Video captioning via a symmetric bidirectional decoder

Shanshan Qi  |  Luxi Yang

School of Information Science and Engineering, Southeast University, Nanjing China

Correspondence
Luxi Yang, School of Information Science and Engineering, Southeast University, Nanjing China. Email: leyang@seu.edu.cn

Abstract
The dominant video captioning methods employ the attentional encoder–decoder architecture, where the decoder is an autoregressive structure that generates sentences from left-to-right. However, these methods generally suffer from the exposure bias issue and neglect the guidance of future output contexts obtained from the right-to-left decoding. Here, the authors propose a new symmetric bidirectional decoder for video captioning. The authors first integrate the self-attentive multi-head attention and bidirectional gated recurrent unit for capturing the long-term semantic dependencies in videos. The authors then apply one single decoder to generate accurate descriptions from left-to-right and right-to-left simultaneously. The decoder in each decoding direction performs two cross-attentive multi-head attention modules to consider both the past hidden states from the same decoding direction and the future hidden states from the reverse decoding direction at each time step. A symmetric semantic-guided gated attention module is specially devised to adaptively suppress the irrelevant or misleading contents in the past or future output contexts and retain the useful ones for avoiding under-description. Experimental evaluations on two widely applied benchmark datasets: Microsoft research video to text and Microsoft video description corpus, demonstrate that the authors’ proposed method obtains substantially state-of-the-art performance, which validates the superiority of the bidirectional decoder.

1  |  INTRODUCTION

The goal of the video captioning task is to automatically describe what is happening in videos with semantically reasonable sentences, which is a crucial and challenging issue in computer vision for closing the gap between vision and language. The generation of accurate and meaningful sentences requires the model to thoroughly understand the rich spatial-temporal semantic information about objects, actions, and their interactions in videos and depict the visual information in a grammatically correct manner. Video captioning has drawn considerable research attention recently for its extensive practical applications in high-level video understanding tasks, including Visual Question Answering (VQA) [1], video retrieval [2] and assisting people with vision problems [3], and so on.

Most dominant video captioning methods adopt the classical attentional encoder–decoder framework [4–13], which has brought significant performance improvements in neural machine translation (NMT) [14, 15] and image captioning tasks [16, 17]. Typically, the 2D or 3D convolution neural networks (CNNs) are exploited to extract visual semantic representations from videos, and the recurrent neural networks (RNNs) are leveraged to decode the video representations into sentences sequentially. Meanwhile, the attention mechanism is incorporated into the decoding process to encourage the decoder to selectively concentrate on the relevant visual contents at each time step, which plays a crucial part in generating diverse and fine-grained language sentences. The video representation dramatically influences the quality of generated sentences, especially for videos with rich semantic information. In the real scenario, the videos may contain complex activity scenes, so there have been long-term semantic dependencies in videos. Each video frame is not only related to adjacent video frames but also related to distant video frames. However, it is hard for existing methods [4, 5] to utilise a simple video representation to capture detailed temporal dynamics over a long time interval without explicitly considering the direct relationships among frames. Therefore, how to efficiently capture long-term semantic dependencies in videos and extract discriminative video features is a pivotal issue in video captioning.
In addition, another primary concern of most video captioning methods is that the RNN-based decoder is an autoregressive structure which predicts each word based on the previously generated words in a left-to-right manner. This unidirectional decoding structure constrains the utilisation of future output contexts generated from the right-to-left (R2L) decoding. By analysing the human cognition process, it can be concluded that the combination of past and future output contexts can help decide what video contents to be described at the current time and avoid under-description. Furthermore, due to the autoregressive property, most existing methods suffer from the exposure bias issue for the discrepancy between training and testing. Concretely, at each time step, the RNN-based decoder is trained to generate each word on the basis of the previous ground-truth words, while in the testing stage, the RNN-based decoder is fed with the previous words predicted by the model itself. Hence, the decoder has never been exposed to the predicted words during training and cannot deal with the errors that never occur in the training stage. Once a wrong word is obtained at the early decoding steps during testing, the errors would be quickly accumulated and propagate alongside the word sequence, which will mislead the word prediction process. This issue becomes more severe as the word sequence becomes longer. It can be observed that the left-to-right (L2R) decoding tends to generate high-quality prefixes and inaccurate suffixes, and the R2L decoding shows the reverse preferences. Moreover, two separate attention modules are adopted to calculate the past or future output contexts on the basis of the current decoding hidden state and the past or future hidden state vectors. There exists a situation where the past or future output contexts are not what the decoder expects, and thus the decoder could be misled to yield false results. In other words, it is probable that the decoder hardly determines whether or how well the past or future hidden state vectors are related to the current decoding hidden state. The extreme case is that no worthy information in the past or future hidden state vectors satisfies the demand of current decoding hidden state, but the past or future output contexts can still be calculated by aggregating all attention weighted past or future hidden state vectors and are utterly irrelevant to the current decoding hidden state.

For addressing the above problems, a new symmetric bidirectional decoding model is proposed for video captioning, which tries to generate the discriminative video representation and employ one single decoder to simultaneously perform decoding from L2R and R2L directions interactively. At first, the self-attentive multi-head attention (MH-Att) and bidirectional gated recurrent unit (BiGRU) are integrated [18] to model the frame-to-frame interactions directly and capture the long-term semantic dependencies in videos. Next, one single decoder is applied to translate the video representations from two reverse decoding directions at the same time for mitigating the exposure bias problem. The two decoding directions are symmetric and are complementary to one another. Then the L2R decoding is used as an instance so that the generation of each word not only depends on the predicted words of the L2R decoding but also relies on the previously generated words of the R2L decoding. This bidirectional decoding structure can facilitate the deep and comprehensive information interaction between the two decoding directions and encourage each decoding direction to embed more semantic knowledge from the reverse one. Besides, we exploit a joint training scheme to make the L2R and R2L decoding interactively improve each other, enhancing the robustness and accuracy of the output predictions. Finally, a symmetric semantic-guided gated attention module (SSGGA) is proposed to measure the relevance between each decoding hidden state and past or future output contexts. Two attention gates are devised to adaptively suppress the irrelevant or misleading contents in the past or future output contexts and retain the useful ones. This module is critical to prevent under-description.

The primary work of the proposed method is outlined as follows:

- We introduce a symmetric bidirectional decoder for video captioning, which employs one single decoder to simultaneously describe the video contents from L2R and R2L decoding directions so as to fully exploit the future contexts and mitigate exposure bias. To the best of our knowledge, our method is the first that tries to investigate the effectiveness of the symmetric bidirectional decoder in video captioning.
- We devise a SSGGA module to determine the relevance between the current decoding hidden state and past or future output contexts and avoid misleading the sentence generation.
- Experimental evaluations on two extensively applied benchmark datasets: Microsoft research video to text (MSR-VTT) and Microsoft video description corpus (MSVD), show that our method obtains substantially state-of-the-art performance compared with several recently proposed methods, which proves the superiority of the bidirectional decoder.

2 | RELATED WORK

This section provides some brief introductions to several recently proposed video captioning methods as well as the extensively applied attention mechanism.

