This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Joint Syntax Representation Learning and Visual Cue Translation for Video Captioning

Jingyi Hou[1], Xinxiao Wu[1]*, Wentian Zhao[1], Jiebo Luo[2], and Yunde Jia[1]
[1]Beijing Laboratory of Intelligent Information Technology, School of Computer Science,
Beijing Institute of Technology, Beijing 100081, China
[2]Department of Computer Science, University of Rochester, Rochester NY 14627, USA

Abstract

Video captioning is a challenging task that involves not only visual perception but also syntax representation learning. Recent progress in video captioning has been achieved through visual perception, but syntax representation learning is still under-explored. We propose a novel video captioning approach that takes into account both visual perception and syntax representation learning to generate accurate descriptions of videos. Specifically, we use sentence templates composed of Part-of-Speech (POS) tags to represent the syntax structure of captions, and accordingly, syntax representation learning is performed by directly inferring POS tags from videos. The visual perception is implemented by a mixture model which translates visual cues into lexical words that are conditional on the learned syntactic structure of sentences. Thus, a video captioning task consists of two sub-tasks: video POS tagging and visual cue translation, which are jointly modeled and trained in an end-to-end fashion. Evaluations on three public benchmark datasets demonstrate that our proposed method achieves substantially better performance than the state-of-the-art methods, which validates the superiority of joint modeling of syntax representation learning and visual perception for video captioning.

1. Introduction

Automatically generating a natural language description of a video has attracted remarkable attention for its important applications, such as semantic video search, visual intelligence in chatting robots and aid for people to perceive the world around them. Previous works [21, 23, 16, 33] describe videos using template-based methods which first manually create fix-structured sentence templates and then fill in the template with the corresponding words. Recently, increasing studies have shown benefits of deep learning on video captioning, owing to the great success of deep neural networks in both computer vision and natural language processing. Many deep learning methods [11, 38, 37, 29] usually build an encoder to compress the input video into a feature representation and a decoder to generate descriptions given the video feature. Most existing methods of video captioning [48, 41, 13, 30, 4, 12, 8, 1] mainly focus on investigating various visual perception models by exploiting informative semantics without considering learning of syntax representation for generating sentences.

In this paper, we propose a novel video captioning approach that takes into account both visual perception and syntax representation learning to generate accurate sentences of videos. In an analogy to natural language understanding, the syntactic structure information is obviously essential for a sentence to interpret the video. For example, the sentences of "A wolf is eating a sheep" and "A wolf and a sheep are eating" have similar semantic primitives, but differ in their meanings with different syntactic structures. Thus, learning syntax representation will benefit a lot to the video captioning. Specifically, we use sentence templates composed of Part-of-Speech (POS) tags to represent the syntax structure of captions, and accordingly, syntax representation learning is performed by directly inferring POS tags from videos. The visual perception is implemented by translating visual cues into lexical words to exploit semantic primitives conditioned on the corresponding POS tags. Therefore, the video captioning task in our method simultaneously performs two sub-tasks: video POS tagging and visual cue translation. To this end, an end-to-end trainable network is built to jointly model and train these two subtasks.

To automatically tag videos with POS, a sequence-to-sequence (S2S) model is employed to generate POS sequences from input videos. The POS sequence can be regarded as an interpretation of syntactic structure of textual description for the video. Since the POS tag categories are much fewer than word categories, it is much easier to use the S2S model for generating a POS sequence than a real

*Corresponding author: Xinxiao Wu

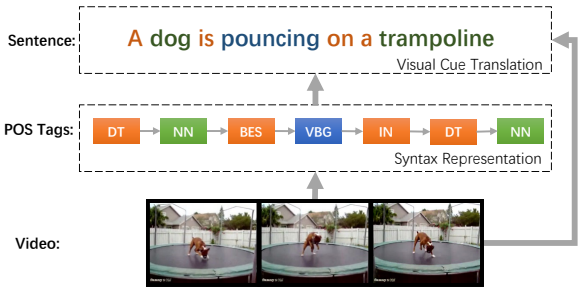

Figure 1. The main process of our method for video captioning. For an input video, we first learn its syntax representation via video POS tagging and then translate the visual cues to words given the inferred POS via a mixture model.

sentence. Generally speaking, a complete sentence should be constituted of various kinds of grammatical elements. Therefore, we introduce a simple yet effective constraint term to diversify the parts of speech in each generated sentence to guarantee the completeness of captions. The constraint term encourages each POS to appear in the generated sentence at least once by using L2 normalization.

In translating visual cues to words, the distribution of word frequencies in captions is extremely imbalanced, i.e., a small fraction of the words appears more frequently than other words, which is known as the Zipfian law of word distribution in nature languages. So directly using imbalanced data to train the decoder of softmax word classifiers will lead to the word bias problem, which degrades the captioning performance. To address this issue, an explicit mixture model is newly proposed to generate the probability distribution of words conditioned on the POS tags, which captures intrinsic semantic primitives by perceiving relevant visual cues for generating accurate words without bias. In the mixture model, each component is sensitive to a specific visual cue for generating words that are semantically related to the visual cue. Thanks to the open-closed characteristic in linguistics, words can be relatively equally divided into multiple subsets according to the POS. In each subset the frequencies of words do not vary dramatically, so the word classifier for each subset will not suffer from the bias problem. The multiple word classifiers are regarded as the visual cue-specific components to compose the mixture model conditioned on the corresponding POS, which enables the most relevant component to dominate the generating of lexical word.

