Guiding the Flowing of Semantics: Interpretable Video Captioning via POS Tag

Xinyu Xiao[1,2], Lingfeng Wang[1], Bin Fan[1], Shiming Xiang[1,2], Chunhong Pan[1]
[1]NLPR, Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
{xinyu.xiao, lfwang, bfan, smxiang, chpan}@nlpr.ia.ac.cn

Abstract

In the current video captioning models, the video frames are collected in one network and the semantics are mixed into one feature, which not only increase the difficulty of the caption decoding, but also decrease the interpretability of the captioning models. To address these problems, we propose an Adaptive Semantic Guidance Network (ASGN), which instantiates the whole video semantics to different POS-aware semantics with the supervision of part of speech (POS) tag. In the encoding process, the POS tag activates the related neurons and parses the whole semantic information into corresponding encoded video representations. Furthermore, the potential of the model is stimulated by the POS-aware video features. In the decoding process, the related video features of noun and verb are used as the supervision to construct a new adaptive attention model which can decide whether to attend to the video feature or not. With the explicit improving of the interpretability of the network, the learning process is more transparent and the results are more predictable. Extensive experiments demonstrate the effectiveness of our model when compared with state-of-the-art models.

1 Introduction

Video captioning, which transforms the semantic information in a video to a natural statement, has received wide attention recently. The series of scenes (both related and unrelated) in video frames bring a huge challenge for the task of video captioning. Therefore, mastering the ability to process the correlated and irrelevant semantic information can improve the performance and interpretability of the model of video captioning.

The classical deep learning based video captioning methods (Venugopalan et al., 2015a,b) incorporate both CNN and LSTM together as an encoder-decoder architecture which extracts all the semantic information in one stream and fuses them in a single feature. In the architecture, the transformation of the semantics to words prediction is uncertainty. Some other recent studies (Dong et al., 2017; Wu et al., 2018) attempt to improve the interpretability of the transformation process. But the flowing of different semantic information in the network streams and the activated neurons in network are still ambiguous. To address these issues, the first is to take the discrimination of different semantic information into consideration. According to (Horoufchin et al., 2018; Fargier and Laganaro, 2015), nouns and verbs typically describe concrete objects and actions which can recruit the canonical neurons system to activate corresponding representation patterns in brain. This indicates that differential neurons in brain are activated for lexical selection of action and object words. Therefore, the part of speech (POS) tag can be applied to guide the flowing of different semantics into the corresponding network streams.

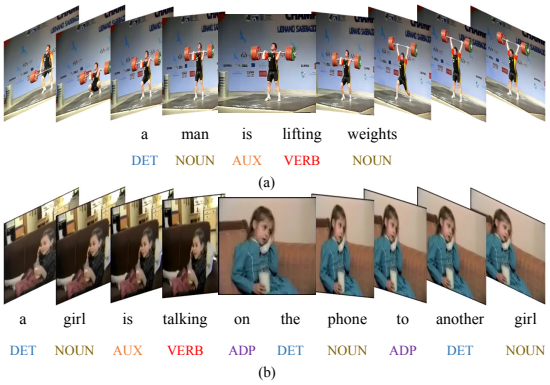

Figure 1: Examples of video description and its POS tags. The definition of POS tag can be seen in Section 3.1. From these examples, it can be seen that the words that are belonging to the same POS tags undergo inflection for similar semantic properties and different POS tags reflect different semantic properties distinctly.

A POS tag is the category of words which has similar grammatical properties. The words assigned to the same POS tag generally reflect similar properties within the grammatical structure of sentence. Fig.1 (a) shows a video and its caption "a man is lifting weights". The nouns "man", "weights" and the verb "lifting" are belonging to different POS tags and referring to different semantic information in video. Moreover, in Fig. 1 (b), the caption of video is "a girl is talking on the phone to another girl". The words of "girl", "talking" and "phone" distinctly have their corresponding visual signals in the video, but the others are uncertain. This example indicates that the nouns of objects and verbs of actions are generally referred to visual words. Another property of the POS tag is that when given a fixed sentence, the POS tags of all the words in the sentence are fixed. It ensures the reliability of the POS tag in helping to extract and guide corresponding semantics.

According to Merity et al. (Merity et al., 2016), the prediction of the next word not always need to attend the visual feature. The gradients from non-visual words could mislead and diminish the effectiveness of the visual signal in guiding caption generation. To this end, Lu et al. (Lu et al., 2016) proposed a "visual sentinel", which applies the hidden state in LSTM as supervision to adaptively decide if it is necessary to input visual feature to the language model when generating the next word. However, the supervision information from the hidden state isn't credible which cannot make sure it contains the corresponding visual decision signals. According to the properties of the POS tag, the nouns and verbs can be applied as the supervision to distinguish whether the remaining words in the sentence are visual words or not.

In this paper, to explicitly improve the interpretability of video captioning model, we propose an Adaptive Semantic Guidance Network (ASGN), which instantiates the whole video semantics by part of speech (POS) tags to different POS-aware semantics. At first, a POS-aware semantic guider is proposed. It predicts the POS tags of the words in descriptions, and guides different POS-aware semantic information of video into corresponding network streams by the predicted POS tags. In this process, the specific CNN neurons are activated under the supervision of POS tags and the whole visual semantics are parsed into POS tags related video features. Moreover, depending on the POS-aware video features, a new adaptive attention operation is introduced. The video features related to noun and verb are used as a supervision to get a sentinel gate, which decides how much the attended feature can be imported into the decoder LSTM when generating the next word. A reinforcement learning (RL) method is applied to optimize our model which further demonstrates the validity of our method on the video captioning task.