2.1 | Video captioning

The primary video captioning methods are mostly template-based methods [19–23], which first define some grammar rules for sentence templates and then align each syntactic component (e.g. subject, verb, object) of the predefined template with semantic words detected from video contents. This type of video captioning is capable of producing grammatically correct video descriptions but loses the flexibility and diversity of sentences. Meanwhile, these methods are insufficient to depict open domain videos with complex semantic contents.

Recently, benefit from the successful development of deep learning techniques (e.g. CNNs and RNNs) and NMT, the
sequence learning-based methods are proposed to depict the visual contents using the accurate sentences of flexible syntactic structures. Venugopalan et al. [24] made the first attempt to adopt the encoder-decoder framework to generate video captions. They applied the mean pooling operation to the video features and obtained the global video representation, which neglected the temporal information among frames. Some latest efforts have been made either in the encoding stage for extracting more discriminative video features or in the decoding stage for generating more robust and meaningful language sentences. The S2VT [25] integrated the Red, Green, Blue (RGB) features and optical flow features for better video representation and leveraged long short-term memory (LSTM) in both the video encoder and sentence decoder. Pan et al. [26] devised the hierarchical RNN, which can excavate the video semantic information with multiple temporal granularity. Ramanishka et al. [27] fused multi-modal source information (i.e., visual, audio, and text) to boost the captioning performance. Baraldi et al. [28] exploited the boundary-aware LSTM cell to discover the discontinuities among the video frames and modified the temporal connectivity of the encoding module accordingly. Wu et al. [29] leveraged a multi-modal circulant fusion (MCF) module to integrate the visual and textual features and fully capture the cross-modal interactions between the two modalities. Chen et al. [30] designed a plug-and-play PickNet module which tried to pick the most discriminative frames and mitigate the interference of duplicated and redundant frames. Aaftaq et al. [31] applied the short fourier transform operation to the CNN-based video representations hierarchically for incorporating the spatio-temporal semantic dynamics into the visual features. Hou et al. [12] learned the syntactic structure of captions by inferring part-of-speech (POS) tags from videos and adopted a mixture model to translate the video features into linguistic words based on the learned syntactic structure. Zhang et al. [31] designed an object relational graph based encoder which attempted to acquire the fine-grained interaction features and incorporated plentiful external language information to guide the caption generation. Moreover, the well-known transformer structure has been applied in the video captioning [32, 33], which focused on learning high-level video semantic features. Also, many methods have applied various attention mechanisms to video captioning and obtained significant performance improvements. Gao et al. [34] exploited the hierarchical attention to mine temporal dynamic information in videos, and Yan et al. [13] presented the spatio-temporal attention mechanism which can adaptively concentrate on the important regions of the most relevant video frames when predicting each word. Wang et al. [5] put forward a reconstruction network developed from an encoder-decoder-reconstructor framework to exploit the bidirectional flows between videos and sentences. Wang et al. [6] introduced a bidirectional proposal framework that leveraged the past and future visual contexts to localise precise activity segments and generated the corresponding descriptions.

As far as we know, the bidirectional encoding structure has been widely applied in video captioning [6,8], but the bidirectional decoding structure has never been explored in video captioning, which has demonstrated its effectiveness in NMT [35–37]. The unidirectional decoder is an autoregressive structure which constrains the utilisation of future output contexts obtained from the R2L decoding. Here, the proposed method is capable of mitigating the exposure bias issue and improving the robustness and accuracy of the output predictions.

2.2 | Attention models

The idea of the attention mechanism derives from the operation mode of human visual system. This technique has been extensively explored in lots of computer vision tasks including video classification [38], visual question answering [39], and image captioning [16], etc. The attention mechanism can help identify the important frames of videos or salient local regions of images for extracting more discriminative features. Recently, various types of attention mechanisms have emerged, such as: adaptive attention [40], self-attention [41], multi-level attention [42], and stacked attention [43].

The self-attention mechanism first converts the input features into queries, keys, and values. Then the attention weight matrices between the queries and keys are calculated through dot product. The final attention features are obtained by aggregating all normalised attention weighted values. Recently, a large number of computer vision tasks using the self-attention mechanism have gained superior performances, which encourages us to employ the self-attention mechanism to our video captioning method for exploring the relationships between different components.

3 | PROPOSED METHOD

The authors propose a symmetric bidirectional decoding architecture for video captioning task, which is illustrated in Figure 2. At first, a video semantic encoder is introduced, which integrates the self-attentive MH-Att module and the BiGRU network to capture the long-term semantic dependencies in videos. Next, the simultaneous bidirectional decoder translates the video representations from two reverse decoding directions (L2R and R2L) interactively. Finally, the symmetric semantic-guided gated attention module adaptively filters out the irrelevant or misleading information in the past or future output contexts and keeps the useful for avoiding under-description.

Notably, gated recurrent unit (GRU) [44] is chosen to be the basic building block of the video encoder and sentence decoder, which is widely used for its relatively fewer parameters required than LSTM [45].

3.1 | Preliminary

Here, the authors briefly present the background knowledge of the Transformer framework [41], which consists of two important attention modules, as shown in Figure 1. We begin
with illustrating the scaled dot-product attention, which is fed with three input features: queries \( Q = (q_1, \ldots, q_l) \in \mathbb{R}^{l \times d_q} \), keys \( K = (k_1, \ldots, k_l) \in \mathbb{R}^{l \times d_k} \) and values \( V = (v_1, \ldots, v_l) \in \mathbb{R}^{l \times d_v} \). Generally, the three input features have the same dimension, denoted as \( d \). The scaled dot-product attention yields the weighted sum of values \( V \), where the inner product is exploited to compute the attention weights based on query \( q_i \) and keys \( K \). Moreover, the attention function applied to each query is calculated at the same time.

\[
\text{Attn}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d}} \right) V
\]

We call the scaled dot-product attention self-attention when the three inputs are set to the same features. Otherwise, it is called cross-attention.

Furthermore, the MH-Att is put forward to strengthen the capacity of attention representations, which contains \( H \) identical scaled dot-product attention modules known as 'head'. Each head attends to the contexts from different representation subspaces and the MH-Att can achieve better results than the single-head attention (SH-Att). We concatenate the output of each head and obtain the final MH-Att results as follows:

\[
\text{MH - Attn}(Q, K, V) = \left[ \text{head}_1, \text{head}_2, \ldots, \text{head}_b \right] W^O
\]

\[
\text{head}_i = \text{Attn}(QW_i^Q, KW_i^K, VW_i^V)
\]

where \( W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d \times d} \) are the parameter projection matrices for the \( i \)-th head, \( i = 1, 2, \ldots, b \) and \( W^O \in \mathbb{R}^{d \times d} \) denotes the linear transformation.

### 3.2 Video semantic encoder

The video semantic encoder aims to build the frame-to-frame interactions and generate discriminative visual features for the
downstream decoder. Recently, CNNs have provided significant performance improvement to various computer vision tasks. The pre-trained CNN networks can be taken as the standard feature extractors and applied to different tasks directly for its strong power of representation learning. Here, given the video frames $X = (x_1, \ldots, x_n)$ of length $n$, we feed each frame into the pre-trained ResNet152 network and get a 2048-dimensional vector from the pool5 layer. Let $V = (v_1, \ldots, v_n)$ denotes the extracted video feature sequence.

