Controllable Video Captioning with POS Sequence Guidance Based on Gated Fusion Network

Bairui Wang[1*]   Lin Ma[2†]   Wei Zhang[1†]   Wenhao Jiang[2]   Jingwen Wang[2]   Wei Liu[2]

[1]School of Control Science and Engineering, Shandong University   [2]Tencent AI Lab

{bairuiwong, forest.linma, cswhjiang, jaywongjaywong}@gmail.com

davidzhang@sdu.edu.cn   wl2223@columbia.edu

Abstract

In this paper, we propose to guide the video caption generation with Part-of-Speech (POS) information, based on a gated fusion of multiple representations of input videos. We construct a novel gated fusion network, with one particularly designed cross-gating (CG) block, to effectively encode and fuse different types of representations, e.g., the motion and content features of an input video. One POS sequence generator relies on this fused representation to predict the global syntactic structure, which is thereafter leveraged to guide the video captioning generation and control the syntax of the generated sentence. Specifically, a gating strategy is proposed to dynamically and adaptively incorporate the global syntactic POS information into the decoder for generating each word. Experimental results on two benchmark datasets, namely MSR-VTT and MSVD, demonstrate that the proposed model can well exploit complementary information from multiple representations, resulting in improved performances. Moreover, the generated global POS information can well capture the global syntactic structure of the sentence, and thus be exploited to control the syntactic structure of the description. Such POS information not only boosts the video captioning performance but also improves the diversity of the generated captions. Our code is at: https://github.com/vsislab/Controllable_XGating.

1. Introduction

Video captioning [18, 56, 46] aims at automatically describing rich content in videos with natural language, which is a meaningful but challenging task for bridging vision and language. This task can be applied for high-level video understanding in a variety of practical applications, such as visual retrieval [26, 37, 48, 24], visual question answer-
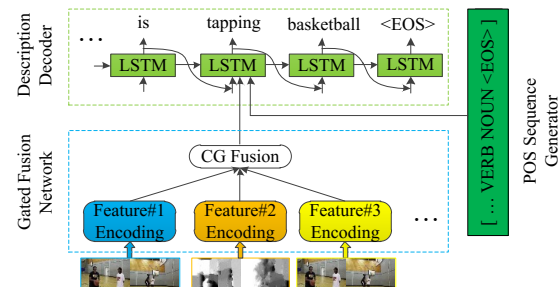

Figure 1. The proposed model for video captioning consists of a gated fusion network, a POS sequence generator, and a description generator. The gated fusion network extracts diverse features from videos, encodes, and fuses them together to generate a more representative video feature. Relying on the global syntactic POS information generated from the POS sequence generator and the fused video feature, the description generator produces one sentence describing the video content.

ing [25, 9], and so on. Video captioning is related to image captioning which describes an image with a sentence, as a video can be regarded as a sequence of images. However, what makes video captioning more challenging than image captioning [12, 44, 7, 17] is not only that the input of video captioning are multiple images, but also that video contains richer semantics, such as spatial/temporal information, content/motion information, and even speech information. Obviously, the existing approaches with one single kind feature [11, 43, 45, 46] are hard to comprehensively exploit the semantic meaning of a video.

Recently, researches on describing videos from diverse representations, such as Inception_ResNet_V2 [38], C3D [40], and I3D [2], have proved that multiple features can improve the video captioning models [42, 53, 27, 28]. It is reasonable as different features can capture video semantic information from different perspectives. However, to the best of our knowledge, the existing methods simply concatenate different representations together, while neglect the relationships among them, which play an important role in fully characterizing the video semantic meaning.

Prior video captioning methods also neglect the syntactic

---

structure of a sentence during the generation process. Analogic to the fact that words are the basic composition of a sentence, the part-of-speech (POS) [10] information of each word in a sentence is the basic structure of the grammar. Therefore, the POS information of the generated sentence is able to act as one prior knowledge to guide and regularize the sentence generation, if it can be obtained beforehand. Specifically, with the obtained POS information, the decoder is aware of the POS information of the word to be generated. As such, it may help reduce the search space of the target word, which is believed to benefit the video captioning. Besides, the changing of POS information, which can be seen as the prior knowledge of the description, is excepted to help generate sentence with more diverse syntax.

In order to fully exploit the relationships among different representations and the POS information, we propose a novel model to describe videos with POS guidance based on the gated fusion results of multiple representations, as shown in Fig. 1. First, a novel gated fusion network relies on a particularly designed cross gating (CG) block to mutually gate diverse features with respect to each other. As such, we can make a comprehensive representation of the video. One POS sequence generator relies on the fused video representation to yield the global POS information. Afterwards, the decoder relies on a gating strategy to dynamically and adaptively incorporate the generated global syntactic POS information for generating each word.

To summarize, the contributions of this work lie in threefold: 1) We propose a novel video captioning model, which relies on a gated fusion network incorporating multiple features information together and a POS sequence generator predicting the global syntactic POS information of the generated sentence. 2) A cross gating (CG) strategy is proposed to effectively encode and fuse different representations. The global syntactic POS information is adaptively and dynamically incorporated into the decoder to guide the decoder to produce more accurate description in terms of both syntax and semantics. 3) Extensive results on benchmark datasets indicate that the proposed fusion strategy can capture the relationships among multiple representations and descriptions with diverse syntax can be obtained by controlling the global POS sequence.

2. Related Work

2.1. Video Captioning

Previous works on video captioning adopt temporal-based methods [21, 15, 34, 33, 52], which define a sentence template with grammar rules. The sentence is parsed into subject, verb and object, each of which is aligned with video content. Obviously, under the predefined template with fixed syntactic structure, those methods are hard to generate flexible language descriptions.

