Visual Consensus Modeling for Video-Text Retrieval

Shuqiang Cao,*[1] Bairui Wang,*[2] Wei Zhang,†[1] Lin Ma†[2]

[1]School of Control Science and Engineering, Shandong University
[2]Meituan
sqiangcao@mail.sdu.edu.cn, {bairuiwong, forest.linma}@gmail.com, davidzhang@sdu.edu.cn

Abstract

In this paper, we propose a novel method to mine the commonsense knowledge shared between the video and text modalities for video-text retrieval, namely visual consensus modeling. Different from the existing works, which learn the video and text representations and their complicated relationships solely based on the pairwise video-text data, we make the first attempt to model the visual consensus by mining the visual concepts from videos and exploiting their co-occurrence patterns within the video and text modalities with no reliance on any additional concept annotations. Specifically, we build a shareable and learnable graph as the visual consensus, where the nodes denoting the mined visual concepts and the edges connecting the nodes representing the co-occurrence relationships between the visual concepts. Extensive experimental results on the public benchmark datasets demonstrate that our proposed method, with the ability to effectively model the visual consensus, achieves state-of-the-art performance on the bidirectional video-text retrieval task. Our code is available at https://github.com/sqiangcao99/VCM.

Introduction

As a meaningful but challenging task for bridging vision and language, the video-text retrieval task aiming to match video and text has been drawing more and more attention under the rapid development of the Internet and the increasing number of videos. It can be applied in various practical applications, such as video search engine that returns the relevant videos by the input text queries. Besides, this task could benefit many downstream tasks, such as video caption (Tan et al. 2020; Zhang et al. 2019; Wang et al. 2018) and video temporal grounding (Wang, Ma, and Jiang 2020).

Although great progress has been made for the video-text retrieval task over the past few years, the semantic gap between video and text still remains a significant challenge. Some methods learn the video and text feature in a common space to enforce the correlated video and text being closer to each other (Dong et al. 2019; Liu et al. 2019; Mithun et al. 2018; Song and Soleymani 2019). To further exploit the complicated relationships between video and text, some methods perform fine-grained matching of the visual and textual features at different representation levels (Chen et al. 2020; Wang, Zhu, and Yang 2021).

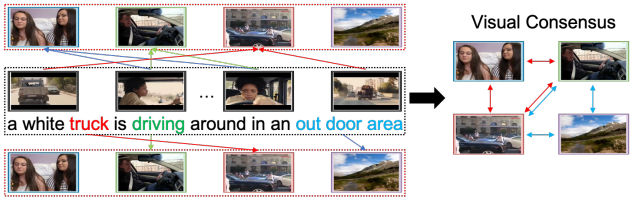

Figure 1: The proposed visual consensus modeling method based on the video-text pair data. The video clips within the red dotted rectangles are the extracted visual concepts from a large scale of video sequences. Given a video-text pair, it can be observed that co-occurrence patterns of the visual concepts (denoted by arrows with different colors) are different in the video (top row) and text (bottom row) modality. With the consideration of visual consensus in both video and text modalities, the constructed visual consensus graph is expected to contain much richer commonsense knowledge and help boost the performance of video-text retrieval task.

Although these methods achieve more effective video-text alignment and competitive performance, only some limited internal information in pairwise video-text data is exploited. The external information such as commonsense knowledge, which is hidden in vision and language, believed as an essential complement for the cross-modality retrieval task, has not been considered. To explore commonsense knowledge, Wang *et al*. (Wang et al. 2020a) propose the CVSE to model the co-occurrence relationships of textual concepts and demonstrate that commonsense knowledge benefits the image-text retrieval task. However, the co-occurrence relationships in the vision domain are ignored.

With the reservoir of commonsense knowledge, humans can handle multi-modal information accurately and efficiently (Wang et al. 2020a). Since vision is the most important channel through which people obtain information, visual commonsense knowledge has become a major component of commonsense knowledge for people (Hutmacher 2019). As a basis of visual commonsense knowledge, the visual consensus describes the potential associations and

*These authors contributed equally.
†Corresponding authors

highly co-occurrence relationships among visual concepts, with which the pairwise video and text can be aligned together even if they contain different information.

Inspired by this, we make the first attempt to build commonsense knowledge from the videos to narrow the semantic gap between vision and language, and a novel visual consensus model for video-text retrieval is proposed in this paper. As shown in Fig. 1, based on the video-text pairs as well as the extracted visual concepts, the visual consensus in both video and text modalities is established. Combining the visual consensus information, our model can produce richer commonsense knowledge, which helps boost the performance of the video-text retrieval task. Specifically, we first employ the spherical K-means method (Hornik et al. 2012) on the video frame representations to yield the centers of all clusters, which are regarded as the visual concept representations and utilized to tokenize the videos and sentences. Afterwards, based on the visual concept representations, a consensus graph for visual consensus modeling is constructed, where the nodes denoted as visual concepts are learnable and initialized by the extracted visual concept representations, and the edges connecting nodes are obtained by analyzing the co-occurrence correlations of visual concepts in all the tokenized videos and sentences. Finally, the visual and textual representations are yielded by referring to the learned consensus graph, which is utilized to perform the video-text retrieval task. The differences between our method and the CVSE (Wang et al. 2020a) lie in three-fold: 1) The concepts from the visual domain are modeled, which provides much richer information than the textual domain in CVSE. 2) A novel approach to exploiting the co-occurrence relationship of visual concepts in different modalities is brought up by replacing word representations with visual concept representations and then calculating co-occurrence frequency. 3) A hybrid graph that incorporates multi-modalities consensus is designed, while only textual consensus is considered in CVSE.

To summarize, the contributions of this work lie in threefold: 1) We propose a novel visual consensus modeling framework, which relies on a learnable multi-model consensus graph exploring the commonsense information in both vision and language domain to narrow the their semantic gap. 2) We model the co-occurrence relationships of visual concepts in vision and textual modalities, respectively, and exploit their co-occurrence relationships to improve the performance of video-text retrieval. 3) Extensive results on benchmark datasets indicate that the proposed visual consensus modeling method can fully utilize the commonsense information to improve the cross-modal retrieval performance.

