Video Captioning with Guidance of Multimodal Latent Topics

Shizhe Chen[*]
Renmin University of China
cszhe1@ruc.edu.cn

Jia Chen[*]
Carnegie Mellon University
jiac@cs.cmu.edu

Qin Jin[†]
Renmin University of China
qjin@ruc.edu.cn

Alexander Hauptmann
Carnegie Mellon University
alex@cs.cmu.edu

Abstract

The topic diversity of open-domain videos leads to various vocabularies and linguistic expressions in describing video contents, and therefore, makes the video captioning task even more challenging. In this paper, we propose an unified caption framework, M&M TGM, which mines multimodal topics in unsupervised fashion from data and guides the caption decoder with these topics. Compared to pre-defined topics, the mined multimodal topics are more semantically and visually coherent and can reflect the topic distribution of videos better. We formulate the topic-aware caption generation as a multi-task learning problem, in which we add a parallel task, topic prediction, in addition to the caption task. For the topic prediction task, we use the mined topics as the teacher to train a student topic prediction model, which learns to predict the latent topics from multimodal contents of videos. The topic prediction provides intermediate supervision to the learning process. As for the caption task, we propose a novel topic-aware decoder to generate more accurate and detailed video descriptions with the guidance from latent topics. The entire learning procedure is end-to-end and it optimizes both tasks simultaneously. The results from extensive experiments conducted on the MSR-VTT and Youtube2Text datasets demonstrate the effectiveness of our proposed model. M&M TGM not only outperforms prior state-of-the-art methods on multiple evaluation metrics and on both benchmark datasets, but also achieves better generalization ability.

Keywords

Video Captioning; Multimodal; Latent Topics; Multi-task

1 Introduction

Videos have become increasingly popular on the Internet, for example, hundreds of hours of video contents are uploaded on YouTube every minute. It is impossible to watch these overwhelming amounts of videos, therefore, automatic techniques to search and analyze video contents are highly desired. Generating natural language descriptions for video contents (a.k.a. video captioning) is one of such important techniques for this challenge. It can benefit a wide range of applications such as assisting the visually impaired people and improving the quality of online video retrieval.

Although there have been significant breakthroughs recently in image captioning [9, 30, 41, 43, 44, 46], video captioning remains very challenging due to the diversity and complexity of video contents. The open-domain videos contain a broad range of topics, such as sports, music, food and so on, which results in very different vocabularies and expression styles to describe video contents across topics. For example, sports terms frequently occur in the sports topic while the typical words in the food topic are about cooking and ingredients. The key contents in descriptions for the sports topic are often the action verbs, while they are object nouns for the food topic and descriptive adjectives for the people topic. Thus, being aware of the topic information can dramatically narrow down general sentence distributions and enable the caption model to focus on the discriminative video contents under the topic, such as the detailed actions in the sports topic.

However, most of the existing caption models ignore the topic information and mainly try to maximize the overall likelihood for videos in all topics, which have a tendency to seek the most common mode in training sentences as indicated by Lee et al. [20]. Such models not only are prone to generate plain descriptions without details, but also unable to distinguish confusing concepts in a topic. Although a few works [8, 13, 33] have exploited the topic information for video description generation, there are still three main challenges in employing the topic information in video captioning. First, there are no direct topic representations in common video caption datasets, so how can we construct and predict the latent topics of videos? Second, what is an effective and efficient way to employ the latent topics to guide the caption model for generating topic-oriented descriptions? And third, how to train the topic-guided caption model to achieve optimal captioning performance?

In this paper, we bring three innovations to deal with the above challenges. First, we propose a multimodal topic mining approach to discover latent video topics from video-sentence pairs. The multimodal latent topics are more semantically coherent and visually consistent than the expert-defined or textual latent topics in previous work. We then use the teacher-student learning strategy to predict latent topics with multimodal video features. Second, in order to exploit the topic effectively, we propose a novel topic-aware language decoder, which implicitly functions as an ensemble of topic-specific decoders for each topic, but is computationally more efficient and requires less training data. The predicted video topic automatically modifies weights in the decoder, enabling to generate topic-oriented video descriptions. Third, to optimize the topic-guided caption performance, we develop a multi-task learning architecture to jointly train the caption system with the topic prediction loss and sentence generation loss in an end-to-end manner.

---

[*]contributed equally.
[†]corresponding author.

It can strengthen the coupling of the two parts to generate better latent topics and sentence descriptions simultaneously. The overall model is called **M&M TGM** (Topic-Guided Model with Multimodal latent topic guidance and Multi-task learning).

The main contributions of this paper are as follows:

- We present an unified framework M&M TGM for video captioning, which can automatically discover and predict the underlying video topics and exploit the topic guidance to generate better topic-oriented video descriptions. The topic prediction and topic-aware sentence generation can be trained jointly with multi-task learning end-to-end.