Moreover, a video may contain complex activity scenes over a long period in the real scenario, which indicates that there have long-term semantic dependencies in the video contexts and each video frame is not only related to the adjacent video frames but also related to the distant frames. We leverage the self-attentive MH-Att module to build direct relationships among different frames and facilitate the adequate comprehension of the complex dependencies. Unlike the traditional RNNs, it is required to incorporate position information into video feature sequence $V$ because the self-attentive structure contains no recurrence and convolution. The sinusoidal time signal is adopted as the absolute positional encoding as follows:

$$\text{PE}_{e, 2j} = \sin\left(\frac{e}{10000^{2j/d}}\right)$$
$$\text{PE}_{e, 2j+1} = \cos\left(\frac{e}{10000^{2j/d}}\right)$$

where $e$ and $j$ indicate the position of video feature sequence and the order along the video dimension, respectively. The positional encoding is then element-wise added to the video feature sequence $V$ directly.

$$\tilde{V} = V + \text{PE}$$

We set the three inputs of MH-Att module to the same video feature sequence: $\tilde{V} = \{\tilde{v}_1, \ldots, \tilde{v}_n\}$, and the final self-attention video feature sequence $\tilde{V} = \{\tilde{v}_1, \ldots, \tilde{v}_n\}$ is given by:

$$M = \text{MH - Attn}(\tilde{V}, \tilde{V}, \tilde{V})$$
$$\tilde{V} = \text{LayerNorm}(\tilde{V} + M)$$

where a residual connection is employed to retain the important information of videos. The subsequent layer is a feed-forward network (FFN), which contains a ReLU function between two linear transformations.

$$\text{FFN}(\tilde{V}) = \text{ReLU}(\tilde{V}W_1 + b_1)W_2 + b_2$$

where $W_1 \in \mathbb{R}^{d \times d}$, $W_2 \in \mathbb{R}^{d \times d}$ denote the weight matrices for the feed-forward layer and $b_1 \in \mathbb{R}^{d}$, $b_2 \in \mathbb{R}^{d}$ are the biases.

We define the framework mentioned above as a sub-encoder. $N$ such identical sub-encoder are stacked to constitute a complete self-attentive MH-Att encoding module, and the output results of the previous sub-encoder are taken as the input to the next sub-encoder.

In addition, the adjacent video frames tend to have strong dependencies, which derive from the fact that the video signal is a highly correlated temporal signal. To strengthen the local dependencies, we feed the self-attentive video feature sequence to a BiGRU network, and obtain the contextual representation of each frame, given by:

$$h_t^- = \text{GRU}^d(\tilde{v}_t, h_{t^-}^-), h_t^+ = \text{GRU}^e(\tilde{v}_t, h_{t^+}^+)$$

where $\text{GRU}^d$ and $\text{GRU}^e$ indicate the L2R and R2L GRU encoding networks, respectively. The contextual representation of each frame is the concatenation of hidden state vectors of L2R and R2L GRU encoding network at each time step. We denote the final video semantic representations as $h = \{[h_t^-, h_t^+]\}_{t=1}^{n}$.

### 3.3 Simultaneous bidirectional decoder

Generally, most video captioning decoders are based on RNNs and generate sentences from left-to-right. These unidirectional decoders are incapable of leveraging the future output contexts obtained from the R2L decoding and always suffer from the exposure bias issue. Intuitively, the future output contexts can provide complementary signals and are crucial for sentence generations.

Thus, to fully exploit the past and future output contexts as well as to mitigate the exposure bias issue, one single decoder is employed to simultaneously perform decoding from L2R and R2L directions interactively. Our bidirectional decoding structure attempts to leverage the complementary nature of the two reverse decoding directions to yield accurate and reasonable language sentences. The decoder in each decoding direction outputs predictions under the semantic guide of visual hidden state vectors generated by the video semantic encoder, the past hidden state vectors obtained from the same decoding direction, and the future hidden state vectors obtained from the reverse decoding direction. Because of the symmetrical characteristic of L2R and R2L decoding, the L2R decoding is considered as an example to explain how the bidirectional decoding model works.

Given the visual hidden state vectors, all past hidden state vectors, and all future hidden state vectors, the bidirectional decoder models how to predict the next word from left-to-right as follows:

$$\tilde{s}_t = \text{GRU}^d\left(\left[E_w(\tilde{y}_{t-1}), \tilde{m}_t, H_t^\text{past}, H_t^\text{future}, \tilde{s}_{t-1}\right]\right)$$

where $\text{GRU}^d$ denotes the L2R GRU decoding network. $\tilde{s}_t$ represents the hidden state vector of L2R decoding at $t$-th time step. $E_w(y_{t-1})$ is the word embedding vector of word $y_{t-1}$. We calculate the visual context information $\tilde{m}_t$ at $t$-th time step by aggregating all attention weighted visual hidden state vectors.
\[ \bar{m}_t = \sum_{i=1}^{n} \bar{a}_i^i h_i \]  

where \( \sum_{i=1}^{n} \bar{a}_i^i = 1 \), and \( \bar{a}_i^i \) denotes the temporal attention weight which measures the relevance between the visual hidden state \( h_i \) and current decoding hidden state \( \bar{s}_i \). The relevance score \( \bar{a}_i^i \) is calculated by a temporal attention function as follows:

\[ \bar{c}_i^i = \bar{w}_a^T \tanh (\bar{w}_a \bar{s}_{t-1} + \bar{U}_d h_i + \bar{b}_a) \]  

where \( \bar{w}_a, \bar{w}_d, \bar{U}_d \) and \( \bar{b}_a \) denote the learnable model parameters.

Then a softmax function is applied to normalise the relevance score:

\[ \bar{a}_i^i = \exp (\bar{c}_i^i) / \sum_{i=1}^{n} \exp (\bar{c}_i^i) \]  

When the bidirectional decoder performs L2R decoding, \( H_{t \text{ past}} \) and \( H_{t \text{ future}} \) are defined as the past output contexts and future output contexts, which can be separately obtained from L2R and R2L decoding at \( t \)-th time step. We employ two separate cross-attentive MH-Att functions to calculate the past and future hidden state vectors that satisfies the demand of current decoding hidden state, and avoid under-description. Likewise, we introduce two different sentences at the same time.

\[ p(\bar{y}_t|\bar{y}_{<t}, \bar{y}_{<t}, \bar{V}; \bar{\theta}) = \text{softmax}(\bar{w}_s \bar{s}_t + \bar{b}_s) \]  

where \( \bar{w}_s \) and \( \bar{b}_s \) denote the learnable model parameters.