Nowadays, benefit from the success of CNN and RNN, the sequence learning methods [42, 53, 27, 28, 8, 45, 49] are widely used to describe video content with flexible syntactic structure. In [43], Venugopalan *et al*. obtained video representation by averaging CNN feature of each frame, which ignored the temporal information. Compare to the average pooling, Yao *et al*. and Yu *et al*. employed the soft attention mechanism to dynamically summarize all frame representations [53, 54]. Recently, to exploit more semantic information, Pan *et al*. modeled the semantic-level correlation of sentence and video with a visual-semantic embedding model [27]. To avoid the negative impact of redundant visual information, Chen *et al*. proposed a PickNet to choose key frames [8].

More recently, different features can help characterizing the video semantic meaning from different perspectives. Many existing works utilize the motion information [42], temporal information [4, 18, 31], and even the audio information [51] to yield competitive performance. However, the diverse features in these works are simply concatenated with each other, which ignores the relationship among them. It is possible to further improve performance with a better fusion strategy. In this paper, we design a gated fusion network to dynamically learn and highlight the correlation between different features, which is expected to fully depict and characterize the video semantic meaning.

2.2. Captioning with POS Information

To the best of our knowledge, the POS tag information of language description has not been introduced in the video captioning task. While in image captioning, Deshpande *et al*. treated the entire POS tag sequence given by benchmark dataset as a sample, and divided them in 1024 categories by a k-medoids cluster [10], which limits the diversity of POS sequence information. He *et al*. controlled the input of image representations based on the predefined POS tag information of each ground-truth word [16], which can hardly obtained in practical scenario. In contrast, we predict POS sequence tag by tag, and embed them as a global POS feature to provide approximate global view on syntactic structure of the sentences. More importantly, the syntactic structure of description is controllable by changing the POS sequence manually.

3. Architecture

Given a video sequence, video captioning aims to generate a natural sentence $S = \{s_1, s_2, \ldots, s_n\}$ to express its semantic meaning, where n denotes the length of a sentence. In this paper, we would like to make a full exploitation of the video sequence by considering diverse video features. Moreover, we also want to predict the syntactic information of the generated sentence, specifically the POS in-
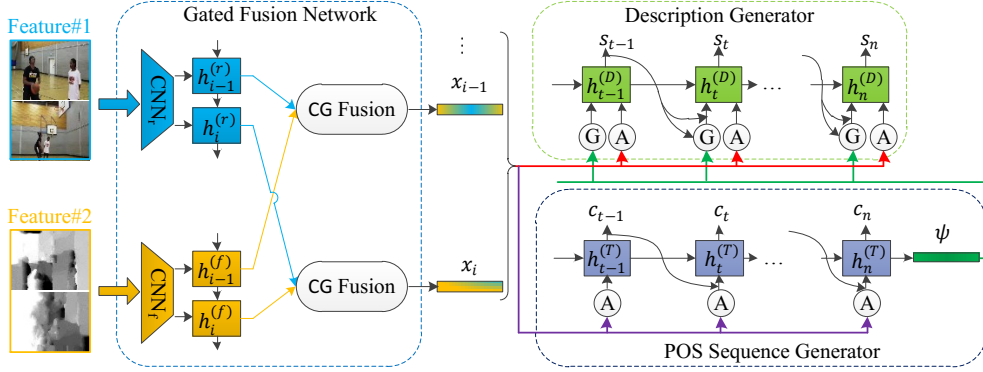
Figure 2. The proposed model for video captioning consists of three components. The gated fusion network encodes and fuses multiple video representations extracted by different CNN networks. The POS generator relies on the fused video representation to predict the global syntactic POS information of the sentence to be generated. The decoder adaptively and dynamically incorporates the global POS information for generating each targeting word. Ⓖ denotes the cross gating mechanism and Ⓐ denotes the soft attention mechanism.

formation $C = \{c_1, c_2, \ldots, c_n\}$, which is thereafter leveraged for guiding the sentence generation.

We propose one model for video captioning, realized in an encoder-decoder architecture, which consists of a gated fusion network, a POS sequence generator, and a description generator, as shown in Fig. 2. The gated fusion network learns to exploit the relationship among different video features to make a comprehensive understanding of the video sequence. The POS sequence generator learns to exploit the relationship between fused representation and POS tags of ground-truth descriptions, and thereby predicts global POS representation for the sentence to be generated. The description generator attentively summarizes the fused representations and generates each word by adaptively integrating the predicted global POS representation.

3.1. Gated Fusion Network

Given the input videos, the gated fusion network first extracts different semantic representations for each frame by multiple CNN networks. For the convenience of expression, we take the visual content feature from RGB frames and motion feature from optical flows as examples in this section, which are denoted as $R = \{r_1, r_2, \ldots, r_m\}$ and $F = \{f_1, f_2, \ldots, f_m\}$, respectively, where r_i and f_i denote the features for the i_{th} frame and optical flow of the input video, respectively. m indicates the total length of the video. Based on the obtained representations R and F, the gated fusion network performs in two stages. First, temporal encoding of each representation is performed, respectively. Afterwards, a cross gating strategy is proposed to fuse the temporally aggregated feature together.

Temporal Encoder. Long short term memory networks (LSTMs) are used to aggregate these representations:

$$h_i^{(r)}, z_i^{(r)} = \text{LSTM}_r^{(E)}\left(r_i, h_{i-1}^{(r)}\right),$$
$$h_i^{(f)}, z_i^{(f)} = \text{LSTM}_f^{(E)}\left(f_i, h_{i-1}^{(f)}\right),$$
(1)

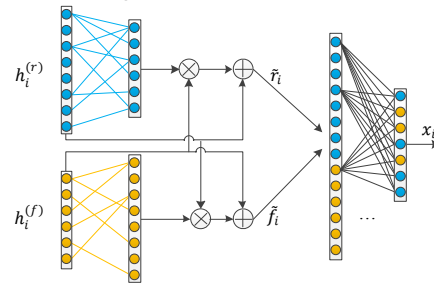where $\text{LSTM}_r^{(E)}$ and $\text{LSTM}_f^{(E)}$ denote the LSTM units for

Figure 3. An illustration of cross gating strategy in the proposed gated fusion network. The cross gating strategy strengthens the information that is related to each other within diverse features, and then fuses them together. ⊗ and ⊕ denote the element-wise multiplication and addition, respectively.

the content and motion features, respectively. $h_i^{(r)}$, $h_i^{(f)}$, $z_i^{(r)}$ and $z_i^{(f)}$ are the corresponding hidden states and memory cells. With LSTM encoding, high-level content and motion feature sequences $\hat{R} = \{h_1^{(r)}, h_2^{(r)}, \ldots, h_m^{(r)}\}$ and $\hat{F} = \{h_1^{(f)}, h_2^{(f)}, \ldots, h_m^{(f)}\}$ are obtained.

