Text with Knowledge Graph Augmented Transformer for Video Captioning

Xin Gu[1,3], Guang Chen[2], Yufei Wang[2], Libo Zhang[3,1]*, Tiejian Luo[1], Longyin Wen[2]
[1]University of Chinese Academy of Sciences, Beijing, China
[2]ByteDance Inc., San Jose, USA
[3]Institute of Software Chinese Academy of Sciences, Beijing, China
guxin21@mails.ucas.edu.cn, guang.chen@bytedance.com, yufei.wang@bytedance.com
libo@iscas.ac.cn, tjluo@ucas.ac.cn, longyin.wen@bytedance.com

Abstract

Video captioning aims to describe the content of videos using natural language. Although significant progress has been made, there is still much room to improve the performance for real-world applications, mainly due to the long-tail words challenge. In this paper, we propose a text with knowledge graph augmented transformer (TextKG) for video captioning. Notably, TextKG is a two-stream transformer, formed by the external stream and internal stream. The external stream is designed to absorb additional knowledge, which models the interactions between the additional knowledge, e.g.*, pre-built knowledge graph, and the built-in information of videos,* e.g.*, the salient object regions, speech transcripts, and video captions, to mitigate the long-tail words challenge. Meanwhile, the internal stream is designed to exploit the multi-modality information in videos (*e.g.*, the appearance of video frames, speech transcripts, and video captions) to ensure the quality of caption results. In addition, the cross attention mechanism is also used in between the two streams for sharing information. In this way, the two streams can help each other for more accurate results. Extensive experiments conducted on four challenging video captioning datasets,* i.e.*, YouCookII, ActivityNet Captions, MSR-VTT, and MSVD, demonstrate that the proposed method performs favorably against the state-of-the-art methods. Specifically, the proposed TextKG method outperforms the best published results by improving* 18.7% *absolute CIDEr scores on the YouCookII dataset.*

1. Introduction

Video captioning aims to generate a complete and natural sentence to describe video content, which attracts much attention in recent years. Generally, most existing methods [21, 38, 41, 58] require a large amount of paired video and description data for model training. Several datasets, such as YouCookII [69], and ActivityNet Captions [19] are constructed to promote the development of video captioning field. Meanwhile, some methods [29, 40, 48, 72] also use the large-scale narrated video dataset HowTo100M [30] to pretrain the captioning model to further improve the accuracy.

Although significant progress has been witnessed, it is still a challenge for video captioning methods to be applied in real applications, mainly due to the long-tail issues of words. Most existing methods [29, 40, 48, 72] attempt to design powerful neural networks, trained on the large-scale video-text datasets to make the network learn the relations between video appearances and descriptions. However, it is pretty tough for the networks to accurately predict the objects, properties, or behaviors that are infrequently or never appearing in training data. Some methods [14, 71] attempt to use knowledge graph to exploit the relations between objects for long-tail challenge in image or video captioning, which produces promising results.

In this paper, we present a text with knowledge graph augmented transformer (TextKG), which integrates additional knowledge in knowledge graph and exploits the multi-modality information in videos to mitigate the long-tail words challenge. TextKG is a two-stream transformer, formed by the external stream and internal stream. The external stream is used to absorb additional knowledge to help mitigate long-tail words challenge by modeling the interactions between the additional knowledge in pre-built knowledge graph, and the built-in information of videos, such as the salient object regions in each frame, speech transcripts, and video captions. Specifically, the information is first retrieved from the pre-built knowledge graphs based on the detected salient objects. After that, we combine the features of the retrieved information, the appearance features of detected salient objects, the features of speech transcripts and

*Corresponding author:libo@iscas.ac.cn, Libo Zhang was supported by Youth Innovation Promotion Association, CAS (2020111).
This work was done during internship at ByteDance Inc.

captions, then feed them into the external stream of TextKG to model the interactions. The internal stream is designed to exploit the multi-modality information in videos, such as the appearance of video frames, speech transcripts and video captions, which can ensure the quality of caption results. To share information between two streams, the cross attention mechanism is introduced. In this way, the two streams can obtain the required modal information from each other for generating more accurate results. The architecture of the proposed method is shown in Figure 1.

Several experiments conducted on four challenging datasets, *i.e*., YouCookII [69], ActivityNet Captions [19], MSR-VTT [56], and MSVD [3] demonstrate that the proposed method performs favorably against the state-of-the-art methods. Notably, our TextKG method outperforms the best published results by improving 18.7% and 3.2% absolute CIDEr scores in the paragraph-level evalution mode on the YouCookII and Activity-Net Captions datasets.

2. Related Work

Video captioning attracts much attention of researchers in recent years. The best practice has been achieved by attention-based methods, which attempts to associate visual components with sentences in videos. Some of them focus on designing powerful network architectures. VLM [55] and VideoBERT [43] take the visual and text modalities as input, and use a shared transformer to construct a task-agnostic video-language model. ViLBERT [28] processes visual and linguistic information separately with two parallel streams, and then use the attention mechanism to model the interactions between visual and language features. Instead of using the separate encoder-decoder architecture, MART [21] designs a shared encoder-decoder network and augments it with the memory module. ActBert [72] uses local regional features to learn better visual-language alignment. WLT [35] takes audio features as an additional input, and uses context fusion to generate multimodal features.

Meanwhile, some methods [10,14,15,62,65,71] focus on exploiting prior knowledge to provide semantic correlations and constraints between objects for image or video captioning, producing promising results. ORG-TRL [64] uses the knowledge information in the language model (BERT) to provide candidate words for video captioning. In contrast, we propose a two-stream transformer for video captioning, with the internal stream used to exploit multi-modality information in videos, and the external stream used to model the interactions between the additional knowledge and the built-in information of videos. These two streams use the cross-attention mechanism to share information in different modalities for generating more accurate results.

Vision-and-language representation learning is a hot topic in recent years. ViLBERT [28], LXMERT [46], UNITER [6], UNIMO [25] and Unified-VL [68] learn the representations between image and text, while Univl [29], VideoBERT [43], ActBERT [72] and MV-GPT [40] learn the representations between videos and transcripts. Notably, most of these methods attempt to learn powerful vision-and-language representations by pre-training the models on the large-scale datasets, *e.g*., Howto100M [31] and WebVid-2M [1], and then finetune them on downstream tasks such as video captioning, video-text retrieval and visual question answering. In contrast, our TextKG method uses the speech transcripts as the text to model the visual and linguistic representations and integrate the additional knowledge in knowledge graph to mitigate long-tail words challenge in video captioning.