Similar to the calculation of sentence generation from left-to-right, the reverse sentence generation process is defined as follows:

\[ \bar{s}_t = \text{GRU}^{\text{R2L}}(E_{\bar{w}}(\bar{y}_{t-1}), \bar{m}_t, \bar{H}_{t \text{ past}}, \bar{H}_{t \text{ future}}, \bar{s}_{t-1}) \]  

\[ p(\bar{y}_t|\bar{y}_{<t}, \bar{y}_{<t}, \bar{V}; \bar{\theta}) = \text{softmax}(\bar{w}_s \bar{s}_t + \bar{b}_s) \]  

where \( \text{GRU}^{\text{R2L}} \) represents the R2L GRU decoding network. \( \bar{w}_s \) and \( \bar{b}_s \) are the trainable parameters of R2L decoding. We initialise the first hidden state vector \( \bar{s}_0 \) of R2L decoding using the first hidden state vector \( h_1 \) of video semantic encoder. \( \bar{m}_t \) (see Equation (20)), \( \bar{H}_{t \text{ past}} \) (see Equation (21)) and \( \bar{H}_{t \text{ future}} \) (see Equation (22)) indicate the reverse visual context information, past output contexts generated from R2L decoding, and future output contexts generated from L2R decoding, respectively.

\[ \bar{m}_t = \sum_{i=1}^{n} \bar{a}_i^i h_i \]  

\[ \bar{H}_{t \text{ past}} = \text{MH} \cdot \text{Attn}(\bar{s}_0, \bar{s}_t, \bar{s}_t) \]  

\[ \bar{H}_{t \text{ future}} = \text{MH} \cdot \text{Attn}(\bar{s}_t, \bar{s}_t, \bar{s}_t) \]  

It can be observed that the bidirectional decoder generates two different sentences at the same time.

### 3.4 Symmetric semantic-guided gated attention

The attention mechanism has been extensively applied in video captioning, which always yields a weighted average attention vector for any input decoding hidden state no matter whether or how well the attended features and the input decoding hidden state are relevant. We have calculated the past and future output contexts by cross-attentive MH-Att. The primary concern is that the past or future output contexts could be not what the bidirectional decoder expects at each time step and would mislead the word prediction process. This would happen when there is no relevant information in the past or future hidden state vectors that satisfies the demand of current decoding hidden state.

Here, we devise a SSGGA module as illustrated in Figure 3, which is designed for measuring the relevance between the past or future output contexts and the current decoding hidden state, and avoiding under-description. Likewise, we introduce...
this module from the perspective of L2R decoding, and the direction arrow sign will be omitted.

The SSGGA module first generates two symmetric attention gates $g_1$ and $g_2$ for the past and future output contexts, respectively. Each attention gate is determined by the semantic guidance factor $s_i$ and the corresponding output contexts $H_i^{\text{past}}$ or $H_i^{\text{future}}$.

$$g_1 = σ(W^1_s s_i + W^1_i H_i^{\text{past}} + b_i)$$

$$g_2 = σ(W^2_s s_i + W^2_i H_i^{\text{future}} + b_i)$$

where $W^1_s, W^2_s, W^1_i, W^2_i \in \mathbb{R}^{d \times d}, b_i, b_i^* \in \mathbb{R}^d$, and $σ$ represents sigmoid function. The value of each channel in the attention gate implies the correlation between the current decoding hidden state and the corresponding channel in the past or future output contexts. In other words, if each channel in $H_i^{\text{past}}$ or $H_i^{\text{future}}$ matches well with $s_i$, it receives a high gate value. Otherwise, it receives a low gate value. We then apply the two attention gates to the past and future output contexts by element-wise multiplication to filter out the irrelevant or misleading information, and obtain the gated past and future output contexts (i.e. $\tilde{H}_i^{\text{past}}$ and $\tilde{H}_i^{\text{future}}$ in Figure 3). Besides, a residual connection is employed to keep the original past or future output contexts.

$$\tilde{H}_i^{\text{past}} = g_1 \odot H_i^{\text{past}} + H_i^{\text{past}}$$

$$\tilde{H}_i^{\text{future}} = g_2 \odot H_i^{\text{future}} + H_i^{\text{future}}$$

where $\odot$ represents the element-wise multiplication. $\tilde{H}_i^{\text{past}}$ and $\tilde{H}_i^{\text{future}}$ denote the final gated attention results, which are refined under the guidance of semantic factor $s_i$.

After that, we use the ‘Combine’ function in Equation (27) to integrate the gated past and future output contexts by element-wise add (see Equation (28)), concatenation (see Equation (29)), low-rank bilinear (see Equation (30)) or gate mechanism (see Equation (31)).

$$H_i = \text{Combine}(\tilde{H}_i^{\text{past}}, \tilde{H}_i^{\text{future}})$$

$$H_i = \tilde{H}_i^{\text{past}} + λ \tilde{H}_i^{\text{future}}$$

$$H_i = W_i [\tilde{H}_i^{\text{past}}, \tilde{H}_i^{\text{future}}] + b_c$$

$$H_i = \tilde{H}_i^{\text{past}} U^T V \tilde{H}_i^{\text{future}}$$

$$r_t = σ(W_i [\tilde{H}_i^{\text{past}}, \tilde{H}_i^{\text{future}}] + b_c)$$

$$H_i = r_t \odot \tilde{H}_i^{\text{past}} + (1 - r_t) \odot \tilde{H}_i^{\text{future}}$$

where $[\cdot]$ denotes the concatenation operation, $W_o, U, V, W_r, b_c$ and $b_c$ are the learnable parameters. We set $λ$ to 0.1. $σ$ denotes the non-linear activation function. The sigmoid function is utilised in the implementation.

### 3.5 Training and Inference

The bidirectional decoding model is trained in two stages. We first optimise the proposed method by minimising the bidirectional cross-entropy (CE) loss defined in Equation (32), which involves both the L2R and R2L likelihood. Let $θ$ and $\bar{θ}$ represent the learnable parameters of L2R and R2L decoding, respectively. $w^*_{1:m}$ indicates the ground-truth caption.

$$L_{BCE}(θ, \bar{θ}) = -(\sum_{i=1}^{m} \log p(\tilde{w}^*_{1:i-1} | w^*_{1:m}, \bar{θ}))$$

$$+ \sum_{i=1}^{m} \log p(\tilde{w}^*_{1:i-1} | \bar{w}^*_{1:m}, θ))$$

In the second stage, the non-differentiable evaluation metrics are optimised with self-critical sequence training (SCST) [46], which can further provide significant performance improvement.

$$L_{RL}(θ) = -E_{w^*_{1:m}} p_t [r(w^*_{1:m})]$$

where the score of metric CIDEr is considered as the reward $r$ [47]. We approximate the gradients as follows:
\[ \nabla_{\theta} L_{RL}(\theta) \approx -(r(\mathbf{w}_{1:m}^t) - r(\overline{\mathbf{w}}_{1:m})) \nabla_{\theta} \log p(\mathbf{w}_{1:m}^t) \] (34)

\( \mathbf{w}^d \) indicates the video caption sampled from the probability distribution, and \( \overline{\mathbf{w}} \) is the video caption of greedy decoding.

4 | EXPERIMENTS

Now, we evaluate the symmetric bidirectional decoding model on two extensively applied benchmark datasets, including MSR-VTT [48] and MSVD [49]. We first give some brief introductions to the datasets and implementation details. Afterwards, our model will be compared to the unidirectional decoding models (i.e. L2R and R2L) and some recently proposed video captioning models.

4.1 | Datasets

4.1.1 | Microsoft Research Video to Text

The MSR-VTT dataset is a benchmark dataset for video captioning which comprises 10,000 open-domain videos classified into 20 different categories. The Amazon mechanical turk (AMT) workers annotate each video with 20 English captions. For fair comparison, we adopt the standard splits in [48] which divides the videos into 6513 for training, 497 for validation, and 2990 for testing.