Cross Gating. A simple concatenation of $\hat{R}$ and $\hat{F}$ can fuse all the different features of a frame. However, such a fusion strategy ignores the relationship between these features. To take full advantage of the related semantic information, we propose a novel cross gating strategy on different features as illustrated in Fig. 3:

$$\tilde{r}_i = \text{Gating}_r^{(E)}\left(h_i^{(f)}, h_i^{(r)}\right),$$
$$\tilde{f}_i = \text{Gating}_f^{(E)}\left(h_i^{(r)}, h_i^{(f)}\right),$$
(2)

where $\tilde{r}_i$ and $\tilde{f}_i$ are the gated results for content and motion representations. We realize the Gating function as follows:

$$\text{Gating}(x, y) = \sigma(wx + b)y + y,$$
(3)

where y denotes the target feature, which is updated under the guidance of the driver feature x. w and b are learnable parameters, and σ is a nonlinear activation function, which is a ReLU function in our implementation. Obviously, in

$\tilde{r}_i$ the content information related to the motion information has been strengthened by the proposed cross gating strategy. And the similar process performs on $\tilde{f}_i$, where the motion information related to content information is strengthened.

Finally, the gated representations of content and motion are fused together by a fully connected layer:

$$x_i = w^{(E)} \left(\left[\tilde{r}_i, \tilde{f}_i \right] + b^{(E)} \right), \quad (4)$$

where $[\cdot]$ denotes the concatenation of inputs. x_i denotes the fused representation for each frame where both content and motion information are included. $w^{(E)}$ and $b^{(E)}$ are the learnable parameters.

3.2. POS Sequence Generator

In addition to natural language descriptions, the POS of each word in sentences is also closely related to the video content. To utilize the POS information, we design a simple POS sequence generation network based on the fused representations. Based on the fused feature sequence $X = \{x_1, x_2, \ldots, x_m\}$, the POS generator predicts POS sequence:

$$h_t^{(T)}, z_t^{(T)} = \text{LSTM}^{(T)} \left(\left[E_{pos}(c_{t-1}), \phi_t \left(X, h_{t-1}^{(T)} \right) \right], h_{t-1}^{(T)} \right),$$

$$P(c_t | c_{<t}, V; \theta_{pos}) = \text{softmax} \left(\mathbf{W}^{(T)} h_t^{(T)} + \mathbf{b}^{(T)} \right),$$
$$(5)$$

where $h_t^{(T)}$ and $z_t^{(T)}$ are hidden state and memory cell of POS generator. c_{t-1} denotes the POS tag predicted at the previous step, E_{pos} is an embedding matrix for POS tags and we denote by $E_{pos}(c_{t-1})$ the embedding vector of POS tag c_{t-1}. θ_{pos}, $\mathbf{W}^{(T)}$ and $\mathbf{b}^{(T)}$ denote the learnable parameters in POS sequence encoder. $P(c_t | c_{<t}, V; \theta_{pos})$ means the probability of predicting correct POS tag c_t given the previous tags $c_{<t} = \{c_1, c_2, \ldots, c_{t-1}\}$ and input video V.

Please note that the symbol $\phi_t(\cdot)$ in Eq. (5) denotes the soft attention process at time step t, which yields a vector representation $\phi_t \left(X, h_{t-1}^{(T)} \right)$ with different weights on X:

$$\phi_t \left(X, h_{t-1}^{(T)} \right) = \sum_{i=1}^{m} \alpha_{t,i} x_i, \quad (6)$$

where $\sum_{i=1}^{m} \alpha_{t,i} = 1$ and $\alpha_{t,i}$ denotes the attention weights computed for the i_{th} fused representation at the t_{th} time step. It encourages the POS sequence encoder to select the useful information related with the POS tag predicted at the current step. The attentive weight α is computed by:

$$e_{t,i} = \mathbf{w}^{(T)\top} \tanh \left(\mathbf{W}^{(T)} h_{t-1}^{(T)} + \mathbf{U}^{(T)} x_i + \mathbf{b}^{(T)} \right),$$

$$\alpha_{t,i} = \exp(e_{t,i}) / \sum_{k=1}^{m} \exp(e_{t,i}),$$
$$(7)$$

where $\mathbf{w}^{(T)\top}$, $\mathbf{W}^{(T)}$, $\mathbf{U}^{(T)}$ and $\mathbf{b}^{(T)}$ are learnable parameters.

When the prediction for the whole POS sequence finishes, the last hidden state $\psi = h_n^{(T)}$ of the LSTM is expected to capture the global information of the POS se-
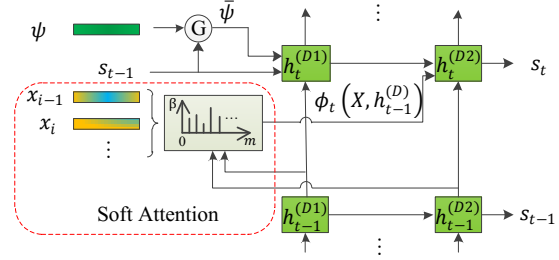

Figure 4. The architecture of the proposed description generator. At each time step, the cross gating $\circledG$ is performed on the embedding vector of predicted word and the predicted global POS feature, by which the POS information related to the current word is dynamically and adaptively incorporated. The soft attention mechanism dynamically summarizes the fused frame features.

quence of the generated sentence, which is further used to guide the description generation and control the syntactic structure to generate the sentence.