Knowledge graph in NLP. Knowledge graph is an useful tool to indicate the real-world entities and their relations, which provides rich structured knowledge facts for language modeling. Large-scale knowledge graphs are used to train knowledge enhanced language models for various natural language processing (NLP) tasks. CoLAKE [44] proposes to inject the knowledge context of an entity, and to jointly learn the contextualized representation for both language and knowledge by a unified structure. ERNIE [63] enhances BERT architecture to better integrate the knowledge information and textual information. KEPLER [52] not only improves language models by integrating factual knowledge, but also generates text-enhanced knowledge representation. JAKET [59] proposes a joint pre-training framework to model knowledge graph and language simultaneously. Inspired by CoLAKE, our method jointly learns the representations of vision, language and knowledge, and enhances the joint visual-language representations by retrieving relevant knowledge in knowledge graphs.

3. Our Approach

As mentioned above, our method aims to integrate additional knowledge using knowledge graph and exploits the multi-modality information in videos to mitigate the long-tail words challenge in the field of video captioning. We design a text with knowledge graph augmented transformer, *i.e*., a two-stream transformer formed by the external and internal streams. Both streams are constructed by N sets of alternately stacked self-attention and cross-attention blocks. Specifically, we first use a detector to generate the salient object regions in each video frame, and use the automatic speech recognition (ASR) method [2] to generate the speech transcripts in videos. After that, the class labels of detected salient objects are used to retrieve the prior knowledge in knowledge graphs. Then, the appearance embeddings of detected salient objects and video frames, and embeddings of retrieved prior knowledge, speech transcripts, and predicted captions are fed into the two-stream transformer to generate subsequent words of the predicted captions. The overall architecture of the proposed TextKG is shown in Figure 1.

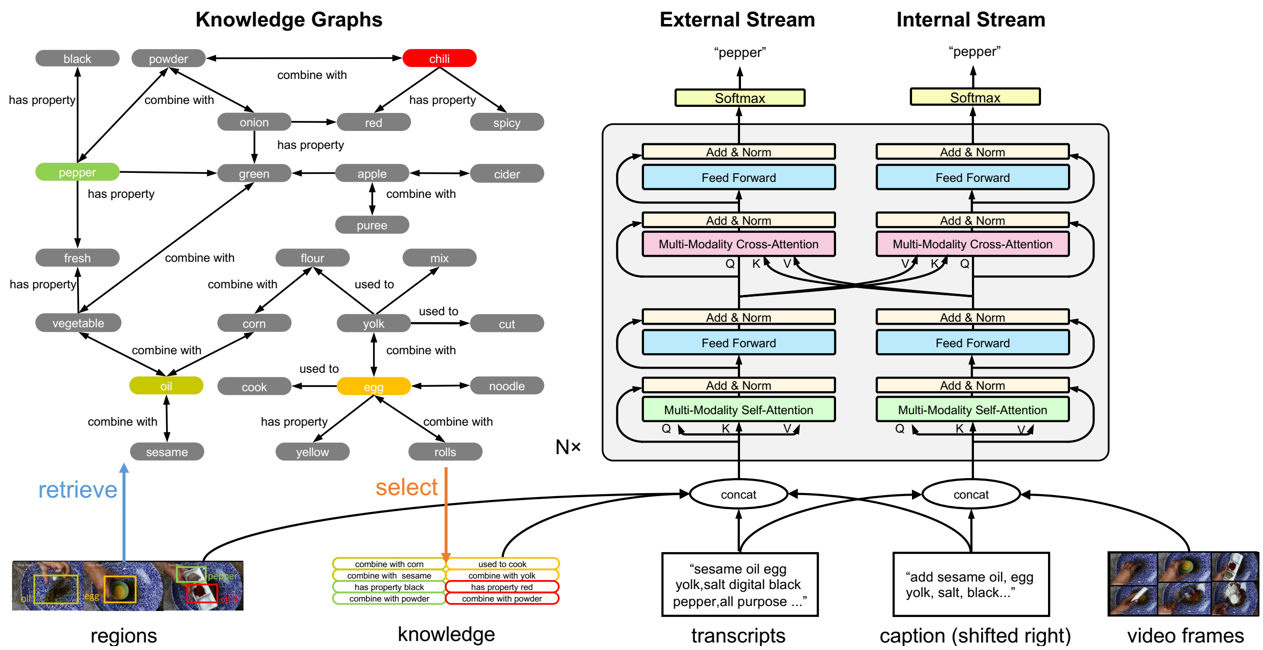

Figure 1. The architecture of our TextKG method, which is formed the external and internal streams. Each stream is stacked with N sets of multi-modality self-attention and cross-attention modules.The cross attention modules are designed to align tokens in different modalities.

3.1. Two-Stream Transformer

As described above, our two-stream transformer is formed by the external stream and internal stream. Several self-attention and cross-attention blocks are interleaved to construct the two streams.

Let $\{y_1, y_2, \cdots, y_l\}$ be the predicted captions, where y_i is the index of the i-th word in the dictionary, $\{p_1, p_2, \cdots, p_l\}$ to be the predicted probabilities of words in the dictionary, where p_i is the probability of the i-th word. The index of the i-th word y_i is computed as $y_i = \arg\max_i p_i$. Meanwhile, let z_i^{ext} and z_i^{int} to be the output probabilities of the external and internal streams of the i-th word. Thus, we have $p_i = \omega_1 z_i^{\text{ext}} + \omega_2 z_i^{\text{int}}$, where ω_1 and ω_2 are the hyperparameters used to balance the outputs of the two-streams. In this way, y_i is computed as

$$y_i = \arg\max_i(\omega_1 z_i^{\text{ext}} + \omega_2 z_i^{\text{int}}). \tag{1}$$

In the following sections, we will describe each module in our two-stream transformer in more details.