The main process of our method for video captioning is demonstrated in Figure 1. Overall, the main contributions are as follows:

- We propose a novel video captioning approach that jointly learns the syntax representation and translates visual cues to generate accurate textual descriptions. An end-to-end trainable network is built to model the joint probability of the POS sequence and the caption words by simultaneously capturing the syntactic structure and exploiting the semantic primitives.

- We design a mixture model of multiple visual cue-specific components to handle the word bias problem caused by imbalanced classes inherent in linguistic data, with the guidance of the interpretable and accessible POS tag.

- Experiments on three public datasets comprehensively verify the superior performance of our method on video captioning compared with the state-of-the-art methods.

2. Related Work

Early methods of video captioning are mainly based on templates, and the sentence templates should be predefined. [21] is one of the first trails to describe human activities in videos by extracting semantic primitives of videos and associating them with components of the template to form a sentence. Krishnamoorthy *et al*. [23] developed a holistic approach to directly select the best subject-verb-object triplet as the video caption. Guadarrama *et al*. [16] built semantic hierarchies and filled in words to generate captions of videos, where the verbs are generated by a zero-shot technique of action recognition. These methods generate fixed-structure sentences with limited diversity of natural language. Different from these template-based methods with fixed syntactic structure, our method automatically infers the human-interpretable POS from the input video, which benefits generating accurate and diverse sentences.

Recently, sequence-to-sequence based methods have become prevalent in video captioning. [38] extracts CNN features to represent input videos and uses an LSTM to generate video descriptions. Some recent efforts have been made on exploring better video representation. The S2VT [37] encodes frame-level features into a global feature of a video and decodes it into a sentence via an encoder-decoder LSTM. [29, 4] exploit the hierarchical structures of the videos for captioning. Concretely, [29] uses a hierarchical LSTM to encode videos along time, and [4] adapts the cells of hierarchical LSTM to be boundary aware for better representing videos. Chen *et al*. [8] used a reinforcement-learning method to select informative frames for video captioning. Several other methods manage to exploit semantic cues or concepts for video captioning. [33] first learns the semantic representations from videos via conditional random field, and then translates them to captions using statistical machine translation. Donahue *et al*. [11] extended [33] by changing the statistical machine translation method

to the LSTM [19] decoder. [48, 41, 13] introduce the attention mechanism for video captioning, where [48, 41] select salient spatial-temporal features to generate sentences and [13] uses hierarchical attention to capture the temporal dynamics for captioning. Wang *et al*. [39] exploited the bidirectional cues between the video and caption by an encoder-decoder-reconstructor to describe videos. [30, 12] use the interpretable cues as the trade-off between the video and natural language. Aafaq *et al*. [1] considered both spatio-temporal dynamics and high-level semantic concepts, and employed Short fourier transform to enrich the visual representation for video captioning. Different from these semantic concept based methods, our method not only learns semantic primitives but also learns the syntax representation from video, which further improves the accuracy of generated sentences.

There are several studies on leveraging POS for image captioning. He *et al*. [18] directly used the POS tag of the current word to locally guide the prediction of the next word, while our method learns the *global* syntax representation (i.e. POS tag sequence) to generate accurate captions. Deshpande *et al*. [10] pre-defined 1,024 POS templates from images by clustering and fed the template with the image into an S2S model to generate the caption, while our method flexibly generates POS tag sequences from videos and exploits multiple visual cues to boost video captioning. Our method predicts sentences conditioned on the POS tags via a mixture model to fully exploit visual cues from videos for precise caption generation.

Our method also differs from the two-stage image captioning methods [24, 27, 42] which first generate sentence templates with slots tied to object entities, and then filling the slots using object detectors. Our method simultaneously learns not only visual concepts but also the syntax representation, thus efficiently exploiting syntactic structure and semantic primitives to generate accurate sentences.

3. Our Approach

Our method contains two key modules: syntax representation learning and visual cue translation, as shown in Figure 2. The syntax representation learning module takes extracted video features and embedded previous words as input and outputs POS sequences using a sequence-to-sequence (S2S) model, also namely video POS tagging. The visual cue translation module is implemented by a mixture of multiple components explicitly conditioned on the inferred POS tags, where each component takes specific visual cue features as input. We build an end-to-end trainable network to jointly model the video POS tagging and visual cue translation for video captioning.