3.3. Description Generator

The description generator produces sentence description for video based on the fused video representation $X = \{x_1, x_2, \ldots, x_m\}$ learned by the video encoder and the predicted global POS representation ψ. We employ a hierarchical decoder, which consists of a two-layer LSTM. The first layer is fed with the generated word s_{t-1} and the global POS feature ψ, while the second one takes the hidden state of first layer and an attentive summary of X as input.

When describing a video V, the word embedding vector is updated by performing the cross gating on the global POS feature ψ generated in Sec. 3.2:

$$\bar{\psi} = \text{Gating}^{(D)} \left(E_{word}(s_{t-1}), \psi \right), \quad (8)$$

where $E_{word}(s_{t-1})$ is the word embedding vector of the word s_{t-1}, which is generated at the previous time step. As such, the global POS information with respect to the predicted word is strengthened.

Afterwards, the process of description generator is as follows:

$$\phi_t(X, h_{t-1}^{(D)}) = \sum_{i=1}^{m} \beta_{t,i} x_i,$$

$$h_t^{(D1)}, z_t^{(D1)} = \text{LSTM}^{(D1)} \left(\left[E_{word}(s_{t-1}), \bar{\psi} \right], h_{t-1}^{(D1)} \right),$$

$$h_t^{(D2)}, z_t^{(D2)} = \text{LSTM}^{(D2)} \left(\left[h_t^{(D1)}, \phi_t(X, h_{t-1}^{(D)}) \right], h_{t-1}^{(D2)} \right),$$

$$P(s_t | s_{<t}, V; \theta_{gen}) = \text{softmax} \left(\mathbf{W}_s^{(D)} h_t^{(D2)} + \mathbf{b}_s^{(D)} \right),$$
$$(9)$$

where LSTM with subscripts $D1$ and $D2$ denote the LSTM units at the first and second layers in decoder. $\mathbf{W}_s^{(D)}$, $\mathbf{b}_s^{(D)}$, and θ_{gen} denote the learnable parameters in the description generator. Once again, we apply the soft attention on X, as in Eq. (7), to dynamically select the high-level fused features, which is denoted as $\phi_t(X, h_{t-1}^{(D)})$. The attentive weights β satisfy $\sum_{i=1}^{m} \beta_{t,i} = 1$. Please note that we use

a hierarchical guidance consisting of $h_t^{(D1)}$ from the first layer and $h_t^{(D2)}$ from the second layer to drive the attention mechanism, which is obtained by:

$$e_{t,i}^{(D)} = \mathbf{w}_a^{(D)\top} \tanh \left(\mathbf{W}_a^{(D)} \left[h_{t-1}^{(D1)}, h_{t-1}^{(D2)} \right] + \mathbf{U}_a^{(D)} x_i + \mathbf{b}_a^{(D)} \right),$$

$$\beta_{t,i} = \exp \left(e_{t,i}^{(D)} \right) / \sum_{j=1}^{m} \exp \left(e_{t,j}^{(D)} \right). \tag{10}$$

3.4. Training

The proposed model is trained in two stages. We first freeze the parameters of the description generator, and train the POS sequence generator in a supervised way with the purpose of obtaining accurate global POS information. The loss function is defined as the negative log probability of each POS sequence:

$$\mathcal{L}_{pos} (\theta_{pos}) = - \sum_{k=1}^{N} \log P \left(C^k | V^k ; \theta_{pos} \right), \tag{11}$$

where N is the total number of training data, and the probability of one POS sequence is defined as:

$$P (C | V ; \theta_{pos}) = \prod_{t=1}^{n} P (c_t | c_{<t}, V ; \theta_{pos}). \tag{12}$$

When the POS generator converges, we predicts the global POS information based on the POS generator for each video sequence. Then, the video encoder and description generator are jointly trained by minimizing the cross-entropy loss, which is similar to Eq. (11) and (12):

$$\mathcal{L}_{gen} (\theta_{gen}) = - \sum_{k=1}^{N} \log P \left(S^k | V^k ; \theta_{gen} \right), \tag{13}$$

where $P (S | V ; \theta_{gen}) = \prod_{t=1}^{n} P (s_t | s_{<t}, V ; \theta_{gen})$.

Besides, we intend to directly train the captioning models guided by evaluation metrics, specifically the CIDEr [41] in this work, instead of the cross-entropy loss. As such an evaluation metric is discrete and non-differentiable, we resort to the self-critical sequence training [32] to further boost the performance of the proposed method. More details about the self-critical strategy can be found in the supplementary material.

3.5. Inference

As the global POS information generated by the proposed POS sequence generator can help improve captioning performance and control the syntactic structures of the generated descriptions, we verify our method in two ways. First, the POS sequence generator generates global POS information without any human intervention. secondly, we control the global POS information by changing the predicted POS tags. For example, we change one or more predicted POS tags, based on which, the corresponding global POS information is then generated.

In both two ways, the global POS information is utilized by the description generator to predict the captions. The first verification is for demonstrating the performance improvements brought by the proposed gated fusion network and global POS sequence guidance, while the other one aims to present the controllablity for the syntactic structure of the video description generation.

4. Experiments

In this section, we evaluate the proposed video captioning method on Microsoft Research video to text (MSR-VTT) [50] and Microsoft Video Description Corpus (MSVD) [3] with the widely-used metrics including BLEU@N [29], METEOR [1], ROUGE-L [22], and CIDEr [41]. They are denoted as B@N, M, R, and C respectively, where N varies from 1 to 4. The codes for these metrics have been released on Microsoft COCO evaluation server [6]. We first briefly describe the datasets used for evaluation, followed by the implementation details. Afterwards, we discuss the experiment results on video captioning.