- The proposed M&M TGM requires no additional labelling for video topics and can exploit the limited video caption data effectively. For the latent topic generation, we propose unsupervised multimodal topic mining for topic discovery and teacher-student learning for topic prediction. For the topic-oriented sentence generation, the topic-aware decoder shares data and parameters among different topics which is efficient in both computation and data.

- Our proposed model is evaluated on two benchmark video caption datasets: MSR-VTT [42] and Youtube2Text [11]. Both quantitative and qualitative analysis show superior performances of using the multimodal latent topic guidance and multi-task learning strategy. The M&M TGM not only outperforms prior state-of-the-art methods, but also has better generalization ability.

The rest of the paper is organized as follows: Section 2 introduces the related work. Section 3 presents the problem formulation. The solution for M&M TGM is described in Section 4. Section 5 presents experimental results and analysis. Section 6 concludes the paper.

2 Related Work

Image Captioning has attracted growing interests recently. Early works are mainly based on rule systems [17, 19] and suffer from generating flexible and accurate descriptions. So more researches have been focusing on the encoder-decoder framework [2, 35] for caption generation. The deep convolutional neural networks (CNNs) [36] function as the encoder to transform image contents into dense vectors. Then the decoder, typically the Long-short Term Memory [12] (LSTM), is utilized to generate sequential words conditioned on image features [41]. Xu et al. [43] propose the spatial attention mechanism based on the basic encoder-decoder. You et al. [46] and Gan et al. [9] propose the semantic attention model and semantic compositional networks (SCN) respectively to exploit detected concepts from the image. Our topic-aware language decoder is inspired by SCN but with different aims of producing topic-oriented video descriptions, which is more suitable for video captioning task as shown in Section 5.4.

Video Captioning is more challenging compared with image captioning due to the temporal dynamics, multiple modalities, and more diverse contents in videos. Yao et al. [45] propose the temporal attention mechanism, and Pan et al. [25] utilize the hierarchical LSTM encoder to explore the temporal structures. Most previous works only focus on the visual modality [26, 40], but recently Jin et al. [13, 14] and Ramanishka et al. [29] have shown improvement of multimodal fusion for video captioning. To promote the video captioning research for open-domain videos, several large-scale

videos with various topic are collected such as MSR-VTT [42] and TGIF [22]. Jin et al. [13] encode expert-defined topics with multimodal features, which results in their winning of the MSR-VTT challenge [31]. Dong et al. [8] utilize the textual mined topics to learn interpretable features. Shen et al. [33] train separate language decoders for each expert-defined topic. Chen et al. [4] explore the guidance from textual minded topics to generate topic-aware sentences. In our work, we address the topic diversity challenge in open-domain videos and propose the novel M&M TGM model to jointly generate the latent video topics and topic-oriented video descriptions with multimodal features in a multi-task framework.

Latent Topic Mining has been a long-standing problem. The latent Dirichlet allocation (LDA) proposed by Blei [3] is one of the most classic models to automatically inference the latent topics for textual documents. Doersch et al. [7] utilize a discriminative clustering approach to discover the visual topics. For multimodal data, Li et al. [21] apply the association rule mining algorithm on image-caption pairs to discover the multimodal topics. In our work, we take the multimodal topic mining perspective on video-sentence pairs by a weighted multimodal clustering method to obtain semantically coherent and visually consistent latent topics.

3 Problem Formulation

Our goal is to leverage the underlying video topic information to make the model more proficient in using vocabularies and expressions within the topic when describing a video.

Suppose we have a video $\mathcal{V}$ along with a set of N_d textual descriptions $Y = \{y_1, y_2, ..., y_{N_d}\}$. Each $y \in Y$ is a sentence with N_s words, denoted as $y = \{w_1, w_2, ..., w_{N_s}\}$, where w is a word from vocabulary W. We encode the video content $\mathcal{V}$ into a fixed-dimensional video representation x. Conditioned on the multimodal video representation x, a caption decoder aims to generate the sentence y with probability

$$\Pr(y|x) = \prod_{t=1}^{N_s} \Pr(w_t|x, w_{<t}) \qquad (1)$$

where $w_{<t}$ denotes the sequential words before the time step t. We abbreviate this as *conditional sentence probability*. We use RNN as decoder and parameterize the probability $\Pr(w_t|x, w_{<t})$ by the recurrent units shared across time steps, which can be expressed as:

$$\begin{bmatrix} \Pr(w_t|x, w_{<t}; \Theta_d) \\ h_t \end{bmatrix} = \psi_d(h_{t-1}, w_{t-1}; \Theta_d) \qquad (2)$$

where ψ_d is the recurrent unit function, Θ_d are the parameters of the recurrent unit and h_t is the hidden state of the recurrent unit at time step t. We define w_0 as the start token and h_0 as the initialized hidden state. For notation simplicity, we abbreviate eq (2) to $\Pr(w_t|x, w_{<t}; \Theta_d) = \psi_d(h_{t-1}, w_{t-1}; \Theta_d)$. Substituting this abbreviation to eq (1), we get:

$$\Pr(y|x; \Theta_d) = \prod_{t=1}^{N_s} \psi_d(h_{t-1}, w_{t-1}; \Theta_d) \qquad (3)$$

