Abstract—Generating consecutive descriptions for videos, that is, video captioning, requires taking full advantage of visual representation along with the generation process. Existing video captioning methods focus on an exploration of spatial–temporal representations and their relationships to produce inferences. However, such methods only exploit the superficial association contained in a video itself without considering the intrinsic visual commonsense knowledge that exists in a video dataset, which may hinder their capabilities of knowledge cognitive to reason accurate descriptions. To address this problem, we propose a simple, yet effective method, called visual commonsense-aware representation network (VCRN), for video captioning. Specifically, we construct a Video Dictionary, a plug-and-play component, obtained by clustering all video features from the total dataset into multiple clustered centers without additional annotation. Each center implicitly represents a visual commonsense concept in a video domain, which is utilized in our proposed visual concept selection (VCS) component to obtain a video-related concept feature. Next, a concept-integrated generation (CIG) component is proposed to enhance caption generation. Extensive experiments on three public video captioning benchmarks: MSVD, MSR-VTT, and VATEX, demonstrate that our method achieves state-of-the-art performance, indicating the effectiveness of our method. In addition, our method is integrated into the existing method of video question answering (VideoQA) and improves this performance, which further demonstrates the generalization capability of our method. The source code has been released at https://github.com/zchoi/VCRN.

Index Terms—Attention mechanism, language generation, video captioning, visual commonsense knowledge.

I. INTRODUCTION

W

ith the widespread use of mobile phones and computers, millions of videos are uploaded daily by users to sharing sites such as TikTok, YouTube, and Netflix. Thus, a powerful video captioning method is essential to automatically generate the appropriate descriptions for user-uploaded videos, which can improve the user experience. Besides, there are other broad application scenarios for video captioning, including visually impaired assistance [1], [2], online video search [3], [4], human–computer interaction [5], [6], and so on. Compared with its twin “image captioning” [7], [8] only dealing with static spatial information, video captioning tends to be more challenging since it involves both consecutive spatial and temporal representations.

The mainstream methods for video captioning follow the paradigm of an encoder–decoder framework, where the encoder employs CNNs to analyze and extract useful visual context features from the source video, and the decoder utilizes RNNs to generate the caption sequentially. One effective solution is to learn representative visual features. Toward this goal, existing methods propose a series of attention mechanisms by learning the temporal relation between video frames [9], [10], the spatial relations between objects in every single frame [11], [12], or spatial–temporal relation using appearance and motion representations [13], [14], [15].

Although the above methods have achieved remarkable progress, they focus on a source video to exploit spatial–temporal relationships to generate caption via a recurrent decoder, which still relies on learning the superficial association contained in a video itself. As a piece of external information, commonsense knowledge is considered a necessary complement to the cross-modal task [16], [17], [18], which remains underexplored. For instance, Zhang et al. [11] designs teacher-recommended learning to take full advantage of the successful external language model (ELM) to integrate rich language knowledge into the captioning model, which only exploits commonsense knowledge in the text domain. However, the commonsense knowledge in a video domain [19] is neglected.

Typically, the generated words or phrases may be described several times across different videos with analogous visual context, for example, the bottom video in Fig. 1 shows visual associations between two video samples. Nonetheless, existing methods may struggle to fully capture visual contexts in the source video due to the diverse words/phrases due to...
insufficient visual details, such as “play instruments” and “a huge crowd.” In reality, when comprehending videos, humans may associate the source video with other videos with similar visual concepts for an analogy to generate more accurate descriptions. Thus, this indicates that video captioning should have the cognitive power of visual commonsense knowledge.

In this article, we design a novel method for video captioning, called a visual commonsense-aware representation network (VCRN). Since directly modeling the relationship between the source video and all other videos will inevitably increase the computational and time cost, we design a video dictionary to summarize the co-occurrence commonsense knowledge of all videos, to explore the association between the source video and visual commonsense knowledge. Our network VCRN comprises the following three major components.

1) **Video dictionary construction (VDC)**, which aims to build commonsense knowledge from a video dataset. Specifically, we employ a K-Means algorithm on the video frame representations derived from all videos to yield a video dictionary consisting of a set of cluster centers. Each center is regarded as a visual concept representing one type of implicit commonsense knowledge.

2) **Visual concept selection (VCS)**, which aims to acquire visual commonsense knowledge related to the source video from the video dictionary. In practice, we adopt a concept-aware multhead attention (MHA) to obtain a video-related concept feature by selecting key concept information from the video dictionary guided by the source video.

3) **Concept-integrated generation (CIG)**, which is designed to enhance caption generation by exploring the relationship between the source video feature and the video-related concept feature. Such a module can provide dynamic control for the propagation of the above two types of features by a gate mechanism.

Fig. 1 shows that the proposed model can successfully generate more fine-grained and diverse words/phrases “play instruments” and “a huge crowd” because our method can capture various relevant visual information corresponding to the source video from the video dictionary.

To evaluate our proposed method, we conduct extensive experiments and analyze it on the three public video captioning benchmarks: MSVD, MSR-VTT, and VATEX. Comprehensive ablation experiments are carried out to prove the effectiveness of each component. Besides, to further improve the generalization of our method, our method is successfully applied to video question answering (VideoQA) tasks. Finally, we qualitatively show that our method can contribute to improved captions through case studies.

To summarize, the contributions of this work lie in threefolds.

1) We propose a simple, yet effective method, namely a VCRN, to explore the effect of visual commonsense information for video captioning, which improves the model’s capability of knowledge cognitive.

2) We design a video dictionary, a plug-and-play component, to model visual commonsense and exploit the association between the source video and commonsense via our proposed VCS component and CIG component to yield a more accurate caption.

3) The extensive experimental results demonstrate the benefits of introducing visual commonsense for the video captioning task. The proposed method VCRN achieves state-of-the-art performance on MSVD and VATEX and competitive performance on MSR-VTT. Besides, our method brings performance gains on the VideoQA task, further demonstrating the generalization of our method.

