Visual Consensus Modeling for Video-Text Retrieval

Shuqiang Cao,⁎1 Bairui Wang,⁎² Wei Zhang,†¹ Lin Ma†²
¹School of Control Science and Engineering, Shandong University
²Meituan
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).

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

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.
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 semantical 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 relationships 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 three-fold: 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 semantical 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 \( V = \{v_1, v_2, ..., v_n\} \) or a sentence \( S = \{s_1, s_2, ..., s_m\} \), 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-
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 

\[ \mathbf{V} = \{v_1, v_2, \ldots, v_n\}, \]

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 

\[ \mathbf{C} = \{c_1, c_2, c_3, \ldots, c_k\}, \]

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 co-similarity between the frame instance representations and the visual concept representations. In this case, a tokenized video can be expressed as 

\[ \mathbf{V}^c = \{c^0_1, c^1_1, \ldots, c^0_n\}, \]

where \( c^0_n \in \mathbf{C} \) 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 \( \mathbf{S} = \{s_1, s_2, \ldots, s_m\} \) by the Transformer and then tokenized as 

\[ \mathbf{S}^c = \{c^0_1, c^1_1, \ldots, c^0_m\} \]

in the same way as the video tokenization mentioned above, where \( c^0_m \in \mathbf{C} \) denotes visual concept representation that replaces the \( m \)-th word instance representation \( s_m \) in a sentence.

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 \( P_{ij}^M \) to model the co-occurrence relationships in visual concepts:

\[
P_{ij}^M = \frac{E^M_{ij}}{N_i^M},
\]

where \( M \in \{V, S\} \) represents the video and text modality and \( i, j \) represent \( c_i \) and \( c_j \) in \( \mathbf{V}^c \) or \( \mathbf{S}^c \), respectively. \( E_{ij}^M \) 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:

\[
\tilde{P}_{ij}^M = f_{CS}(P_{ij}^M),
\]

where \( \tilde{P}_{ij}^M \) 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 \( \tilde{P}_{ij}^M \).
\[ V = \{v_1, v_2, ..., v_n\} \]

\[ Y = \{y_1, y_2, ..., y_k\} \]

\[ S = \{s_1, s_2, ..., s_m\} \]

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.

\[ G^M_{ij} = \begin{cases} 0 & \text{if } P^M_{ij} < \epsilon^M, \\ 1 & \text{if } P^M_{ij} \geq \epsilon^M, \end{cases} \]

where \( G^M_{ij} \in \{G^V, G^S\} \) denotes the binary co-occurrence probability of visual concepts in video corpus \( V \) and text corpus \( S \), \( \epsilon^M \) is a hyper-parameter and used to filter out unreliable co-occurrence relationships whose \( P^M_{ij} \) is too small to achieve the \( \epsilon^M \). In this work, \( \epsilon^M \) 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 \( C \) and \( G^M \), 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 \( Y = \{y_1, y_2, y_3, ..., y_k\} \) 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 \( V = \{v_1, v_2, ..., v_n\} \) in a video and obtain the video instance representation:

\[ V^I = \frac{1}{n} \sum_{j=1}^{n} v_j, \]

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 \( \alpha^M_i \) with the consensus knowledge representations \( Y \). Hence a knowledge constructed representation \( M^C \) can be calculated via weighted summation of cross-modal consensus representations as follows:

\[ \alpha^M_i = \frac{\exp (\theta M^I W^M y^I_i)}{\sum_{j=1}^{k} \exp (\theta M^I W^M y^I_j)}, \]

\[ M^C = \sum_{i=1}^{k} \alpha^M_i \cdot y_i, \]

where \( M^I \in \{V^I, S^I\} \) represents the video or sentence instance representation, \( W^M \) denotes the learnable parameter matrix, \( \theta \) controls the smoothness of the softmax function. It’s worth noting that \( M^C \) contains the knowledge 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:

\[ M^F = \gamma M^I + (1 - \gamma) M^C, \]

where \( M^F \in \{V^F, S^F\} \) is the knowledge enhance representation and \( \gamma \) 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 \( M^F \).