4.1. Datasets

MSR-VTT. The MSR-VTT is a large-scale dataset for video captioning, which covers the most diverse visual contents so far. It contains 10,000 video clips from 20 categories and 200,000 video-caption pairs with 29,000 unique words in total. Each video clip corresponds to 20 English sentence descriptions. Following the existing work, we use the public splits for training and testing, where 6,513 for training, 497 for validation, and 2,990 for testing.

MSVD. There are 1,970 short video clips collected from YouTube, with each one depicts a single activity in 10 seconds to 25 seconds. Each clip has roughly 40 English descriptions. Similar to the prior work [27, 53], we take 1200 video clips for training, 100 clips for validation and 670 clips for testing.

4.2. Implementation Details

For the sentences in the benchmark datasets motioned above, we first remove the punctuation and convert all words into lowercase. The sentences are truncated at 28 words and tokenized. The size of word embedding size for each word is set to 468. The POS tags of words in ground-truth are processed by Stanford Log-linear Part-Of-Speech Tagger [39], which are then divided into 14 categories for training the POS sequence generator: verb(VERB), noun(NOUN), adjective(ADJ), adverb(ADV), conjunction(CONJ), pronoun(PRON), preposition(PREP), article(ART), auxiliary(AUX), participle(PRT), number qualifier(NUM), symbol(SYM), unknown(UNK) and the end-of-sentence (EOS). Each of them also corresponds to an embedding vector with 468 dimensions.

For videos, we use TVL1-flow [30] to compute the optical flows in both horizontal and vertical directions for adjacent frames. Then an Inflated 3D ConvNet (I3D) [2] trained on Kinetics action classification dataset [19] extracts 8 1024-dimensional feature vectors representing the motion features for each continuous 64 optical flow frames. To extract the content features, we feed static frames to Inception-Resnet-v2 [38], which is pre-trained on ILSVRC-2012-CLS image classification dataset [35], and obtain a 1536-dimensional feature for each frame. We also extract the spatiotemporal features by C3D network [40]. We take equally-spaced 30 features of a video, respectively, and pad them with zero vectors if the number of features is less than 30.

In our model, all LSTMs have a 512-dimensional hidden size, while the input dimension of LSTMs in video encoder is 1536 and 1024, which are equal to the size of content features and motion features, respectively. The input size of LSTMs in POS sequence generator and the first layer in description generator are set to 468 . The input size of the second layer of description generator is 512.

During the training, the model is optimized by the AdaDelta [55]. When no better CIDEr score appears in the following 30 successive validations, the training stops and the optimal model is obtained. In the testing, we use the beam search with size 5 for the final description generation.

4.3. Performance Comparisons

In this subsection, we compare our method with the state-of-the-art methods with multiple features on benchmark datasets, including SA [53], M3 [47], v2t_navigator [18], Aalto [36], VideoLab [31], MA-LSTM [51], M&M-TGM [4], PickNet [8], LSTM-TSA$_{IV}$ [28], SibNet [23], MGSA [5], and SCN-LSTM [14], most of which fuse different features by simply concatenating.

We first show the quantitative results on MSR-VTT in Table 1. Since the methods use different CNN features or combinations of features, it is hard to make an absolutely fair comparison. Trained by cross-entropy loss, our proposed method is better than most of the competing models, including PickNet trained by RL, which demonstrates the benefits of the proposed gated fusion network as well as the incorporated global syntactic structure POS information. It is worth noticing that our model performs inferiorly to M&M-TGM and v2t_navigator on some metrics, such as CIDEr and METEOR. The reason is that the features of some modalities are of great differences with different expressive abilities. For example, the v2t_navigator applies audio (A) and topic (Ca) features providing by MSR-VTT, while M&M-TGM use a multi-task to predict the topic (Ca) features. These modalities can provide strong prior knowledge for captioning generation, which can however not be

directly obtained by content or motion features. Better performances are expected if these strong features are also fused in our methods. Besides, same modalities are utilized in MGSA(IR+C3D) and Ours(IR+C3D), based on which our method performs better, mainly attributed to the proposed cross gating strategy and the introduced POS information.

We also train our model by RL, specifically the self-critical sequence training [32], which is denoted as Ours_RL(IR+M) in Table 1. Obviously, self-critical strategy generates better performances than training with the traditional cross-entropy loss on all metrics except BLEU@4. It is reasonable as we mainly focus on optimizing the CIDEr metric with RL. Comparing with other competing models, Ours_RL(IR+M) obtains the state-of-the-art performances on both ROUGE-L and CIDEr, which achieve 62.1 and 53.4, respectively. The superior performances further demonstrate the benefits of the proposed gated fusion network and the incorporation of global POS information.

Besides, we also verify our work on MSVD as shown in Table 2. Once again, our methods outperforms other competitors, especially on CIDEr. Without self-critical, Ours(IR+M) has obtained superior scores. When trained with self-critical, Ours_RL(IR+M) significantly improves all the metric scores and achieves the new state-of-the-art results on BLEU@4, METEOR, ROUGE-L and CIDEr. It is worth noticing that SibNet(G) is an excellent method that achieves the state-of-the-art on MSVD using only GoogleNet feature, as additional content and semantic branches are introduced in SibNet and significantly boost the performances. Trained with the decoder loss alone, the SibNet-DL(G) performs slightly inferiorly to our method. Our model is orthogonal to SibNet, which can be incorporated into SibNet for further boosting the performances.

4.4. Ablation Studies

To demonstrate the effectiveness of the proposed components, we design several baseline models with different structures by removing certain components, which are listed as follows:

- EncDec+F: This is the basic model where the video encoder fuses diverse features as one video representation by simple concatenating.
- EncDec+CG: This model employs the proposed gated fusion network to effectively fuse different features, but without the POS sequence generator.
- Ours (EncDec+CG+POS): It is the proposed model, where the gated fusion network is employed and the global POS tag information generated by the proposed POS sequence generator are considered for video captioning.