II. RELATED WORKS

A. Video Captioning

Video captioning is one of the mainstays in the multimodal domain, which has received extensive interest and made rapid development. With the advent of the encoder–decoder framework, recent research mainly focuses on the sequential-learning-based methods for the generation process [12], [20], [21], [22], [23], [24]. Technically, these methods employ an encoder to refine the video representation from a group of fixed video feature maps, and then a language-based decoder integrates textual descriptions with the refined video features to learn a modality-aligned representation for caption generation. As one of the precedents that adopt such an encoder–decoder structure, Venugopalan et al. [25] generated captions by LSTM with mean pooled video representation overall frame features. Yao et al. [9] proposed temporal attention to dynamically select video frames based on the current decode step. To further align the semantic information between video and language modalities and improve the performance, extensive methods with elaborate structure [13], [24], [26], [27], [28] have been proposed. For instance, Ryu et al. [13] encoded a video into semantic groups by aligning frames around the phrases of partially decoded caption and described the video by exploiting the semantic groups as information units. Chen and Jiang [29] utilized optical flow to guide spatial attention, which can capture the pattern of apparent motion between video and language modalities.
consecutive video frames. To improve caption quality, Yang et al. [24] proposed an alternative paradigm to decompose the captioning procedure into two stages. More recently, there are some methods [11], [12], [30] that have drawn attention to object-level information. Zhang and Peng [30] and Bai et al. [31] adopted a bidirectional temporal graph to capture fine-grained dynamic flow for salient objects in a video. Tan et al. [12] performed visual reasoning over both space and time domains and then located regions over the video by spatial–temporal attention.

Unlike these methods, our method does not introduce extra visual features or pretrained end-to-end architectures but mines the underlying semantic knowledge hidden in the datasets, which aims to provide high-level visual concepts for the model reasoning.

B. Knowledge-Based Learning

To further move toward a cognitive understanding of models, many knowledge-based methods have been proposed [32], [33], [34], [35], [36]. In general, most of the existing methods can be categorized into two types. The first one focuses on the structured knowledge base (e.g., DBpedia [37] and WordNet [38]) to perform knowledge inference and assist model reasoning. For instance, Wang et al. [39] applied a large-scale knowledge base as visual concepts, that is, ConceptNet [40], for explainable visual question answering (VQA). Zhou et al. [41] leveraged structured concept graphs to improve the performance of image captioning. Wen and Peng [35] proposed multilevel commonsense knowledge-based learning for visual commonsense reasoning. The other one focuses on the unstructured knowledge base, which explicitly represents knowledge from the linguistic corpus or vision modality. Compared with the structured one, it is regularly acquired through elaborate designs such as pretrained language (LMs) or retrieval models. For instance, Salaberria et al. [36] hypothesized that a system that relies exclusively on text will allow LMs to better leverage their implicit knowledge and then utilized it on VQA tasks. Zhang et al. [42] proposed a pluggable retriever to retrieve sentences as prior hints into the video captioning model. Additionally, some works aim to exploit relationships between objects via prior knowledge. Hou et al. [43] proposed a commonsense and reasoning method that explores prior knowledge in objects without detector dependency. Wu et al. [44] introduced a knowledge-aware method to enhance generation quality with deep-level object relationships during the training stage.

Different from previous methods that exploit consensus knowledge from the external source, our method aims to explore latent association in video sets and mine intrinsic commonsense knowledge between videos from the inside.

C. Video Question Answering

VideoQA is another fundamental multimodal task, which aims to predict an accurate answer according to a video and a corresponding question. The benefit to the success of deep learning, various techniques, for example, the attention mechanism [45], [46], [47], memory network [48], [49], and graph neural network [50], [51], have been proposed to build the relationship between vision and language to answer questions. For instance, Li et al. [45] proposed temporal attention to focus on the key information through questions as guidance. Gao et al. [48] applied a co-memory network to learn the important cues from both motion and appearance and obtain the multilevel contextual facts to infer the answer. Le et al. [46] introduced a hierarchical conditional relation network to construct more sophisticated relations across video and questions, which obtains diverse modalities and contextual information. Seo et al. [50] proposed a motion-appearance synergistic network for action-oriented cross-modal joint representations between motion and appearance by the graph neural network. In this article, our proposed method is applied to the task of VideoQA to verify its effectiveness.

III. Method

In this section, we present the proposed video captioning method based on VCRN in detail, which follows the paradigm of the encoder–decoder framework. As shown in Fig. 2, our VCRN consists of three components.

1) VDC: We first extract motion and appearance features for all videos to present visual information. Next, we construct a video dictionary to capture and store visual commonsense knowledge in the video domain with an unsupervised method (see Section III-A).

2) VCS: Based on the video dictionary, we perform VCS to obtain video-related concept features via a concept-aware MHA module (see Section III-B).

3) CIG: In the decoding stage, we treat the source video feature and the video-related concept feature as visual representations and send them into the proposed CIG for word prediction (see Section III-C).

A. Video Dictionary Construction

As discussed above, directly operating at a source video to generate a description leads to insufficient visual details. A plain idea is to introduce other similar videos to compensate for the deficiency. However, recklessly building the relationships between the source video and videos in the dataset leads to unaffordable computational costs. Intuitively, if we implicitly summarize the co-occurrence visual concept that exists in video datasets as visual commonsense knowledge, it would be more feasible. Motivated by this, we construct a video dictionary C to obtain intrinsic visual commonsense knowledge in an unsupervised way, containing multiple representative visual concepts.

Concretely, we first employ the 2D-CNN and 3D-CNN to extract appearance feature \( V^a = \{v^a_i\}_{i=1}^L \) and motion feature \( V^m = \{v^m_i\}_{i=1}^L \), where \( L \) denotes the number of video frames, and \( v^a_i \) and \( v^m_i \) denote the feature of the \( i \)th frame. Then, we concatenate \( V^a \) and \( V^m \) together as the source video representation \( V = [V^a; V^m] = \{v_i\}_{i=1}^L \), where \( [\cdot] \) means the concatenate operation. Based on the above process, we extract the features of all videos \( V_{all} \) in the dataset. Afterward, we utilize the K-Means algorithm to cluster overall video representations \( V_{all} \) and obtain \( M \) cluster centers, denoted as

**Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.**
$C = \{c_j\}_{j=1}^M$, where $c_j$ is regarded as the $j$th visual concept representation. We define $C$ as a video dictionary, which will be used to assist the source video in obtaining additional visual commonsense knowledge.