For each input video v, our model generates its corresponding POS sequence $t = (t_1, \cdots, t_N)$ and word sequence $s = (s_1, \cdots, s_N)$, where N is the number of words

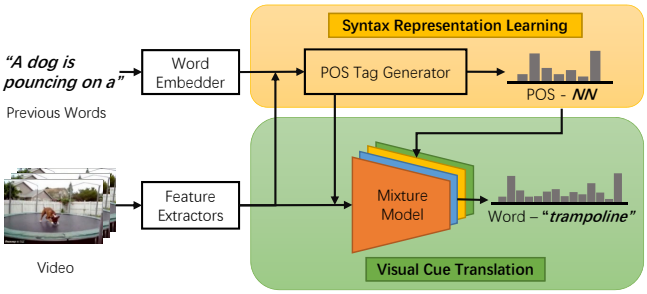

Figure 2. The overall architecture of our method. There are two key modules: syntax representation learning and visual cue translation. The syntax representation learning module takes video features extracted by feature extractors and embedded previous words as input and outputs POS sequences. The visual cue translation module is implemented by a mixture of multiple components where each component takes specific visual cue features as input to output lexical words.

in a sentence. t_i indicates the i-th POS tag, belonging to a pre-defined POS tag set $\mathcal{T}$ which includes 26 POS tags and 1 tag denoting the end of sentence, i.e., $t_i \in \mathcal{T}$. s_i represents the i-th word, belonging to a fixed vocabulary $\mathcal{S}$, i.e., $s_i \in \mathcal{S}$.

A probabilistic directed acyclic graph is built to learn the joint probability of the POS sequence t and the caption s. The joint probability of t and s given a video v is formulated by

$$p(t, s|v; \theta) = \prod_{i}^{N} p(t_i|s_{<i}, v; \theta_t) p(s_i|t_i, v; \theta_s), \tag{1}$$

where $\theta = \theta_t \cup \theta_s$ represents the model parameters and $s_{<i}$ indicates all the previous words at time step i. $p(t_i|s_{<i}, v; \theta_t)$ means that the generation of POS tag t_i at time step i is conditioned on the input video v and all the previous words $s_{<i}$ to determine its contextual location in a sentence. $p(s_i|t_i, v; \theta_s)$ represents that the generation of word s_i is conditioned on its corresponding POS tag t_i as well as the video v. To learn the optimal parameters θ^*, we maximize the likelihood function over the training data:

$$\begin{aligned} \theta^* = \arg\max_{\theta} \sum_{v \in \mathcal{V}} \sum_{s \in \mathcal{S}^v} \sum_{i=1}^{N} \big(& \log p(t_i|s_{<i}, v; \theta_t) \\ & + \log p(s_i|t_i, v; \theta_s) \big), \end{aligned} \tag{2}$$

where $\mathcal{S}^v \subseteq \mathcal{S}$ is the caption set of the video v. Accordingly, the overall optimization problem can be regarded as jointly optimizing two terms in Eq. (2), corresponding to video POS tagging and visual cue translation, respectively.

3.1. Video POS Tagging

An S2S model is introduced to infer POS sequences from input videos. The loss function for video POS tagging is formulated as

$$L_t = \sum_{v\in\mathcal{V}} \sum_{s\in\mathcal{S}^v} \sum_{i=1}^{N} -\log p(t_i|s_{<i}, v; \theta_t). \tag{3}$$

To guarantee the diversity of the parts of the speech in each sentence. We encourage each POS tag to appear in a sentence at least once. This constraint is given by

$$L_c = \sum_{v\in\mathcal{V}} \frac{1}{\sqrt{N}} ||\mathrm{sgn}(\sum_{i=1}^{N} \boldsymbol{y}_i) - \mathbf{1}||_2, \tag{4}$$

where $\boldsymbol{y}_i$ is a one-hot vector indicating the class of the POS tag, $\mathbf{1}$ is an all-one vector with the same dimension as $\boldsymbol{y}_i$, N is the number of POS tags in a sentence, and $\mathrm{sgn}(\cdot)$ is the sign function. Obviously, the gradient of the function is discontinuous, so we reduce the impact on the final loss function by multiplying with a small coefficient to avoid the collapse during training.

In training, the POS of each training video can be easily obtained via the existing language POS tagging method, such as NLTK tool [45], without manual annotation.

3.2. Visual Cue Translation

The goal of visual cue translation is to classify the visual cues into words conditioned on the POS tags for generating video captions. The loss function for this sub-task is formulated as

$$L_s = \sum_{v\in\mathcal{V}} \sum_{s\in\mathcal{S}^v} \sum_{i=1}^{N} -\log p(s_i|t_i, v; \theta_s). \tag{5}$$

To avoid word bias problem of representing various semantics in natural language when using a single softmax classifier, a mixture model is designed to formulate the conditional probabilities of words $p(s_i|t_i, v; \theta_s)$:

$$p(s_i|t_i, v; \theta_s) = \sum_{j=1}^{J} \alpha_j p(s_i|t_i, v; \theta_s^j), \ \sum_{j=1}^{J} \alpha_j = 1, \tag{6}$$

where J is the number of mixture components, α_j is the mixture weight characterizing the prior probability of the j-th component, and θ_s^j is the parameters of j-th component, *i.e.*, $\theta_s = \cup_{j=1}^{J} \theta_s^j$. Each mixture component describes a categorical distribution over the sample space generated by a non-linear model. These components are designed to be sensitive to different visual cues in videos, and we call them visual cue-specific components.