As mentioned earlier, videos in different topics have quite different vocabularies and expression styles in their video descriptions, and different topics can coexist in one video. Based on this observation, we introduce the latent topic variables $\{z_i\}_{i=1,...,K}$, where K is the number of topics. Then the conditional sentence probability,

Pr(y|x) is actually the marginal probability of the joint sentence and topic distribution conditioned on video content:

$$\Pr(y|x) = \sum_{z_1, \ldots, z_K} \Pr(y, z_1, \ldots, z_K|x)$$

$$= \sum_{z_1, \ldots, z_K} \underbrace{\Pr(y|z_1, \ldots, z_K, x)}_{\substack{\text{topic-aware} \\ \text{sentence distribution}}} \underbrace{\Pr(z_1, \ldots, z_K|x)}_{\substack{\text{latent topic} \\ \text{distribution}}} \quad (4)$$

In the second step of eq (4), we factorize the joint distribution to latent topic distribution and topic-aware sentence distribution.

We come to parameterize the probability in eq (4). For latent topic distribution, we parameterize it by *topic predictor* ψ_z as follows:

$$\Pr(z_1, \ldots, z_K|x) = \psi_z(x; \Theta_z) \quad (5)$$

where Θ_z are parameters in the topic predictor. To parameterize the topic-aware sentence distribution by RNN, we need to introduce the topic-related parameter Θ'_z in addition to Θ_d in eq (3) as now the probability we are modelling is also conditioned on topics $z_1, \ldots, z_K$:

$$\Pr(y|z_1, \ldots, z_K, x; \Theta_d, \Theta'_z) = \prod_{t=1}^{N_s} \Pr(w_t|w_{<t}, x, z_1, \ldots, z_K; \Theta_d, \Theta'_z)$$

$$= \prod_{t=1}^{N_s} \psi_d(h_{t-1}, w_{t-1}, z_1, \ldots, z_K; \Theta_d, \Theta'_z) \quad (6)$$

Putting all these together, we get the parameterization for *topic-aware conditional sentence probability* as:

$$\Pr(y|x) = \Pr(y|x; \Theta_d, \Theta'_z, \Theta_z)$$

$$= \sum_{z_1, \ldots, z_K} \psi_z(x; \Theta_z) \prod_{t=1}^{N_s} \psi_d(h_{t-1}, w_{t-1}, z_1, \ldots, z_K; \Theta_d, \Theta'_z) \quad (7)$$

Contrasting with *conditional sentence probability* parameterization in eq (3), we see that we have additional parameters Θ_z for topic prediction and Θ_d, Θ'_z for topic-aware sentence modelling in the parameterization of *topic-aware conditional sentence probability*.

The standard loss function for caption generation task is to maximize the log probability of the *conditional sentence probability*:

$$L_{caption}(y, \tilde{y}) = -\sum_{t=1}^{N_s} \sum_{\tilde{w}_t \in W} \delta_{w_t, \tilde{w}_t} \log \Pr(\tilde{w}_t|x, w_{<t}; \Theta_d) \quad (8)$$

where $\delta_{w_t, \tilde{w}_t}$ is the indicator function. We could directly use this loss to train our topic-aware parameterized model in eq (4) as it is a special form of the general *conditional sentence probability*. However in training, it requires large amount of data for the model to discover latent topics from scratch and to learn topic-aware sentence distribution simultaneously. The training data of caption task is limited, and as the labelling cost of such training data is very high, the amount of the training data is not likely to grow very fast in the near future. To solve this problem, we introduce an auxiliary task, topic prediction, to guide the caption task. In this task, we leverage the existing topic mining approaches to generate a set of topics as the teacher topics z. The details of the topic mining approach are given in Section 4.2. We use the teacher topics to guide the latent topic learning by an additional topic loss function $L_{topic}(z, \tilde{z})$ on the topic predictor $\psi_t(x; \Theta_z)$, which penalizes the latent topic predictions that are far away from the teacher z. The details of *dist* function in $L_{topic}(z, \tilde{z})$ will be discussed in section 4.1. The two losses are combined by the trading off hyper-parameter
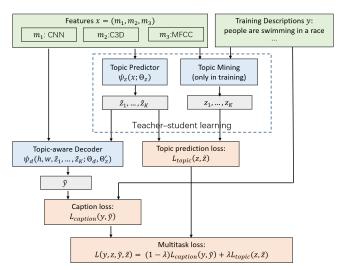

Figure 1: Framework of M&M TGM. The green boxes are inputs; blue boxes are modules; gray boxes are outputs; red boxes are loss functions. The topic mining module is only used in the training stage while other modules are used in both training and testing.