### B. Visual Concept Selection

Typically, a sequence of video frames may contain diverse visual entities, thus the ability to dynamically select different visual concepts from the video dictionary for each frame may be critical. To this end, we propose a VCS module. For each video frame, VCS aims to adaptively select the relevant commonsense knowledge from the video dictionary and aggregate them with relevant scores. After iterating over all frames, we finally obtain a video-related concept representation $\hat{C} = \{\hat{c}_i\}_{i=1}^I$, where $\hat{c}_i$ is the selected visual concepts by the $i$th frame.

As shown in Fig. 2, the main architecture of VCS applies a concept-aware multihead cross-attention (dubbed C-MCA). As a widely used mechanism, cross-attention can help select video-related commonsense knowledge in parallel and generate rich contextual representation during long sequential modeling, which is suitable for consecutive video frames.

Formally, we apply the source video representation $V$ as queries ($Q_v$) and video dictionary $C$ as keys ($K_c$) and values ($V_c$). Then, a scaled dot-product is adopted to calculate the similarity between $V$ and $C$

$$S = \text{sim}(Q_v, K_c) = \text{SOFTMAX} \left( \frac{Q_v K_c^T}{\sqrt{d}} \right).$$

(1)

Here, the similarity $S_{i,j}$ indicates the score that the $j$th concept $c_j$ should attend to the $i$th video frame $v_i$, and $d$ is a scaling factor.

Moreover, instead of calculating $S$ with a single attention function, we use MHA to linearly project the queries, keys, and values $h$ times and calculate the similarity in (1) with several subspaces. Subsequently, the similarity $S^{(h)}$ of head $h$ is calculated by (1) to aggregate multiple semantic information between query $Q_v^{(h)}$ and key $K_c^{(h)}$. All similarity heads are concatenated together and fused with the learnable projection $W^O$

$$C^{(h)} = \text{DROPOUT}(S^{(h)} V_c^{(h)}), \quad \text{for } h = 1, 2, \ldots, H$$

$$C_s = [C^{(1)}_s, C^{(2)}_s, \ldots, C^{(H)}_s] W^O$$

(2)

where $H$ is the number of heads and $C^{(h)}_s$ is the output of the $h$th head. Finally, $C_s$ is normalized via Layer Normalization and added to a source video feature to produce a video-related concept feature $C_i$

$$C_i = V + \text{LAYERNORM}(C_s).$$

(3)

We stack $N$ C-MCA blocks to obtain a more refined video-related concept representation and take the output of the last C-MCA block as a final video-related concept feature $\hat{C}$.

### C. Concept-Integrated Generation

At the decoding stage, we design a novel CIG to generate captions based on the source video feature and video-related concept video feature. The CIG is composed of three parts: an Attention-LSTM, a Gated Controller, and a Language-LSTM. We describe the proposed generator in detail as follows.

1) **Attention-LSTM**: At the $t$th time step, the Attention-LSTM (LSTM$_A$) aims to obtain the semantics of the current state $h^t$ according to the previous hidden state $h^t_{t-1}$ of the Language-LSTM, concatenated with global video feature $\bar{v}$ and the previous word $w_{t-1}$

$$h^t = \text{LSTM}_A ([h^t_{t-1}; \bar{v}; W^e w_{t-1}], h^t_{t-1})$$

$$\bar{v} = \frac{1}{L} \sum_{i=1}^L v_i$$

(4)

where $[;]$ means the operation of concatenation and $W^e$ denotes the word embedding matrix.

2) **Gated Controller**: The designed Gated Controller is adopted for the aggregation of video representations $V$ and $\hat{C}$ according to the current hidden state $h^t$ of the Attention-LSTM, it enables which information flows (i.e., $V$ and $\hat{C}$) should play a more important role in the language-LSTM.

Fig. 2. Overview of the proposed VCRN for video captioning. It consists of three main components: 1) VDC, which models the visual commonsense knowledge extracted from all videos; 2) VCS, which aims to learn associations between the current video and the commonsense knowledge; and 3) CIG, which generates linguistic descriptions by time step. $C$, $\hat{C}$, and $V$ denote video dictionary, video-related concept feature, and video feature, respectively.
Specifically, we first apply multiplicative attention mechanism to aggregate video feature $V$ with the current hidden state $h^t$ at frame-level to obtain the attended video feature $V'$:

$$V' = \sum_{i=1}^{L} \alpha_i V_i$$

$$\alpha_i = \text{SOFTMAX}(W_1 \tanh(W_2 V_i + W_3 h^t))$$

(5)

where $\oplus$ is element-wise addition, $W_\ast$ is the learnable matrices. $V_i$ means the $i$th frame-level vector in a video feature $V$. To simplify the process of $V'$ extraction, we formulate it as

$$V' = \text{ATT}_V(V, h^t).$$

(6)

Similar to the operation of $\text{ATT}_V$, we integrate current hidden state $h^t$ with video-related concept feature to produce the attended concept feature $C'$. The extraction module of $C'$ is defined as

$$C' = \text{ATT}_C(\hat{C}, h^t).$$

(7)

As illustrated in Fig. 2, the context gate $\lambda$ controls the propagation of the information of $V'$ and $C'$ to Language-LSTM. In practice, the value of $\lambda$ is based on $V'$, $C'$, and $h^t$ via a nonlinear layer

$$\lambda = \sigma(W_\lambda \cdot [V'; C'; h^t])$$

(8)

where $W_\lambda$ is a learnable parameter and $\sigma(\cdot)$ denotes the sigmoid function. Next, we utilize this gated controller in a bilateral scheme, where $\lambda$ determines the flow of $V'$ and the complementary part $1 - \lambda$ governs the amount of $C'$, to get conceptual integrated video feature $e'$

$$e' = \lambda \odot f(V') + (1 - \lambda) \odot f(C')$$

(9)

where $\odot$ is the Hadamard product and $f(\cdot)$ can be represented as fully-connected layer or identity mapping.

