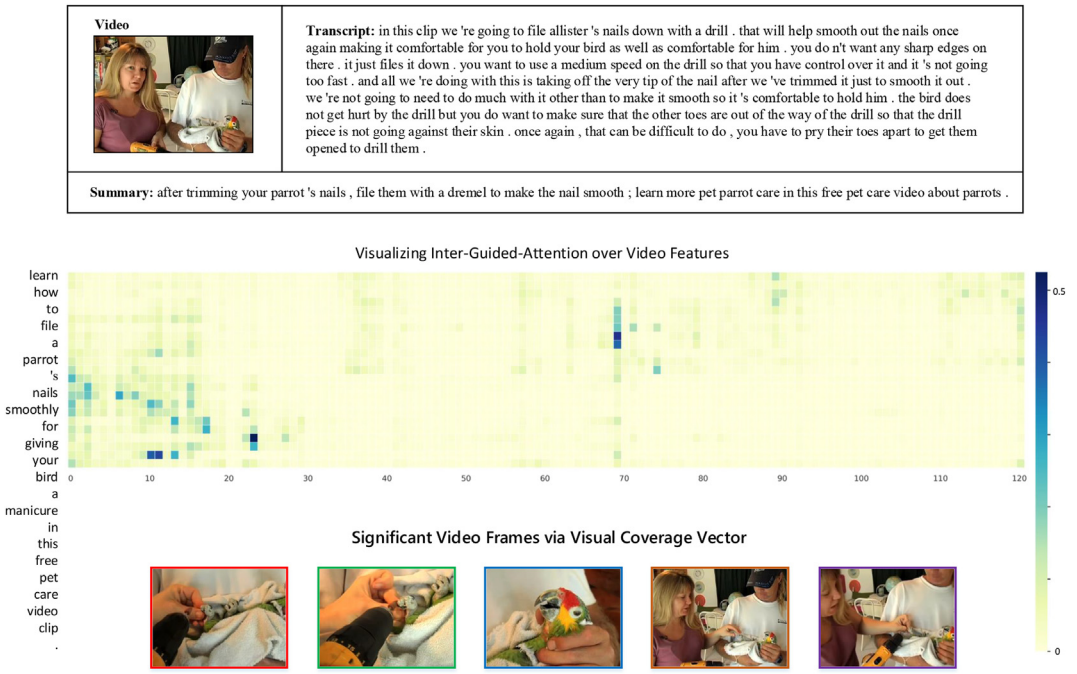

Fig. 4. This case is taken from the How2 test set, which displays the video first figure, audio transcription, and reference summary. The abstractive summary generated by the D-MmT is arranged vertically on the left. In the attention map, the depth of the color block indicates the degree of attention of the text modality to video modality at each decoding step. Significant video frames derived from the visual coverage vector are displayed at the bottom.

The color blocks in the attention map indicate the degree of attention of the text modality (source transcript and generated summary words) to the video modality in each decoding time step. We use the visual coverage vector, which is the accumulation of historical attention at all the decoding time steps, to calculate the saliency of the video frame. In this case, we select the top five keyframes in the visual coverage vector for visualization.

As shown in Fig. 4, we can observe that the extracted keyframes reflect the salient content of the video. Moreover, the keyframes provide more detailed information compared to text summary output. For example, we can observe from the picture that it is “use a drill” to trim the nails of a parrot, which is not mentioned in the text summary. This helps users understand video content more easily. In this study, we briefly explore and visualize the extraction of keyframes based on the attention distribution. In future work, we will continue to focus on the study of unsupervised extraction of video keyframes to generate a multi-modal output and establish evaluation metrics.

5. Conclusions and future work

In this work, we propose the D-MmT: a concise and effective decoder-only multi-modal transformer framework for multi-modal abstractive summarization for videos. We proposed an in-out shared multi-modal decoder to make the multi-source and target fully interact and couple in the shared feature space, which also reduces model parameters. We design a concise CXMI module in the multi-modal decoder that generates joint fusion representations and establishes a fine-grained intra- and inter- association between multi-modalities. In addition, we introduce a joint in-out loss to make full use of the ultra-long transcript input information to enhance the contextual language feature representations for generating language summaries. Experimental results on the How2 dataset confirm the superiority of our approach in performance and parameter size.

In the future, we will extend our work to generate multi-modal summaries (text and keyframes), and adopt the pretraining strategy on the D-MmT to further improve model performance.