The main contributions of this paper are:

- Designing a POS-aware semantic guider to predict the POS tags of words and guide different semantic information of video into corresponding network streams. Under the supervision of POS tags, the CNN neurons are selectively activated and aware of the related POS tags, so that the whole video semantics are parsed into the corresponding POS-aware video features.

- Depending on the noun and verb instantiated video features, a new adaptive attention model is constructed to decide how much the visual feature is imported into the decoder.

- Due to the guiding of POS tags, the flowing of the type of semantic in which network stream can be easily clarified. With the supervision of the noun and verb related features, the judgment of the predicted word is visual word or not is more reasonable. These make the learning process more interpretable while achieving state-of-the-art performance.

2 Related Works

Here, we first review the recent implements of the POS tag in computer vision, then review the most relevant works on video description task like attention-based methods and interpretable improved models.

The POS tag has been received attention in some computer vision tasks, like visual question answering (Wang et al., 2018b) and image captioning (A et al., 2019). He et al. (He et al., 2017) utilized the POS tag of each word to determine whether it is essential to input image representation into the word generator. Wang et al. (Wang et al., 2018b) exploited the POS tag guided attention model in VQA to put more emphasis on the important words such as nouns, verbs and adjectives. All these methods realized the importance

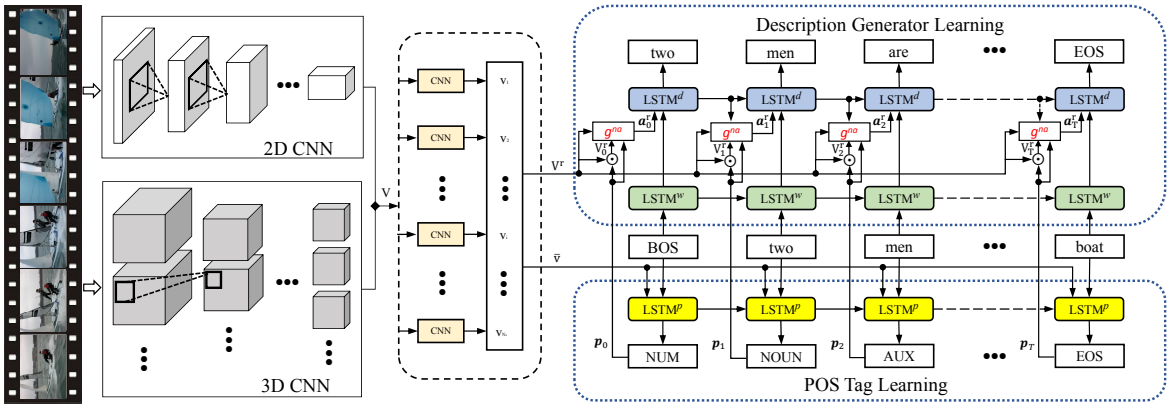

Figure 2: The architecture of our Adaptive Semantic Guidance Network (ASGN) which can shunt the required semantic information into different video features when generating every word in caption. The features $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{N_s}\}$ are extracted by the CNN modules. $\odot$ is the hadamard product module. g^{na} is the new adaptive attention model. A single-layer LSTM is set as the POS tag decoder and a two-layer stacked LSTM is set as the language model. The BOS and EOS denote the begin-of-sentence and end-of-sentence, respectively.

of the POS tag in the linguistic computer vision. However, they ignore the important property of POS tag, which relates to different visual semantics with different types of POS tag.

Attention-based methods have been widely used in visual captioning models. Yao et al. (Yao et al., 2015) considered the temporal structure of video and proposed a temporal attention mechanism to generate descriptions. Lu et al. (Lu et al., 2016) proposed an adaptive attention model in image captioning which can decide either to look at the image or to rely on the context of sentence to generate the next word.

Because of the highly nonlinearity and unclear working mechanism of neural networks, the operational processes of neural networks are always treated as black-box processes. For the video description task, some researchers (Dong et al., 2017; Wu et al., 2018) attempted to improve the interpretability of video description models. Dong et al. (Dong et al., 2017) interpreted the learned features of each neuron by a wide range of visual concepts in the video description task. Wu et al. (Wu et al., 2018) considered both the motion information and the sentence semantic structure with an attentive structured localization mechanism to enhance the captioning model's interpretability.

In this paper, we find the POS tag can be employed as a supervision to process irrelevant or relevant semantic information in video description task. Under the supervision of POS tag, an Adaptive Semantic Guidance Network (ASGN) is proposed to guide different POS-aware semantic information of video into corresponding network streams. Moreover, the video features related to noun and verb are used to get a sentinel gate which can decide how much the attended feature can be imported to the decoding process.

3 Our Method

In this section, we describe our ASGN in detail. First, the POS-aware semantic guider with the POS tag learning model is introduced. Next, an adaptive attention model which constructs a new sentinel gate will be presented. Finally, the description generator and its learning methods are introduced.