Multi-modality Self-Attention Module. As shown in Figure 1, we use the self-attention module [49] in both the external and internal streams to model the interactions among multi-modality information. Specifically, for the external stream, the concatenation of feature embeddings of detected objects $\mathcal{F}_{\text{r}}$, retrieved prior knowledge $\mathcal{F}_{\text{k}}$, speech transcripts $\mathcal{F}_{\text{s}}$, and predicted video captions $\mathcal{F}_{\text{c}}$ are fed into the self-attention module to model the interactions, *i.e.*, $\mathcal{X}_{\text{ext}} = \text{Concat}(\mathcal{F}_{\text{r}}, \mathcal{F}_{\text{k}}, \mathcal{F}_{\text{s}}, \mathcal{F}_{\text{c}})$. Meanwhile, for the internal stream, the concatenation of feature embeddings of speech transcripts $\mathcal{F}_{\text{s}}$, predicted video captions $\mathcal{F}_{\text{c}}$, and video frame $\mathcal{F}_{\text{v}}$ are fed into the self-attention model to exploit the interactions, *i.e.*, $\mathcal{X}_{\text{int}} = \text{Concat}(\mathcal{F}_{\text{s}}, \mathcal{F}_{\text{c}}, \mathcal{F}_{\text{v}})$. The self-attention module computes the interactions as follows,

$$\Phi(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}} + M)V, \tag{2}$$

where d is the feature dimension of queries Q and keys K. We have $Q = \mathcal{X}_{(\cdot)}W^{\text{Q}}$, $K = \mathcal{X}_{(\cdot)}W^{\text{K}}$, $V = \mathcal{X}_{(\cdot)}W^{\text{V}}$, where W^{Q}, W^{K}, and W^{V} are learnable parameters, and $\mathcal{X}_{(\cdot)} \in \{\mathcal{X}_{\text{ext}}, \mathcal{X}_{\text{int}}\}$ are the concatenated feature embeddings for the external and internal streams, respectively.

Following [21], we introduce a mask matrix M in attention function (2) of both the external and internal streams to prevent the model from seeing future words. That is, for the i-th caption token, we set $M_{i,j} = 0$ for $j = 1, \cdots, i$, and set $M_{i,j} = -\inf$ for $j > i$. Meanwhile, for the external stream, we set $M_{i,j} = -\inf$ to prevent the associations between irrelevant retrieved prior knowledge and the detected salient object regions.

Multi-modality Cross-Attention Module. Besides self-attention module, we use the cross-attention module to align the interactions modeled by both the external and internal streams. Specifically, we compute the affinity matrix to guide the alignments between the modeled interactions by injecting the information. Similar to the self-attention module, we also introduce the mask matrix into the attention function to compute the cross-attention. Intuitively, the retrieved prior knowledge should not have direct effect on the predicted captions, but through the detected salient objects.

Thus, we set the corresponding elements to $-\inf$ to prevent the retrieved prior knowledge directly influencing the predicted captions.

3.2. Optimization

The cross entropy loss function is used to guide the training of our TextKG method. Given the ground-truth indices of previous $(i-1)$ words and the concatenated of two-streams $\mathcal{X}_{\text{ext}}$ and $\mathcal{X}_{\text{int}}$, we can get the predictions of the current i-th word. After that, the training loss of our method is computed as

$$\mathcal{L} = -\sum_{i=1}^{l}(\lambda_1 \log z_i^{\text{ext}} + \lambda_2 \log z_i^{\text{int}}), \quad (3)$$

where z_i^{ext} and z_i^{int} are the i-th output word of the external and internal streams, λ_1 and λ_2 are the preset parameters used to balance the two streams, and l is the total length of predicted captions. The Adam optimization algorithm [17] with an initial learning rate of $1e-4$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $L2$ weight decay of 0.01 is used to train the model. The Warmup strategy is adopted in training phase by linearly increasing the learning rate from 0 to the initial learning rate for the first 10% of training epochs, and linearly decreasing to 0 for the remaining 90% of epochs.

3.3. Multimodal Tokens

The appearance embeddings of detected salient objects and video frames, and embeddings of retrieved prior, speech transcripts, and predicted captions are fed into the two-stream transformer to generate subsequent captions after tokenization. We will describe the feature extraction and tokenization processes of the aforementioned embeddings.

Video tokens encode the appearance and motion information of video frames. Specifically, the aligned appearance and motion features form the video tokens. In each second of the video, we sample 2 frames to extract features. Following [21], we use the "Flatten-673" layer in ResNet-200 [12] to extract appearance features, and the "global pool" layer in BN-Inception [16] to extract the motion features on the YouCookII [69] and ActivityNet Captions [19] datasets. Meanwhile, for fair comparison, we use the InceptionResNetV2 [45] and C3D [11] models to extract appearance and motion features on the MSR-VTT [56] and MSVD [3] datasets, following the methods [58, 64, 66].

Region tokens are used to describe the visual information of the detected salient objects. We use the Faster R-CNN method [37] to extract visual features for detected object regions, which is pretrained on the Visual Genome dataset [20], and the N_{r} detected objects with highest confidence score in each frame are reserved for video captioning.

Transcript tokens are used to encode the content of speech transcripts, which contains some key information in videos. The GloVe model [34] is used to extract features of each word in transcripts to form the transcript tokens.

Caption tokens encode the information of predicted captions. Similar to transcript tokens, GloVe [34] is adopted to extract features of each word in the generated captions.

Knowledge tokens model the content of retrieved knowledge, which is important to mitigate the long-tail words challenge in video captioning. Our TextKG method uses a triple structure to encode the retrieved knowledge, formed by two object items and the relations between them, *i.e.*, $(\alpha_{\text{head}}, \alpha_{\text{tail}}, \rho_{\text{rel}})$, where α_{head} and α_{tail} are the two nodes in knowledge graph, and ρ_{rel} indicates the edge encoding the relations between them. For example, α_{head} ="`knife`", α_{tail} ="`hard`", and ρ_{rel} ="`has property`". For each detected object α_{head}, we aim to integrate prior context information for more accurate caption results. That is, we use the GloVe model to extract linguistic features of the node α_{tail} and set the learnable embeddings for the edge ρ_{rel}. After that, we compute the summation of them to get the knowledge embeddings of the knowledge tokens.

For each detected object, we use its category name to retrieve knowledge (*i.e.*, triples containing the category name of the detected object) from the knowledge graph. Notably, only a portion of the retrieved knowledge are highly related to the video. The rests are potentially harmful to the performance of caption generation as it may distract the training and inference and also increase computational cost. For example, to generate caption of a cooking tutorial video, the knowledge "`knife - used to -cut`" is more useful rather than "`knife - has property - hard`". Because cutting food may be an important procedure for food preparation. Thus, we design a primary rank mechanism to sort the retrieved knowledge based on their semantic similarity to the video. The cosine similarity between the emebeddings of knowledge tokens and video transcripts is computed as the semantic similarity for ranking. The pretrained SBERT [36] model is used to extract feature embeddings of transcripts. Finally, N_{k} knowledge items with the highest similarity scores are reserved for video captioning.