Specifically, the set of POS tags $\mathcal{T}$ is explicitly divided into four subsets $\mathcal{T}^j|_{j=1}^4$ which correspond to the object, motion, property and context cues of videos, respectively. Table 1 shows the detailed division of the set of POS tags. For instance, the POS "NN" (noun) corresponds to the object cues of videos. Thus, the number of mixture components is set to $J = 4$, and the corresponding mixture weight α_j is given by

Subsets	POS tags
Object	NN, NNS
Motion	VB, VBD, VBG, VBN, VBP, VBZ
Property	CD, JJ, JJR, RB
Context	CC, DT, EX, IN, MD, PRP, PRP$, RP, TO, WDT, WP, WRB, BES, EOS

Table 1. The 27 tags are divided into four subsets. The words are classified into POS by the NLTK tool [45]. We choose the top 24 most frequent POS tags in the video captioning corpus, and other POS tags as the UNK tag. Because linking verbs do not contain the motion information, we classify the verb "be" and its various forms as the BES which represents contextual information.

$$\alpha_j = \sum_{t_i\in\mathcal{T}^j} p(t_i|s_{<i}, v; \theta_t), \ j = 1, 2, 3, 4. \tag{7}$$

The inputs of the four visual cue-specific components are related to their corresponding visual cues. For example, the "motion"-specific component takes the motion-related video feature as input, and the "object"-specific component takes the object-related feature as input. Existing generic pre-trained CNN models, such as ResNets [17], Inceptions [34] and C3D [35]), are readily adopted as feature extractors for extracting robust features to represent specific visual cues of the video.

3.3. Training and Inference

The whole model is trained by the following loss function:

$$L = L_t + L_s + \gamma L_c, \tag{8}$$

where γ is the coefficient of the constraint term L_c. Empirically, at the early stage of training, the POS tags of the generated sentence are often lacking of diversity. The effectiveness of the constraint term could be obvious, and L_c would guide the model to converge to a better solution. Since the term is too restrictive, that is, not every sentence contains all kinds of POS tags, L_c is suppressed later in the training procedure. Accordingly, the coefficient of L_c is set to:

$$\gamma = \exp(-\beta k), \tag{9}$$

where k denotes the number of training epochs, and β is the coefficient which is empirically determined by the learning rate.

During inference, given the video and a start token <S> to the trained model, the first word is generated and fed together with the former input into the model to generate the second word. The process is repeated until the ending token is predicted or the maximum length is reached. We use beam search with a size of 5 to generate the final sentences.

3.4. Analysis of Mixture Model

In this subsection, we provide theoretical analysis of the mixture model in Section 3.2 from the perspective of the imbalanced classes. In linguistics, there is an open-closed distinction among the words with different POS. Function words (e.g., conjunctions) appear frequently but their number is finite and relative small (about 150 in the English language), while content words (e.g., nouns) are just the opposite. Benefiting from this characteristic, our mixture model can address the imbalanced classes problem by dividing the words into four subsets. Generally speaking, during the back-propagation of training using imbalanced data, the majority classes would produce major contribution to the parameter update, which causes the model more sensitive to the data in majority classes. In the proposed model, the gradient of the loss function L_s is

$$\nabla_\theta L_s = \sum_{v\in\mathcal{V}} \sum_{s\in\mathcal{S}^v} \sum_{i=1}^{N} -\frac{1}{\nabla_{\theta_s} p(s_i|t_i, v; \theta_s)}, \tag{10}$$

where

$$\nabla_{\theta_s} p(s_i|t_i, v; \theta_s) = \sum_{j=1}^{J} \left(\alpha_j \nabla_{\theta_s^j} p(s_i|t_i, v; \theta_s^j) + p(s_i|t_i, v; \theta_s^j) \nabla_{\theta_t} \alpha_j\right). \tag{11}$$

As inferred from the first term in Eq. (11), the impact of a word on the model parameters will be emphasized when the POS tag identifies the word as content words. The amount of content words in video captioning dataset is larger than that of function words, so the overall impact of the content words will be larger. The second term in Eq. (11) guarantees that the probabilities of the POS will be updated only when the word is correctly predicted.

We also find that the proposed mixture model can alleviate the softmax bottleneck problem in natural language which has been revealed by [46, 20]. The softmax bottleneck is about the case that a single softmax function is used at the top of the network to obtain the probability distribution over word categories, which is known as the linear-softmax layer. When the number of output category, i.e., all the words of the vocabulary, is much larger than the representation dimension, the linear-softmax layer will limit representation power. The detailed explanation of the softmax bottleneck comes from the classical matrix factorization theory, and the low-rank property of the log-probability matrix prevents linear-softmax layer to exactly catch all the appropriate words from the large vocabulary set. Our mixture model in Eq. (6) integrates four softmax together, which provides a non-linear function, *i.e.*, log-sum-exp, during calculating the log-probability matrix, and the matrix can be arbitrarily high-rank. In this way, the softmax bottleneck problem is alleviated.

4. Experiments

4.1. Datasets

MSVD [6] comprises 1,970 video clips collected from Youtube, each annotated with roughly 40 captions. Following [37], we split the videos into three sets, consisting of 1,200 training, 100 validation and 670 testing videos, respectively.

MSR-VTT [44] contains 10K video clips, each of which has 20 captions. As in [44], we take 6,513 videos for training, 497 for validation, and 2,990 for testing.

ActivityNet Captions [22] contains 20K videos annotated with 100K temporally localized sentences. We use the ground-truth proposals and the corresponding captions in this dataset to evaluate our method following the split standard in [22].