Related Works

Knowledge Based Learning

The key to human wisdom is the absorption and utilization of knowledge, based on which many knowledge-based approaches have been proposed for various deep learning tasks (Deng et al. 2014; Gu et al. 2019; Marino, Salakhutdinov, and Gupta 2016; Wang et al. 2017; Yu et al. 2019; Shi et al. 2019; Wang et al. 2020a,b; Fang et al. 2020). Commonsense can be expressed in many ways. Wang *et al.* (Wang et al. 2020a) propose a framework to mine the co-occurrence relationships among words as consensus-aware concept embeddings for image-text retrieval. Furthermore, Wang *et al.* (Wang et al. 2020b) utilized causal reasoning to solve object detection. This work also belongs to knowledge-based learning but mines the consensus knowledge in both visual and textual domains, which is different from the existing methods and gains semantically richer commonsense knowledge.

Video-Text Retrieval

Most existing video-text retrieval frameworks (Wang, Zhu, and Yang 2021; Portillo-Quintero, Ortiz-Bayliss, and Terashima-Marín 2021; Luo et al. 2021; Liu et al. 2021a; Chen et al. 2020; Mithun et al. 2018; Wang, Zhu, and Yang 2021; Liu et al. 2019; Dzabraev et al. 2021; Lei et al. 2021) focus on constructing meaningful representations for video and text, which contain essential information in their respective modalities, such as motion information for video and the internal relevance of part-of-speech for text. These representations are embedded in a shared space and matched according to their similarity metric. Dong *et al.* (Dong et al. 2021) encode videos by CNN and texts by bi-GRU, and employ mean pooling to get multi-levels representations. Chen *et al.* (Chen et al. 2020) propose the hierarchical graph reasoning model, which solves the video-text retrieval task using a global-local method by decomposing the texts into events, actions, and entities. Some methods introduce multi-modal features extracted from videos for efficient retrieval, such as motion and audio features (Liu et al. 2019; Mithun et al. 2018; Wang, Zhu, and Yang 2021). Recently, the pre-trained models (Amrani et al. 2020; Luo et al. 2020; Lei et al. 2021; Dzabraev et al. 2021; Liu et al. 2021b; Portillo-Quintero, Ortiz-Bayliss, and Terashima-Marín 2021; Luo et al. 2021) bring significant performance improvements over previous models. Portillo *et al.* (Portillo-Quintero, Ortiz-Bayliss, and Terashima-Marín 2021) adopt CLIP (Radford et al. 2021), an image-text pre-trained model, for zero-shot video-text retrieval, and Luo *et al.* (Luo et al. 2021) build an end-to-end model based on CLIP and explore several similarity calculation methods.

Method

Given a video $\mathrm{V} = \{\mathrm{v}_1, \mathrm{v}_2, ..., \mathrm{v}_n\}$ or a sentence $\mathrm{S} = \{\mathrm{s}_1, \mathrm{s}_2, ..., \mathrm{s}_m\}$, where n and m respectively denote the number of frames in a video and the number of words in a sentence, video-text retrieval task aims to find the most relevant sentence or video. In this paper, we propose the visual consensus modeling(VCM) framework which consists of a cross-modal knowledge learning(CKL) module and a knowledge integration(KI) module to exploit the commonsense knowledge hidden in videos and sentences and narrow the semantic gap between vision and language, as shown as Fig. 2. The CKL module learns to extract visual concepts in videos and incorporates their co-occurrence relationships in videos and sentences to obtain visual consensus representa-

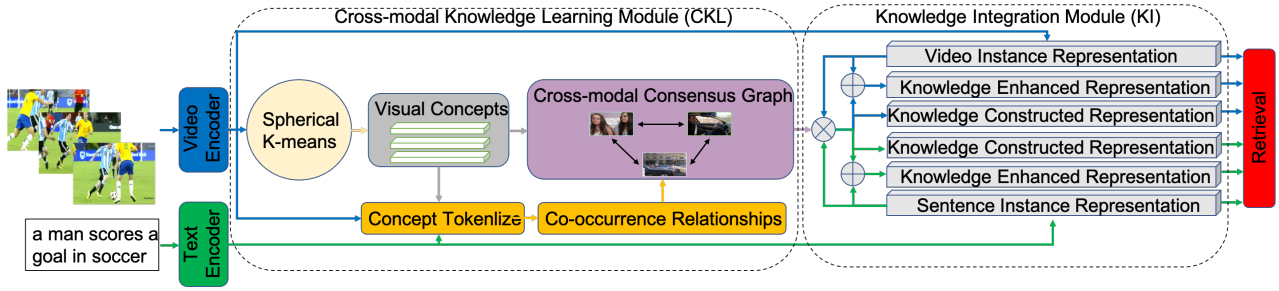

Figure 2: Overview of the proposed VCM method for video-text retrieval. The VCM consists of CKL module and KI module. Firstly, the instance representations of frames and words are extracted by the CLIP model. Then the CKL module learns the visual concepts from frame instance representations and embeds the consensus information of visual and textual modalities into the consensus representations by a cross-modal consensus graph. Finally, the instance representations and the consensus representations are integrated together by the KI module for the video-text retrieval, where $\otimes$ denotes the attention module and $\oplus$ denotes the weighted sum.

tions which have been embedded in the consensus information of both visual and textual modalities. The KI module aims to incorporate commonsense knowledge into the video and text representations for the final similarity calculation.