$\lambda \in [0, 1)$:

$$L(y, z, \tilde{y}, \tilde{z}) = (1 - \lambda)L_{caption}(y, \tilde{y}) + \lambda L_{topic}(z, \tilde{z}) \quad (9)$$

$$L_{caption}(y, \tilde{y}) = -\sum_{t=1}^{N_s} \sum_{\tilde{w}_t \in W} \delta_{w_t, \tilde{w}_t} \log \Pr(\tilde{w}_t|x, w_{<t}; \Theta_d, \Theta'_z, \Theta_z) \quad (10)$$

$$L_{topic}(z, \tilde{z}) = dist(z, \tilde{z}) \quad (11)$$

In this way, we weave the video latent topics into video description generation as guidance.

4 M&M Topic-Guided Model

In this section, we present our solutions for the topic-guided video captioning. Figure 1 illustrates the overall framework. Our proposed M&M TGM consists of three components: topic mining, topic predictor and topic-aware decoder, which is optimized by multi-task training. We will first present the overall multi-task training scheme and then introduce the three modules in details.

4.1 Multi-task Training Scheme

It is intractable to integrate the *topic-aware conditional sentence probability* in eq (7) over the latent topic space. But since the latent topic is highly dependent on the video content, usually the latent topic distributions of the video could be very skew with massive likelihood on the peak point. Therefore, it is reasonable that we use the maximum latent topic likelihood to approximate the summation over the latent topic space to gain efficiency without losing much accuracy. The approximation for eq (7) is as follows:

$$\Pr(y|x) = \sum_{z_1, \ldots, z_K} \Pr(y|z_1, \ldots, z_K, x)\Pr(z_1, \ldots, z_K|x) \quad (12)$$

$$\approx \Pr(y|z_1 = \tilde{z}_1, \ldots, z_K = \tilde{z}_K, x)\Pr(z_1 = \tilde{z}_1, \ldots, z_K = \tilde{z}_K|x)$$

where $\tilde{z}_1, \ldots, \tilde{z}_K = \arg\max_{z_1, \ldots, z_K} \Pr(z_1, \ldots, z_K|x)$. Our experiments show such approximation leads us to good results.

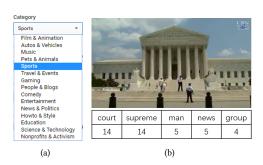| court	supreme	man	news	group
14	14	5	5	4

(a)   (b)

Figure 2: (a) Expert-defined topic schema on the YouTube website. (b) An example video in MSR-VTT dataset with its top frequent caption words.

The multi-task training scheme is conducted as follows:

1. We first pretrain the topic predictor using the topic loss L_{topic}, and use it to initialize Θ_z in the M&M TGM (as shown in eq (7)). For the *dist* in topic loss, we choose two widely used distances: l_2-distance $l_2(z, \tilde{z}) = \| z - \tilde{z} \|_2^2$ and KL-divergence $KL(z, \tilde{z}) = \sum_{k=1}^{K} \tilde{z}_k log(\tilde{z}_k/z_k)$.

2. We then pretrain the topic-aware decoder using the caption loss $L_{caption}$ (as shown in eq (10)) with the topic guidance from the fixed topic predictor in the above step, and use it as a good initialization for Θ'_z and Θ_d in the M&M TGM (as shown in eq (7)).

3. Finally, based on the parameters initialization in the above two steps, we use the multi-task loss $L(y, z, \tilde{y}, \tilde{z})$ (as shown in eq (9)) to further train both models jointly, which fine-tunes all the parameters Θ_z, Θ'_z and Θ_d to optimize the caption performance.

4.2 Topic Mining and Prediction

We first briefly present two topic structures adopted in previous works, and then propose our multimodal latent topic mining approach and the topic prediction model.

Expert-defined Topics: The online video websites often provide an expert-defined topic schema as shown in Figure 2(a). Video uploaders can select one of the topics to better organize their videos. Such topics can reflect the topic variety to some extent, but have the following drawbacks: 1) The user assigned labels are noisy with labelling mistakes; 2) The topic distributions are exclusive which ignores the topic diversity inside the video; and 3) There might exist different semantic meanings and visual appearances within a topic, which is suboptimal as latent topics.