4.2. Experimental Setup

Evaluation metrics. We use the metrics of BLEU-4 (B@4) [31], METEOR [9], ROUGE-L [26], and CIDEr [36] for evaluations by the MSCOCO toolkit [7]. For all the metrics, higher values indicate better performances.

Feature representations. The input visual cue representations of the mixture model consist of four different types of video features, i.e., *context*, *RGB*, *motion* and *local features*. The context feature is extracted from the S2S model which is applied for tagging the POS. The other three types of features are extracted by several existing CNNs from videos. The details of extracting these features can be found in Section 4.3. The input visual cue representation for each component in the mixture model can thus be obtained by combining the four types of video features in different ways.

- Object cue representation as the input to the object-specific component is calculated by soft-assigning the local features depending on the context feature at the current time step via a soft attention operation, since an object often locates in a local region of a video.
- Motion cue representation as the input to the motion-specific component is obtained by concatenating the motion and context features.
- Property cue representation as the input to the property-specific component is obtained by concatenating the RGB and context features. RGB features represent the global information within the video

Feature Extractors	Methods	MSVD				MSR-VTT			
		B@4	METEOR	ROUGE-L	CIDEr	B@4	METEOR	ROUGE-L	CIDEr
ResNet-152	HRL [41]	-	-	-	-	41.3	28.7	61.7	48.0
	PickNet [8]	**52.3**	33.3	69.6	76.5	41.3	27.7	59.8	44.1
	Ours	52.1	**33.7**	**69.8**	**80.6**	**41.4**	**28.9**	**62.0**	**48.1**
ResNet-152+C3D	SCN [12]	51.1	33.5	-	77.7	-	-	-	-
	Ours	**52.4**	**33.7**	**70.2**	**81.3**	40.7	28.9	61.7	48.3
Incpetion-v4	RecNet [39]	52.3	34.1	69.8	80.3	39.1	26.6	59.3	42.7
	Ours	**52.5**	**34.4**	**70.3**	**83.0**	**40.7**	**28.3**	**60.4**	**45.3**
IRv2+C3D+YOLO	GRU-EVE [1]	47.9	35.0	71.5	78.1	38.3	28.4	60.7	48.1
IRv2+C3D	Ours	**52.8**	**36.1**	**71.8**	**87.8**	**42.3**	**29.7**	**62.8**	**49.1**

Table 2. Performance evaluation of our method using the same features with the recent state-of-the-art methods on the MSVD and MSR-VTT datasets.

frames. We apply the RGB features as the property cue representation because some properties such as the comparative and cardinal number might be better described by referring the global videos.

- Context cue representation as the input of the context-specific component is comprised of the context feature.

Implementation details. The ConvCap network with the default parameter settings in [2] is used as the S2S model for video POS tagging, and the embedding size of each input word is set to 512. The top-most layer output of the ConvCap network without the last attention operation is used as the context feature to fully represent the semantic relationship between words. In the mixture model, each visual cue-specific component is constructed by a fully connected layer with an RReLU activation layer [43] and a softmax classifier. To spatially align the features, the layer normalization operation [3] is applied before all the concatenations. The proposed method is implemented with PyTorch [32] on a Titan X GPU with 12G memory. The RMSprop [15] is employed to optimize our model, and the learning rate is set to $1\mathrm{e}^{-4}$.

4.3. Comparison with the State-of-the-Art

To evaluate the effectiveness of joint modeling of syntax representation learning and visual perception in our method for video captioning, several recently proposed methods [41, 8, 12, 39, 1] that are closely related to our method are employed for comparison. For fair comparison, the same features with these methods are used as the input of our method, which are detailed as follows.

- ResNet-152: The RGB and local features are extracted from the average pooling layer and the res5b layer of the ResNet-152 [17], respectively. The motion features are generated by using a temporal attention operation on the RGB features.
- ResNet-152+C3D: The RGB and local features are the same with (1). The motion features are extracted from the pool5 layer of the C3D [35].
- Inception-v4: The RGB and local features are derived from the average pooling layer and the Reduction-B layer of the Inception-v4 [34], respectively. The motion features are calculated by using a temporal attention operation on the RGB features.
- IRv2+C3D: The RGB features and local are derived from the average pooling layer and the Reduction-B layer of the IRv2 [34], respectively. The motion features are extracted from the pool5 layer of the C3D.

Concretely, we sample frames of each video at $3fps$ to obtain the RGB features. The motion features extracted by the C3D are obtained using 16-frame clips as input with an 8-frame overlap. For local features, we randomly sample 4 RGB frames as the input to the feature extractors for the attention operation to reduce the computational cost.

Table 2 shows the comparison results on both MSVD and MSR-VTT datasets. In is evident that our method achieves satisfactory performances when compared with other methods using the same features. Note that in comparison with GRU-EVE [1], for computation simplicity, we do not use the YOLO model to detect objects for extracting better representations and our method still achieves significantly better results than GRU-EVE. This substantially validates the superiority of our method on simultaneously exploring syntatic structure of sentences and perceiving semantic primitives for video captioning.