3.4. Knowledge Graphs Construction

As mentioned above, we use the knowledge graph to retain the prior knowledge. The nodes in the graph could corresponding to a noun, an adjective, a verb or an adverb, etc. The edges in the graph are used to encode the relations between different nodes. Apparently, it is important to include all key information of videos in knowledge graphs for accurate caption results. Thus, besides general knowledge graph covering most key knowledge in general scenarios, we also introduce the specific knowledge graph for each dataset covering the key knowledge in specific scenarios.

General knowledge graph is designed to include most key information in general scenarios that we are interested, such as cooking and activity. Specifically, the general knowledge graph is exported from the public available giant knowledge graph ConceptNet [42] by extracting key words[1] in ConceptNet with the connected edges and neighboring nodes.

Specific knowledge graph. Besides the general knowledge graph, we also construct the specific knowledge graph to cover key information in specific scenarios. We believe that the speech transcripts of videos contain most of the crucial information and use the speech transcripts as the source to construct the specific knowledge graph. We first use the automatic speech recognition (ASR) model [2] to convert the speech in all videos to transcripts. After that, we use the Stanford NLP library[2] to analyze the components in each sentence in transcripts to form a structured grammer tree. We collect the "adjective and noun", "noun and noun" and "adverb and verb" phrases from the grammar trees to construct the specific knowledge graph.

4. Experiments

4.1. Datasets

YoucookII contains $2,000$ long untrimmed videos from 89 cooking recipes, including $1,333$ videos for training, 457 videos for testing. The average length of videos is 5.3 minutes and each video is divided into 7.7 video clips on average with the manually labeled caption sentences.

ActivityNet Captions is a large-scale challenging dataset for video captioning, with $10,024$ videos for training, $4,926$ for validation, and $5,044$ for testing. The average length of each video is 2.1 minutes and is divided into 3.65 clips on average. Each clip is assigned with a manually labeled caption sentence. Since the testing set is not publicly available, following MART [21], we split the validation set into two subsets, *i.e*., ae-val with $2,460$ videos for validation and ae-test with $2,457$ videos for testing.

MSR-VTT includes $10,000$ video clips with 15 seconds on average. Each video clip contains 20 human annotated captions and the average length of each sentence is 9.28 words. Following [5, 38, 58], $6,513$, 497, and $2,990$ videos are used for training, validation, and testing, respectively.

MSVD consists of 1970 video clips collected from YouTube and each clip is approximately 10 seconds long. Each clip is labeled with 40 sentences. Following [5,38,58], we split the dataset into $1,200$, 100, and 670 video clips for training, validation and testing, respectively.

4.2. Evaluation metrics

To compare the performance of the proposed method with other methods, five model-free automatic evaluation metrics are used, including BLEU@4 [33] (abbreviated to B) that are precision-based metric, METEOR [8] (abbreviated to M) that computes sentence-level scores, CIDEr [50] (abbreviated to C) that is consensus-based metric, ROUGE [26] (abbreviated to R) that uses longest common subsequence to compute the similarities between sentences, and Rep@4 [54] (abbreviated to Rep) that computes the redundancy score for every captions. Lower Rep indicates better model. The CIDEr score is used as the primary metric in evaluation. Notably, both micro-level and paragraph-level evaluation modes are used. The micro-level mode evaluates the predicted captions of each video clip independently, while the paragraph-level mode considers the relations between captions of different clips. Following [9, 21], the paragraph-level evaluation mode is used in the YouCookII and ActivityNet datasets, which contains multiple clips in each video.

4.3. Implementation Details

We sample every 2 frames per second in each video and use the pre-trained models to extract video features. For the YouCookII and ActivityNet Captions datasets, similar to [21], the ResNet-200 and BN-Inception models are used to extract the appearance and motion features, respectively. Following [58, 64, 66], the InceptionResNetV2 and C3D models are used to extract the appearance and motion features for the MSR-VTT and MSVD dataset, respectively. Meanwhile, the Faster R-CNN method using ResNet-101 as the backbone is adopted to detect object in each video frame, which is trained on the Visual Genome dataset [20]. Notably, only $N_{\mathrm{r}} = 6$ classes of the detected objects with top scores are used to form region features in each frame.

The caption sentences, speech transcripts, and retrieved knowledge in knowledge graph are split into individual words using the NLTK toolkit[3], and the GloVe [34] algorithm is used to extract the word embeddings for each word to construct the transcript, caption, and knowledge tokens. The dimension of word embeddings is set to 300. The captions are truncated to 20 words and the maximum text length of transcripts is set to 300 words. For each detected object, we retrieve up to $N_k = 5$ pieces of knowledge items in knowledge graph depending on the relevant scores.

Notably, the proposed method is implemented using Pytorch. 3 blocks of multi-modality self-attention and cross-attention modules are used in the MSR-VTT [56] and MSVD [3] datasets, and 2 are used in the YouCookII [69] and ActivityNet Captions [19] datasets, to form the two-stream transformer, respectively. The feature dimension in each block is set to 768, and the number of heads in multi-head architecture is set to 12. The proposed method is trained using the Adam algorithm [17] with the initial learning rate of $1e-4$. The batch size in training phase is set to 6. The hyper-parameters ω_1, ω_2, λ_1, and λ_2 are set to 0.8, 0.2, 0.5 and 0.5, empirically.

[1] The key words are exported from the speech transcripts and annotated ground-truth captions in the training sets of the caption datasets.

[2] https://stanfordnlp.github.io/CoreNLP/

[3] https://github.com/nltk/nltk

Method	B	M	C	Rep
Van-Trans [70]	7.6	15.7	32.3	7.8
Trans-XL [7]	6.6	14.8	26.4	6.3
Trans-XLRG [22]	6.6	14.7	25.9	6.0
MART [21]	8.0	16.0	35.7	4.4
Van-Trans+COOT [9]	11.1	19.8	55.6	5.7
COOT [9]	11.3	19.9	57.2	6.7
TextKG	**14.0**	**22.1**	**75.9**	**2.8**
Human	-	-	-	1.0

Table 1. Evaluation results on the YouCookII val subset in the paragraph-level evaluation mode.