4.1.2 | Microsoft Video Description Corpus

MSVD is another extensively applied dataset in video captioning which consists of 1970 videos collected from the YouTube website. Each video depicts an activity scene and is annotated with 40 English captions by AMT. Likewise, we follow the standard splits in [5, 25, 50] to separate the dataset with 1200 videos for training, 100 for validation, and 670 for testing, respectively.

4.2 | Metrics

There are four standard metrics employed to evaluate the experimental performance, including BLEU@n [51], ROUGE-L [52], METEOR [53], and CIDEr [47].

The BLEU metric estimates the precision of n-grams (\( n = 1, 2, 3, 4 \)) between predicted captions and ground-truth captions. This metric highly agrees with human assessment results. The ROUGE-L metric calculates the mean of accuracy and the recall of the longest common subsequence between the compared captions. The METEOR metric measures the word correspondence between the ground-truth and predicted captions by generating the alignment. The CIDEr metric computes the mean cosine similarity of n-grams between the compared captions, weighted by Term Frequency-Inverse Document Frequency (TF-IDF). Generally, METEOR and CIDEr are more semantically accurate, and CIDEr is specially designed for video captioning.

In addition, these evaluation metrics have been implemented and supplied by the Microsoft COCO evaluation server. Therefore, we can directly make use of these evaluation functions to compare the experimental results of various methods. For all the metrics, higher values mean better results.

4.3 | Implementation details

4.3.1 | Video processing

We resize the video frames to the standard size 224 x 224 and then feed the video frames into the pre-trained ResNet152 [54], which returns a sequence of 2048-dimensional feature vectors. We sample 40 equally-spaced features from the sequence of feature vectors. If the total number of video features is less than 40, we pad them with zero vectors.

4.3.2 | Sentence processing

We first convert the ground-truth sentence into lowercase, tokenize each sentence into words using the word_tokenize function in the NLTK toolbox, and eliminate the punctuations. Moreover, we add the starting position of each sentence with a label <SOS> and the ending position with a label <EOS>. Unknown words are set to the label <UNK>. Afterwards, a vocabulary is built with 8488 words for MSVD and 16,848 words for MSR-VTT.

4.3.3 | Training details

We embed each word into a 512-dimensional word vector. For the video encoder, the self-attentive MH-Att module has \( N = 3 \) identical sub-encoders, \( H = 8 \) attention heads, and the BiGRU has a 1024-dimensional hidden size. For the bidirectional decoder, the cross-attentive MH-Att module has \( H = 8 \) attention heads, and the GRU has 1024 hidden units. Moreover, the regularisation technique Dropout [55] with a rate value 0.5 is employed to the input and output of GRUs for preventing overfitting. We initialise all the model parameters from a distribution on \([-0.1, 0.1]\), and fix the batch size to 64. Besides, the Adam [56] algorithm is leveraged to optimise the bidirectional decoding model by minimising the negative CE loss with a learning rate of \( 10^{-4} \). We train the bidirectional decoder for 500 epochs and the learning rate is decayed by 0.5 for every 50 epochs. The model gradients are clipped between \(-5\) and \(+5\) for avoiding gradient explosion. The proposed model is executed using PyTorch by an NVIDIA RTX 2080Ti GPU with 11G memory. The beam size is five for the final sentence generation at the testing time.
4.4 | Performance comparison

To validate the effectiveness of the bidirectional decoding structure, the proposed method is compared with the unidirectional decoding methods (L2R and R2L) and several recently proposed video captioning methods.

- **MP-LSTM [24]**: The mean pooling-LSTM (MP-LSTM) method applies the mean pooling operation to all the sampled features to obtain the global video representation, which is taken as the input to an LSTM-based sentence decoder.

- **S2VT [25]**: The sequence to sequence video to text (S2VT) method extracts both RGB and optical flow features and applies the LSTM to the video encoder and sentence decoder.

- **SA-LSTM (L2R) [50]**: The SA-LSTM method adopts the soft attention (SA) to selectively concentrate on the most relevant visual contents at each time step. We take SA-LSTM as the unidirectional decoding model, which generates the descriptions from left-to-right, and our model is designed based on the SA-LSTM method.

- **SA-LSTM (R2L) [50]**: The SA-LSTM (R2L) method is a variant of SA-LSTM (L2R), which generates the sentence in a right-to-left manner.

- **h-RNN [59]**: The h-RNN method introduces a hierarchical-RNN (h-RNN) framework, which leverages the inter-sentence dependency to generate multiple video descriptions.

- **M3 [4]**: The M3 method designs a shared memory structure that selectively reads/writes both visual contents and sentences for video captioning.

- **hLSTMat [57]**: The hLSTMat method proposes an adjusted temporal mechanism that determines the usage of visual and textual information.

- **RecNet [5]**: The RecNet method designs an encoder-decoder-reconstructor structure to make use of the bidirectional flow between videos and sentences.

- **BAE [28]**: The boundary-aware encoder (BAE) introduces a boundary-aware LSTM cell that recognises the discontinuities of videos by extracting the hierarchical structures of videos.

- **aLSTMs [34]**: The aLSTMs method utilises the hierarchical attention mechanism to concentrate on the discriminative video frames and then explores the correlation between the words and video contents.

- **PickNet [30]**: The PickNet method puts forward a plug-and-play reinforcement-learning-based network that selects the most informative video frames for avoiding content noise.

- **MCNN [29]**: The multi-stage CNN (MCNN) method designs a Multi-modal circulant fusion (MCF) module to exploit the interactions among visual-textual features.

- **JSRVC [12]**: The joint syntax representation and visual cue (JSRVC) method jointly learns the syntactic structure of caption through inferring the POS tags from videos and utilises a mixture model to translate the video features into words based on the learned syntax structure.

- **GRU-EVE [11]**: The GRU-EVE method devises a novel video encoding module that enriches the video representations with spatio-temporal dynamical information and high-level semantic concepts.

- **CAM-RNN [7]**: The CAM-RNN method introduces a co-attention model based RNN to extract the most relevant visual and textual features to the generated sentences.

- **ASM [60]**: The ASM method puts forward an attribute selection mechanism (ASM) to automatically focus on the useful attributes.

- **STAT [13]**: The STAT method designs a spatial-temporal attention (STAT) mechanism that considers the spatial and temporal structures of videos simultaneously and catches the significant regions of the most relevant video frames when predicting each word.

- **ConvRS [58]**: The ConvRS method presents a convolutional reconstruction-to-sequence framework, which uses the inter-frame difference to generate discriminative video representation and then adopts the multiple dilated convolution layers to generate sentences.