Textual Latent Topics: The annotated descriptions in video caption datasets provide rich and accurate information about the video content, which can also reflect more appropriate latent topic distributions. Thus, our previous work [4] mined topics from the annotated video captions on the training set. The main idea is to cluster the video captions and each cluster can represent a latent topic. We use the kernel K-means algorithm [6] for clustering. We group all the groundtruth captions of a video as one data sample. Stopwords are removed and the bag-of-words are used as our text features. The cosine kernel is adopted to generate nonlinear cluster separations. Since a video might contain several topics, we utilize the soft assignment according to the distance between samples and clusters to generate topic distributions.

Multimodal Latent Topics: Although the textual latent topics may better fit with the videos' underlying topic distributions, the mined topics purely based on textual data still suffer from the following problems:

1) Polysemy phenomena: Words can convey several different meanings, which might lead to improper assignment of the video. As shown in Figure 2(b), the most frequent word *court* can either mean tribunal or sports field, so the video could mistakenly peak at some sports related topics.

2) Certain messy clusters: The unsupervised clustering method cannot perfectly generate meaningful latent topics, which could produce some semantically unclear topics with dissimilar visual appearances. This brings great harm for the topic prediction model to learn from those messy topics.

To address these issues, we further propose to combine the description texts and visual features to learn the multimodal latent topic representation. The textual features are preprocessed in the same way as in textual latent topics and the visual features are elaborated in Section 5.2. We use the weighted kernel K-means algorithm to fuse the textual and visual features with weights of 1 and 0.2 respectively, because we consider the textual features are more accurate than the visual features to unveil an eligible latent topic structure.

Topic Predictor: We take the teacher-student perspective [1] to train the topic predictor ψ_z. The distributions of the above mined latent topics serve as the teacher z to guide the ψ_z to generate latent topic predictions with the topic prediction loss L_{topic}. In this work, we adopt a two-layer perception as the topic predictor based on inputs of multimodal video features m_1, m_2, m_3:

$$\psi_z(x; \Theta_z) = softmax(W_2(\varphi(W_1[m_1; m_2; m_3] + b_1) + b_2) \quad (13)$$

where $W_i, b_i(i = 1, 2)$ are parameters, $[\cdot]$ denotes feature concatenation and φ is the RELU nonlinear function [10].

4.3 Topic-aware Decoder

In this section, we first briefly introduce the standard LSTM, and then describe our proposed topic-aware decoder based on the LSTM model to employ the topic guidance during sentence generation.

The LSTM model addresses the vanishing gradients problem in traditional RNN by employing a memory cell and three gates to control the information flow in the network. The formulas of the LSTM cell $\psi_d(h_t, w_t; \Theta_d)$ at timestep t are given below:

$$input\ gate : i_t = \sigma(W_{iw}w_t + U_{ih}h_{t-1}^{hidden} + b_i) \quad (14)$$

$$forget\ gate : f_t = \sigma(W_{fw}w_t + U_{fh}h_{t-1}^{hidden} + b_f) \quad (15)$$

$$output\ gate : o_t = \sigma(W_{ow}w_t + U_{oh}h_{t-1}^{hidden} + b_o) \quad (16)$$

$$cell\ input : g_t = \phi(W_{gw}w_t + U_{gh}h_{t-1}^{hidden} + b_g) \quad (17)$$

$$cell\ state : h_t^{cell} = i_t \odot g_t + f_t \odot h_{t-1}^{cell} \quad (18)$$

$$cell\ output : h_t^{hidden} = o_t \odot \phi(h_t^{cell}) \quad (19)$$

where h_t^{hidden} is the hidden state, h_t^{cell} is the cell state, σ is sigmoid function, ϕ is tanh function, $\odot$ denotes element-wise production, and $\Theta_d = \{W_{*w}, U_{*h}, b_*\}$ are parameters.

Recall that in Section 3, when we use LSTM cell to parameterize $Pr(y|z_1, \ldots, z_K, x)$, we need to change the LSTM cell function to $\psi_d(w_t, h_t, z_1, \ldots, z_K; \Theta_d, \Theta'_z)$ to employ the topic dependency.

Inspired by Gan et al. [9], we extend each weight matrix of the conventional LSTM to be an ensemble of a set of topic-dependent weight matrices.