4.4. Comparison to the State of the Art Methods

YouCookII dataset. To validate the effectiveness of the proposed method, we compare it to the state-of-the-art (SOTA) methods using the paragraph-level evaluation mode, reported in Table 1. As shown in Table 1, our TextKG method achieves the best results on the YouCookII dataset. Specifically, TextKG improves 18.7% absolute CIDEr score compared to the SOTA method COOT [9]. COOT [9] focuses on leveraging the hierarchy information and modeling the interactions between different levels granularity and different modalities, such as frames and words, clip and sentences, and videos and paragraphs, which is pre-trained on the large-scale HowTo100M [31] dataset. In contrast, our TextKG method aims to exploit additional knowledge in knowledge graph and multi-modality information in videos to improve the caption results.

Beside the paragraph-level evaluation mode, we also report the evaluation results based on the micro-level mode on the YouCookII dataset in Table 2. As shown in Table 2, TextKG significantly outperforms other methods, including Univl [29] and AT [13], both of which use transcript information to enhance model performance by directly concatenating visual and transcript embedding. We argue that with the use of knowledge graphs, TextKG gives a better understanding of the content included in the speech transcripts and thus yields favorable results.

ActivityNet Caption dataset. We also evaluate our TextKG method on the challenging ActivityNet Caption dataset in Table 3. As shown in Table 3, TextKG achieves the best results on 3 out of 4 metrics, *i.e.*, BLEU, METOR, and CIDEr. The ActivityNet Caption dataset [19] contains a series of complex events with annotated caption sentences describing the events that occur. These events may occur over various periods of time and may co-occur. In contrast to existing methods [5, 38, 58, 66], our TextKG understands complex events more comprehensively with the help of exploiting information from transcripts and integrating the relevant knowledge in knowledge graph to enhance the common-sense information for video captioning.

MSR-VTT dataset. The evaluation results on the MSR-VTT dataset are reported in Table 4. For the fair comparison, we use the same visual features as the existing methods [5, 58, 64, 66]. In contrast to existing methods focusing on network architecture design to exploit visual information, our TextKG aims to exploit mutli-modality information in external knowledge graph and original videos, achieving the state-of-the-art performance on 3 out of 4 metrics, *i.e.*, BLEU, METEOR, and CIDEr.

Method	Input	B	M	R	C
Masked Trans [70]	V	3.8	11.6	27.4	38
S3D [53]	V	3.2	9.5	26.1	31
VideoAsMT [18]	V	5.3	13.4	-	-
SwinBERT [27]	V	9	15.6	37.3	109
VideoBERT [43]	V	4.0	11.0	27.5	49
VideoBERT+S3D [43]	V	4.3	11.9	28.8	50
ActBERT [72]	V	5.4	14.3	30.6	65
AT [13]	V+S	9.0	17.8	36.7	112
DPC [41]	V+S	2.8	18.1	-	-
VALUE [24]	V+S	12.4	**18.8**	**40.4**	130
Univl [29]	V+S	9.5	16.3	37.4	115
MV-GPT [40]	V+S	**13.3**	17.6	35.5	103
TextKG	V+S	11.7	18.4	40.2	**133**

Table 2. Evaluation results on the YouCookII val subset in the micro-level evaluation mode. 'V' indicates the methods use video appearance information, and 'S' indicates the methods use the speech information.

Method	B	M	C	Rep
HSE [60]	9.8	13.8	18.8	13.2
GVD [67]	11.0	15.7	22.0	8.8
GVDsup [67]	**11.3**	16.4	22.9	7.0
Van-Trans [70]	9.8	15.6	22.2	7.8
Trans-XL [7]	10.4	15.1	21.7	8.5
Trans-XLRG [22]	10.2	14.8	20.4	8.9
MART [21]	10.3	15.7	23.4	**5.2**
TextKG	**11.3**	**16.5**	**26.6**	6.3

Table 3. Evaluation results on the Activity-Net Captions *ae-val* subset in the paragraph-level evaluation mode.

Meanwhile, we also compare the TextKG method to the SOTA methods focusing on multi-modality pretraining models in Table 5. As shown in Table 5, our TextKG method performs favorably against existing methods, such as DECEM [48], UniVL [29] and MV-GPT [40], that are pretrained on the large-scale datasets (*e.g.*, Howto100M [31]), *i.e.*, improve 0.7 absolute CIDEr score compare to the SOTA method LAVENDER [23].

MSVD dataset. We also evaluate the proposed TextKG method on the MSVD dataset [3] that doesn't have speech and repoprt the results in Table 6. As shown in Table 6, our TextKG achieves the best results without speech transcripts by improving 1.2% and 3.8% CIDEr scores compared to the SOTA method HMN [58] and the JCRR method [14]

Method	B	M	R	C
OA-BTG [61]	41.4	28.2	-	46.9
POS-CG [51]	42	28.2	61.6	48.7
MGSA [4]	42.4	27.6	-	47.5
STG-KD [32]	40.5	28.3	60.9	47.1
ORG-TRL [64]	43.6	28.8	62.1	50.9
SGN [38]	40.8	28.3	60.8	49.5
MGRMP [5]	41.7	28.9	62.1	51.4
HMN [58]	43.5	29	**62.7**	51.5
TextKG	**43.7**	**29.6**	62.4	**52.4**

Table 4. Evaluation results on the MSR-VTT test subset in the micro-level evaluation mode.

Method	Input	Features	B	M	R	C
SWINBERT [27]	V	VidSwin	45.4	30.6	64.1	55.9
CLIP4C [47]	V	CLIP	46.1	30.7	63.7	57.7
CMVC [57]	V	CLIP	**48.2**	**31.3**	**64.8**	58.7
LAVENDER [23]	V	VidSwin	-	-	-	60.1
DeCEM [48]	V+S	BERT	45.2	29.7	64.7	52.3
UniVL [29]	V+S	S3D	41.8	28.9	60.8	50.0
MV-GPT [40]	V+S	ViViT	48.9*	38.7*	64.0	60.0
TextKG (CLIP)	V+S	CLIP	46.6	30.5	**64.8**	**60.8**

Table 5. Comparison to the methods focusing on model pretraining on the MSR-VTT dataset in the micro-level evaluation mode. Notably, following CLIP4C, we use the pre-trained CLIP model on LAION-400M [39] to extract video appearance features. $*$ The authors use a different library to compute BLEU and METOR. Thus, the results on BLEU and METOR are not directly comparable to other methods.