Table 1 shows the comparison results between the unidirectional decoding (L2R, R2L) and bidirectional decoding models on MSR-VTT and MSVD dataset, respectively. For fair comparison, we remove the self-attentive MH-Att module from the full model and extract the video features merely with a BiGRU network, the same with the unidirectional decoding models. We can observe that the bidirectional decoding model makes a significant improvement over the unidirectional decoding (L2R) model by 6.1% and 10.2% (CIDEr) on MSR-VTT and MSVD dataset when only trained by CE loss. These results strongly indicate the effectiveness of bidirectional decoding structure in mitigating the exposure bias issue and regularising the output predictions to be more consistent with the visual semantics. Moreover, we train our model on the basis of the reinforcement learning (RL), which further provides a noticeable gain to the captioning performance in most evaluation metrics, except BLEU@4.

| TABLE 1 | Compared results of L2R-unidirectional, R2L-unidirectional, and Bidirectional decoding structures with BLEU@$n(n=1,4)$, METEOR, ROUGE-L, CIDEr on the MSR-VTT and MSVD dataset (%) |
|----------------|-----------------|-----------------|-----------------|-----------------|
| **Methods** | **MSR-VTT** | **MSVD** | **MSR-VTT** | **MSVD** |
| | **B@1** | **B@4** | **M** | **R** | **C** | **B@1** | **B@4** | **M** | **R** | **C** |
| L2R | 76.5 | 36.4 | 26.2 | 57.9 | 40.9 | 75.3 | 40.1 | 30.9 | 66.6 | 68.6 |
| R2L | 75.7 | 35.3 | 25.8 | 57.1 | 40.5 | 73.4 | 38.9 | 29.3 | 64.5 | 65.3 |
| Proposed (CE) | 78.7 | 38.5 | 27.8 | 59.1 | 43.4 | 77.8 | 44.3 | 32.8 | 67.9 | 75.6 |
| Proposed (CE + RL) | 79.1 | 37.9 | 28.4 | 59.5 | 45.1 | 78.3 | 43.6 | 33.3 | 68.5 | 76.9 |

*Note: CE and RL denote cross entropy and reinforcement learning, respectively.*
In Tables 2 and 3, we illustrate the quantitative comparison results of several recently proposed video captioning methods on the MSR-VTT and MSVD dataset, respectively. It should be noted that not all the compared methods are trained by RL, so our method will be compared to other state-of-the-art methods merely under the CE loss for fair comparison. We can see that our symmetric bidirectional decoding model obtains the best performance in terms of METEOR and CIDEr and

**Table 2** Compared results with various video captioning methods of MSR-VTT dataset (%)

| Methods          | B@4 | M   | R   | C    |
|------------------|-----|-----|-----|------|
| S2VT [25]        | 31.4| 25.7| 55.9| 35.2 |
| MP-LSTM [24]     | 34.8| 24.7| -   | -    |
| SA-LSTM [50]     | 36.6| 25.9| -   | -    |
| aLSTMs [34]      | 38.0| 26.1| -   | 43.2 |
| M3 [4]           | 38.1| 26.6| -   | -    |
| hLSTM at [57]    | 38.4| 26.3| -   | 43.4 |
| RecNet local [5] | 39.1| 26.6| 59.3| 42.7 |
| PickNet [30]     | 38.9| 27.2| 59.5| 42.1 |
| MCNN [29]        | 38.1| 27.2| -   | 42.1 |
| STAT [13]        | 39.3| 27.1| -   | 43.8 |
| ConvRS [58]      | 38.1| 27.2| -   | 42.1 |
| GRU-EVE [11]     | 32.9| 26.4| 57.2| 39.2 |
| Proposed         | 39.2| 27.8| 59.4| 44.5 |

Note: The symbol "-" means that this metric is unreported.

**Table 3** Compared results with various video captioning methods of MSVD dataset (%)

| Methods          | B@4 | M   | R   | C    |
|------------------|-----|-----|-----|------|
| SA-LSTM [50]     | 41.9| 29.6| -   | 51.7 |
| h-RNN [59]       | 49.9| 32.6| -   | 65.8 |
| aLSTMs [34]      | 50.8| 33.3| -   | 74.8 |
| M3 [4]           | 49.3| 31.1| -   | -    |
| BAE [28]         | 42.5| 32.4| -   | 63.5 |
| hLSTM at [57]    | 53.0| 33.6| -   | 73.8 |
| RecNet local [5] | 52.3| 34.1| 69.8| 80.3 |
| PickNet [30]     | 46.1| 33.1| 69.2| 76.0 |
| MCNN [29]        | 46.5| 33.7| -   | 75.5 |
| STAT [13]        | 52.0| 33.3| -   | 73.8 |
| JSRVC [12]       | 52.1| 33.7| 69.8| 80.6 |
| CAM-RNN [7]      | 42.4| 33.4| 69.4| 54.3 |
| ASM [60]         | 50.6| 34.2| -   | 76.1 |
| GRU-EVE [11]     | 45.6| 33.7| 69.8| 74.2 |
| Proposed         | 47.7| 34.3| 69.4| 81.6 |

Note: The symbol "-" means that this metric is unreported.

4.5 | Ablation study

This section investigates the impact of different components in the proposed method, and for that we devise several ablated models with different settings by removing the specific component, as illustrated in Table 4. We first define a ‘base’ model, which employs the BiGRU network to learn the video contextual features and drops the SSGGA component in the

**Table 4** Setting and results of ablation study of MSR-VTT dataset (%)

| Methods                | B@4 | M   | R   | C    |
|------------------------|-----|-----|-----|------|
| Base(Enc:BiGRU + Dec:Sh-Att) | 37.2| 27.1| 58.1| 41.5 |
| +Enc:MH-Att            | 37.8| 27.3| 58.7| 42.0 |
| +Dec:MH-Att            | 37.4| 27.4| 58.5| 42.2 |
| +Dec:SSGGA             | 37.7| 27.6| 58.7| 42.4 |
| +Dec:MH-Att + SSGGA    | 38.5| 27.8| 59.1| 43.4 |
| Full                   | 39.2| 27.8| 59.4| 44.5 |

**Table 5** Compared results with various settings of video semantic encoder on the MSR-VTT dataset (%)

| Methods               | B@4 | M   | R   | C    |
|-----------------------|-----|-----|-----|------|
| Encoder (N = 3, H = 8) | 39.2| 27.8| 59.4| 44.5 |
| Encoder (N = 1, H = 8) | 38.7| 27.7| 59.1| 43.5 |
| Encoder (N = 2, H = 8) | 38.9| 27.9| 59.2| 43.7 |
| Encoder (N = 4, H = 8) | 39.1| 27.9| 59.3| 44.2 |
| Encoder (N = 5, H = 8) | 38.9| 27.8| 59.3| 43.9 |
| Encoder (N = 6, H = 8) | 38.8| 27.7| 59.2| 43.6 |
| Encoder (N = 3, H = 1) | 38.6| 27.6| 58.8| 43.9 |
| Encoder (N = 3, H = 4) | 38.9| 27.7| 59.2| 44.1 |

Note: N denotes the number of layers and H denotes the number of attention heads.
bidirectional decoder as well as exploits the single-head attention (SH-Att) to compute the past and future output contexts. Considering the restricted time and computation resource, we only implement the ablation experiments on the MSR-VTT dataset.

**Effect of video semantic encoder** To validate the effectiveness of the self-attentive MH-Att mechanism in capturing long-term semantic dependencies in videos, we compare our video encoding scheme with the merely BiGRU-based encoding. The self-attentive MH-Att mechanism is suited to the modelling of frame-to-frame interactions. From Table 4, it can be observed that the self-attentive video representations have positive impacts and improve the CIDEr score by 0.5.