CRediT authorship contribution statement

Nayu Liu: Conceptualization, Methodology, Writing - original draft, Software. **Xian Sun:** Writing - review & editing, Validation, Supervision. **Hongfeng Yu:** Data curation, Formal analysis. **Wenkai Zhang:** Visualization, Investigation. **Guangluan Xu:** Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks, in: Proceedings of the International Conference on Neural Information Processing Systems (NIPS), 2014, pp. 3104–3112.

[2] A. See, P.J. Liu, C.D. Manning, Get to the point: summarization with pointer-generator networks, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 2017, pp. 1073–1083.

[3] A.M. Rush, S. Harvard, S. Chopra, J. Weston, A neural attention model for sentence summarization, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2015, pp. 379–389.

[4] Y.-C. Chen, M. Bansal, Fast abstractive summarization with reinforce-selected sentence rewriting, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 2018, pp. 675–686.

[5] X. Zhang, F. Wei, M. Zhou, Hibert: document level pre-training of hierarchical bidirectional transformers for document summarization, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 2019, pp. 5059–5069.

[6] Y. Liu, M. Lapata, Hierarchical transformers for multi-document summarization, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 2019, pp. 5070–5081.

[7] S. Xu, H. Li, P. Yuan, Y. Wu, X. He, B. Zhou, Self-attention guided copy mechanism for abstractive summarization, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 1355–1362.

[8] L. Lebanoff, J. Muchovej, F. Dernoncourt, D.S. Kim, L. Wang, W. Chang, F. Liu, Understanding points of correspondence between sentences for abstractive summarization, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 191–198.

[9] R. Sanabria, O. Caglayan, S. Palaskar, D. Elliott, L. Barrault, L. Specia, F. Metze, How2: a large-scale dataset for multimodal language understanding, in: Proceedings of the Workshop on the International Conference on Neural Information Processing Systems (NIPS), 2018.

[10] J. Libovický, S. Palaskar, S. Gella, F. Metze, Multimodal abstractive summarization of open-domain videos, in: Proceedings of the Workshop on the International Conference on Neural Information Processing Systems (NIPS) , 2018.

[11] S. Palaskar, J. Libovický, S. Gella, F. Metze, Multimodal abstractive summarization for how2 videos, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 2019, pp. 6587–6596.

[12] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, D. Parikh, Vqa: visual question answering, in: Proceedings of the IEEE International Conference on Computer Vision (CVPR), 2015, pp. 2425–2433.

[13] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, W. Xu, Are you talking to a machine? Dataset and methods for multilingual image question answering, in: Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS), 2015, pp. 2296–2304.

[14] J. Lu, J. Yang, D. Batra, D. Parikh, Hierarchical question-image co-attention for visual question answering, in: Proceedings of the International Conference on Neural Information Processing Systems (NIPS), 2016, pp. 289–297.

[15] D.-K. Nguyen, T. Okatani, Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6087–6096.

[16] J. Song, P. Zeng, L. Gao, H.T. Shen, From pixels to objects: cubic visual attention for visual question answering, in: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2018, pp. 906–912.

[17] Z. Yu, J. Yu, Y. Cui, D. Tao, Q. Tian, Deep modular co-attention networks for visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6274–6283.

[18] D. Wang, T. Li, M. Ogihara, Generating pictorial storylines via minimum-weight connected dominating set approximation in multi-view graphs, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2012, pp. 683–689.

[19] W.Y. Wang, Y. Mehdad, D.R. Radev, A. Stent, A low-rank approximation approach to learning joint embeddings of news stories and images for timeline summarization, in: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2016, pp. 58–68.

[20] J. Bian, Y. Yang, T.-S. Chua, Multimedia summarization for trending topics in microblogs, in: Proceedings of the ACM International Conference on Information & Knowledge Management, ACM, 2013, pp. 1807–1812.

[21] J. Bian, Y. Yang, H. Zhang, T.-S. Chua, Multimedia summarization for social events in microblog stream, IEEE Trans. Multimedia 17 (2) (2014) 216–228.

[22] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, Y. Avrithis, Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention, IEEE Trans. Multimedia 15 (7) (2013) 1553–1568.

[23] H. Li, J. Zhu, C. Ma, J. Zhang, C. Zong, Multi-modal summarization for asynchronous collection of text, image, audio and video, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2017, pp. 1092–1102.

[24] M. Schinas, S. Papadopoulos, G. Petkos, Y. Kompatsiaris, P. A. Mitkas, Multimodal graph-based event detection and summarization in social media streams, in: Proceedings of the ACM International Conference on Multimedia, ACM, 2015, pp. 189–192..

[25] H. Li, J. Zhu, T. Liu, J. Zhang, C. Zong, Multi-modal sentence summarization with modality attention and image filtering, in: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2018, pp. 4152–4158.