Let us take one of the input gate weight matrices W_{iw} as an example, and transformation for other parameters in LSTM model are alike. We define the ensemble 3D weight matrix $W_{i\tau} \in \mathbb{R}^{n_h \times n_w \times K}$, where n_h is the number of hidden units and n_w is the dimension of input vectors. $W_{i\tau}[k]$ denotes the k-th slice of $W_{i\tau}$, which represents the 2D weight matrix belonging to the LSTM model for the k-th topic. Therefore, we explicitly specify K language decoders, which is, for each topic there is a pair of topic-specific LSTM weights. Given the latent topic $\tilde{z}$, we can define the mixture topic LSTM weight matrix $W_i(\tilde{z}) \in \mathbb{R}^{n_h \times n_w}$ as

$$W_i(\tilde{z}) = \sum_{k=1}^{K} \tilde{z}_k W_{i\tau}[k] \qquad (20)$$

where $\tilde{z}_k$ is the k-th topic in $\tilde{z}$. When $\tilde{z}$ is not the exclusive one-hot topic distribution, the video data is shared among different topics. However, the parameters still grow linearly with the number of topics K and no parameters are shared among different video topics which can easily result in over-fitting. So the 3-way factorization method [16, 24] is used to share parameters. We re-represent $W_{i\tau}$ in terms of three matrices $W_{ia} \in \mathbb{R}^{n_h \times n_f}$, $W_{ib} \in \mathbb{R}^{n_f \times K}$ and $W_{ic} \in \mathbb{R}^{n_f \times n_w}$, where n_f is the number of factors, such that

$$W_i(z) = W_{ia} \cdot diag(W_{ib}\tilde{z}) \cdot W_{ic} \qquad (21)$$

W_{ia} and W_{ic} are shared among all topics, while W_{ib} can be viewed as the latent vectors of topics.

Therefore, the transformation from input vectors to input gates is changed from $W_{iw} \cdot w_t$ to $W_{ia} \cdot (W_{ib}\tilde{z} \odot W_{ic}w_t)$. The topic distribution $\tilde{z}$ affects the LSTM parameters associated to the video when decoding, which implicitly works as an ensemble of K topic-aware language decoders.

5 Experiments

5.1 Datasets

To validate the effectiveness, robustness and generalization of our proposed methods, we conduct extensive experiments on two benchmark video captioning datasets: MSR-VTT [42] and Youtube2Text [11].

MSR-VTT: The MSR-VTT corpus [42] is currently the largest open-domain video captioning dataset with a wide variety of video topics. It consists of 10,000 video clips with 20 human annotated captions per clip. Each clip also contains a noisy expert-defined topic label crawled from YouTube, the distribution of which is shown in Figure 3. Following the standard data split in the MM2016 challenge [31], we use 6,513 videos for training, 497 videos for validation and the remained 2,990 for testing.

Youtube2Text: The Youtube2Text corpus [11] contains 1970 Youtube video clips with around 40 human annotated sentences per clip. For fair comparison, we crawl the expert-defined topics from YouTube for the whole dataset as shown in Figure 3, which are also noisy. We adopt the same data splits as provided in Yao et al. [45], with 1,200 videos for training, 100 videos for validation and 670 videos for testing.

Description Preprocessing: We convert all descriptions to lower case and remove all punctuations. We add begin-of-sentence
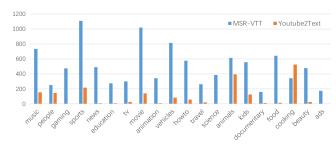

Figure 3: The distribution of noisy expert-defined topics in MSR-VTT and Youtube2Text datasets.

tag <BOS> and end-of-sentence tag <EOS> to our vocabulary. We preserve words that appear more than twice for MSR-VTT, resulting in a vocabulary size of 10,868, and words that appear more than once for Youtube2Text, resulting in a vocabulary size of 7,245.

5.2 Implementation Details

Multi-modality Features: We extract features from image, motion and aural modalities. For image features, we extract activations from the penultimate layer of the inception-resnet [36] pre-trained on the ImageNet, the dimensionality of which is 1,536. For motion features, we extract activations from the last 3D convolution layer of the C3D model [37] pre-trained on the Sports-1M dataset. We perform max-pooling on the spatial dimension (width and height), resulting in 512 dimensional features. For aural features, We extract the Mel-Frequency Cepstral Coefficients (MFCCs) [5] and use Bag-of-Audio-Words [27] and Fisher Vector [32] encoding methods to generate video-level features, with dimensionality of 1,024 and 624 respectively. We simply pad zeros for videos without the sound track.

Training Settings: We empirically set the hidden layer of the topic prediction model with 512 units. The dimension of the LSTM hidden size is set to be 512. The output weights to predict the words are the transpose of the input word embedding matrix. We apply dropout with rate of 0.5 on the input and output of LSTM and use ADAM algorithm [15] with learning rate of 10^{-4}. Beam search with beam width of 5 is used to generate sentences during testing process.

Evaluation Metrics: We evaluate the caption results comprehensively on all major metrics, including BLEU [28], METEOR [18], ROUGE-L [23] and CIDEr [38].