\[ \mathcal{L}_{v2t} = -\frac{1}{B} \sum_{i}^{B} \log \frac{\exp (\text{sim} (V^I_i, S^N_i))}{\sum_{j=1}^{B} \exp (\text{sim} (V^I_i, S^N_j))}, \]

\[ \mathcal{L}_{t2v} = -\frac{1}{B} \sum_{i}^{B} \log \frac{\exp (\text{sim} (V^N_i, S^I_i))}{\sum_{j=1}^{B} \exp (\text{sim} (V^N_i, S^I_j))}, \]

\[ \mathcal{L}^N = \mathcal{L}^N_{v2t} + \mathcal{L}^N_{t2v}, \]

where \( B \) denotes the batch size, \( N \in \{I, C, F\} \) represents instance representation, knowledge constructed representation and knowledge enhanced representation, and \( V^N_i \) and \( S^N_i \) 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 similarity.

170
function in this work. The losses for video-to-text and text-to-video retrieval task are denoted as $\mathcal{L}_{\text{v2t}}$ and $\mathcal{L}_{\text{t2v}}$, respectively, which are summed as $\mathcal{L}_{\text{N}}$ for the convenience of expression, where $\mathcal{L}_{\text{N}} \in \{\mathcal{L}_{\text{t}}, \mathcal{L}_{\text{c}}, \mathcal{L}_{\text{v}}\}$.

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 $a^V = \{a^V_1, a^V_2, \ldots, a^V_k\}$ and $a^S = \{a^S_1, a^S_2, \ldots, a^S_k\}$ to be close to each other:

$$\mathcal{L}_{KL} (a^V \| a^S) = \sum_{i=1}^{k} \alpha^V_i \log \left( \frac{\alpha^V_i}{\alpha^S_i} \right),$$

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

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

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

where $\lambda_1$ to $\lambda_4$ 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 $\text{Sim}^I$, $\text{Sim}^C$ and $\text{Sim}^F$, respectively. The final similarity confidence $\text{Sim}$ is obtained as follows:

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

where $\beta_1$, $\beta_2$ and $\beta_3$ are trade-off parameters. The higher the $\text{Sim}$ 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).

**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 $e^{w_i}$ in Eq. 3 to filter out unreliable relationships in video concepts.

---

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Text $\rightarrow$ Video (T2V)</th>
<th>Video $\rightarrow$ Text (V2T)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@1</td>
<td>R@5</td>
</tr>
<tr>
<td>CE (Liu et al. 2019)</td>
<td>20.9</td>
<td>48.8</td>
</tr>
<tr>
<td>MMT-Pretrained (Gabeur et al. 2020)</td>
<td>26.6</td>
<td>57.1</td>
</tr>
<tr>
<td>AVLNet (Rouditchenko et al. 2020)</td>
<td>27.1</td>
<td>55.6</td>
</tr>
<tr>
<td>SUPPORT-SET (Patrick et al. 2020)</td>
<td>27.4</td>
<td>56.3</td>
</tr>
<tr>
<td>FROZEN (Bain et al. 2021)</td>
<td>31.0</td>
<td>59.5</td>
</tr>
<tr>
<td>CLIP (Portillo-Quintero et al. 2021)</td>
<td>31.2</td>
<td>53.7</td>
</tr>
<tr>
<td>TT-CE+ (Croitoru et al. 2021)</td>
<td>29.6</td>
<td>61.6</td>
</tr>
<tr>
<td>HIT-pretrained (Liu et al. 2021b)</td>
<td>30.7</td>
<td>60.9</td>
</tr>
<tr>
<td>MDMMT (Dzabraev et al. 2021)</td>
<td>38.9</td>
<td>69.0</td>
</tr>
<tr>
<td>CLIP4Clip-seqTransf (Luo et al. 2021)</td>
<td>43.1</td>
<td>70.4</td>
</tr>
<tr>
<td>CLIP4Clip-seqTransf (Luo et al. 2021)</td>
<td>44.5</td>
<td>71.4</td>
</tr>
<tr>
<td>VCM(V+i)</td>
<td>43.8</td>
<td>71.0</td>
</tr>
</tbody>
</table>
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 ActivityNet 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 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 $\beta_1$, $\beta_2$ and $\beta_3$ 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 $\beta_2$ and $\beta_3$ 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.
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.