**Effect of symmetric semantic-guided gated attention** We compare the experimental performance of leveraging different attention mechanisms to calculate the past or future output contexts. We perform experiments with cross-attentive SH-Att and cross-attentive MH-Att. It is notable from Table 4 that replacing SH-Att with MH-Att can obtain slightly better results. Moreover, applying the SSGGA module to the bidirectional decoder can further improve the performance by 2.84% (CIDEr), which substantially implies the effectiveness of the SSGGA module in suppressing the irrelevant or misleading contents in the past or future output contexts and keeping the useful ones. This module can also effectively avoid under-description. Besides, the gated attention module is lightweight that requires little computation.

**Effect of various settings of video semantic encoder** The self-attention layer number and attention heads control the

| Methods            | B@1 | B@4 | M   | R   | C   |
|--------------------|-----|-----|-----|-----|-----|
| Concatenation      | 78.4| 37.9| 27.6| 58.6| 42.4|
| Element-wise add   | 78.7| 38.5| 27.8| 59.1| 43.4|
| Bilinear pooling   | 78.7| 38.3| 27.9| 58.8| 43.2|
| Gate               | 78.5| 37.6| 27.5| 58.7| 42.1|

**FIGURE 4** Visualisation of the qualitative examples on the MSR-VTT dataset with different video captioning models. For each video, six frames are sampled and four captions are given, including the captions generated by the proposed model, L2R decoding model and R2L decoding model, and the ground-truth caption (GT). Compared to the unidirectional decoding models, our proposed model is capable of yielding more descriptive and fine-grained words highlighted in red, for example ‘eyes’, ‘audience’, ‘commentating’, and ‘guitar’.
contribution of the self-attentive MH-Att module in capturing long-term semantic dependencies from video contexts and excavating discriminative visual semantic features. Table 5 illustrates the experimental results of various settings with $N$ varying from one to six and $H$ chosen from one, four or eight. We can observe that increasing $N$ can boost the performance improvement, and the performance (BLEU@4, ROUGE-L, CIDEr) drops when $N$ is larger ($N \geq 3$). It is probable that stacking more self-attention layers could make the video encoder easier to lose low-level semantic information. Moreover, the experimental results clearly show that a larger $H$ provides significant gains to the captioning performance and is helpful to generate more discriminative video features. Thus, the 3-layer, 8-head setting is the best choice in our implementation.

**Effect of combination methods** We investigate the influence of several fusion schemes on the captioning performance, including concatenation, element-wise add, bilinear pooling, and gate mechanism. The performance comparison results are illustrated in Table 6. We can observe that element-wise add provides slightly gains than other combination methods in BLEU@4, ROUGE-L, and CIDEr. Thus, combining the past and future output contexts appropriately can provide robust guidance to the word prediction process.

4.6 Qualitative analysis

Although some standard evaluation metrics have validated the superiority of the symmetric bidirectional decoding structure, it is still not intuitive to see the quality of the generated video descriptions. Figure 4 illustrates the qualitative examples of our video captioning method from eight videos, and six frames are selected to represent each video. It can be concluded that our method has the ability to generate descriptions with more accurate semantic meanings than the unidirectional decoding methods (L2R, R2L). The examples (a), (b), (e), (f) demonstrate that our method can not only express the general meanings of the whole video, but also is capable of identifying more detailed objects ('eyes', 'brush', 'suit', 'audience'), attributes ('red and blue dress') and actions ('playing guitar in stage'). Our method can also ground larger-scale semantic concepts, such as 'A group of people' in example (b). In addition, in example (d) and (g), our model can recognise the video game scenes rather than a normal environment. In particular, example (g) is able to discern that a 'man' is playing a video game and 'commentating' on the game without any person appearing on the screen. According to the above observations, we can conclude that the bidirectional decoding structure can contribute to generating more robust and fine-grained video captions.

5 Conclusion

The dominant video captioning methods usually adopt the attentional encoder-decoder framework, so they might suffer from the exposure bias problem and leave the future output contexts generated from the R2L decoding unexplored. Here, the authors introduce a new symmetric bidirectional decoder for video captioning. Unlike previous methods, the past and future output contexts are both taken into account, and our method can learn the capability of employing one single decoder to generate the video descriptions from the L2R and R2L directions simultaneously. We also devise an effective symmetric semantic-guided gated attention module to suppress the irrelevant or misleading contents in the past or future output contexts, which provides the accurate guidance to the word prediction process. The L2R and R2L likelihood losses are integrated together to optimise the full model. Extensive experiments conducted on two widely-used datasets verify the superiority of our proposed method compared to the unidirectional decoding (L2R, R2L) methods and most recently proposed methods. We provide some qualitative examples to intuitively visualise the effectiveness of the symmetric bidirectional decoder. Consequently, our proposed method can generate accurate and robust video captions.