Cross-modal Knowledge Learning Module

Visual Concept Extracting To extract the visual concept representations, we first employ the CLIP model (Radford et al. 2021) that pretrained on the image-text retrieval task as the video encoder to extract the frame instance representations for all frames in a video, which are denoted as $\mathbf{V} = \{v_1, v_2, ..., v_n\}$, where n denotes the number of frames in this video and v_n denotes the instance representation of the n-th frame. The CLIP model is text-dependent and considers not only the visual appearance of frames but also the semantic meanings of textual descriptions, which helps to further enrich the consensus information in visual concepts by mining the co-occurrence relationships in words. Afterwards, we apply the spherical K-means method (Dhillon and Modha 2001) on all of the frame instance representations, and k cluster center vectors which are considered as the visual concept representations are generated and denoted as $\mathbf{C} = \{c_1, c_2, c_3, ..., c_k\}$, where c_k denotes the k-th visual concept representation.

Consensus Knowledge Extracting As shown in Fig. 3, to obtain the consensus relationships of visual concepts in vision modality, we tokenize each video by calculating the cos-similarity between the frame instance representations and the visual concept representations. In this case, a tokenized video can be expressed as $\mathbf{V^c} = \{c_v^0, c_v^1, \ldots, c_v^n\}$, where $c_v^n \in \mathbf{C}$ denotes visual concept representation that replaces the n-th frame instance representation v_n in a video.To further narrow the semantic gap in video and text, the sentences are first embedded to word instance representations $\mathbf{S} = \{s_1, s_2, ..., s_m\}$ by the Transformer and then tokenized as $\mathbf{S^c} = \{c_s^0, c_s^1, \ldots, c_s^m\}$ in the same way as the video tokenization mentioned above, where $c_s^m \in \mathbf{C}$ denotes visual concept representation that replaces the m-th word instance representation s_m in a sentence.

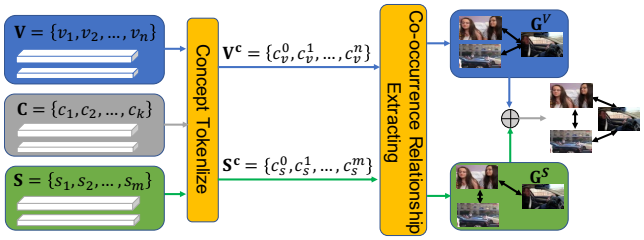

Figure 3: Illustration of the Consensus Knowledge Extracting. First, the videos ($\mathbf{V}$) and sentences ($\mathbf{S}$) are tokenized by the visual concept representations ($\mathbf{C}$). Afterwards, the co-occurrence relationships in video and text corpora are counted, with which a cross-modal consensus graph containing the consensus knowledge is constructed.

Afterwards, we formalize the co-occurrence relationship of visual concepts in video and text corpora as a graph to obtain the cross-modal commonsense knowledge. Specifically, we utilize conditional probabilities $\mathbf{P}_{ij}^M$ to model the co-occurrence relationships in visual concepts:

$$\mathbf{P}_{ij}^M = \frac{\mathbf{E}_{ij}^M}{N_i^M}, \tag{1}$$

where $M \in \{\mathbf{V}, \mathbf{S}\}$ represents the video and text modality and i, j represent c_i and c_j in $\mathbf{V^c}$ or $\mathbf{S^c}$, respectively. $\mathbf{E}_{ij}^M$ represents co-occurrence times of visual concept representation c_i and c_j in video or text corpus, and N_i^M indicates the occurrence times of visual concept c_i. Subsequently, to overcome the bias caused by the long-tail distribution of video concepts, we adopt the scale function (Wang et al. 2020a) denoted as $f_{CS}(\cdot)$ to rescale the probability:

$$\tilde{\mathbf{P}}_{ij}^M = f_{CS}\left(\mathbf{P}_{ij}^M\right), \tag{2}$$

where $\tilde{\mathbf{P}}_{ij}^M$ is the rescaled probability. Besides, to extenuate the aligning errors of the CLIP between the video and text, the binary operation (Chen et al. 2019) is employed to the rescaled co-occurrence probability $\tilde{\mathbf{P}}_{ij}^M$:

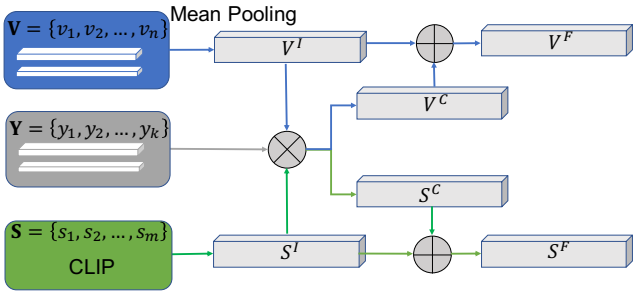

Figure 4: Illustration of the Knowledge Integration module(KI), where the superscripts I, F, and C of videos(V) and sentences(S) are denote as instance representation, knowledge enhanced representation and knowledge constructed representation, respectively. The KI module explores the consensus information related to instance information of the video or sentence, and integrates them with the instance information together.

$$\mathbf{G}_{ij}^{M} = \begin{cases} 0 & \text{if } \tilde{\mathbf{P}}_{ij}^{M} < \epsilon^{M} \\ 1 & \text{if } \tilde{\mathbf{P}}_{ij}^{M} \geq \epsilon^{M} \end{cases}, \tag{3}$$

where $\mathbf{G}_{ij}^{M} \in \{\mathbf{G}^{\mathbf{V}}, \mathbf{G}^{\mathbf{S}}\}$ denotes the binary co-occurrence probability of visual concepts in video corpus $\mathbf{V}$ and text corpus $\mathbf{S}$, ϵ^{M} is a hyper-parameter and used to filter out unreliable co-occurrence relationships whose $\tilde{\mathbf{P}}_{ij}^{M}$ is too small to achieve the ϵ^{M}. In this work, ϵ^{M} is set as 0.3.

Thus, an initialized cross-modal graph with shared visual consensus information in video and text domains can be constructed by $\mathbf{C}$ and $\mathbf{G}^{M}$, where the nodes are visual concept representations and the edges connecting nodes are the co-occurrence relationships in both video and text.