3) Language-LSTM: The Language-LSTM feeds the hidden conceptual integrated video feature $e'$ to generate the current hidden state $h'^t$. The logits distribution of the caption model $p_t$ is acquired via a single linear function and the softmax operation at the decoding step $t$

$$h'^t = \text{LSTM}_t([h^t; e'], h'_{t-1})$$

$$p_t = \text{SOFTMAX}(W_h h'^t + b_v)$$

(10)

where $p_t$ is a vector of the vocabulary size and $W_h$ and $b_v$ are learnable parameters.

Following the standard objective of video captioning, we adopt cross-entropy loss to optimize our model:

$$L_{CE} = - \sum_{t=1}^{T} \log(p_t(w^t|w^t_{<t})))$$

(11)

where $w^t_{<t}$ is the target ground-truth sequences and $\theta$ is the parameters of our captioning model.

IV. EXPERIMENTS

A. Datasets and Metrics

1) Datasets: Following the previous works [11], we evaluate our method VCRN on three publicly available datasets: MSVD, MSR-VTT, and VATEX, for video captioning.

MSVD [57] is a collection of 1970 short clip videos downloaded from the YouTube website. Each clip has 35 captions annotated by humans. To be consistent with the previous works, we use standard splits, namely 1200 clips for training, 100 clips for validation, and 670 clips for testing.

MSR-VTT [58] consists of 10000 open domain videos from YouTube with 20 human descriptions for each video clip. We follow the standard split with 6573 videos for training, 497 videos for validation, and the remaining 2900 for testing.

VATEX [59] is a recently released large-scale multilingual video description dataset, which reuses the video source from Kinetics-600. It contains over 41 250 videos, where each video clip is annotated with ten English and Chinese descriptions, respectively. In this article, we only utilize English captions for our experiments. According to the official split, the dataset is divided into 25 991 for training, 3000 for validation, and 6000 for public testing.

Besides, to verify the generalization of our method, we also conduct experiments on two VideoQA datasets: MSVD-QA and MSRVT-QA.

MSVD-QA [60] is derived from the existing MSVD dataset with the same video data, containing 1970 short clips and 50505 question–answer pairs. These question–answer pairs are split into five types according to question purpose: what, where, when, how, and who.

MSRVT-QA [58] is composed of 10k videos from the MSR-VTT dataset and 243k annotated question–answer pairs, where the questions are also of five types. Compared to MSVD-QA, the video length of MSRVT-QA is much longer, roughly around 10–30 s with more complex scenes.

2) Evaluation Metrics: For the captioning task, we employ the standard captioning evaluation metrics, including BLEU-4 [61], METEOR [62], ROUGE-L [63], and CIDEr [64], to evaluate our method. For VideoQA, accuracy is adopted as the evaluation metric.

B. Implementation Details

1) Feature Extraction: For the visual features, we use ResNet [65] as 2-D CNN and ResNeXt [66] as 3-D CNN from the MXNet library [67] to extract appearance feature and motion feature, respectively. The above features are extracted from 26 keyframes of videos by equal interval sampling.

For captions, we remove punctuations, convert all words to lowercase, and keep the words that occur more than 2$\times$ for MSR-VTT and MSVD (five for VATEX) to a word vocabulary. Descriptions longer than 26 words (30 for VATEX) will be truncated for the convenience of training. Besides, we add three special tokens (“bos”), “(eos),” and “(pad)” to the word vocabulary. GloVe [68] is utilized to initialize the word embedding.

2) Training Details: We adopt the Adam [69] optimizer with the learning rate of $1e^{-4}$ to train our model. We choose
TABLE I

PERFORMANCE COMPARISONS ON MSVD AND MSR-VTT DATASETS. ALL METHODS ARE DIVIDED INTO TWO CATEGORIES. TOP ROWS USE THREE VISION FEATURES, INCLUDING APPEARANCE FEATURE, MOTION FEATURE, AND OBJECT FEATURE TO TRAIN THE MODEL, WHILE BOTTOM ROWS ONLY USE APPEARANCE FEATURE AND MOTION FEATURE TO TRAIN THE MODEL.

| Models       | Features     | MSVD  | MSR-VTT |
|--------------|--------------|-------|---------|
|              | Appearance   | BLEU-4 | METEOR | ROUGE-L | CIDEr | BLEU-4 | METEOR | ROUGE-L | CIDEr |
| OA-BTG [30]  | ✓            | 56.9  | 36.2   | -     | 90.6  | 41.4  | 28.2   | -      | 46.9  |
| MGSA [29]    | ✓            | 53.4  | 33.0   | -     | 86.7  | 42.4  | 27.6   | -      | 47.5  |
| STG [52]     | ✓            | 52.2  | 36.9   | 73.9  | 93.0  | 40.5  | 28.3   | 60.9   | 47.1  |
| SAAT [53]    | ✓            | 46.5  | 33.5   | 69.4  | 81.0  | 40.5  | 28.2   | 60.9   | 49.1  |
| RMN [12]     | ✓            | 54.6  | 36.5   | 73.4  | 94.4  | 42.5  | 28.4   | 61.6   | 49.6  |
| MGPMP [54]   | ✓            | 53.2  | 35.4   | 73.5  | 90.7  | 42.1  | 28.8   | 61.4   | 50.1  |
| ORG-TRL [11] | ✓            | 54.3  | 36.4   | 73.9  | 95.2  | 43.6  | 28.8   | 62.1   | 50.9  |
| MARN [55]    | ✓            | 48.6  | 35.1   | 71.9  | 92.2  | 40.4  | 28.1   | 60.7   | 47.1  |
| M3 [56]      | ✓            | 52.8  | 33.3   | -     | -     | 38.1  | 26.6   | -      | -     |
| POS-CG [27]  | ✓            | 52.5  | 34.1   | 71.3  | 88.7  | 42.0  | 28.2   | 61.6   | 48.7  |
| MDT [14]     | ✓            | 49.0  | 35.3   | 72.2  | 92.5  | 40.2  | 28.2   | 61.1   | 47.3  |
| SGN [13]     | ✓            | 52.8  | 33.5   | 72.9  | 94.3  | 40.8  | 28.3   | 60.8   | 49.5  |
| HRNAT [15]   | ✓            | 55.7  | 36.8   | 74.1  | 98.1  | 42.1  | 28.0   | 61.6   | 48.2  |
| VCRN (ours)  | ✔            | 59.1  | 37.4   | 74.6  | 100.8 | 41.5  | 28.1   | 61.2   | 50.2  |