Moreover, we provide the video captioning results of other state-of-the-art methods for comprehensive comparisons on the MSVD and MSR-VTT datasets in Table 3 and Table 4, respectively. In this experiment, we use IRv2+C3D as the feature extractors to obtain the input feature representations. It is obvious that our method consistently performs better than the state-of-the-art methods on both MSVD and MSR-VTT datasets for most evaluation metrics. Note that

Methods	B@4	METEOR	ROUGE-L	CIDEr
SA-LSTM [47]	41.9	29.6	-	51.7
HRNE [29]	46.7	33.9	-	-
h-RNN [48]	49.9	32.6	-	65.8
BAE [4]	42.5	32.4	-	63.5
TSA [30]	**52.8**	33.5	-	74.0
aLSTMs [13]	50.8	33.3	61.1	74.8
SCN [12]	51.1	33.5	-	77.7
M^3 [40]	**52.8**	33.1	-	-
RecNet [39]	52.3	34.1	69.8	80.3
PickNet [8]	52.3	33.3	69.6	76.5
GRU-EVE [1]	47.9	35.0	71.5	78.1
Ours	**52.8**	**36.1**	**71.8**	**87.8**

Table 3. Comparison with the state-of-the-art methods on the MSVD dataset.

Methods	B@4	METEOR	ROUGE-L	CIDEr
SA-LSTM [47]	37.1	28.4	-	-
aLSTMs [13]	38.0	26.1	-	43.2
M^3 [40]	38.1	26.6	-	-
RecNet [39]	39.1	26.6	59.3	42.7
HRL [41]	41.3	28.7	61.7	48.0
PickNet [8]	41.3	27.7	59.8	44.1
GRU-EVE [1]	38.3	28.4	60.7	48.1
Ours	**42.3**	**29.7**	**62.8**	**49.1**

Table 4. Comparison with the state-of-the-art methods on the MSR-VTT dataset.

Methods	B@4	METEOR	ROUGE-L	CIDEr
DCE [22]	1.6	8.9	-	25.1
DVC [25]	1.6	10.3	-	25.2
SDVC [28]	1.3	**13.1**	-	43.5
Ours	**1.9**	11.3	**22.4**	**44.2**

Table 5. Comparison with the state-of-the-art methods on the ActivityNet Captions dataset.

the improvement of our method under B@4 is not as remarkable as other evaluation metrics. The probable reason is that our method aims at learning the syntactic structure representation to generate sentences and B@4 is a metric based on lexical rather than syntactic matching [14]. This phenomenon that integrating syntactic information often fail to improve the BLEU score is also found and explained in the previous work [5].

We also show comparison results on the validation set of ActivityNet Captions in Table 5. For fair comparison, our input features are segment-level C3D features, the same with the compared methods. The RGB features in our model are derived from the average pooling of the C3D features. We apply self-attention operations on the C3D features to calculate the local features. The motion feature of each proposal is obtained by encoding the C3D features using LSTM. From Table 5, it is obvious that our method generally outperforms the state-of-the-art methods on a more challenging dataset.

Methods	B@4	METEOR	ROUGE-L	CIDEr
Baseline S2S	50.1	34.3	69.4	77.0
w/o L_s	51.7	34.3	70.7	82.1
w/o L_t	52.5	35.1	71.2	85.6
w/o L_c	51.3	34.7	70.7	83.3
Ours	**52.8**	**36.1**	**71.8**	**87.8**

Table 6. Results of ablation experiments on the MSVD dataset.

4.4. Ablation Studies

To go deeper with each component of our method, we compare our method with four variations: without POS tagging and mixture model (baseline S2S), without mixture model (w/o L_s), without POS tagging (w/o L_t) and without the constraint term L_c (w/o L_c).

- Baseline S2S uses the ConvCap [2], which has the same network architecture with the POS tag generator in our method except that the last 1D convolutional layer uses the attention mechanism to directly generate captions. It has the same word embedder and feature extractors with our method.
- w/o L_s uses softmax classifier instead of the mixture model to generate words with the guidance of inferred POS tags.
- w/o L_t directly uses the mixture model to generate sentences given the input feature presentations without video POS tagging.
- w/o L_c removes the constraint term L_c from the loss function.

These ablation studies are conducted on the MSVD dataset using the IRv2+C3D as feature extractors. The results are reported in Table 6. We can have the following observations: (1) our method achieves the best result, which obviously validate the importance of each individual component in our method; (2) The improvement of our method compared with "w/o L_s" proves that the proposed mixture model can effectively solve the word bias problem and further boost the performance by leveraging the learned syntactic structures; (3) Our method outperforms "w/o L_t" which clearly shows that learning syntax representation of sentences via video POS tagging is helpful to generating accurate descriptions of videos. (4) The performance drops when removing the constraint L_c, which validates the importance of encouraging the variety in POS of each sentence.