Cross-modal Commonsense Knowledge Learning The Graph Convolutional Network(GCN) (Kipf and Welling 2016; Bruna et al. 2013) is employed on the initialized cross-modal graph to embed the co-occurrence relationships into the visual concepts. Afterwards, the nodes, that are the visual concept representations, are updated by propagating information along the edges of the graph. We take the output nodes of the GCN as the cross-modal consensus representations $\mathbf{Y} = \{y_1, y_2, y_3, ..., y_k\}$ which incorporate the interactions between visual concepts in video and text, where y_k denotes the k-th visual consensus representation.

Knowledge Integration Module

In order to integrate the instance information and related consensus information together, a KI module is introduced in this work, as shown in Fig. 4. For videos, we employ mean pooling on all the frame instance representations $\mathbf{V} = \{v_1, v_2, ..., v_n\}$ in a video and obtain the video instance representation:

$$V^{I} = \frac{1}{n}\sum_{j=1}^{n} v_j, \tag{4}$$

where V^I denotes the video instance representation of a video. For sentences, the sentence instance representation S^I is obtained from "CLS token" of the CLIP model.

Afterwards, we take video and sentence instance representations $M^I \in \{V^I, S^I\}$ as the query to gain the attention weights α_i^M with the consensus knowledge representations $\mathbf{Y}$. Thence a knowledge constructed representation M^C can be calculated via weighted summation of cross-modal consensus representations as follows:

$$\alpha_i^{M} = \frac{\exp\left(\theta M^{I} \mathbf{W}^{M} y_i^{\top}\right)}{\sum_{j=1}^{k} \exp\left(\theta M^{I} \mathbf{W}^{M} y_j^{\top}\right)}, \tag{5}$$

$$M^{C} = \sum_{i=1}^{k} \alpha_i^{M} \cdot y_i, \tag{6}$$

where $M^I \in \{V^I, S^I\}$ represents the video or sentence instance representation, W^M denotes the learnable parameter matrix, θ controls the smoothness of the softmax function. It's worth noting that M^C contains the knowledge that related to the video or sentence instance representation, even if the knowledge did not appear in them. This external knowledge can further enhance the alignment between video and text by merging instance representation and the knowledge constructed representation as the knowledge enhanced representation:

$$M^{F} = \gamma M^{I} + (1-\gamma) M^{C}, \tag{7}$$

where $M^F \in \{V^F, S^F\}$ is the knowledge enhance representation and γ is a hyper-parameter that controls the proportion of instance representation and knowledge constructed representation.

Training

During the training process, we employ the symmetric cross entropy loss (Wang et al. 2019) to train video and sentence instance representation M^I, knowledge constructed representation M^C, and knowledge enhanced representation M^F.

$$\mathcal{L}_{v2t}^{\mathcal{N}} = -\frac{1}{B}\sum_{i}^{B} \log \frac{\exp\left(\text{sim}\left(V_i^{\mathcal{N}}, S_i^{\mathcal{N}}\right)\right)}{\sum_{j=1}^{B} \exp\left(\text{sim}\left(V_i^{\mathcal{N}}, S_j^{\mathcal{N}}\right)\right)}, \tag{8}$$

$$\mathcal{L}_{t2v}^{\mathcal{N}} = -\frac{1}{B}\sum_{i}^{B} \log \frac{\exp\left(\text{sim}\left(V_i^{\mathcal{N}}, S_i^{\mathcal{N}}\right)\right)}{\sum_{j=1}^{B} \exp\left(\text{sim}\left(V_j^{\mathcal{N}}, S_i^{\mathcal{N}}\right)\right)}, \tag{9}$$

$$\mathcal{L}^{\mathcal{N}} = \mathcal{L}_{v2t}^{\mathcal{N}} + \mathcal{L}_{t2v}^{\mathcal{N}}, \tag{10}$$

where B denotes the batch size, $\mathcal{N} \in \{I, C, F\}$ represents instance representation, knowledge constructed representation and knowledge enhanced representation, and $V_i^{\mathcal{N}}$ and $S_i^{\mathcal{N}}$ denotes the representation mentioned above of the i-th video and the i-th sentence, respectively. The sim denotes the function for calculating similarity, which is the cosine

	Text → Video (T2V)					Video → Text (V2T)				
Model	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
CE (Liu et al. 2019)	20.9	48.8	62.4	6.0	28.2	20.6	50.3	64.0	5.3	25.1
MMT-Pretrained (Gabeur et al. 2020)	26.6	57.1	69.6	4.0	24.0	27.0	57.5	69.7	3.7	21.3
AVLnet (Rouditchenko et al. 2020)	27.1	55.6	66.6	4.0	-	28.5	54.6	65.2	4.0	-
SUPPORT-SET (Patrick et al. 2020)	27.4	56.3	67.7	3.0	-	26.6	55.1	67.5	3.0	-
FROZEN (Bain et al. 2021)	31.0	59.5	70.5	3.0	-	-	-	-	-	-
CLIP (Portillo-Quintero et al. 2021)	31.2	53.7	64.2	4.0	-	27.2	51.7	62.6	5.0	-
TT-CE+ (Croitoru et al. 2021)	29.6	61.6	74.2	3.0	-	32.1	62.7	75.0	3.0	-
HIT-pretrained (Liu et al. 2021b)	30.7	60.9	73.2	2.6	-	32.1	62.7	74.1	3.0	-
MDMMT (Dzabraev et al. 2021)	38.9	69.0	79.7	**2.0**	16.5	-	-	-	-	-
CLIP4Clip-meanP (Luo et al. 2021)	43.1	70.4	80.8	**2.0**	16.2	43.1	70.5	81.2	**2.0**	12.4
CLIP4Clip-seqTransf (Luo et al. 2021)	**44.5**	**71.4**	**81.6**	**2.0**	15.3	42.7	70.9	80.6	**2.0**	11.6
VCM(v+t)	43.8	71.0	80.9	**2.0**	**14.3**	**45.1**	**72.3**	**82.3**	**2.0**	**10.7**

Table 1: Comparisons of experimental results on the testing split of the MSR-VTT dataset (%).

function in this work. The losses for video-to-text and text-to-video retrieval task are denoted as $\mathcal{L}^{\mathcal{N}}_{v2t}$ and $\mathcal{L}^{\mathcal{N}}_{t2v}$, respectively, which are summed as $\mathcal{L}^{\mathcal{N}}$ for the convenience of expression, where $\mathcal{L}^{\mathcal{N}} \in \{\mathcal{L}^{I}, \mathcal{L}^{C}, \mathcal{L}^{F}\}$.