3.1 POS-Aware Semantic Guider

The key of the proposed POS-aware semantic guider is the POS tag, which is used to guide different semantic information of video into corresponding network streams. Supposing that we have a video described by a textual sentence $\mathcal{S}$, which consists of T words. The POS tag set of each word in $\mathcal{S}$ is defined as $\mathcal{P}$. To achieve this, a POS tag learning model is designed to predict the POS tags of the words in description.

We refer to (Al-Rfou et al., 2013) and annotate the POS tags of captions in training set by the polyglot toolkit[1]. Polyglot toolkit is a natural language pipeline that supports massive multilingual applications, including the POS tag identification. According to the setting of the polyglot toolkit, there are seventeen categories of POS tags: noun (NOUN), verb (VERB), adjective (ADJ),

[1] http://polyglot.readthedocs.org

adposition (ADP), adverb (ADV), auxiliary verb (AUX), coordinating conjunction (CONJ), determiner (DET), interjection (INTJ), numeral (NUM), particle (PART), pronoun (PRON), proper noun (PROPN), punctuation (PUNCT), subordinating conjunction (SCONJ), symbol (SYM) and other unknown types (X). Different POS tags have different descriptions or embellishments. Following the universal set of this toolkit, the number of POS tag categories is defined as $N_s = 17$.

Based on these POS tag categories, an encoder-decoder framework is proposed to predict the specific POS tag of each word in sentences, which can be seen in Fig. 2. Specifically, the video feature $\mathbf{V}$ is concatenated from the extractions of the 2D CNN and 3D CNN. Then, $\mathbf{V}$ is flowed into N_s semantic guiding CNN modules. Each CNN module relates to a corresponding POS tag category. The outputs of these CNN modules are N_s video features $\mathbf{v}_i$, where $i = 1, \ldots, N_s$. The mean-pooled feature $\bar{\mathbf{v}} = \frac{1}{N_s}\sum_{i=1}\mathbf{v}_i$ and the sentence $\mathcal{S} = \{\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_T\}$ are taken as the inputs to the POS tag decoder LSTM^p, where $\boldsymbol{s}_0$ is defined as the begin-of-sentence (BOS). All the POS tags of words in caption are sequentially generated by LSTM^p. The hidden state $\boldsymbol{h}_t^p$ is updated at time step $t \in \{0, \ldots, T\}$ through:

$$\boldsymbol{h}_t^p = \text{LSTM}^p(\mathbf{W}_e^p \boldsymbol{s}_t, \bar{\mathbf{v}}, \boldsymbol{h}_{t-1}^p), \tag{1}$$

where $\mathbf{W}_e^p \in \mathbb{R}^{N_h \times V}$ is the word embedding matrix, N_h is the length of hidden state and V denotes the vocabulary size of the corresponding dataset's text library.

Given the ground-truth POS tags which are annotated to the corresponding sentence. Therefore, the associated POS tags of the k-th video $\mathbf{V}$ are $\mathcal{P}^k = \{p_0^k, \ldots, p_T^k\}$. We define the POS tag learning loss as:

$$L_p = -\frac{1}{N}\sum_{k=1}^{N}\sum_{t=0}^{T}\log(p(p_t^k | p_{0:t-1}^k, \bar{\mathbf{v}}^k)), \tag{2}$$

where N is the number of training examples. In the encoder-decoder framework, the POS tag probability vector $\boldsymbol{p}_t$ is predicted at time step t. The mapping function is $\boldsymbol{p}_t = f(\boldsymbol{h}_t^p) = \text{Softmax}(\text{FC}(\boldsymbol{h}_t^p))$, where $f : \boldsymbol{h}_t^p \rightarrow \boldsymbol{p}_t$ is a Softmax mixed function and FC is a full connection layer. The predicted POS tag has the maximum probability in $\boldsymbol{p}_t$. Similar to visual captioning, the learned POS-aware semantic guider applies each predicted word to predict the POS tag of the next word in testing.

The learned $\boldsymbol{p}_t$ is used to guide the flowing of semantics from the N_s CNN modules at time step t . The i-th CNN module is related to the i-th POS tag category. The outputs of the CNN modules are concatenated as $\mathbf{V}^r = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{N_s}]$, where $\mathbf{v}_i$ is generated by the specific neuron of the i-th CNN module. The POS tag vector $\boldsymbol{p}_t$ is applied to activate the specific CNN module at the t-th step. The operation is implemented by a hadamard product module on the channel level:

$$\mathbf{V}_t^r = \mathbf{V}^r \odot \boldsymbol{p}_t, \tag{3}$$

where $\mathbf{V}_t^r$ represents the video representation at time step t. Then the video feature $\mathbf{V}_t^r$ which contains relevant visual semantic information will be inputted into the language model at time step t.

Generally, the learned POS tag representations can be used to activate the specific neurons of the CNN encoder and parse the whole video representation to guide semantic information into corresponding feature representations at each time step. Compared with the normal CNN+LSTM model, our POS-aware semantic guider constructs a correlation between different types of POS-aware semantic and the corresponding CNN modules.

3.2 Adaptive Attention Model

Although, Lu et al. (Lu et al., 2016) proposed a sentinel to decide whether the predicted words are visual words or not. Their model is learning from the gradient of back propagation, which is still an ambiguous process. The video features related to different POS tags have different properties. For example, the video features related to noun and verb always have sufficient visual signals. According to the properties of POS tags, we propose a more credible adaptive spatial attention model to predict next word. Specifically, the related video features of noun and verb are applied as the supervision to get a sentinel gate. Through the sentinel gate, when generating the next word, we can distinguish its visual word or not and decide how much the attended feature can be imported into the decoder LSTM.