Method	B	M	R	C
OA-BTG [61]	56.9	36.2	-	90.6
POS-CG [51]	52.5	34.1	71.3	88.7
MGSA [4]	53.4	35	-	86.7
STG-KD [32]	52.2	36.9	73.9	93.0
ORG-TRL [64]	54.3	36.4	73.9	95.2
SGN [38]	52.8	35.5	72.9	94.3
MGRMP [5]	55.8	36.9	74.5	98.5
JCRR [14]	57.0	36.8	-	96.8
HMN [58]	59.2	37.7	**75.1**	104.0
TextKG	**60.8**	**38.5**	**75.1**	**105.2**

Table 6. Evaluation results on the MSVD test subset in the micro-level evaluation mode.

exploiting knowledge graph to model the relations between objects, demonstrating the effectiveness of the knowledge graph usage in our method for video captioning.

4.5. Ablation study

Influence of different modules. To validate the effectiveness of different modules in TextKG, we conduct several experiments on the YouCookII dataset in Table 7. As

V-F	R-F	Text	G-KG	S-KG	K-S	B	M	C	Rep
✓						7.4	15.7	32.1	4.1
✓	✓					9.5	17.7	45.9	5.2
✓	✓		✓		✓	9.7	17.8	48.9	4.2
✓	✓			✓	✓	9.7	18.0	48.5	3.3
✓	✓		✓	✓	✓	9.6	17.7	49.8	5.5
✓		✓				13.0	21.2	62.5	2.7
✓	✓	✓				13.9	**22.1**	71.3	2.9
✓	✓	✓	✓		✓	13.7	22.0	73.5	**2.0**
✓	✓	✓		✓	✓	13.5	21.9	74.8	2.1
✓	✓	✓	✓	✓		13.7	21.8	72.0	2.8
✓	✓	✓	✓	✓	✓	**14.0**	**22.1**	**75.9**	2.8

Table 7. Ablation study on the YouCookII val subset in the paragraph-level evaluation mode. "V-F" and "R-F" indicate the video and region features, "Text" indicates the speech transcripts features, "G-KG" and "S-KG" indicate general and specific knowledge graph, and "K-S" indicates the knowledge selection mechanism.

shown in Table 7, the baseline method considers the information of video features and predicted captions, achieving 32.1 CIDEr score. After integrating the speech transcripts, the CIDEr score is improved to 62.5, which ensures the key information of videos is considered in generating video captions. After integrating the region features, TextKG achieves 71.3 CIDEr score. The CIDEr score can be further improved 4.6 CIDEr score to 75.9 by exploiting additional knowledge in knowledge graph. Notably, even without the information of speech transcripts, the knowledge graphs are still crucial for video captioning, *i.e.*, improving 3.9 CIDEr score (49.8 *vs.* 45.9). Meanwhile, we also demonstrate the effectiveness of knowledge selection mechanism in Table 7. As shown in Table 7, without the knowledge selection mechanism, the CIDEr score significantly drops 3.9 to 72.0. We believe that the noisy knowledge is harmful to the accuracy of video captions.

Influence of transcript quality. To explore the effect of transcript quality on the knowledge graph, we add some noises to the speech transcripts (*i.e.*, mask some amount of speech transcripts) and evaluate the performance of our models on the YouCookII dataset in Figure 3. We find that the method using both the general and specific knowledge graphs performs favorable against other methods with different degrees of masked speech transcripts. Meanwhile, as the amount of masked speech transcripts increasing, the accuracy of the method using specific knowledge graph drops sharply, while the accuracy of the method using general knowledge graph drops at a much slower pace. The aforementioned results demonstrate the importance of general knowledge graph for the video captioning task.

4.6. Qualitative Results

We present the qualitative results of the proposed method on the YouCookII dataset in Figure 2, including the key

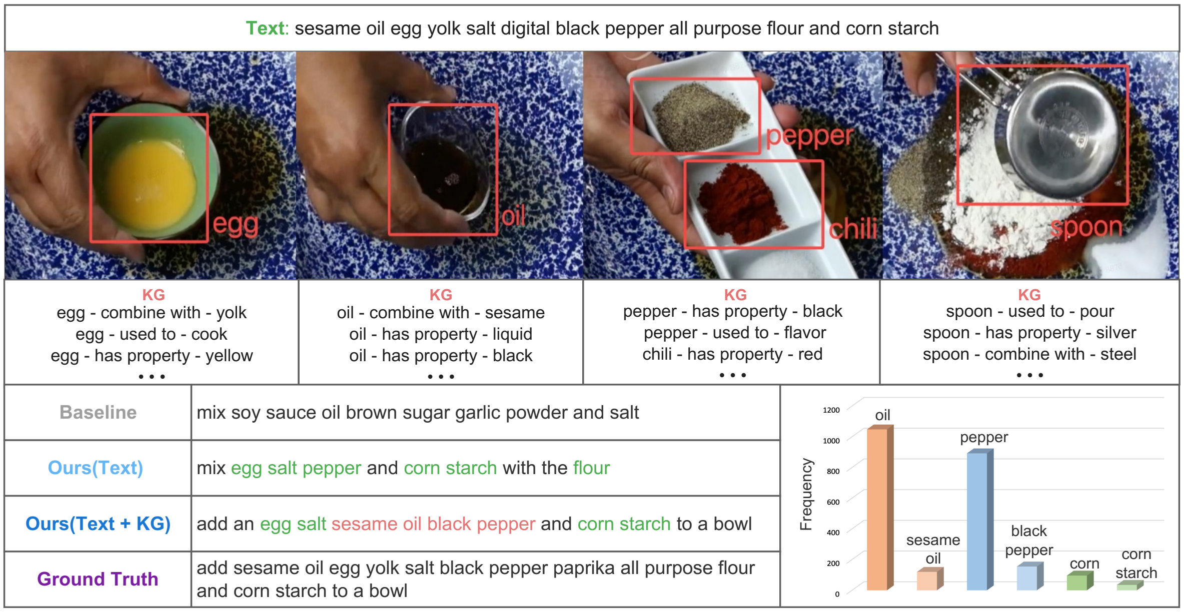

Figure 2. Qualitative results of the proposed method on the YouCookII dataset. The information in the speech transcripts and knowledge graph is not considered in the baseline.