5.3 Evaluation of M&M TGM

In this subsection, we introduce two baselines to show the effectiveness of topic guidance and our multi-task learning for video captioning. The first baseline, *Vanilla*, consists of the multimodal encoder and the standard LSTM decoder. This baseline doesn't involve any topic information. The second baseline is *TGM*, a trimmed version of our M&M TGM. We discard the multi-task joint optimization step (last step) in section 4.1.

To evaluate the effectiveness of topic guidance, we compare the caption performance of the Vanilla and TGMs using different latent topic guidance in Table 1. The numbers of textual and multi-modal latent topics are optimized on the validation set (50 topics for MSR-VTT and 5 topics for Youtube2Text). We can see that TGMs outperform Vanilla model consistently on all four metrics

Table 1: Caption performance comparison between Vanilla (no topic guidance) and TGM (using different latent topic guidance) on the MSR-VTT and Youtube2Text datasets. The best results are marked in bold and the second best results with <u>underline</u>.

model	topic guidance	MSR-VTT				Youtube2Text			
		BLEU4	METEOR	ROUGE$_l$	CIDEr	BLEU4	METEOR	ROUGE$_l$	CIDEr
Vanilla	no topic	42.23	28.68	61.44	46.06	46.05	32.97	69.94	72.12
TGM	pred expert-defined topic	42.36	28.51	61.57	46.37	47.42	33.51	69.61	77.36
	pred textual latent topic	<u>43.38</u>	<u>29.08</u>	<u>62.11</u>	48.38	<u>47.45</u>	<u>34.11</u>	**70.59**	<u>79.45</u>
	pred multimodal latent topic	**43.81**	**29.26**	**62.13**	<u>48.50</u>	**47.56**	**34.21**	<u>70.45</u>	**79.57**
	assigned expert-defined topic	42.75	29.07	61.77	**48.59**	46.47	32.99	69.20	74.24

Table 2: Caption performance of TGM (single-task learning) and M&M TGM (multi-task learning) on MSR-VTT and Youtube2Text datasets.

dataset	model	BLEU4	METEOR	CIDEr
MSR-VTT	TGM	43.81	29.26	48.50
	M&M TGM	**44.33**	**29.37**	**49.26**
Youtube2Text	TGM	47.56	34.21	79.57
	M&M TGM	**48.76**	**34.36**	**80.45**

Table 3: Caption performance of M&M TGM and state-of-the-art methods on MSR-VTT and Youtube2Text datasets.

dataset	model	BLEU4	METEOR	CIDEr
MSR-VTT	M&M TGM	**44.33**	**29.37**	**49.26**
	Aalto [34]	39.80	26.90	45.70
	v2t_navigator [13]	40.80	28.20	44.80
	dense caption [33]	41.40	28.30	48.90
Youtube2Text	M&M TGM	48.76	**34.36**	**80.45**
	LSTM-YT [40]	33.30	29.10	-
	S2VT [39]	-	29.80	-
	LSTM-I [8]	44.60	31.10	-
	SA [45]	41.92	29.60	51.67
	LSTM-E [26]	45.30	31.00	-
	h-RNN decoder [47]	49.90	32.60	65.80
	h-RNN encoder [25]	43.80	33.10	-
	SCN-LSTM [9]	**50.20**	33.40	77.70

across MSR-VTT and Youtube2Text datasets, which demonstrates that exploiting topic information is beneficial to generate video descriptions.

The first three rows in the TGM block in Table 1 shows a fair comparison of the impact from different predicted latent topics on caption performance. We can see that the automatically mined topics (both textual and multimodal latent topics), outperform expert-defined topics by a large margin on all four metrics and two datasets, such as over 2% absolute gain on the CIDEr score. The multimodal latent topics further achieve better performance than textual latent topics consistently on multiple metrics across datasets. The improvement is also proved to be significant in the Student's t-test such as p-value of 0.03 on the BLEU@4 score. These results suggest that the multimodal topic mining approach can discover better underlying topics as guidance for TGM, and such improvement is prominent and robust for different evaluation metrics and datasets.

Since video uploaders can manually assign expert-defined topic labels on online websites, we also compare our automatically predicted multimodal latent topics with the user assigned expert-defined topics in the last two rows of Table 1. However, the assigned expert-defined topics are very noisy, because the annotations are labelled on the entire videos but the caption generation is only applied on the short segments and the users may also make labelling mistakes. For almost all metrics, the performances of our predicted latent topics are superior to the noisy assigned expert-defined topics on both MSR-VTT and Youtube2Text datasets, which shows that the latent topics learned in an unsupervised approach are competitive or even better than the assigned expert-defined topics with noise.

To validate the influence of joint training strategy, we compare the caption performance of TGM and M&M TGM. As shown in Table 2, M&M TGM consistently improves TGM on all datasets and metrics, with significance p-value < 0.05 for different metrics in Student's t-test, which proves the benefits brought by the multi-task learning.