REFERENCES

1. Zeng, K., et al.: Leveraging video descriptions to learn video question answering. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, pp. 4334–4340. San Francisco (2017)
2. Yu, Y., et al.: End-to-end concept word detection for video captioning, retrieval, and question answering. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3261–3269.CVPR 2017, Honolulu (2017)
3. Vorkinska, V., et al.: How blind people interact with visual content on social networking services. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, pp. 1582–1593.CSCW 2016, San Francisco (2016)
4. Wang, J., et al.: M3: multimodal memory modelling for video captioning. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, pp. 7512–7520.IEEE Computer Society, Salt Lake City (2018)
5. Wang, B., et al.: Reconstruction network for video captioning. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, pp. 7622–7631.IEEE Computer Society, Salt Lake City (2018)
6. Wang, J., et al.: Bidirectional attentive fusion with context gating for dense video captioning. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, pp. 7190–7198.IEEE Computer Society, Salt Lake City (2018)
7. Zhao, B., Li, X., Lu, X.: CAM-RNN: co-attention model based RNN for video captioning. IEEE Trans. Image Process. 28(11), 5552–5565 (2019)
8. Bin, Y., et al.: Describing video with attention-based bidirectional LSTM. IEEE Trans. Cybern. 49(7), 2631–2641 (2019)
9. Zhang, J., Peng, Y.: Object-aware aggregation with bidirectional temporal graph for video captioning. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, pp. 8327–8336. Long Beach (2019)
10. Pei, W., et al.: Memory-attended recurrent network for video captioning. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, pp. 8347–8356 (2019)
11. Aafaq, N., et al.: Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, pp. 12487–12496. Long Beach (2019)
12. Hou, J., et al.: Joint syntax representation learning and visual cue translation for video captioning. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, pp. 8917–8926.IEEE, Seoul (2019)
13. Yan, C., et al.: STAT: spatial-temporal attention mechanism for video captioning. IEEE Trans. Multim. 22(1), 229–241 (2020)
14. Chung, J., Cho, K., Bengio, Y.: A character-level decoder without explicit segmentation for neural machine translation (2016). arXiv preprint arXiv:160306147
15. Johnson, M., et al.: Google’s multilingual neural machine translation system: enabling zero-shot translation. Trans. Assoc. Comput. Linguist. 5, 339–351 (2017)
16. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. In: Proceedings of the 32nd international conference on machine learning, ICML 2015, Lille. Vol. 37 of JMLR Workshop and Conference Proceedings, pp. 2048–2057. JMLR.org (2015)
17. Huang, L., et al.: Attention on attention for image captioning. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, pp. 4633–4642. IEEE, Seoul (2019)
18. Luo, X., et al.: Attention-based relation extraction with bidirectional gated recurrent unit and highway network in the analysis of geological data. IEEE Access, 6, 5705–5715 (2018)
19. Kojima, A., Tamura, T., Fukunaga, K.: Natural language description of human activities from video images based on concept hierarchy of actions. Int. J. Comput. Vis. 50(2), 171–184 (2002)
20. Rohrbach, M., et al.: Translating video content to natural language descriptions. In: IEEE International Conference on Computer Vision, ICCV 2013, pp. 433–440. IEEE Computer Society, Sydney (2013)
21. Hancke, M., Schutte, K., Burghouts, G.J.: Automated textual descriptions for a wide range of video events with 48 human actions. In: Computer Vision - ECCV 2012. Workshops and Demonstrations, Florence, Proceedings, Part I. vol. 7583 of Lecture Notes in Computer Science, pp. 372–380. Springer (2012)
22. Krishnamoorthy, N., et al.: Generating natural-language video descriptions using text-mined knowledge. In: Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, July 14-18, 2013, Bellevue, AAAI Press (2013)
23. Guadarrama, S., et al.: YoutTube2text: recognizing and describing arbitrary human activities from video images based on deep recurrent neural networks (2014). arXiv preprint arXiv:14124729
24. Venugopalan, S., et al.: Translating videos to natural language using deep recurrent neural networks (2014). arXiv preprint arXiv:14124729
25. Venugopalan, S., et al.: Sequence to sequence - video to text. In: IEEE International Conference on Computer Vision, ICCV 2015, Santiago, pp. 4534–4542. IEEE Computer Society (2015)
26. Pan, P., et al.: Hierarchical recurrent neural encoder for video representation with application to captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1029–1038. IEEE Computer Society, Las Vegas (2016)
27. Ramanishka, V., et al.: Multimodal video description. In: Proceedings of the 2016 ACM Conference on Multimedia Conference, pp. 1092–1096. MM 2016, ACM, Amsterdam (2016)
28. Baraldi, L., Grana, C., Cucchiara, R.: Hierarchical boundary-aware neural encoder for video captioning. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pp. 3185–3194. IEEE Computer Society, Honolulu (2017)
29. Wu, A., Han, Y.: Multi-modal circulant fusion for video-to-language and backward. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, pp. 1029–1035. IJCAI.org, Stockholm (2018)
30. Chen, Y., et al.: Less is more: picking informative frames for video captioning of Lecture Notes in Computer Science. In: Computer Vision – ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII, vol. 11217, pp. 367–384. Springer (2018)
31. Zhang, Z., et al.: Object relational graph with teacher-recommended learning for video captioning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, pp. 13275–13285. IEEE, Seattle (2020)
32. Sun, C., et al.: Video2obert: a joint model for video and language representation learning (South). In: IEEE/CVF International Conference on Computer Vision, ICCV 2019, pp. 7463–7472. IEEE, Seoul (2019)
33. Zhou, L., et al.: End-to-end dense video captioning with masked transformer. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, pp. 8739–8748. IEEE Computer Society, Salt Lake City (2018)
34. Gao, L., et al.: Video captioning with attention-based lstm and semantic consistency. IEEE Trans. Multimed. 19(9), 2045–2055 (2017)
35. Zhang, X., et al.: Asynchronous bidirectional decoding for neural machine translation. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), pp. 5068–5070. AAAI Press, New Orleans (2018)
36. Zhou, L., Zhang, J., Zong, C.: Synchronous bidirectional neural machine translation. Trans. Assoc. Comput. Linguist., 7, 91–105 (2019)
37. Zheng, Y., et al.: Forward-backward decoding for regularising end-to-end tts (2019). arXiv preprint arXiv:190709006
38. Wang, X., et al.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7794–7803. IEEE Computer Society, Salt Lake City (2018)
39. Lu, J., et al.: Hierarchical question-image co-attention for visual question answering. An. In: Conference on Neural Information Processing Systems, vol. 29, pp. 289–297 (2016)
40. Lu, J., et al.: Knowing when to look: adaptive attention via a visual sentinel for image captioning. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pp. 3242–3250. IEEE Computer Society, Honolulu (2017)
41. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, pp. 5998–6008. Long Beach (2017)
42. Yu, D., et al.: Multi-level attention networks for visual question answering. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, pp. 21–29 (2016)
43. Cho, K., et al.: Learning phrase representations using rnn encoder-decoder for statistical machine translation (2014). arXiv preprint arXiv:14061078
44. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. 9(8), 1735–1780 (1997)
45. Rennie, S.J., et al.: Self-critical sequence training for image captioning. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pp. 1179–1195. IEEE Computer Society, Honolulu (2017)
46. Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: consensus-based image description evaluation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, pp. 4566–4575. IEEE Computer Society, Boston (2015)
47. Xu, J., et al.: Msr-vtt: A large video description dataset for bridging video and language. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5288–5296. IEEE Computer Society, Las Vegas (2016)
48. Chen, D.L., Dolan, W.B.: Collecting highly parallel data for paraphrase evaluation. In: The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 190–200. The Association for Computer Linguistics, Portland (2011)
49. Yao, L., et al.: Describing videos by exploiting temporal structure. In: IEEE International Conference on Computer Vision, ICCV 2015, Santiago, pp. 7794–7803. IEEE Computer Society, Salt Lake City (2018)
50. Papineni, K., et al.: Bleu: a method for automatic evaluation of machine translation. Trans. Assoc. Comput. Linguist. 7, 91–105 (2019)
51. Papineni, K., et al.: Bleu: a method for automatic evaluation of machine translation. Trans. Assoc. Comput. Linguist. 7, 91–105 (2019)
52. Lin, C.Y.: Rouge: a package for automatic evaluation of summaries. In: Workshop on Text Summarization Branches Out at ACL (2004)
53. Banerjee, S., Lavie, A.: METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the
54. He, K., et al.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, pp. 770–778. IEEE Computer Society (2016)

55. Srivastava, N., et al.: Dropout: a simple way to prevent neural networks from overfitting. J. Machine Learn. Res. 15(1), 1929–1958 (2014)

56. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization (2014). arXiv preprint arXiv:1412.6980

57. Song, J., et al.: Hierarchical lstm with adjusted temporal attention for video captioning (2017). arXiv preprint arXiv:1706.01231

58. Wu, A., et al.: Convolutional reconstruction-to-sequence for video captioning. IEEE Trans. Circuits Syst. Video Technol. 30(11), 4299–4308 (2020)

59. Yu, H., et al.: Video paragraph captioning using hierarchical recurrent neural networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, pp. 4584–4593. IEEE Computer Society, Las Vegas (2016)

60. Xiao, H., Shi, J.: A novel attribute selection mechanism for video captioning. In: 2019 IEEE International Conference on Image Processing, pp. 619–623. ICIP, IEEE, Taipei (2019)

How to cite this article: Qi S, Yang L. Video captioning via a symmetric bidirectional decoder. IET Comput. Vis. 2021;15:283–296. https://doi.org/10.1049/cvi2.12043