The ablation experimental results of aforementioned models with different components on testing split of MSR-

Training Strategy	Model	B@1	B@2	B@3	B@4	M	R	C
Cross-Entropy	SA(V+C3D) [50]	82.3	65.7	49.7	36.6	25.9	-	-
	M3(V+C3D) [47]	73.6	59.3	48.3	38.1	26.6	-	-
	MA-LSTM(G+C3D+A) [51]	-	-	-	36.5	26.5	59.8	41.0
	VideoLab(R-152+C3D+A+Ca) [31]	-	-	-	39.1	27.7	60.6	44.1
	v2t_navigator(C3D+A+Ca) [18]	-	-	-	42.6	28.8	61.7	46.7
	M&M-TGM(IR+C3D+A) [4]	-	-	-	44.3	29.4	-	49.3
	SibNet-DL(G) [23]	-	-	-	39.4	26.9	59.6	45.3
	SibNet(G) [23]	-	-	-	40.9	27.5	60.2	47.5
	MGSA(IR+C3D) [5]	-	-	-	42.4	27.6	-	47.5
	MGSA(IR+C3D+A) [5]	-	-	-	45.4	28.6	-	50.1
Reinforcement Learning	PickNet(R-152+Ca) [8]	-	-	-	41.3	27.7	59.8	44.1
Cross-Entropy	Ours(C3D+M)	78.5	65.3	52.6	41.2	27.7	60.9	46.7
	Ours(I3D+M)	79.3	65.8	53.3	41.7	27.8	61.2	48.5
	Ours(IR+C3D)	79.2	66.5	53.7	42.3	28.1	61.3	48.6
	Ours(IR+I3D)	79.1	66.0	53.3	42.0	28.1	61.1	49.0
	Ours(IR+M)	78.4	66.1	53.4	42.0	28.2	61.6	48.7
Reinforcement Learning	Ours_RL(IR+M)	81.2	67.9	53.8	41.3	28.7	62.1	53.4

Table 1. Performance comparisons with different competing models on the testing set of the MSR-VTT in terms of BLEU@1∼4, METEOR, and ROUGE-L, CIDEr scores (%). V, G, C3D, R-N, IR, I3D, A and Ca denote VGG19, GoogleNet, C3D, N-layer ResNet, Inception_ResNet-v2, I3D, Audio, and Category features, respectively. M denotes the motion features from optical flow extracted by I3D.

Model	B@4	M	R	C
MA-LSTM(G+C3D)	52.3	33.6	-	70.4
LSTM-TSA$_{IV}$(V+C3D)	52.8	33.5	-	74.0
SCN-LSTM(R-152+C3D)	50.2	33.4	-	77.0
M&M-TGM(IR+C3D+A)	48.8	34.4	-	80.5
SibNet-DL(G)	51.9	33.1	69.9	81.9
SibNet(G)	**54.2**	34.8	71.7	88.2
Ours(IR+M)	52.5	34.1	71.3	88.7
Ours_RL(IR+M)	53.9	**34.9**	**72.1**	**91.0**

Table 2. Performance comparisons with different baseline methods on the testing set of the MSVD dataset (%).

Model	B@4	M	R	C
EncDec+F(IR+M)	39.7	26.8	59.3	45.4
EncDec+CG(IR+M)	41.7	27.9	61.0	48.4
Ours(IR+M)	**42.0**	**28.2**	**61.6**	**48.7**
EncDec+F(I3D+M)	39.7	26.6	58.8	45.1
EncDec+CG(I3D+M)	41.4	27.7	61.0	47.7
Ours(I3D+M)	**41.7**	**27.8**	**61.2**	**48.5**

Table 3. Performance comparisons with different baseline methods on the testing set of the MSR-VTT dataset(%). Methods of the same name but different text in the brackets indicates the same method with different feature inputs.

Model	B@4	M	R	C
MCB(IR+M)	41.2	27.5	60.6	46.3
MLB(IR+M)	41.4	27.6	60.9	47.6
Element-wise adding(IR+M)	40.2	27.0	60.2	46.3
EncDec+CG(IR+M)	**41.7**	**27.9**	**61.0**	**48.4**

Table 4. Performance comparisons with different fusion strategies on the testing set of the MSR-VTT(%).

MLB [20], and element-wise adding, which used compact bilinear pooling, low-rank bilinear pooling, and feature vector adding to exploit the relationship of each element in different modal features, respectively. The performance comparisons are illustrated in Table 4. The cross gating mechanism enhances the relevant part of different modalities and uses a residual structure to retain information that may not be relevant but unique, which can not be modeled by MCB, MLB, or element-wise adding. As such, the relationships between different modalities can be more comprehensively exploited to further benefit the video captioning.

Our proposed method, considering both the gated fusion network and the global POS information, takes a further step on the EncDec+CG and is trained by adaptively and dynamically incorporating POS to guide the generation of each word. As illustrated in Table 3, the proposed model yields the highest performance on the four metrics with the same features pair, namely IR and M. The same observations can be observed if using (I3D, M) pair as the feature representations in Table 3, which proves that the POS information provides a global view on the potential syntactic structure of its language descriptions and thereby further improves the performance of video captioning.

4.5. Qualitative Analysis

Besides, some qualitative examples are shown in Fig. 5. it can be observed that the proposed model, with the cross gating strategy and incorporating global POS information, can generate more accurate descriptions than the baseline model. For example, in the first example our model realizes that it is not related to cooking and correctly predicts the action '*mixing*' and object '*ingredients*' under the guid-

VTT are shown in Table 3. We use different feature combinations, *e.g.*, (IR, M) pair and (I3D, M) pair, each of which consists of one content and one motion features.