[26] J. Chen, H. Zhuge, Abstractive text-image summarization using multi-modal attentional hierarchical rnn, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2018, pp. 4046–4056.

[27] J. Zhu, H. Li, T. Liu, Y. Zhou, J. Zhang, C. Zong, Msmo: multimodal summarization with multimodal output, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2018, pp. 4154–4164.

[28] J. Zhu, Y. Zhou, J. Zhang, H. Li, C. Zong, C. Li, Multimodal summarization with guidance of multimodal reference, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2020, pp. 9749–9756.

[29] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training, Tech. rep., OpenAI (2018)..

[30] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, Tech. rep., OpenAI (2019)..

[31] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, arXiv preprint arXiv:2005.14165, 2020..

[32] P.J. Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, N. Shazeer, Generating wikipedia by summarizing long sequences, in: Proceedings of the International Conference on Learning Representations (ICLR), 2018.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the International Conference on Neural Information Processing Systems (NIPS), 2017, pp. 6000–6010.

[34] J.L. Ba, J.R. Kiros, G.E. Hinton, Layer normalization, arXiv preprint arXiv:1607.06450, 2016..

[35] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[36] X. Zhou, S. Luo, Y. Wu, Co-attention hierarchical network: generating coherent long distractors for reading comprehension, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2020, pp. 9725–9732.

[37] D. Hendrycks, K. Gimpel, Bridging Nonlinearities and Stochastic Regularizers with Gaussian Error Linear Units, arXiv preprint arXiv:1606.08415, 2016..

[38] K. Hara, H. Kataoka, Y. Satoh, Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp 6546–6555.

[39] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al., The kinetics human action video dataset, arXiv preprint arXiv:1705.06950, 2017..

[40] P.-T. de Boer, D. Kroese, S. Mannor, R. Rubinstein, A tutorial on the cross-entropy method, Ann. Oper. Res. 134 (1) (2005) 19–67.

[41] M.-T. Luong, H. Pham, C.D. Manning, Effective approaches to attention-based neural machine translation, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2015, pp. 1412–1421.

[42] Y. Liu, M. Lapata, Text summarization with pretrained encoders, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3721–3731.

[43] L. Zhou, Y. Zhou, J.J. Corso, R. Socher, C. Xiong, End-to-end dense video captioning with masked transformer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8739–8748.

[44] N. Liu, X. Sun, H. Yu, W. Zhang, G. Xu, Multistage fusion with forget gate for multimodal summarization in open-domain videos, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 1834–1845.

[45] A. Khullar, U. Arora, Mast: multimodal abstractive summarization with trimodal hierarchical attention, in: Proceedings of the Workshop on the Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 60–69..

[46] D. P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2014..

[47] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al., Google's neural machine translation system: bridging the gap between human and machine translation, arXiv preprint arXiv:1609.08144, 2016..

[48] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL), Association for Computational Linguistics, 2002, pp. 311–318..

[49] C.-Y. Lin, ROUGE: a package for automatic evaluation of summaries, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 2004, pp. 74–81.

[50] S. Banerjee, A. Lavie, Meteor: an automatic metric for mt evaluation with improved correlation with human judgments, in: Proceedings of the

Workshop on the Annual Meeting on Association for Computational Linguistics (ACL), 2015, pp. 65–72.
[51] R. Vedantam, C. Lawrence Zitnick, D. Parikh, Cider: consensus-based image description evaluation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4566–4575.

Nayu Liu received the bachelor's degree from Xidian University, Xian, China, in 2018.
He is currently a Ph.D. candidate in the Aerospace Information Research Institute, University of Chinese Academic of Sciences, Beijing, China. His research interests include text generation, summarization, and multimodal information processing.

Xian Sun received the B.Sc. degree from Beijing University of Aeronautics and Astronautics, Beijing, China, in 2004. He received the M.Sc. and Ph.D. degrees from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2009.
He is currently a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include computer vision, geospatial data mining, and remote sensing image understanding.

Hongfeng Yu received the B.Sc. degree and M.Sc. degree from Peking University, Beijing, China, in 2013 and 2016 respectively.
He is currently a Research Assistant at the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include deep learning and multimodal information mining.

Wenkai Zhang received the B.Sc. degree from China University of Petroleum, Shandong, China, in 2013, and the Ph.D. degree from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2018.
He is currently a Research Assistant at the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include remote sensing image semantic segmentation and multi-media information processing.

Guangluan Xu received the B.Sc. degree from the Beijing Information Science and Technology University, Beijing, China, in 2000, and the M.Sc. and Ph.D. degrees from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2005.
He is currently a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include computer vision and remote sensing image understanding.