First, the video feature $\mathbf{V}_t^r$ is reshaped to $\mathbf{V}_t^r = [\boldsymbol{v}_{t1}^r, \boldsymbol{v}_{t2}^r, \ldots, \boldsymbol{v}_{tm}^r]$, where m is the value of the width times height of $\mathbf{V}_t^r$. Normal attention model is defined as $g^a(\mathbf{V}_t^r, \boldsymbol{h}_{t-1}^d)$. $\boldsymbol{h}_{t-1}^d$ is generated by the description decoder LSTM^d which is shown in

Fig. 2. Formally, for the t-th time step, the attention part of the model g^a is defined as follows:

$$\begin{aligned} \boldsymbol{z}_t^r &= \tanh((\mathbf{W}^r \mathbf{V}_t^r + \boldsymbol{b}^r) \oplus \mathbf{W}^{hr} \boldsymbol{h}_{t-1}^d), \\ \boldsymbol{\alpha}_t^r &= softmax(\mathbf{W}^{\alpha r} \boldsymbol{z}_t^r + \boldsymbol{b}^{\alpha r}), \end{aligned} \tag{4}$$

where $\mathbf{W}^r \in \mathbb{R}^{l \times (N_c * N_s)}$, $\mathbf{W}^{hr} \in \mathbb{R}^{l \times N_h}$, $\mathbf{W}^{\alpha r} \in \mathbb{R}^{l \times 1}$ are the transformation matrices that map the CNN feature and hidden state to the same size; N_c is the channel size of $\mathbf{v}_i$; $\boldsymbol{b}^r \in \mathbb{R}^l$ and $\boldsymbol{b}^{\alpha r} \in \mathbb{R}^1$ are the model biases. $\boldsymbol{\alpha}_t^r$ is the attention weight related to $\mathbf{V}_t^r$.

Second, to improve the structure of g^a, our model learns to use the related video features of noun and verb as a supervision to distinguish whether the words in a sentence are visual words or not. All of the N_s video features are extracted after the ReLU operation, and thus the values in these feature maps are not less than zero. It indicates that the value of the feature map can reflect the response activation of the corresponding CNN module. We assign the related feature maps of noun and verb as the reference and obtain a value $R = \text{mean}[\frac{1}{2m} \sum_{j=1}^{m} (\mathbf{v}_n + \mathbf{v}_v)]$, where $\mathbf{v}_n$ and $\mathbf{v}_v$ are the related features of noun and verb from $\mathbf{V}^r$. Depending on the attended visual feature, a formulae is introduced to ascertain whether the current t word is a visual word or not:

$$c_t^v = \sum_{i=1}^{N_s} \{p_t^i \quad \text{if mean}(\sum_{j=1}^{m} \alpha_{tj}^r \boldsymbol{v}_j^{ri}) \geq R\}, \tag{5}$$

where c_t^v is the weight of visual word at time step t, p_t^i is i-th value in the t-th POS tag representation $\boldsymbol{p}_t$. Therefore, the weight of non-visual word is $c_t^{nv} = 1 - c_t^v$. In our method, the sentinel gate is defined as:

$$\beta_t = \frac{1}{1 + \exp(-log(\frac{c_t^v}{c_t^{nv}}))}. \tag{6}$$

The design of the sentinel gate can avoid the gate value of β_t being too small or too large. Based on β_t, the new adaptive attention model g^{na} can decide how much the attended feature can be imported into the decoder LSTM as follows:

$$\boldsymbol{a}_t^r = \beta_t g^a = \beta_t \sum_{j=1}^{m} \alpha_{tj}^r \boldsymbol{v}_{tj}^r. \tag{7}$$

3.3 Description Generator

In the description generation stage, we adopt a stacked two-layer LSTM to generate captions, namely LSTM^w and LSTM^d. Following (Donahue et al., 2015), the first LSTM layer LSTM^w is applied to encode the inputted sentence to enhance the textual context information of each word vector $\boldsymbol{h}_t^w$.

In the decoding stage, the encoded word vector $\boldsymbol{h}_t^w$ and the processed video feature $\boldsymbol{a}_t^r$ are taken as the inputs at t time step. The updating procedure from 0 to T of LSTM^d is written below:

$$\boldsymbol{h}_t^d = \text{LSTM}^d(\boldsymbol{h}_t^w, \boldsymbol{a}_t^r, \boldsymbol{h}_{t-1}^d), \tag{8}$$

where $\boldsymbol{h}_t^d$ is the current output of LSTM^d at time step t, the $\boldsymbol{h}_{-1}^d$ can be set as a null vector.

3.3.1 Description Generator Learning

Maximum Likelihood Estimation: Given the k-th video $\mathbf{V}^k$ and the associated sentence $\mathcal{W}^k = \{w_0^k, \ldots, w_T^k\}$, the generator loss can be formulated as follows with the optimization of maximum likelihood estimation (MLE):

$$L_s = -\frac{\lambda}{N} \sum_{k=1}^{N} \sum_{t=0}^{T} \log(p(w_t^k | w_{0:t-1}^k, \mathbf{V}^k)), \tag{9}$$

where $p(w_t^k | w_{0:t-1}^k, \mathbf{V}^k)$ is obtained from a Softmax mixed function; λ is the hyper-parameter.