Compared with the basic model EncDec+F(IR+M) that simply concatenates the features, we can observe a significant improvement on all the evaluation metrics by incorporating the proposed gated fusion network in EncDec+CG(IR+M). The similar performance improvements also appear in EncDec+CG(I3d+M). The better scores on BLUE@4, METEOR, ROUGE-L and CIDEr of EncDec+CG in Table 3 demonstrate that: 1) There exist complicated semantic relationships between different features, specifically the content feature and motion features, which a simple concatenation can not capture. 2) By performing the gated fusion network on the content and motion features, a more representative video feature for video captioning can be obtained.

To further demonstrate the effectiveness of the proposed fusion strategy, we also compare the gated fusion network with more complex fusion algorithms, such as MCB [13],

EncDec+F: A person is cooking.
Ours: a woman is mixing some ingredients in a bowl.
[POS]: ART NOUN VERB VERB ART NOUN PREP ART NOUN.
GT: A woman is **mixing** nailpolish and putting an egg into it.

EncDec+F: People are playing baseball.
Ours: A baseball player is hitting the ball.
[POS]: ART NOUN VERB VERB ART NOUN.
GT: A baseball Player **hits** a ball to the back of the field
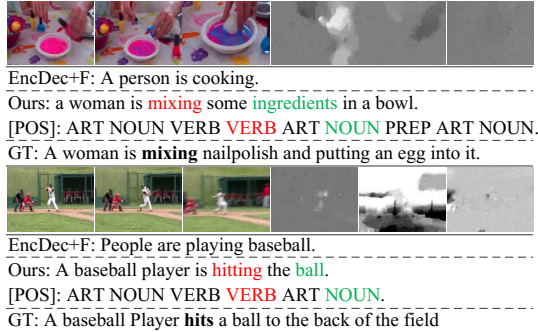
Figure 5. Visualization of some video captioning examples on the MSR-VTT with the basic model and the proposed model. Due to the page limit, only one ground-truth (GT) sentence is given as reference. Also we illustrate the generated POS sequence. Compared to the base model EncDec+F, the proposed model yields more accurate sentence descriptions.

ance of POS tag '*VERB*' highlighted in red and '*NOUN*' highlighted in green, respectively. In the second example, compared to the general description of the EncDec+F, our model accurately captures the detail '*hitting the ball*', which makes a more specific and vivid description.

Original Description:
[POS]: ART NOUN VERB VERB NOUN
Ours: A man is explaining something.
Controlling Description: add adjectives
[POS]: ART ADJ NOUN VERB VERB ART NOUN
Ours: A man in a pink shirt is giving a presentation.

Original Description:
[POS]: ART NOUN VERB VERB ART AUX ART VERB
Ours: A person is cooking food in a pan.
Controlling Description: generate 'THERE BE'
[POS]: ADV VERB ART NOUN VERB VERB ART NOUN
Ours: There is a woman is making a dish.

Original Description:
[POS]: ART NOUN VERB VERB AUX ART NOUN
Ours: A group of men are playing soccer in the ground.
Controlling Description: change quantity
[POS]: NUM NOUN VERB VERB ART NOUN
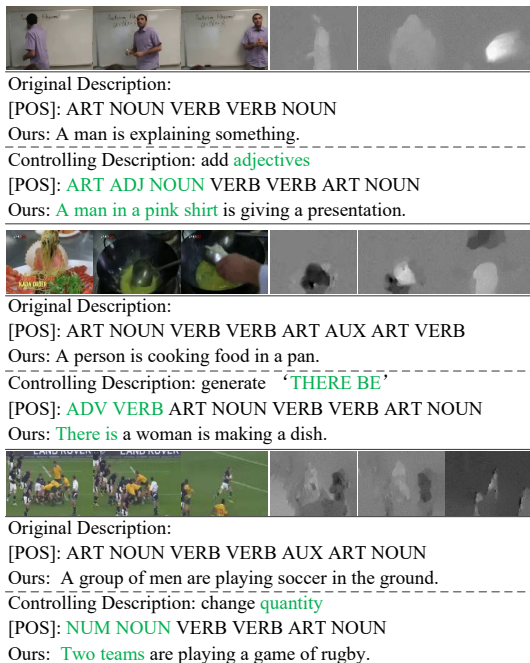Ours: Two teams are playing a game of rugby.

Figure 6. Visualization of some video captioning examples on the MSR-VTT by controlling the captioning generation with modifying the generated POS tag sequence. The POS tags in green denote the human modified ones, while the words in green are generated under the guidance of the modified global POS information.

4.6. Controllable of Syntax for Video Captioning

Finally, we show the controllable of the syntactic structure for generating captions by manually modifying the generated POS tag sequence in the inference stage. For exam-

ple, when we expect an adjective on the current time step, the predicted POS tag will be replaced by '*ADJ*' tag manually, whatever the predicted result is. The changed '*ADJ*' tag is subsequently fed to the POS sequence generator and the next POS tag is predicted, meanwhile the hidden state of POS sequence generator is modified. As such, the global POS information can be modified to control the overall syntactic structure of the generated sentence.

Some examples are illustrated in Fig. 6. For the first sample, we add the '*ADJ*' in the front of the subject, so that we can describe the event with more details. With the changed POS information, the description generator predicts "*a man in a pink shirt*", which is in line with our expectations. In the second sample, we would like to generate a sentence with "*there be*" as the beginning. Our approach once again meet the requirement. The most interesting thing is when we replace the article ('*ART*') with number ('*NUM*'), the generator provides "*two teams*", instead of "*two men*", to replace "*a group of men*". These results demonstrate that the proposed gated fusion network effectively and fully captures semantic meaning of the video by understanding the relationship between different features. Therefore, even though the global POS information is changed, it can accurately generate the reliable descriptions. Meanwhile, the global POS information can indeed control the overall syntactic structure of the generated sentence. More experimental results can be referred to the supplementary material[1].

5. Conclusions

In this paper, we proposed a novel model for controllable video captioning using a gated fusion network and a POS sequence generator. This model can fuse diverse information with a cross-gating strategy and produce a global syntactic structure as the guidance for addressing video captioning. The proposed model achieves competitive performances on both MSR-VTT and MSVD datasets, which indicates the superiority of our model. Moreover, the generated global POS information can be further leveraged to control the syntactic structure of the generated caption, thereby improving the corresponding diversity.