TABLE II

PERFORMANCE COMPARISONS ON THE VATEX TESTING SET. NOTE THAT THE ORG-TRL [11] EMPLOY THREE FEATURES (I.E., APPEARANCE FEATURE, MOTION FEATURE, AND OBJECT FEATURE) AND EXPLOITS LINGUISTIC KNOWLEDGE, WHILE THE REST METHODS INCLUDING OURS ONLY ADOPT THE APPEARANCE FEATURE AND MOTION FEATURE. B@4, M, R, AND C INDICATE BLEU-4, METEOR, ROUGE-L, AND CIDEr, RESPECTIVELY.

| Model        | B@4 | M   | R   | C   |
|--------------|-----|-----|-----|-----|
| ORG-TRL [11] | 32.1 | 22.2 | 48.9 | 49.7 |
| Shared Base  | 28.1 | 21.6 | 46.9 | 44.3 |
| Shared Enc   | 28.4 | 21.7 | 47.0 | 45.1 |
| Shared Enc-Dec | 27.9 | 21.6 | 46.8 | 44.2 |
| HRNAT [15]   | 32.1 | 21.9 | 48.4 | 48.5 |
| VCRN (ours)  | 32.4 | 22.4 | 48.9 | 49.9 |

TABLE III

ABLATION STUDIES OF THE PROPOSED VDC, VCS, AND CIG. B@4, M, R, AND C INDICATE BLEU-4, METEOR, ROUGE-L, AND CIDEr, RESPECTIVELY.

| Methods     | MSVD | MSR-VTT | Param. |
|-------------|------|---------|--------|
|             | B@4 | M   | C    | B@4 | M   | C    | B@4 | M   | C   |
| Baseline (B)| 57.9 | 16.7 | 74.3 | 94.6 | 40.8 | 27.7 | 60.6 | 46.0 | 19.0 |
| B = VDC + VCS | 59.5 | 16.8 | 74.2 | 90.1 | 41.2 | 27.3 | 61.1 | 64.0 | 20.8 |
| B = VDC + VCS + CIG | 59.1 | 37.4 | 74.6 | 100.8 | 41.2 | 28.1 | 43.1 | 59.2 | 25.9 |

C. Performance Comparisons

1) Compared Methods: In this section, we compare our method VCRN with the state-of-the-art methods on the MSVD, MSR-VTT, and VATEX datasets. These methods can be divided into two categories: 1) the first category employs appearance feature, motion feature, and object feature to train their model, including OA-BTG [30], MGSA [29], STG [52], SAAT [53], RMN [12], MGPMP [54], and ORG-TRL [11] and 2) the second category only utilizes appearance feature and motion feature without the help of object feature, including MARN [55], M3 [56], POS-CG [27], MDT [14], SGN [13], HRNAT [15], Shared Base [59], Shared Rnc [59], and Shared Enc-Dec [59]. Here, our model belongs to the second category.

2) Comparisons on MSVD: The results of the comparison on MSVD are reported in Table I. We can find that our VCRN model exceeds all previous models in all metrics (BLEU-4, METEOR, ROUGE-L, and CIDEr). Compared with the second category methods, our method outperforms the best counterpart HRNAT, especially with an increase of 3.4% and 2.7% in terms of BLEU-4 and CIDEr, respectively. Compared with the first category methods, our model can still significantly outperform them by a large margin and, in particular, increases BLEU-4 and CIDEr by 4.8% and 5.6%, respectively. It clearly demonstrates the effectiveness of our method.

3) Comparisons on MSR-VTT: Table I also shows the results of the comparison on the MSR-VTT dataset. We can see that our model maintains relatively comparable performance compared to the existing methods. Although the improvement in the MSR-VTT dataset is not as obvious as in the MSVD dataset, our method gets second place with a CIDEr of 50.2%. The possible reason may be that compared to the MSVD dataset, MSR-VTT retains longer videos and more complex visual scenarios, which increases the difficulty of reasoning. The SOTA methods belonging to the first category introduce additional regional information to build object relationships, which brings a more fine-grained visual representation for caption generation. Specifically, our method achieves better performance in CIDEr, compared to the second category methods, in particular obtaining 2.0% relative gains. Moreover, compared to the best counterpart ORG-TRL belonging to the first category, our method is only slightly degraded in performance. The possible reason

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
may be that the best method ORG-TRL belonging to the first category uses additional language models, in addition to introducing object features, which is more helpful for reasoning on the MS-R-VTT dataset.

4) Comparisons on VATEX: To further verify the robustness of our method, we provide quantitative results on the VATEX dataset in Table II. From the table, we can observe that our method shows better superiority over all compared methods in all metrics. In particular, in terms of CIDEr, our method brings an increase of 0.2% and 1.4% compared to the second (ORG-TRL) and third (HRNAT) methods, respectively. These results demonstrate the effectiveness of our method.

D. Ablation Study

In this section, we elaborate on a series of ablation studies in the following Q&As to better prove the validity of our model. All experimental results are conducted on MSVD and MSR-VTT.