Policy Gradient Optimization: For a fair comparison with recent works, the policy gradient (PG) technique is adopted as the optimizer to training our model. The objective in learning is to minimize the negative expected reward of the complete sampled sentence $\mathcal{W}^s = \{w_0^s, \ldots, w_T^s\}$:

$$L_\theta = E_{\mathcal{W}^s \sim p_\theta}[r(\mathcal{W}^s)], \tag{10}$$

where $r(\mathcal{W}^s)$ is calculated by comparing sampled caption with the reference caption in the specified evaluation metric. Following the implementation in (Rennie et al., 2017), we apply a single Monte-Carlo sample to calculate the relative reward $\Delta r(\mathcal{W}^s)$, which is computed by a baseline reward b. b is obtained by performing greedy decoding:

$$\begin{aligned} b = r(\hat{\mathcal{W}}),\ \hat{\mathcal{W}} = \ arg\ max\ p(w_t | \boldsymbol{h}_t^d), \\ \nabla_\theta L_\theta \approx -(r(\mathcal{W}^s) - r(\hat{\mathcal{W}})) \nabla_\theta \log[p_\theta(\mathcal{W}^s)]. \end{aligned} \tag{11}$$

Table 1: Ablation studies of our proposed video captioning model on Youtube2Text and MSR-VTT datasets. The (RL) indicates this model is optimized by the reinforcement learning. The best results are in bold.

Model	Youtube2Text				MSR-VTT			
	B-4	M	R	C	B-4	M	R	C
ASGN	0.494	0.325	0.690	0.733	0.271	0.266	0.587	0.439
ASGN+L	0.501	0.329	0.699	0.762	0.391	0.268	0.595	0.449
ASGN+LA	0.514	0.331	0.696	0.758	0.384	0.269	0.602	0.448
ASGN+LNA	0.518	0.333	0.700	0.776	0.395	0.274	0.609	0.465
ASGN+LNA (RL)	**0.521**	**0.333**	**0.703**	**0.803**	**0.405**	**0.278**	**0.615**	**0.490**

4 Experiments

4.1 Dataset and Evaluation

We report the results of our method on the Youtube2Text (Guadarrama et al., 2013) and MSR-VTT (Xu et al., 2016) datasets. The Youtube2Text dataset contains 1970 YouTube video clips. According to the publicly provided splits (Venugopalan et al., 2015b), 1200 videos are used for training, 100 videos for validation and the rest are used for testing. MSR-VTT is the largest public dataset for video captioning up to now. We follow the public splits (Venugopalan et al., 2015a) and divide them into 6,513, 497 and 2,990 samples for training, validation and testing, respectively. We reserve the words that appear in the training set and yield two vocabularies which contain 12,182 and 16,630 words for Youtube2Text and MSR-VTT datasets, respectively.

In evaluation, we report the following metrics: B-N (N=2,3,4) (Papineni et al., 2002), Meteor (Banerjee and Lavie, 2005), Rouge-L (Lin, 2004) and CIDEr (Vedantam et al., 2015). All the metrics are computed by the MS-COCO caption evaluation tool[2].

4.2 Training Details

CNN Encoder: For the video representations, we use a 2D CNN and a 3D CNN as the CNN encoder collectively. The 3D CNN can operate all video frames as a whole, which ensures the extracted visual features contain all the semantic information. The 2D CNN has more efficient learning and representation capacity. The details of these two CNNs can be seen as follows:

- **2D CNN** We use ResNet-152 (He et al., 2016) as the 2D CNN model. The feature map is taken from the *res5c* layer (2,048-dim).
- **3D CNN** The C3D (Tran et al., 2015) is applied as the 3D CNN model. The feature map is extracted from the *conv5b* layer.

The equally-spaced 16 and 32 frames are sampled from one video clip for Youtube2Text and MSR-VTT, respectively. We perform a mean operation among all the 2D CNN features. The representation $\mathbf{V}$ of each video is composed by a concatenation of the 3D CNN feature and the 2D CNN feature. Then, the feature map $\mathbf{V}$ is processed by N_s semantic separating CNN modules. A residual block is adopted as the CNN module in our method. The hidden state dimension of the LSTM units is 1,024.

The Adam optimizer is adopted in training. We first train the POS tag learning model in the POS-aware semantic guider. In the later training, the parameters of the POS tag learning model are fixed. The other parameters of our model is learned with MLE. λ is set to be 1. The maximum number of epochs of the MLE training is 30. The RL method is applied to optimize the MLE trained model with the CIDEr metric. At each epoch, the validation set is used to evaluate the training model, and the best CIDEr score model is selected for the final testing. All of our experiments are implemented with Pytorch (Paszke et al., 2017) and are conducted on a Titan X GPU with 12G memory.

In caption testing, the beam search is adopted for caption generation. The search size is set to be 5 in experiments.