As the video and text in one pair usually contain similar information, the attention weights on the cross-modal consensus representations obtained by the video and text in one pair should have a similar distribution. Thus, the KL divergence is employed to force the information distribution of weight vectors $\mathbf{a}^V = \{\alpha_1^V, \alpha_2^V, \ldots, \alpha_k^V\}$ and $\mathbf{a}^S = \{\alpha_1^S, \alpha_2^S, \ldots, \alpha_k^S\}$ to be close to each other:

$$\mathcal{L}_{\mathcal{KL}}\left(\mathbf{a}^V \| \mathbf{a}^S\right) = \sum_{i=1}^{k} \alpha_i^V \log\left(\frac{\alpha_i^V}{\alpha_i^S}\right), \tag{11}$$

where $\mathcal{L}_{\mathcal{KL}}$ denotes the degree of information divergence. The smaller the $\mathcal{L}_{\mathcal{KL}}$ is, the more similar information distributions the $\mathbf{a}^V$ and $\mathbf{a}^S$ have.

In conclusion, the final training loss $\mathcal{L}$ is defined as follows:

$$\mathcal{L} = \lambda_1\mathcal{L}^{I} + \lambda_2\mathcal{L}^{C} + \lambda_3\mathcal{L}^{F} + \lambda_4\mathcal{L}_{\mathcal{KL}}, \tag{12}$$

where λ_1 to λ_4 are used to control the learning pace of different modules.

Inference

During the inference stage, three kinds of similarity confidences between video and sentence are calculated by instance representation, knowledge constructed representation and knowledge enhanced representation, which are denoted as Sim^I, Sim^C and Sim^F , respectively. The final similarity confidence Sim is obtained as follows:

$$Sim = \beta_1 Sim^I + \beta_2 Sim^C + \beta_3 Sim^F, \tag{13}$$

where β_1, β_2 and β_3 are trade-off parameters. The higher the Sim is, the more relevant the video and the sentence are.

Experiments

Datasets

We perform experiments on two public benchmark datasets for the video-text retrieval task, including the MSR-VTT (Xu et al. 2016) and the ActivityNet (Krishna et al. 2017).

- The **MSR-VTT** (Xu et al. 2016) dataset contains 10,000 videos and 200,000 descriptions, where each video is annotated with 20 sentences. Following the setting from (Liu et al. 2019; Miech et al. 2019; Gabeur et al. 2020; Luo et al. 2021), we use 9,000 videos for training and report results on the other 1,000 videos.
- The **ActivityNet** (Krishna et al. 2017) dataset consists of 20,000 Youtube videos with 100,000 densely annotated descriptions. Following the setting from (Zhang, Hu, and Sha 2018; Gabeur et al. 2020), we perform a video-paragraph retrieval task by concatenating all the descriptions of a video as a paragraph. Performances are reported on the "val1" split of the ActivityNet.

Metrics

We employ the standard retrieval metrics, including recall at rank *K* (R@*K*), median rank (MdR), and mean rank (MnR), to evaluate our method. The R@*K* measures the fraction of queries for which the matched samples are found among the top *K* retrieved results. The higher the R@*K* is, the better the model performs. We report R@1, R@5, and R@10 for the MSR-VTT and R@50 for the ActivityNet. The MdR/MnR measures the median/mean positions of the ground-truth results in the ranking. The lower the MdR and MnR are, the better the model performs.

Implementation Details

We employ the pretrained CLIP(ViT-B/32) as the visual and textual encoders and initialize the other parameters randomly. The length of the input sequence varies according to the average length of the sequences of the dataset. For the MSR-VTT, the frame sequence length is set to 12, and the word sequence length is set to 32, while for the ActivityNet, both the frame and word sentence lengths are set to 64. The dimensions of instance representations, visual concept representations, knowledge constructed representations and knowledge enhanced representations are set to 512. In the CKL module, the number of visual concept representations k for building the cross-modal consensus knowledge graph is set to 300, and 0.3 is assigned to ϵ^M in Eq. 3 to filter out unreliable relationships in video concepts. In

the KI module, we set θ in Eq. 5 to 10 and $\gamma = 0.85$ in Eq. 7. Besides, we set λ_1, λ_2, λ_3 and λ_4 to 1.0, 0.25, 0.0125 and 0.4 in the loss function Eq. 12, respectively. The hyper-parameters β_1, β_2 and β_3 for different types of similarity in Eq.13 are set to 0.35, 0.25, 0.40, respectively. In the experiments, the encoders are optimized by Adam (Kingma and Ba 2014), and the rest components of the model are trained by AdaDelta (Zeiler 2012). During the training stage, the batch size is set to 128, the learning rate is set to 1e-4, and the max training epoch is set to 10. All of the experiments are conducted on 4 NVIDIA Tesla V100 GPUs.