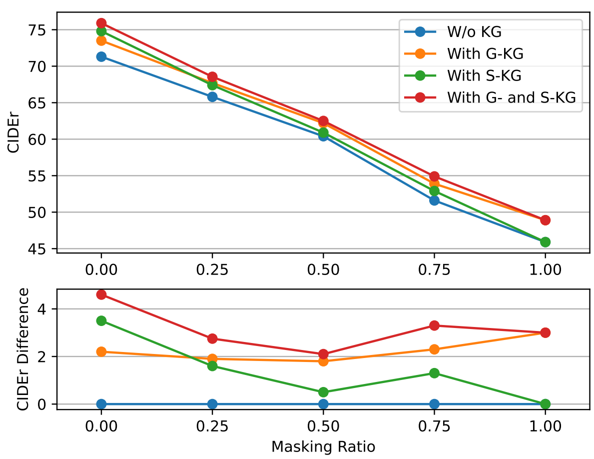

Figure 3. Influence of the transcripts qualities. "w/o KG" indicates that the method does not use the general and specific knowledge graphs. "with G-KG" and "with S-KG" indicate that only the general and specific knowledge graphs are used, respectively. "with G- and S-KG" indicates that both the general and specific knowledge graphs are used. The CIDEr differences between the "with G-KG", "with S-KG", "with G- and S-KG" methods and the "w/o KG" method are shown at the bottom corner.

frames, detected salient objects, speech transcripts, retrieved knowledge items, caption results predicted by our model, ground-truth captions, and key words distributions in the training set. As shown in Figure 2, we observe that the transcript contains key information in the video, *e.g.*, egg, salt, pepper, com starch and flour, to ensure the quality of the generated video captions. However, the model with only the speech transcripts fails to predict the long-tail phrases, such as sesame oil and black pepper. With the help of knowledge graph providing the relations among these phrases, *e.g.*, sesame oil, black pepper and corn starch, our TextKG model can predict the phrases more accurately. This indicates that the knowledge graph is crucial to mitigate the long-tail word challenges in video captioning.

5. Conclusion

In this paper, we present a text with knowledge graph augmented transformer for video captioning, which aims to integrate external knowledge in knowledge graph and exploit the multi-modality information in video to mitigate long-tail words challenge. Externsive experiments conducted on four challenging datasets demonstrate the effectiveness of the proposed method.

In the future, we plan to improve the proposed method in two directions, *i.e.*, (1) optimizing the knowledge retrieve strategy by considering semantic context information of detected objects and corresponding actions in videos; (2) constructing multi-modal knowledge graph (*e.g.*, the nodes in knowledge graph formed by text, speech, and images or videos) to improve the accuracy of video captioning.

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for

end-to-end retrieval. In *IEEE/CVF International Conference on Computer Vision*, pages 1708–1718, 2021. 2

[2] William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. Listen, attend and spell. 2015. 2, 5

[3] David L. Chen and William B. Dolan. Collecting highly parallel data for paraphrase evaluation. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *Annual Meeting of the Association for Computational Linguistics*, 2011. 2, 4, 5, 6

[4] Shaoxiang Chen and Yu-Gang Jiang. Motion guided spatial attention for video captioning. In *Association for the Advancement of Artificial Intelligence*, pages 8191–8198, 2019. 7

[5] Shaoxiang Chen and Yu-Gang Jiang. Motion guided region message passing for video captioning. In *IEEE/CVF International Conference on Computer Vision*, pages 1523–1532, 2021. 5, 6, 7

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: learning universal image-text representations. *CoRR*, 2019. 2

[7] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc Viet Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, 2019. 6

[8] Michael J. Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, 2014. 5

[9] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. COOT: cooperative hierarchical transformer for video-text representation learning. In *Advances in Neural Information Processing Systems*, 2020. 5, 6

[10] Xin Gu, Hanhua Ye, Guang Chen, Yufei Wang, Libo Zhang, and Longyin Wen. Dual-stream transformer for generic event boundary captioning. 2022. 2

[11] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 4

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4

[13] Jack Hessel, Bo Pang, Zhenhai Zhu, and Radu Soricut. A case study on combining ASR and visual features for generating instructional video captions. In *Conference on Computational Natural Language Learning*, pages 419–429, 2019. 6

[14] Jingyi Hou, Xinxiao Wu, Xiaoxun Zhang, Yayun Qi, Yunde Jia, and Jiebo Luo. Joint commonsense and relation reasoning for image and video captioning. In *Association for the Advancement of Artificial Intelligence*, pages 10973–10980, 2020. 1, 2, 6, 7

[15] Feicheng Huang, Zhixin Li, Shengjia Chen, Canlong Zhang, and Huifang Ma. Image captioning with internal and external knowledge. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 535–544. ACM, 2020. 2

[16] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis R. Bach and David M. Blei, editors, *International Conference on Machine Learning*, pages 448–456, 2015. 4

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *International Conference on Learning Representations*, 2015. 4, 5

[18] Bruno Korbar, Fabio Petroni, Rohit Girdhar, and Lorenzo Torresani. Video understanding as machine translation. *CoRR*, abs/2006.07203, 2020. 6

[19] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *IEEE/CVF International Conference on Computer Vision*, pages 706–715, 2017. 1, 2, 4, 5, 6

[20] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 4, 5

[21] Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara L. Berg, and Mohit Bansal. MART: memory-augmented recurrent transformer for coherent video paragraph captioning. In *Annual Meeting of the Association for Computational Linguistics*, pages 2603–2614, 2020. 1, 2, 3, 4, 5, 6

[22] Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara L. Berg, and Mohit Bansal. MART: memory-augmented recurrent transformer for coherent video paragraph captioning. In *Annual Meeting of the Association for Computational Linguistics*, pages 2603–2614, 2020. 6

[23] Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. LAVENDER: unifying video-language understanding as masked language modeling. 2022. 6, 7

[24] Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohit Pillai, Yu Cheng, Luowei Zhou, Xin Wang, William Yang Wang, Tamara L. Berg, Mohit Bansal, Jingjing Liu, Lijuan Wang, and Zicheng Liu. VALUE: A multi-task benchmark for video-and-language understanding evaluation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. 6

[25] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. UNIMO: towards unified-modal understanding and generation via cross-modal contrastive learning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *ACL-IJCNLP*, pages 2592–2607, 2021. 2

[26] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*, pages 74–81, 2004. 5

[27] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. Swinbert: End-to-end transformers with sparse attention for video captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17928–17937, 2022. 6, 7

[28] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems*, pages 13–23, 2019. 2