Acknowledgments

---

[1]https://github.com/vsislab/Controllable_XGating/blob/master/supplementary.pdf

References

[1] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005.

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[3] David L Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *ACL*, 2011.

[4] Shizhe Chen, Jia Chen, Qin Jin, and Alexander Hauptmann. Video captioning with guidance of multimodal latent topics. In *ACM MM*, 2017.

[5] Shaoxiang Chen and Yu-Gang Jiang. Motion guided spatial attention for video captioning. In *AAAI*, 2019.

[6] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv:1504.00325*, 2015.

[7] Xinpeng Chen, Lin Ma, Wenhao Jiang, Jian Yao, and Wei Liu. Regularizing rnns for caption generation by reconstructing the past with the present. In *CVPR*, 2018.

[8] Yangyu Chen, Shuhui Wang, Weigang Zhang, and Qingming Huang. Less is more: Picking informative frames for video captioning. *ECCV*, 2018.

[9] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *CVPR*, 2018.

[10] Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander Schwing, and David A Forsyth. Diverse and controllable image captioning with part-of-speech guidance. *CVPR*, 2019.

[11] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.

[12] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. Unsupervised image captioning. In *CVPR*, 2019.

[13] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*, 2016.

[14] Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. Semantic compositional networks for visual captioning. In *CVPR*, 2017.

[15] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *ICCV*, 2013.

[16] Xinwei He, Baoguang Shi, Xiang Bai, Gui-Song Xia, Zhaoxiang Zhang, and Weisheng Dong. Image caption generation with part of speech guidance. *Pattern Recognition Letters*, 119:229–237, 2019.

[17] Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. Recurrent fusion network for image captioning. In *ECCV*, 2018.

[18] Qin Jin, Jia Chen, Shizhe Chen, Yifan Xiong, and Alexander Hauptmann. Describing videos using multi-modal fusion. In *ACM MM*, 2016.

[19] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv:1705.06950*, 2017.

[20] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. *ICLR*, 2017.

[21] Atsuhiro Kojima, Takeshi Tamura, and Kunio Fukunaga. Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision*, 50(2):171–184, 2002.

[22] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004.

[23] Sheng Liu, Zhou Ren, and Junsong Yuan. Sibnet: Sibling convolutional encoder for video captioning. In *ACM MM*, 2018.

[24] Lin Ma, Wenhao Jiang, Zequn Jie, Yu-Gang Jiang, and Wei Liu. Matching image and sentence with multi-faceted representations. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[25] Lin Ma, Zhengdong Lu, and Hang Li. Learning to answer questions from image using convolutional neural network. In *AAAI*, 2016.

[26] Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. Multimodal convolutional neural networks for matching image and sentence. In *ICCV*, 2015.

[27] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly modeling embedding and translation to bridge video and language. In *CVPR*, 2016.

[28] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. Video captioning with transferred semantic attributes. In *CVPR*, 2017.

[29] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

[30] Javier Sánchez Pérez, Enric Meinhardt-Llopis, and Gabriele Facciolo. Tv-l1 optical flow estimation. *Image Processing On Line*, 2013:137–150, 2013.

[31] Vasili Ramanishka, Abir Das, Dong Huk Park, Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, and Kate Saenko. Multimodal video description. In *ACM MM*, 2016.

[32] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017.

[33] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. Coherent multi-sentence video description with variable level of detail. In *GCPR*, 2014.

[34] Marcus Rohrbach, Wei Qiu, Ivan Titov, Stefan Thater, Manfred Pinkal, and Bernt Schiele. Translating video content to natural language descriptions. In *ICCV*, 2013.

[35] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[36] Rakshith Shetty and Jorma Laaksonen. Frame-and segment-level features and candidate pool evaluation for video caption generation. In *ACM MM*, 2016.

[37] Jingkuan Song, Lianli Gao, Li Liu, Xiaofeng Zhu, and Nicu Sebe. Quantization-based hashing: a general framework for scalable image and video retrieval. *Pattern Recognition*, 75:175–187, 2018.

[38] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, 2017.

[39] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL*, 2003.

[40] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015.

[41] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015.

[42] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *ICCV*, 2015.

[43] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. Translating videos to natural language using deep recurrent neural networks. *NAACL*, 2015.

[44] Bairui Wang, Lin Ma, Wei Zhang, Wenhao Jiang, and Feng Zhang. Hierarchical photo-scene encoder for album storytelling. In *AAAI*, 2019.

[45] Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. Reconstruction network for video captioning. In *CVPR*, 2018.

[46] Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. Bidirectional attentive fusion with context gating for dense video captioning. In *CVPR*, 2018.

[47] Junbo Wang, Wei Wang, Yan Huang, Liang Wang, and Tieniu Tan. M3: Multimodal memory modelling for video captioning. In *CVPR*, 2018.

[48] Jingdong Wang, Ting Zhang, Nicu Sebe, Jingkuang Song, and Heng Tao Shen. A survey on learning to hash. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):769–790, 2018.

[49] Xian Wu, Guanbin Li, Qingxing Cao, Qingge Ji, and Liang Lin. Interpretable video captioning via trajectory structured localization. In *CVPR*, 2018.

[50] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016.

[51] Jun Xu, Ting Yao, Yongdong Zhang, and Tao Mei. Learning multimodal attention lstm networks for video captioning. In *ACM MM*, 2017.

[52] Ran Xu, Caiming Xiong, Wei Chen, and Jason J Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *AAAI*, 2015.

[53] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. Describing videos by exploiting temporal structure. In *ICCV*, 2015.

[54] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *CVPR*, 2016.

[55] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *aarXiv:1212.5701*, 2012.

[56] Wei Zhang, Bairui Wang, Lin Ma, and Wei Liu. Reconstruct and represent video contents for captioning via reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.