Comparison to the State of the Art

In this subsection, we compare our VCM with the state-of-the-art methods on the video-text retrieval task on the MSR-VTT and ActivityNet datasets, the experimental results of which are listed in Table 1 and Table 2, respectively, where VCM(v+t) represents that the consensus graph is constructed with the co-occurrence relationships in both video and text modalities. It can be observed from Table 1 that the proposed VCM(v+t) performs better than the other methods on most metrics. For example, compared to CLIP4Clip-meanP that transforms frame features to video features in the same way with the VCM, that is mean pooling, the VCM(v+t) increases the R@1 by 0.7 and decreases the MnR by 1.9 in T2V task. Moreover, an overall improvement on all the metrics in V2T task is achieved by VCM(v+t) on the MSR-VTT. The improvements motioned above demonstrate that the visual consensus of vision and language derive semantically rich representations for video-text retrieval task. It is worth noting that the VCM(v+t) is slightly inferior to the CLIP4Clip-seqTransf on R@1, R@5 and R@10 in the T2V task. The reason is that a transformer module is utilized in CLIP4Clip-seqTransf to further explore the temporal information of the video domain, which is a strong feature for this task but the VCM lacks. With these strong temporal features, VCM is expected to perform better. Additionally, VCM(v+t) achieves the SOTA on AcitvityNet on both T2V and V2T tasks as shown in Table 2, which indicates that VCM is more effective in boosting the retrieval performance on a larger dataset.

Ablation Studies

Study of the Consensus Information In this section, we first discuss the impact of different configurations of visual consensus graphs on the retrieval performance. We design three types of visual consensus graphs with the co-occurrence relationships in video modality only (VCM(v)), text modality only (VCM(t)), and both video and text modalities (VCM(v+t)), respectively. Besides, a graph without any co-occurrence relationship in video or text modality is also constructed (VCM(w/o)). Experimental results are shown in Table 3. Compared with CLIP4Clip-meanP, VCM(w/o) presents a slight performance improvement from 43.1 to 43.3 on R@1(T2V) and from 70.4 to 70.6 on R@5(T2V), which benefits from some commonsense information in video and text brought by the CLIP model. When we introduce the co-occurrence information in video or text modality to VCM, i.e., VCM(v) or VCM(t), the performance on both

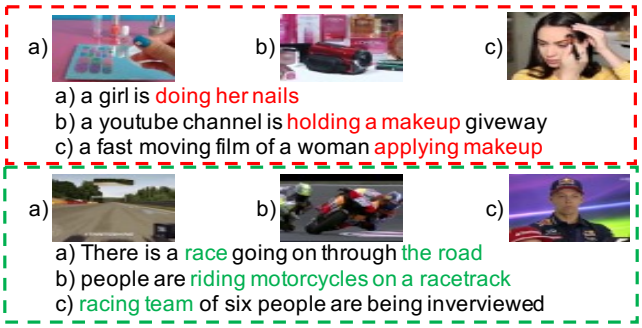

Figure 5: Visualization of the visual concepts in MSR-VTT. The red and green dashed boxes denote different visual concepts generated by employing the clustering algorithm on the CLIP-based frame instance representations. Images in red and green dashed boxes are selected from the clusters randomly, and the sentence below each image is the textual description of the video to which the image belongs.

T2V and V2T tasks are further improved. Finally, when the co-occurrence relationships of the visual concepts in video and text modalities are considered by VCM(v+t) simultaneously, a significant improvement is achieved, especially for the V2T task where R@1 is improved from 43.6 to 45.1. Such experimental results demonstrate two facts: 1) Consensus knowledge in video and text modalities are complementary, using which at the same time helps to narrow the semantic gap between the two modalities. 2) Besides the CLIP features, the proposed VCM consisting of consensus modeling and visual information clustering plays a key role in performance improvements.

Study of the Trade-off Parameters We further explore how the different hyper-parameters β_1, β_2 and β_3 in Eq. 13 affect our performance. As shown in Table 4, the results from No. (1) to (4) and No. (5) to (7) show the appropriate β_2 and β_3 that balance the effectiveness of consensus information embedded in knowledge constructed representation and knowledge enhanced representation will generate optimal performance. However, it can be observed from No. (2), (8), (9) that ignoring any one of the instance representation, knowledge constructed representation and knowledge enhanced representation will make the performance worse, which demonstrates the necessity of consensus information from VCM.

Qualitative Analysis

The Visualization of Visual Concepts To demonstrate the effectiveness of visual concepts, in this section, we visualize the visual concept structure. It can be observed in Fig. 5 that although different appearances are presented, their semantic meanings are similar, which are closely related to *makeup* and *racing*. The reason is that the CLIP model considers the visual appearance and the textual semantic meanings simultaneously, which further enriches the consensus information of visual concepts by mining the co-occurrence relationships in words.

Model	Text → Video (T2V)					Video → Text (V2T)				
	R@1	R@5	R@50	MdR	MnR	R@1	R@5	R@10	MdR	MnR
FSE (Zhang, Hu, and Sha 2018)	18.2	44.8	89.1	7.0	-	16.7	43.1	7.0	-	
CE (Liu et al. 2019)	18.2	47.7	91.4	6.0	23.1	17.7	46.6	-	6.0	24.4
HSE (Zhang, Hu, and Sha 2018)	20.5	49.3	-	-	-	18.7	48.1	-	-	-
MMT (Gabeur et al. 2020)	28.7	61.4	94.5	3.3	16.0	28.9	61.1	-	4.0	17.1
SSB (Patrick et al. 2020)	29.2	61.6	94.7	3.0	-	28.7	60.8	-	2.0	-
ClipBERT (Lei et al. 2021)	21.3	49.0	-	6.0	-	-	-	-	-	-
HiT (Liu et al. 2021b)	29.6	60.7	95.6	3.0	-	-	-	-	-	-
TT-CE+ (Croitoru et al. 2021)	23.5	57.2	96.1	4.0	-	-	-	-	-	-
CLIP4Clip-meanP (Luo et al. 2021)	40.5	72.4	98.1	**2.0**	7.4	42.5	74.1	85.8	**2.0**	6.6
CLIP4Clip-seqLSTM (Luo et al. 2021)	40.1	72.2	98.1	**2.0**	**7.3**	**42.6**	73.4	85.6	**2.0**	6.7
CLIP4Clip-seqTransf (Luo et al. 2021)	40.5	72.4	**98.2**	**2.0**	7.5	41.4	73.7	85.3	**2.0**	6.7
CLIP4Clip-tightTransf (Luo et al. 2021)	19.5	47.6	93.1	6.0	17.3	18.9	49.6	65.8	6.0	16.3
VCM(v+t)	**40.8**	**72.8**	**98.2**	**2.0**	**7.3**	**42.6**	**74.9**	**86.2**	**2.0**	**6.4**

Table 2: Comparisons of experimental results on the "val1" split of the ActivityNet dataset (%).