[29] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Xilin Chen, and Ming Zhou. Univilm: A unified video and language pre-training model for multimodal understanding and generation. *CoRR*, abs/2002.06353, 2020. 1, 2, 6, 7

[30] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In *IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019. 1

[31] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019. 2, 6

[32] Boxiao Pan, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Juan Carlos Niebles. Spatio-temporal graph for video captioning with knowledge distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10867–10876, 2020. 7

[33] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 5

[34] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014. 4, 5

[35] Tanzila Rahman, Bicheng Xu, and Leonid Sigal. Watch, listen and tell: Multi-modal weakly supervised dense event captioning. In *IEEE/CVF International Conference on Computer Vision*, pages 8907–8916, 2019. 2

[36] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *EMNLP-IJCNLP*, pages 3980–3990, 2019. 4

[37] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. pages 1137–1149, 2017. 4

[38] Hobin Ryu, Sunghun Kang, Haeyong Kang, and Chang D. Yoo. Semantic grouping network for video captioning. In *Association for the Advancement of Artificial Intelligence*, pages 2514–2522, 2021. 1, 5, 6, 7

[39] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: open dataset of clip-filtered 400 million image-text pairs. *CoRR*, 2021. 7

[40] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17938–17947, 2022. 1, 2, 6, 7

[41] Botian Shi, Lei Ji, Yaobo Liang, Nan Duan, Peng Chen, Zhendong Niu, and Ming Zhou. Dense procedure captioning in narrated instructional videos. In *Annual Meeting of the Association for Computational Linguistics*, pages 6382–6391, 2019. 1, 6

[42] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Association for the Advancement of Artificial Intelligence*, pages 4444–4451, 2017. 5

[43] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *IEEE/CVF International Conference on Computer Vision*, pages 7463–7472, 2019. 2, 6

[44] Tianxiang Sun, Yunfan Shao, Xipeng Qiu, Qipeng Guo, Yaru Hu, Xuanjing Huang, and Zheng Zhang. Colake: Contextualized language and knowledge embedding. In *International Conference on Computational Linguistics*, pages 3660–3670, 2020. 2

[45] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In Satinder Singh and Shaul Markovitch, editors, *Association for the Advancement of Artificial Intelligence*, 2017. 4

[46] Hao Tan and Mohit Bansal. LXMERT: learning cross-modality encoder representations from transformers. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *EMNLP-IJCNLP*, pages 5099–5110, 2019. 2

[47] Mingkang Tang, Zhanyu Wang, Zhenhua Liu, Fengyun Rao, Dian Li, and Xiu Li. Clip4caption: CLIP for video caption. In *ACM International Conference on Multimedia*, pages 4858–4862, 2021. 7

[48] Zineng Tang, Jie Lei, and Mohit Bansal. Decembert: Learning from noisy instructional videos via dense captions and entropy minimization. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 2415–2426, 2021. 1, 6, 7

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 3

[50] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, 2015. 5

[51] Bairui Wang, Lin Ma, Wei Zhang, Wenhao Jiang, Jingwen Wang, and Wei Liu. Controllable video captioning with POS sequence guidance based on gated fusion network. In *IEEE/CVF International Conference on Computer Vision*, pages 2641–2650, 2019. 7

[52] Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. KEPLER: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194, 2021. 2

[53] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning for video understanding. *CoRR*, abs/1712.04851, 2017. 6

[54] Yilei Xiong, Bo Dai, and Dahua Lin. Move forward and tell: A progressive generator of video descriptions. In *European Conference on Computer Vision*, pages 468–483, 2018. 5

[55] Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. VLM: task-agnostic video-language model pre-training for video understanding. In *Findings of the Association for Computational Linguistics*, pages 4227–4239, 2021. 2

[56] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5288–5296, 2016. 2, 4, 5

[57] Bang Yang, Tong Zhang, and Yuexian Zou. CLIP meets video captioning: Concept-aware representation learning does matter. In *Pattern Recognition and Computer Vision - 5th Chinese Conference, PRCV 2022, Shenzhen, China, November 4-7, 2022, Proceedings, Part I*, 2022. 7

[58] Hanhua Ye, Guorong Li, Yuankai Qi, Shuhui Wang, Qingming Huang, and Ming-Hsuan Yang. Hierarchical modular network for video captioning. pages 17939–17948, 2022. 1, 4, 5, 6, 7

[59] Donghan Yu, Chenguang Zhu, Yiming Yang, and Michael Zeng. JAKET: joint pre-training of knowledge graph and language understanding. In *Association for the Advancement of Artificial Intelligence*, pages 11630–11638, 2022. 2

[60] Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and hierarchical modeling of video and text. In *European Conference on Computer Vision*, pages 385–401, 2018. 6

[61] Junchao Zhang and Yuxin Peng. Object-aware aggregation with bidirectional temporal graph for video captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8327–8336, 2019. 7

[62] Yu Zhang, Xinyu Shi, Siya Mi, and Xu Yang. Image captioning with transformer and knowledge graph. *Pattern Recognition Letters*, 143:43–49, 2021. 2

[63] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. ERNIE: enhanced language representation with informative entities. In *Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, 2019. 2

[64] Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. Object relational graph with teacher-recommended learning for video captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13275–13285, 2020. 2, 4, 5, 6, 7

[65] Wentian Zhao, Yao Hu, Heda Wang, Xinxiao Wu, and Jiebo Luo. Boosting entity-aware image captioning with multi-modal knowledge graph. *CoRR*, 2021. 2

[66] Qi Zheng, Chaoyue Wang, and Dacheng Tao. Syntax-aware action targeting for video captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13093–13102, 2020. 4, 5, 6

[67] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J. Corso, and Marcus Rohrbach. Grounded video description. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6578–6587, 2019. 6

[68] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and VQA. In *Association for the Advancement of Artificial Intelligence*, pages 13041–13049, 2020. 2

[69] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In *Association for the Advancement of Artificial Intelligence*, pages 7590–7598, 2018. 1, 2, 4, 5

[70] Luowei Zhou, Yingbo Zhou, Jason J. Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8739–8748, 2018. 6

[71] Yimin Zhou, Yiwei Sun, and Vasant G. Honavar. Improving image captioning by leveraging knowledge graphs. In *IEEE Winter Conference on Applications of Computer Vision*, pages 283–293, 2019. 1, 2

[72] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8743–8752, 2020. 1, 2, 6