1) Does Each Component of VCRN Affect the Results? We evaluate the effectiveness of each component by gradually taking VDC, VCS, and CIG into the baseline model, where the baseline only adopts appearance features and motion features and is based on a vanilla encoder–decoder with temporal attention. The results are shown in Table III. Overall, all the proposed components contribute significantly to the overall performance. Specifically, the baseline model first performs the worst. By integrating the VDC and VCS into the baseline, the performance obtains larger improvement, particularly increased by 1.7% and 1.6% in terms of CIDEr on MSVD and MSR-VTT, respectively. It reveals the importance of visual commonsense knowledge, which provides additional visual information to help model reasoning. Then, the CIG is added to the model B + VDC + VCS, which in turn further enhances the performance, indicating that our CIG can effectively integrate the original video feature and video-related concept feature from VCS. Besides, we also report the parameters of each component. As can be seen, the proposed VDC and VCS (note that VDC is an offline clustering process and is not involved in the model training) increase the parameters by 6.2 M when compared to the baseline model. Furthermore, integrating CIG into the model brings a slight increase of parameters by 1.3 M. This demonstrates the effectiveness of our method which can achieve significant performance with small computational costs.

2) Does the Number of Clustered Centers $M$ in Visual Dictionary Affect the Results? We exploit how the number of clustered centers in the visual dictionary affects the performance of our VCRN. In experiments, we select different clustered centers for retraining our VCRN, where $M \in [0, 300, 500, 1000, 2000, 3000]$. Here, $M = 0$ is treated as the baseline without including VCS and CIG, and we pick CIDEr as the metric of caption performance as it reflects the generation relevant to video content. Fig. 3 shows the experimental results. We can see that the performance is best when $M$ is set to 1000. When the number of clusters is greater than 1000 or less than 1000, there exists degradation in model performance. An intuitive explanation is that too many or too few clustered centers can lead to the introduction of redundant or insufficient visual commonsense information, respectively. Thus, we set $M = 1000$ in the final model.

3) Which Is Better, Fixed or Jointly Trained Visual Dictionary? Fig. 4 displays the real-time test results of the fixed visual dictionary and jointly trained visual dictionary during training, where we also choose CIDEr as the main metric, which has three settings.

1) TrainVD (w/ Rand.): Jointly trained visual dictionary initialized by random parameters.
2) TrainVD (Ours): Jointly trained visual dictionary initialized by our proposed VDC method.
3) FixedVD (Ours): Fixed visual dictionary initialized by our proposed VDC method.

From Fig. 4, it is observed that FixedVD (ours) is better than TrainVD (ours) and TrainVD (w/ Rand.). This may be because the fixed video dictionary can retain more original visual commonsense knowledge that is more helpful for generation compared to a jointly trained video dictionary.

4) Does the Quality of the Video Dictionary Affect the Results? We analyze the effect of randomly selecting different proportions of video in the train set to simulate different quality video dictionaries. The experimental results are shown in Table IV. Table lines 1–4 illustrate that higher-quality video dictionaries facilitate the generation of fine-grained descriptions. It can be explained by the fact that higher-quality video dictionaries contain more visual commonsense knowledge and provide more hints related to video content for generation. In addition, we construct the video dictionary by using the video data from the validation and test sets rather than the training set, as shown in Table IV lines 5 and 6. It can be seen that the performance drops slightly but is still comparable. The possible reason is that the validation and test sets contain fewer videos compared to the training set, which makes the video
dictionary store lower commonsense knowledge and visual clues for sentence generation.

To explore the impact of a larger-scale video dictionary for the proposed method, we further utilize various combinations of training, validation, and test sets to construct a video dictionary. The results are summarized in Table IV lines 7–9. We observe that larger-scale concept sources can significantly benefit our model. It achieves the best performance when using the train + val + test as the concept source. This further proves that a high-quality video dictionary can help the generation of descriptions.

5) Does Cross-Dataset Video Dictionary Affect the Results? We conduct this experiment by using different video dictionaries constructed from other video datasets to demonstrate the generalization of the model. In our experiments, we use the video dictionary from the MSR-VTT dataset for training and testing on the MSVD dataset, and vice versa. As shown in Table V, the model can maintain competitive performance with only a slight drop when using other video dictionaries. It proves that our model has a strong learning ability, which can be extended by changing different video dictionaries that are not strongly correlated even with the test data. Moreover, when we combine the MSVD and MSR-VTT as the concept source, it achieves better performance. This indicates that our model can have some influence on model performance. Compared to other clustering algorithms, using Agglomerative and Birch achieves the best performance on MSVD and MSR-VTT, respectively. The reason may be that Agglomerative is typically optimal for small-scale data distribution, while Birch has stronger scalability for large-scale data distribution. Note that in other experiments we choose K-Means to cluster videos in the dataset since it is the most general clustering algorithm that in other experiments we choose Birch achieves the best performance on MSVD and MSR-VTT.

6) Does the Control Strategies in CIG Affect the Results? We compare the effect of different control strategies in CIG, including the following:

1) ADD: Adopting an element-wise addition to aggregate $V'$ and $C'$.
2) MLP: Adopting a multilayer perceptron to aggregate $V'$ and $C'$.
3) MHA: Adopting MHA to aggregate $V'$ and $C'$.
4) GATE: Adopting our proposed gate controller to aggregate $V'$ and $C'$.

The results are summarized in Table VI. Compared with other control strategies, our method can achieve better performance on all metrics by a large margin. It indicates the effectiveness of our proposed content gate in CIG.

7) Does Different Clustering Algorithms Affect the Results? We conduct experiments to test the impact of various clustering algorithms for the proposed method. The results are summarized in Table VII. We observe that different clustering algorithms can have some influence on model performance. Compared to other clustering algorithms, using Agglomerative and Birch achieves the best performance on MSVD and MSR-VTT, respectively. The reason may be that Agglomerative is typically optimal for small-scale data distribution, while Birch has stronger scalability for large-scale data distribution. Note that in other experiments we choose K-Means to cluster videos in the dataset since it is the most general clustering algorithm and easy to cover.