Model	Text → Video (T2V)					Video → Text (V2T)				
	R@1	R@5	R@50	MdR	MnR	R@1	R@5	R@10	MdR	MnR
CLIP4Clip-meanP (Luo et al. 2021)	43.1	70.4	80.8	**2.0**	16.2	43.1	70.5	81.2	**2.0**	12.4
VCM(w/o)	43.3	70.6	80.5	**2.0**	14.9	43.6	71.7	81.9	**2.0**	10.9
VCM(v)	**44.0**	70.3	80.4	**2.0**	14.9	44.3	**72.3**	82.2	**2.0**	10.8
VCM(t)	43.5	70.1	**80.9**	**2.0**	15.3	44.2	71.9	**82.6**	**2.0**	10.8
VCM(v+t)	43.8	**71.0**	**80.9**	**2.0**	**14.3**	**45.1**	**72.3**	82.3	**2.0**	**10.7**

Table 3: Effect of different configurations of consensus graph on the MSR-VTT dataset (%).

No.	β_1	β_2	β_3	R@1	R@5	R@10	MdR	MnR
(1)	0.35	0.00	0.00	43.3	70.5	79.8	**2.0**	15.5
(2)	0.35	0.00	0.25	43.3	70.5	**80.0**	**2.0**	**15.4**
(3)	0.35	0.00	0.40	**43.4**	**70.6**	**80.0**	**2.0**	**15.4**
(4)	0.35	0.00	0.45	**43.4**	70.5	79.9	**2.0**	15.5
(5)	0.35	0.20	0.40	43.5	**71.0**	80.6	**2.0**	14.5
(6)	0.35	0.25	0.40	**43.8**	**71.0**	**80.9**	**2.0**	**14.3**
(7)	0.35	0.30	0.40	43.6	70.6	**80.9**	**2.0**	**14.3**
(8)	0.35	0.25	0.00	42.5	70.2	80.7	2.0	14.5
(9)	0.00	0.25	0.40	43.0	70.7	80.3	2.0	14.4

Table 4: Experimental results of VCM(v+t) with different β_1, β_2, and β_3 on the test split of MSR-VTT for T2V task.

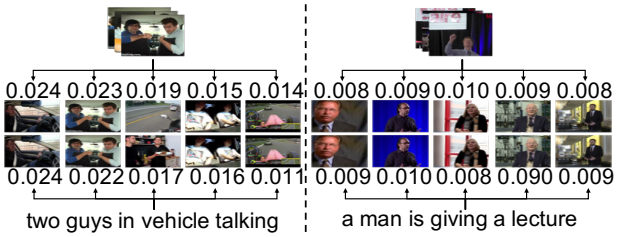

Figure 6: The visualization of the attention distributions on the top 5 visual concepts focused by video and sentence. Each frame arranged vertically represents a visual concept.

The Visualization of Attention Distributions on Visual Concepts Furthermore, we visualize the attention distributions on the visual concepts related to the video and sentence instance information. As shown in the left part of Fig. 6, even though *drive* does not appear in the textual description, concepts related to *drive* are captured by the sentence, such as the first and the fifth visual concepts on the text side. Besides, both the video and the sentence in the right part of Fig. 6 are able to capture the same relevant visual concepts, which demonstrates that the proposed VCM can narrow the semantic gap between video and text.

Conclusion

In this paper, a visual consensus modeling method is proposed to extract the visual consensus knowledge which consists of visual concepts and their consensus in both video and text modalities. The individual instance information of each video and sentence is combined with the related consensus information for video-text matching. Our proposed model achieves competitive performances on both MSR-VTT and ActivityNet datasets, which indicates the superiority of integrating consensus information for video-text retrieval task.

Acknowledgments

We gratefully acknowledge the support of the National Natural Science Foundation of China under Grants 61991411 and U1913204, the National Key Research and Development Plan of China under Grant 2018AAA0102504, the Natural Science Foundation of Shandong Province for Distinguished Young Scholars under Grant ZR2020JQ29, the Shandong Major Scientific and Technological Innovation Project 2019JZZY010428. This work is partially supported by Meituan.

References

Amrani, E.; Ben-Ari, R.; Rotman, D.; and Bronstein, A. 2020. Noise estimation using density estimation for self-supervised multimodal learning. *arXiv preprint arXiv:2003.03186*, 8.

Bain, M.; Nagrani, A.; Varol, G.; and Zisserman, A. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. *arXiv preprint arXiv:2104.00650*.

Bruna, J.; Zaremba, W.; Szlam, A.; and LeCun, Y. 2013. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*.

Chen, S.; Zhao, Y.; Jin, Q.; and Wu, Q. 2020. Fine-grained Video-Text Retrieval with Hierarchical Graph Reasoning. *CVPR*.

Chen, Z.-M.; Wei, X.-S.; Wang, P.; and Guo, Y. 2019. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5177–5186.

Croitoru, I.; Bogolin, S.-V.; Liu, Y.; Albanie, S.; Leordeanu, M.; Jin, H.; and Zisserman, A. 2021. TEACHTEXT: CrossModal Generalized Distillation for Text-Video Retrieval. *arXiv preprint arXiv:2104.08271*.

Deng, J.; Ding, N.; Jia, Y.; Frome, A.; Murphy, K.; Bengio, S.; Li, Y.; Neven, H.; and Adam, H. 2014. Large-scale object classification using label relation graphs. In *European conference on computer vision*, 48–64. Springer.

Dhillon, I. S.; and Modha, D. S. 2001. Concept decompositions for large sparse text data using clustering. *Machine learning*, 42(1): 143–175.