4.3 Compared Approaches

Our method is compared with **HRNE** (Pan et al., 2016), **VideoLAB** (Ramanishka et al., 2016), **SA-LSTM** (Wang et al., 2018a) and attention-based methods like **MA** (Hori et al., 2017), **SCN** (Gan et al., 2016) and recently proposed **PickNet** (Chen et al., 2018) and state-of-the-art RL optimized methods **Weakly** (Shen et al., 2017), **CIDEnt-RL**

[2] https://github.com/tylin/coco-caption

Table 2: Performance compared with state-of-the-art methods on Youtube2Text dataset. The (−) is an unknown metric.

Model	B-4	M	R	C
LSTM-I	0.446	0.297	–	–
HRNE	0.438	0.331	–	–
MA	0.504	0.318	–	0.699
SCN	0.511	0.335	–	0.777
TSA	0.517	0.340	–	0.749
SA-LSTM	0.523	0.341	0.698	0.803
PickNet	0.523	0.333	0.696	0.765
ASGN+LNA	**0.547**	**0.342**	**0.717**	**0.813**

Table 3: Performance compared with state-of-the-art methods on MSR-VTT dataset.

Model	B-4	M	R	C
VideoLAB	0.391	0.277	0.606	0.441
SA-LSTM	0.391	0.266	0.593	0.427
Weakly	0.414	0.283	0.611	0.489
CIDEnt-RL	0.405	0.284	0.614	**0.517**
HRL	0.413	**0.287**	0.617	0.480
PickNet	0.413	0.277	0.598	0.441
ASGN+LNA	**0.420**	0.282	**0.621**	0.505

(Pasunuru and Bansal, 2017), **HRL** (Wang et al., 2017). Moreover, we compare our method with interpretable improvement methods **LSTM-I** (Dong et al., 2017) and **TSA** (Wu et al., 2018).

4.4 Ablation Study

We perform the ablation studies on the Youtube2Text and MSR-VTT datasets for our video captioning model. The results are shown in Table 1. **ASGN** which imports video feature into different network streams without the POS tag guidance is adopted as the baseline model. **ASGN+L** predicts the POS tag of each word and applies the POS tag to guide the semantic separation. "L" denotes the POS-aware semantic guider. **ASGN+LA** adds an attention model proposed by Lu et al. (Lu et al., 2016). As a comparison, our proposed new adaptive attention model is introduced to **ASGN+LNA**.

It can be seen that the ASGN+LNA achieves the best performances in all metrics, which indicates our proposed sentinel gate is more effectiveness and reasonable to decide the quantity of attended feature to the decoder LSTM. Compared with the baseline model, the introduction of POS tag in ASGN+L, which activates the specific neurons and parses the whole visual representation of video, can assign appropriate POS-aware semantic information and achieve better performance. Comparing with the results of MLE-based and RL-based methods, the RL method can improve the performance of MLE-based model by significant margins across all metrics.

Table 4: Human evaluation between ASGN and ASGN+L models.

Indifference	ASGN+L Wins	ASGN Wins
0.418	0.329	0.253

Table 5: The results under different supervisions of POS tags on Youtube2Text dataset.

Model	B-4	M	R	C
A+A	0.489	0.325	0.695	0.756
V	0.506	0.330	0.699	0.762
N	0.509	0.330	0.701	0.773
N+V	0.518	0.333	0.700	0.776

4.5 Quantitative Analysis

In Table 2 and Table 3, we compare our ASGN+LNA model with the state-of-the-art models on the Youtube2Text and MSR-VTT datasets. Following the operation of (Gan et al., 2016; Pasunuru and Bansal, 2017), ASGN+LNA is the average ensemble of 5 ASGN+LNA (RL) models trained with different initializations. From the results, our method achieves the competitive performance on the two datasets. Compared with the other interpretable improvement methods (Dong et al., 2017; Wu et al., 2018), interpretability of our neural network is explicitly improved, and the performance of our model is more competitive.

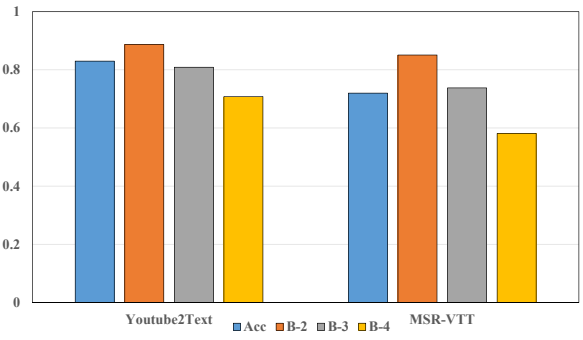

Figure 3: The performance of the POS tag prediction.