8) Does Different Attending Strategies in VCS Affect the Performance? To investigate different attending strategies in the VCS component, we conduct experiments by replacing cross-attention with two alternative mechanisms, including mean pooling and co-attention. As shown in Table VIII, adopting mean pooling and co-attention both bring performance degradation on MSVD and MSR-VTT datasets. Moreover,

### Table IV

| Concept Source | MSVD | MSR-VTT |
|----------------|------|---------|
| Train Set + | 57.2 | 74.0 | 98.6 | 40.9 | 27.8 | 60.7 | 48.6 |
| Train Set + | 57.1 | 74.0 | 97.4 | 41.1 | 28.0 | 61.1 | 49.4 |
| Train Set + | 58.8 | 71.4 | 99.8 | 41.3 | 28.1 | 61.2 | 49.6 |
| Train Set + | 59.1 | 74.6 | 100.8 | 41.5 | 28.1 | 61.2 | 50.3 |
| Val Set + | 59.6 | 71.4 | 97.9 | 41.2 | 27.8 | 60.8 | 48.8 |
| Test Set + | 59.0 | 70.4 | 98.7 | 41.2 | 28.1 | 61.2 | 49.6 |
| Train+Val Set + | 61.5 | 83.4 | 103.6 | 40.8 | 27.8 | 60.6 | 49.3 |
| Train+Test Set + | 60.1 | 73.4 | 102.4 | 41.2 | 28.1 | 61.2 | 50.3 |
| Train+Val+Test Set + | 60.6 | 82.2 | 103.8 | 41.4 | 28.3 | 61.3 | 50.7 |

### Table V

| Concept Source | MSVD | MSR-VTT |
|----------------|------|---------|
| MSVD | 59.1 | 73.4 | 100.8 | 40.4 | 27.7 | 60.5 | 48.9 |
| MSR-VTT | 57.9 | 73.5 | 97.8 | 41.5 | 28.1 | 61.2 | 50.2 |
| MSVD+MSR-VTT | 59.3 | 77.7 | 94.6 | 41.6 | 28.2 | 61.5 | 50.5 |

### Table VI

| Fusion strategy | MSVD | MSR-VTT |
|-----------------|------|---------|
| ADD | 59.5 | 36.8 | 74.2 | 98.1 | 41.2 | 27.9 | 61.1 | 49.6 |
| MLP | 57.5 | 37.0 | 74.3 | 96.4 | 40.8 | 28.0 | 61.1 | 49.2 |
| MHA | 55.9 | 36.6 | 73.6 | 99.2 | 40.2 | 27.9 | 60.7 | 48.7 |

### Table VII

| Clustering Algorithms | MSVD | MSR-VTT |
|-----------------------|------|---------|
| K-Means | 59.1 | 37.4 | 74.6 | 100.8 | 41.5 | 28.1 | 61.2 | 50.2 |
| Bootstrapping K-Means | 61.9 | 38.4 | 75.3 | 103.3 | 41.2 | 28.0 | 61.2 | 50.3 |
| Birch | 58.1 | 36.3 | 74.9 | 99.3 | 41.5 | 28.4 | 61.6 | 50.5 |
| Agglomerative | 61.1 | 38.3 | 75.2 | 103.4 | 40.9 | 27.8 | 60.8 | 49.6 |
| Spectral | 58.4 | 36.2 | 74.3 | 99.0 | 40.7 | 27.8 | 60.8 | 49.5 |
| MeanShift | 62.1 | 37.9 | 74.4 | 102.3 | 41.5 | 28.3 | 61.0 | 49.3 |

### Table VIII

| Methods | MSVD | MSR-VTT |
|---------|------|---------|
| Mean Pooling | 56.2 | 36.7 | 74.0 | 98.2 | 40.9 | 28.0 | 60.5 | 49.1 |
| Co-attention | 58.2 | 36.8 | 74.1 | 99.4 | 41.1 | 28.1 | 60.8 | 49.5 |
| Cross attention | 59.1 | 37.4 | 74.6 | 100.8 | 41.5 | 28.1 | 61.2 | 50.2 |
Fig. 5. Fig. 5. Visualization of the baseline model and our proposed VCRN on MSR-VTT. Each example consists of a raw video, a ground-truth description, and the generated descriptions by baseline and VCRN (ours).

9) Does the Model Have the Ability to Generate Diverse Captions? To analyze the diversity of generated captions, we calculate three metrics following [71], including Novel (the percentage of captions that are not observed in the training set), Unique (the percentage of unique captions among all generated captions), and Vocab Usage (the percentage of words in the vocabulary that are used to generate captions). The results are summarized in Table IX. Compared to the baseline model, VCRN achieves the best performance in all evaluation metrics on MSVD and MSR-VTT. This demonstrates the superior ability of the proposed method which can generate more diverse captions.

E. Generalization on VideoQA

To further prove the generalization of our method, we apply the proposed video dictionary to VideoQA. Table X shows the experimental results. Practically, we first choose three popular VideoQA methods, including HME [70], MASN [50], and HCRN [46], as our baseline models. We reproduced their results by running the available code. Then, we simply integrate the video dictionary into the three models via the same operation of VCS. As we can see, our method gains a certain level of improvement on the current VideoQA models. For instance, when equipped with our video dictionary, the current SOTA method HCRN can boost the accuracy by around 1.2% on MSVD-QA and 0.7% on MSRVT-QA, respectively. Hence, it demonstrates that our proposed visual commonsense has a strong generalization ability in other video-related tasks.

F. Qualitative Results

Fig. 5 illustrates the generated captions on the MSR-VTT dataset. Overall, it is observed that the context of captions
We also illustrate the qualitative results of VideoQA on the MSRVTT-QA dataset, as shown in Fig. 7. We selected HCRN as our baseline model due to the available code and reproducible results. In comparison with the baseline, integrating with the proposed video dictionary can generate more accurate answers. These examples further demonstrate the great generalization ability for visual commonsense reasoning in the VideoQA task.