4.5. Qualitative Analysis

Figure 3 shows some qualitative results of video captioning from six videos. For each video, three frames are

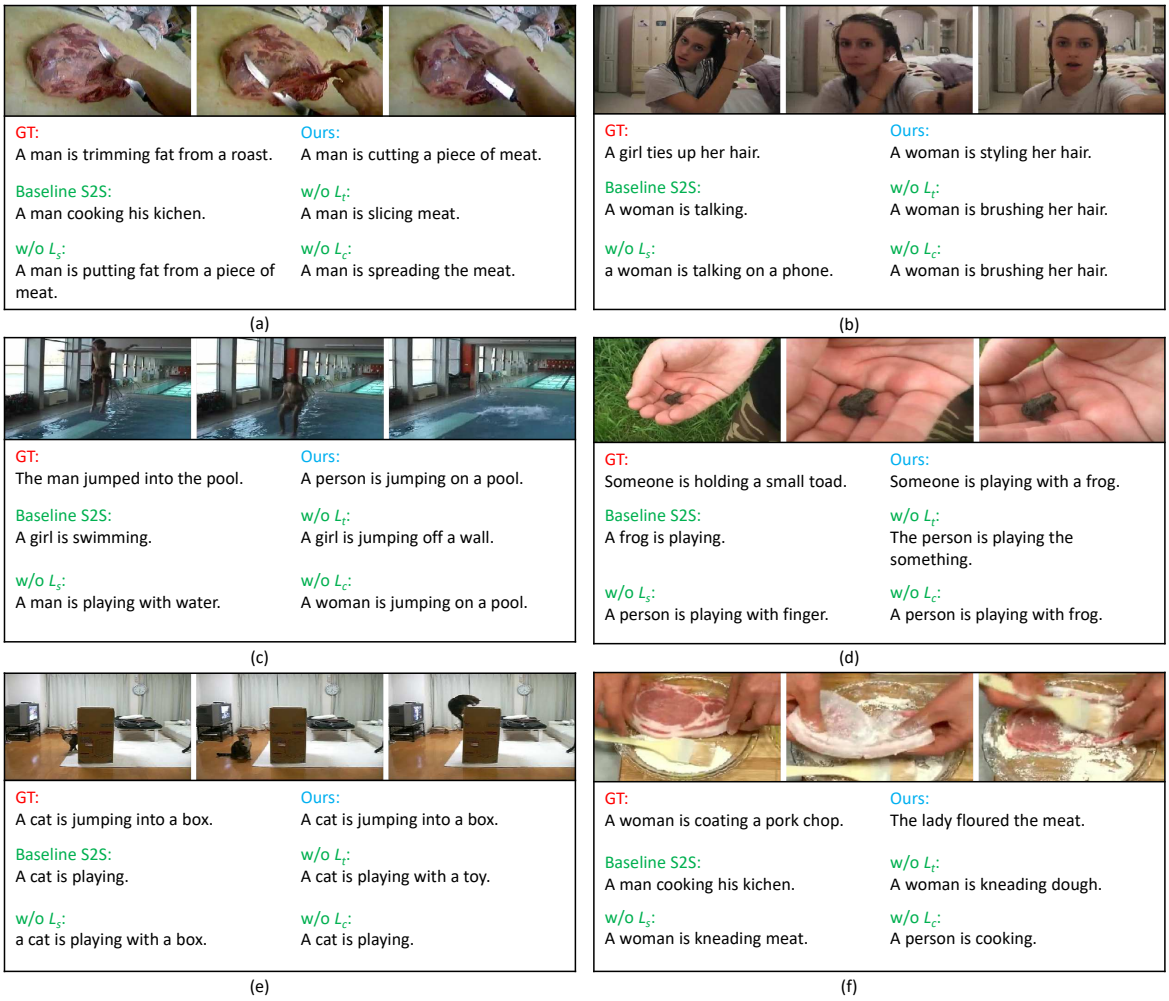

Figure 3. Qualitative results for video captioning. There are six videos. For each video, three frames are selected for illustration and six sentences are shown, including the ground truth (GT) sentence, the sentence generated by our method (Ours), and the other four sentences generated by four variants of our method in the ablation studies (baseline S2S, w/o L_s, L_t, and L_c).

selected for illustrations. It is interesting to observe that our method can generate sentences with more accurate semantics and syntax for describing the videos. Compared with the method w/o L_s in (c), sentences generated by our method express more precise semantic meanings via effectively solving the word bias problem. (d) indicates that learning syntax representation of sentences using L_t is indispensable to our method. The effect of the constraint L_c can be observed from (e) and (f) where sentences generated by the method without L_c lacks the diversity of syntactic structures. According to these observations, we conclude that all the proposed modules in our method contribute to generating accurate video captions.

5. Conclusion

We have presented a novel approach of jointly learning syntax representation and translating visual cues for video captioning. It can simultaneously capture the syntactic structure of sentences via video POS tagging and perceive intrinsic semantic primitives via a new mixture model. The mixture model can successfully address the word bias problem inherent in natural language data. An end-to-end trainable network is built to model the joint probability of the POS sequence and the lexical words, which is capable of generating accurate and diverse descriptions of videos. Extensive experiments on three public datasets demonstrate that our method outperforms the state-of-the-art methods on video captioning.

Acknowledgments This work was supported in part by the Natural Science Foundation of China (NSFC) under Grant No.61673062.

References

[1] Nayyer Aafaq, Naveed Akhtar, Wei Liu, Syed Zulqarnain Gilani, and Ajmal Mian. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In *CVPR*, pages 12487–12496, 2019.