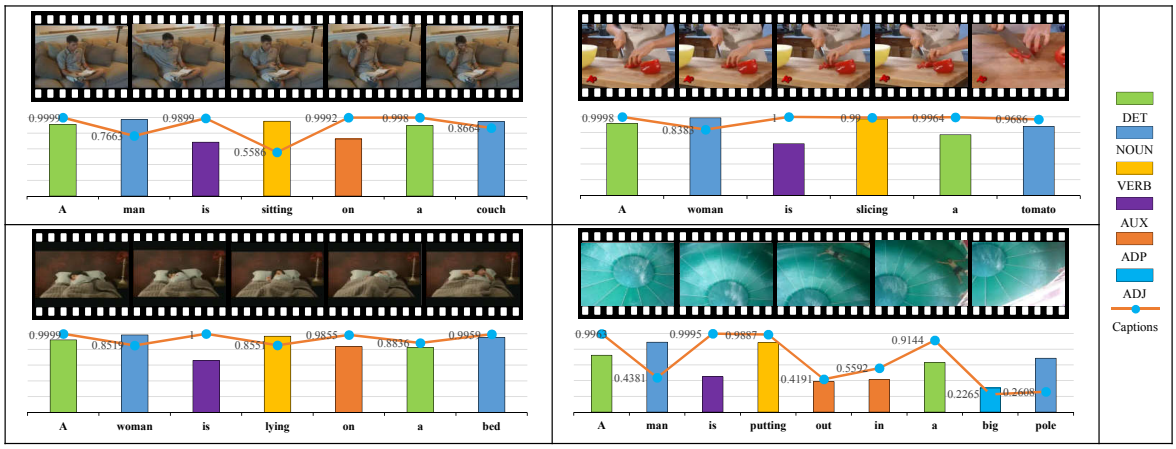

Figure 4: Visualizations of the probabilities of each word and corresponding POS tag in caption.

We introduce the human evaluation from (Pasunuru and Bansal, 2017) for comparison between ASGN and ASGN+L models. The relevance measures how related is the generated caption w.r.t, to the video content is adopted as the metric. In Table 4, the results of 150 samples from the Youtube2Text test set are studied. It can be found that the proposal of the semantic guider significantly improves the semantic extraction ability of network.

Moreover, to better verify the reliability of the supervision of nouns and verbs, we add comparisons by adjusting the supervision to single nouns (N), single verbs (V), adjectives and adverbs (A+A), nouns and verbs (N+V), respectively. The results on the Youtube2Text are presented in Table 5. It can be found that the model under the supervision of V+N achieves the best performance. Compared to verbs, the model under the supervision of nouns is more reliable. Under the supervision of A+A, the results of the model indicate the words of adjectives or adverbs are not always related to visual signals.

4.6 Visualized Analysis

To examine the reliability of the POS tag prediction, the performance of the POS tag learning model is measured using accuracy (Acc) and B-N (N=2,3,4) over the test datasets of Youtube2Text and MSR-VTT. The results are shown in Fig. 3. The metric of Acc is to test the total predict performance and the metric of B-N is to test the continuity of prediction. To better illustrate the effectiveness of the model, we set the beam search size to be 1 which is the same with the training process. These results indicate that the POS tag learning model can provide reliable POS tag representations to guide the semantics' separation to corresponding network streams.

Fig. 4 gives some results of generated captions and the corresponding probability of words and POS tags. In Fig. 4, we can see that the POS-aware semantic guider successfully guides the POS-aware semantic to the generation of captions. The POS-aware neurons have higher probability are activated to extract corresponding semantics to predict the relevant words at each time step. In the first example, the POS tags of the generated words with the highest probability have high probability as well. It demonstrates the improvement of interpretability of our model.

To further present the interpretability of our model, the neuron activations associated with the POS tags, the weights of the sentinel gate, the generated video captions, and the real POS tags of the captions are visualized in Fig. 5. Each element of these samples is illustrated along with the word prediction in sequence. Through these examples, the process of the POS-aware semantic information extraction and guidance in corresponding network stream can be visualized explicitly. From the illustration between the activated neuron and the truth POS tags of the captions. It can be found that the corresponding POS-aware neurons have high activations through time. For example, in Fig. 5 (b), the "DET", "NOUN" and "VERB" are precisely pointed to the POS tag of the words in "a man is flying into the water". It illustrates that the corresponding neurons of the POS tags are successfully activated at each time step. Moreover, the illustrations between the weights of the sentinel gate and the generated captions reveal that our adaptive attention model can effectively capture both the visual words and non-visual words.