Dong, J.; Li, X.; Xu, C.; Ji, S.; He, Y.; Yang, G.; and Wang, X. 2019. Dual encoding for zero-example video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9346–9355.

Dong, J.; Li, X.; Xu, C.; Yang, X.; Yang, G.; Wang, X.; and Wang, M. 2021. Dual Encoding for Video Retrieval by Text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Dzabraev, M.; Kalashnikov, M.; Komkov, S.; and Petiushko, A. 2021. Mdmmt: Multidomain multimodal transformer for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3354–3363.

Fang, Z.; Gokhale, T.; Banerjee, P.; Baral, C.; and Yang, Y. 2020. Video2commonsense: Generating commonsense descriptions to enrich video captioning. *arXiv preprint arXiv:2003.05162*.

Gabeur, V.; Sun, C.; Alahari, K.; and Schmid, C. 2020. Multi-modal transformer for video retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, 214–229. Springer.

Gu, J.; Zhao, H.; Lin, Z.; Li, S.; Cai, J.; and Ling, M. 2019. Scene graph generation with external knowledge and image reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1969–1978.

Hornik, K.; Feinerer, I.; Kober, M.; and Buchta, C. 2012. Spherical k-means clustering. *Journal of statistical software*, 50: 1–22.

Hutmacher, F. 2019. Why is there so much more research on vision than on any other sensory modality? *Frontiers in psychology*, 10: 2246.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Carlos Niebles, J. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, 706–715.

Lei, J.; Li, L.; Zhou, L.; Gan, Z.; Berg, T. L.; Bansal, M.; and Liu, J. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7331–7341.

Liu, A. H.; Jin, S.; Lai, C.-I. J.; Rouditchenko, A.; Oliva, A.; and Glass, J. 2021a. Cross-Modal Discrete Representation Learning. *arXiv preprint arXiv:2106.05438*.

Liu, S.; Fan, H.; Qian, S.; Chen, Y.; Ding, W.; and Wang, Z. 2021b. Hit: Hierarchical transformer with momentum contrast for video-text retrieval. *arXiv preprint arXiv:2103.15049*.

Liu, Y.; Albanie, S.; Nagrani, A.; and Zisserman, A. 2019. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*.

Luo, H.; Ji, L.; Shi, B.; Huang, H.; Duan, N.; Li, T.; Li, J.; Bharti, T.; and Zhou, M. 2020. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*.

Luo, H.; Ji, L.; Zhong, M.; Chen, Y.; Lei, W.; Duan, N.; and Li, T. 2021. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*.

Marino, K.; Salakhutdinov, R.; and Gupta, A. 2016. The more you know: Using knowledge graphs for image classification. *arXiv preprint arXiv:1612.04844*.

Miech, A.; Zhukov, D.; Alayrac, J.-B.; Tapaswi, M.; Laptev, I.; and Sivic, J. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2630–2640.

Mithun, N. C.; Li, J.; Metze, F.; and Roy-Chowdhury, A. K. 2018. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, 19–27.

Patrick, M.; Huang, P.-Y.; Asano, Y.; Metze, F.; Hauptmann, A.; Henriques, J.; and Vedaldi, A. 2020. Support-set bottlenecks for video-text representation learning. *arXiv preprint arXiv:2010.02824*.

Portillo-Quintero, J. A.; Ortiz-Bayliss, J. C.; and Terashima-Marín, H. 2021. A straightforward framework for video retrieval using clip. In *Mexican Conference on Pattern Recognition*, 3–12. Springer.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Rouditchenko, A.; Boggust, A.; Harwath, D.; Chen, B.; Joshi, D.; Thomas, S.; Audhkhasi, K.; Kuehne, H.; Panda, R.; Feris, R.; et al. 2020. Avlnet: Learning audio-visual language representations from instructional videos. *arXiv preprint arXiv:2006.09199*.

Shi, B.; Ji, L.; Lu, P.; Niu, Z.; and Duan, N. 2019. Knowledge Aware Semantic Concept Expansion for Image-Text Matching. In *IJCAI*, volume 1, 2.

Song, Y.; and Soleymani, M. 2019. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1979–1988.

Tan, G.; Liu, D.; Wang, M.; and Zha, Z.-J. 2020. Learning to discretely compose reasoning module networks for video captioning. *arXiv preprint arXiv:2007.09049*.

Wang, B.; Ma, L.; Zhang, W.; and Liu, W. 2018. Reconstruction network for video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7622–7631.

Wang, H.; Zhang, Y.; Ji, Z.; Pang, Y.; and Ma, L. 2020a. Consensus-aware visual-semantic embedding for image-text matching. In *European Conference on Computer Vision*, 18–34. Springer.

Wang, J.; Ma, L.; and Jiang, W. 2020. Temporally grounding language queries in videos by contextual boundary-aware prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12168–12175.

Wang, P.; Wu, Q.; Shen, C.; Dick, A.; and Van Den Hengel, A. 2017. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10): 2413–2427.

Wang, T.; Huang, J.; Zhang, H.; and Sun, Q. 2020b. Visual commonsense r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10760–10770.

Wang, X.; Zhu, L.; and Yang, Y. 2021. T2vlad: global-local sequence alignment for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5079–5088.

Wang, Y.; Ma, X.; Chen, Z.; Luo, Y.; Yi, J.; and Bailey, J. 2019. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 322–330.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5288–5296.

Yu, J.; Yang, C.; Qin, Z.; Yang, Z.; Hu, Y.; and Shi, Z. 2019. Semantic modeling of textual relationships in cross-modal retrieval. In *International Conference on Knowledge Science, Engineering and Management*, 24–32. Springer.

Zeiler, M. D. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Zhang, B.; Hu, H.; and Sha, F. 2018. Cross-modal and hierarchical modeling of video and text. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 374–390.

Zhang, W.; Wang, B.; Ma, L.; and Liu, W. 2019. Reconstruct and represent video contents for captioning via reinforcement learning. *IEEE transactions on pattern analysis and machine intelligence*, 42(12): 3088–3101.