[2] Jyoti Aneja, Aditya Deshpande, and Alexander G. Schwing. Convolutional image captioning. In *CVPR*, pages 5561–5570, June 2018.

[3] Lei Jimmy Ba, Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016.

[4] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. Hierarchical boundary-aware neural encoder for video captioning. In *CVPR*, pages 3185–3194, 2017.

[5] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluation the role of bleu in machine translation research. In *EACL*, 2006.

[6] David L. Chen and William B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *ACL*, pages 190–200, 2011.

[7] 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. *CoRR*, abs/1504.00325, 2015.

[8] Yangyu Chen, Shuhui Wang, Weigang Zhang, and Qingming Huang. Less is more: Picking informative frames for video captioning. In *ECCV*, pages 367–384, September 2018.

[9] Michael J. Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *ACL*, pages 376–380, 2014.

[10] Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander G. Schwing, and David A. Forsyth. Fast, diverse and accurate image captioning guided by part-of-speech. In *CVPR*, pages 10695–10704, 2019.

[11] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Trevor Darrell, and Kate Saenko. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, pages 2625–2634, 2015.

[12] 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*, pages 1141–1150, 2017.

[13] Lianli Gao, Zhao Guo, Hanwang Zhang, Xing Xu, and Heng Tao Shen. Video captioning with attention-based LSTM and semantic consistency. *IEEE Trans. Multimedia*, 19(9):2045–2055, 2017.

[14] Jesús Giménez and Lluís Màrquez. Linguistic features for automatic evaluation of heterogenous MT systems. In *WMT@ACL*, 2007.

[15] Alex Graves. Generating sequences with recurrent neural networks. *CoRR*, abs/1308.0850, 2013.

[16] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond J. Mooney, Trevor Darrell, and Kate Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *ICCV*, pages 2712–2719, 2013.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[18] 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*, 2017.

[19] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[20] Sekitoshi Kanai, Yasuhiro Fujiwara, Yuki Yamanaka, and Shuichi Adachi. Sigsoftmax: Reanalysis of the softmax bottleneck. In *NeurIPS*, pages 284–294, 2018.

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

[22] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, pages 706–715, 2017.

[23] Niveda Krishnamoorthy, Girish Malkarnenkar, Raymond J. Mooney, Kate Saenko, and Sergio Guadarrama. Generating natural-language video descriptions using text-mined knowledge. In *AAAI*, pages 541–547, 2013.

[24] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE TPAMI*, 2013.

[25] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Jointly localizing and describing events for dense video captioning. In *CVPR*, pages 7492–7500, 2018.

[26] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *WAS*, 2004.

[27] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *CVPR*, pages 7219–7228, 2018.

[28] Jonghwan Mun, Linjie Yang, Zhou Ren, Ning Xu, and Bohyung Han. Streamlined dense video captioning. In *CVPR*, pages 6588–6597, 2019.

[29] Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, and Yueting Zhuang. Hierarchical recurrent neural encoder for video representation with application to captioning. In *CVPR*, pages 1029–1038, 2016.

[30] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. Video captioning with transferred semantic attributes. In *CVPR*, pages 984–992, 2017.

[31] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318, 2002.

[32] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

[33] Marcus Rohrbach, Wei Qiu, Ivan Titov, Stefan Thater, Manfred Pinkal, and Bernt Schiele. Translating video content to natural language descriptions. In *ICCV*, pages 433–440, 2013.

[34] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-resnet and the

impact of residual connections on learning. In *AAAI*, pages 4278–4284, 2017.

[35] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015.

[36] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, 2015.

[37] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond J. Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence - video to text. In *ICCV*, pages 4534–4542, 2015.

[38] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond J. Mooney, and Kate Saenko. Translating videos to natural language using deep recurrent neural networks. In *NAACL-HLT*, pages 1494–1504, 2015.

[39] Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. Reconstruction network for video captioning. In *CVPR*, pages 7622–7631, 2018.

[40] Junbo Wang, Wei Wang, Yan Huang, Liang Wang, and Tieniu Tan. M3: multimodal memory modelling for video captioning. In *CVPR*, pages 7512–7520, 2018.

[41] Xin Wang, Wenhu Chen, Jiawei Wu, Yuan-Fang Wang, and William Yang Wang. Video captioning via hierarchical reinforcement learning. In *CVPR*, pages 4213–4222, 2018.

[42] Yu Wu, Linchao Zhu, Lu Jiang, and Yi Yang. Decoupled novel object captioner. In *ACM MM*, pages 1029–1037, 2018.

[43] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *CoRR*, abs/1505.00853, 2015.

[44] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296, 2016.

[45] Nianwen Xue. Steven bird, evan klein and edward loper. *Natural Language Processing with Python*. o'reilly media, inc 2009. ISBN: 978-0-596-51649-9. *Natural Language Engineering*, 17(3):419–424, 2011.

[46] Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W. Cohen. Breaking the softmax bottleneck: A high-rank RNN language model. In *ICLR*, 2018.

[47] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher J. Pal, Hugo Larochelle, and Aaron C. Courville. Describing videos by exploiting temporal structure. In *ICCV*, pages 4507–4515, 2015.

[48] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *CVPR*, pages 4584–4593, 2016.