V. CONCLUSION

In this article, we present a novel VCRN for video captioning, which aims to mine the cognitive power of the model’s visual commonsense knowledge. By constructing a video dictionary from all videos in the dataset, we can obtain effective visual commonsense representation for captioning. Furthermore, our proposed VCS and CIG can capture video-related commonsense information and generate more accurate captions, respectively. Our proposed model achieves state-of-the-art performance on both MSVD and VATEX datasets and comparable results on the MSR-VTT dataset. Extensive experiments and qualitative results have demonstrated the effectiveness of each module. Besides, we also demonstrate the strong generalization of our method by transferring to the VideoQA task.

In future work, investigating more alternative models (e.g., Transformer [72] and GNN [73]) with commonsense knowledge to complete high-quality description generation might be important. Then, we will explore larger-scale concept sources (e.g., ActivityNet [74] and Conceptual Caption [75]) and verify their effectiveness. In addition, we can test the generalization ability of VCRN on more video-related tasks, such as video grounding, video recognition, and cross-modal retrieval.

REFERENCES

[1] J. Y. F. Lee, N. Rajeev, and A. Bhojan, “Goldeye: Enhanced spatial awareness for the visually impaired using mixed reality and vibrotactile feedback,” in *Proc. ACM Multimedia Asia*, 2021, pp. 1–7.
[2] C. Rane, A. Lashkare, A. Karande, and Y. Rao, “Image captioning based smart navigation system for visually impaired,” in *Proc. ICCICT*, 2021, pp. 1618–1626.
[3] N. Han, J. Chen, G. Xiao, H. Zhang, Y. Zeng, and H. Chen, “Fine-grained cross-modal alignment network for text-video retrieval,” in *Proc. ICCV*, 2021, pp. 3826–3834.
[4] C. Jiang et al., “Learning segment similarity and alignment in large-scale content based video retrieval,” in *Proc. ACM MM*, 2021, pp. 1618–1626.
[5] A. Dix, J. Finlay, G. D. Abowd, and R. Beale, “Human–computer interaction,” in *Proc. Harlow UA*, 2000, pp. 1–64.
[6] A. Das, S. Kottur, J. M. Moura, S. Lee, and D. Batra, “Learning cooperative visual dialog agents with deep reinforcement learning,” in *Proc. ICCV*, 2017, pp. 2951–2960.
[7] X. Zhang et al., “RSTNet: Captioning with adaptive attention on visual and non-visual words,” in *Proc. CVPR*, 2021, pp. 15465–15474.
[64] R. Vedantam, C. L. Zitnick, and D. Parikh, “CIDEr: Consensus-based image description evaluation,” in Proc. CVPR, 2015, pp. 4566–4575.
[65] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. CVPR, 2016, pp. 770–778.
[66] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” in Proc. CVPR, 2017, pp. 1492–1500.
[67] T. Chen et al., “MXNet: A flexible and efficient machine learning library for heterogeneous distributed systems,” 2015, arXiv:1512.01274.
[68] J. Pennington, R. Socher, and C. D. Manning, “GloVe: Global vectors for word representation,” in Proc. EMNLP, 2014, pp. 1532–1543.
[69] D. P. Kingma and J. Ba, “ADAM: A method for stochastic optimization,” in Proc. ICLR, 2015, pp. 1–15.
[70] C. Fan, X. Zhang, S. Zhang, W. Wang, C. Zhang, and H. Huang, “Heterogeneous memory enhanced multimodal attention model for video question answering,” in Proc. CVPR, 2019, pp. 1999–2007.
[71] B. Dai, S. Fidler, and D. Lin, “A neural compositional paradigm for image captioning,” in Proc. NeurIPS, vol. 31, 2018, pp. 1–11.
[72] A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., vol. 30, 2017, pp. 1–11.
[73] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, “A comprehensive survey on graph neural networks,” IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 1, pp. 4–24, Jan. 2021.
[74] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, “ActivityNet: A large-scale video benchmark for human activity understanding,” in Proc. CVPR, 2015, pp. 961–970.
[75] P. Sharma, N. Ding, S. Goodman, and R. Sorour, “Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning,” in Proc. ACL, 2018, pp. 2556–2565.

Pengpeng Zeng received the B.E. degree from the Xi’an University of Technology, Xi’an, China, in 2016, and the M.E. and Ph.D. degrees from the University of Electronic Science and Technology of China, Chengdu, China, in 2019 and 2023, respectively. He is currently a Post-Doctoral Research Fellow with the Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China. His current research interests include visual understanding, machine learning, and reinforcement learning.

Haonan Zhang received the B.E. degree in computer science and technology from Xidian University, Xi’an, China, in 2020. He is currently pursuing the Ph.D. degree in computer science and technology with the University of Electronic Science and Technology of China, Chengdu, China. His research interests include computer vision and natural language processing.

Lianli Gao (Member, IEEE) received the Ph.D. degree in information technology from The University of Queensland (UQ), Brisbane, QLD, Australia, in 2015. She is currently a Professor with the School of Computer Science and Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China. She has been focusing on integrating natural language for visual content understanding.

Dr. Gao was the Winner of the IEEE TRANSACTIONS ON MULTIMEDIA 2020 Prize Paper Award; the Best Student Paper Award in the Australian Database Conference, Australia, in 2017; the IEEE TCMC Rising Star Award in 2020; and the ALIBABA Academic Young Fellow.

Xiangpeng Li received the Ph.D. degree from the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China, in 2022. He is currently working on image understanding, image/video captioning, and visual dialogue.

Jin Qian is currently pursuing the bachelor’s degree with the SWJTU-Leeds Joint School, Southwest Jiaotong University (SWJTU), Chengdu, China. Her research interests include data compression and machine learning.

Heng Tao Shen (Fellow, IEEE) received the B.Sc. (Hons.) and Ph.D. degrees from the Department of Computer Science, National University of Singapore, Singapore, in 2000 and 2004, respectively. He is currently a Professor of the National Thousand Talents Plan and the Dean of the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China, where he is also the Director of the Center for Future Media. He is also an Honorary Professor with The University of Queensland, Brisbane, QLD, Australia. His research interests include multimedia search, computer vision, artificial intelligence, and big data management.

Dr. Shen is an ACM Distinguished Member and an Optical Society of America (OSA) Fellow.