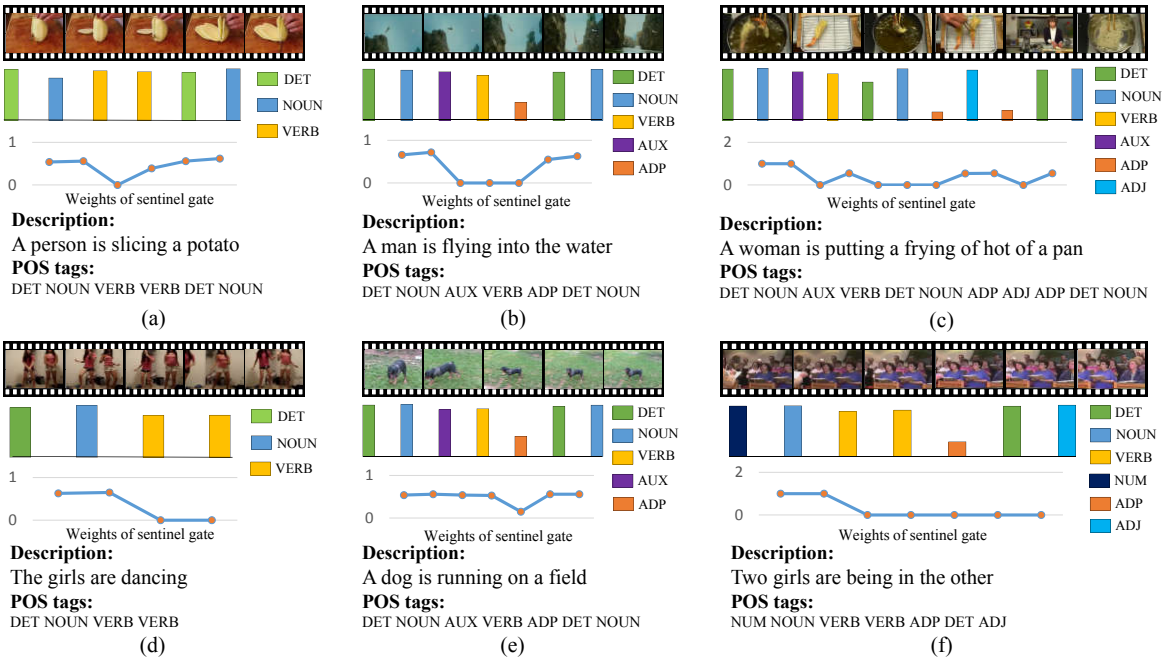

Figure 5: Visualization of the sampled video frames, the neuron activations associated with the POS tags, the weights of the sentinel gate, the generated captions, and the real POS tags of the captions.

From Fig. 5 (a), the words of "person", "slicing" and "potato" are obviously visual words which have related visual signals in video, and our model successfully extracts their visual information.

5 Conclusion

In this paper, we have proposed an Adaptive Semantic Guidance Network (ASGN), which extracts as well as guides the POS-aware semantic information into the corresponding encoded visual representations under the supervision of POS tag. Moreover, a new sentinel gate is introduced to determine how much the attended feature can be imported into the decoding process. It indicates that interpretable improvement not only makes the learning process more transparent, but also gives model more space to explore. The promising performance and interpretability improved merits of our method demonstrate the effectiveness of the POS tag. Furthermore, the proposed ASGN has a good flexibility which can be employed to the other language and vision fields, such as image captioning, visual question answering, and so on.

6 Acknowledgment

We thank the anonymous reviewers for their helpful comments. This work was supported by the National Natural Science Foundation of China under Grants 91646207, 61773377, and 61573352, the Young Elite Scientists Sponsorship Program by CAST (2018QNRC001), and the Beijing Natural Science Foundation under Grant L172053.

References

Deshpande A, Aneja J, and Wang L. 2019. Diverse and accurate image captioning guided by part-of-speech. In *CVPR*, pages 10695–10704.

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual NLP. In *CoNLL*, pages 183–192.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL Workshop*, volume 29, pages 65–72.

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

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

Yinpeng Dong, Hang Su, Jun Zhu, and Bo Zhang. 2017. Improving interpretability of deep neural networks with semantic information. In *CVPR*, pages 975–983.

Raphaël Fargier and Marina Laganaro. 2015. Neural dynamics of object noun, action verb and action noun production in picture naming. *Brain and language*, 150:129–142.

Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. 2016. Semantic compositional networks for visual captioning. *Computing Research Repository (CoRR)*, abs/1611.08002.

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

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

Xinwei He, Baoguang Shi, Xiang Bai, Gui-Song Xia, Zhaoxiang Zhang, and Weisheng Dong. 2017. Image caption generation with part of speech guidance. *Pattern Recognition Letters*.

Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John Hershey, Tim Marks, and Kazuhiko Sumi. 2017. Attention-based multimodal fusion for video description. In *ICCV*, pages 4203–4212.

Houpand Horoufchin, Danilo Bzdok, Giovanni Buccino, Anna Borghi, and Ferdinand Binkofski. 2018. Action and object words are differentially anchored in the sensory motor system - a perspective on cognitive embodiment. 8.

Chin Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL Workshop*, volume 8.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2016. Knowing when to look: Adaptive attention via A visual sentinel for image captioning. *Computing Research Repository (CoRR)*, abs/1612.01887.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *Computing Research Repository (CoRR)*, abs/1609.07843.

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

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

Ramakanth Pasunuru and Mohit Bansal. 2017. Reinforced video captioning with entailment rewards. In *EMNLP*, pages 979–985.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. In *NeruIPS Workshop*.

Vasili Ramanishka, Abir Das, Dong Park, Subhashini Venugopalan, Lisa Hendricks, Marcus Rohrbach, and Kate Saenko. 2016. Multimodal video description. In *the ACM International Conference on Multimedia*, pages 1092–1096.

Steven Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *CVPR*, pages 1179–1195.

Zhiqiang Shen, Jianguo Li, Zhou Su, Minjun Li, Yurong Chen, Yu-Gang Jiang, and Xiangyang Xue. 2017. Weakly supervised dense video captioning. In *CVPR*, pages 5159–5167.

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

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

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

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

Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. 2018a. Reconstruction network for video captioning. *Computing Research Repository (CoRR)*, abs/1803.11438.

Xin Wang, Wenhu Chen, Jiawei Wu, Yuan-Fang Wang, and William Yang Wang. 2017. Video captioning via hierarchical reinforcement learning. *Computing Research Repository (CoRR)*, abs/1711.11135.

Zhe Wang, Xiaoyi Liu, Limin Wang, Yu Qiao, Xiaohui Xie, and Charless Fowlkes. 2018b. Structured triplet learning with pos-tag guided attention for visual question answering. In *WACV*, pages 1888–1896.

Xian Wu, Guanbin Li, Qingxing Cao, Qingge Ji, and Liang Lin. 2018. Interpretable video captioning via trajectory structured localization. In *CVPR*, pages 6829–6837.

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

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