5.4 Comparison with the State-of-the-art

Table 3 presents our M&M TGM with several state-of-the-art methods applied on the two video caption datasets. Our method achieves significant better performance than prior works on MSR-VTT dataset, for example, the BLEU@4 achieves 7.08% relative improvement than the previous best performance. For Youtube2Text dataset, the METEOR and CIDERr scores improved significantly but our performance on BLEU@4 is lower than the SCN-LSTM. The BLEU@4 metric focuses on the syntactic agreement, while METEOR and CIDEr concern more about semantic meanings. Therefore, we consider the sentence generated from M&M TGM are more semantically relevant with the video content.

To be noted, the semantic concepts used in SCN-LSTM are trained with additional image data because they claim Youtube2Text corpus is too small to train reliable concept classifiers, but our M&M TGM is purely trained on Youtube2Text for latent topic prediction and sentence generation. Hence, we argue that for video captioning, the guidance from latent topics might be superior than detected semantic concepts for the following reasons: 1) videos contain more objects than images but many might be irrelevant to the description; 2) topics contain additional information besides concepts such as actions from motion modality and words from speech modality; and 3) the prediction accuracy is important for the decoder and topics are easier to be classified than concepts.

---

winner of the MM16 VTT challenge.

Table 4: Comparison of generalization ability of the vanilla model and our proposed M&M TGM model. The models are trained on MSR-VTT training set and evaluated on the Youtube2Text testing set.

	BLEU4	METEOR	ROUGE$_l$	CIDEr
Vanilla	32.92	30.16	62.51	51.02
M&M TGM	**34.67**	**30.68**	**63.68**	**55.39**

Table 5: Caption performance comparison of different topic prediction losses.

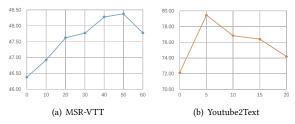| dataset	loss	BLEU4	METEOR	ROUGE$_l$	CIDEr
MSR-VTT	l_2	**43.81**	**29.26**	**62.13**	48.50
	KL	43.40	29.11	62.01	**48.64**
Youtube2Text	l_2	**47.56**	**34.21**	**70.45**	**79.57**
	KL	45.98	33.65	69.76	78.06

(a) MSR-VTT (b) Youtube2Text

Figure 4: CIDEr scores with number of topics on (a) MSR-VTT and (b) Youtube2Text datasets.

5.5 Experimental Analysis

Generalization Ability: To evaluate the generalization ability of our proposed method, we conduct the cross dataset experiment. We train M&M TGM on MSR-VTT dataset and test its performance on the Youtube2Text dataset. Results are presented in Table 4. We can see that the proposed M&M TGM works significantly better than Vanilla model for cross datasets evaluation on all four metrics, which demonstrates that our method not only can improve the caption performance for in-domain videos but also generalize well on videos in the wild.

Topic Loss Selection: We compare the caption performance using l_2-distance or KL-divergence as L_{topic} in the TGM model. Table 5 presents the results. The l_2-distance consistently surpasses the KL-divergence across datasets and metrics. So unless otherwise specified, we use l_2-distance as the topic prediction loss.

The Number of Topics K: We also explore different number of topics on each dataset and the results are shown in Figure 4. The best number of topics for MSR-VTT is 50 and for Youtube2Text is 5, which shows that the more diverse the dataset is, the more amount of topics is required.

The Multi-task Parameter λ: Since caption generation is our main goal, we consider that the weight on topic prediction loss should not surpass the weight on caption generation loss. Therefore, we search the best multi-task hyper parameter $\frac{\lambda}{1-\lambda}$ in range [0.1, 1] with step of 0.1. To save space, we present the sum of all the metrics in Figure 5 and the trends are similar for different metrics. The multi-task learning with different $\lambda > 0$ all achieves better performance
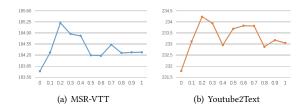

(a) MSR-VTT (b) Youtube2Text

Figure 5: The caption performance (sum of all metrics) with different multi-task hyper-parameter λ.

Vanilla Baseline: a group of people are dancing.
M&M TGM : a group of people are fighting in a war.
Random GT: fighting scenes are shown from a movie.

Vanilla Baseline: people are playing soccer in a stadium.
M&M TGM : two teams are playing a game of rugby in a stadium.
Random GT: two teams of rugby players are playing a rugby game.

Figure 6: Examples on MSR-VTT testing set. M&M TGM can generate more accurate descriptions than Vanilla model.

than single-task learning ($\lambda = 0$), which proves the robustness of our multi-task training with respect to the hyper-parameter.

5.6 Qualitative Analysis

To gain an intuition of the improvement on generated video descriptions from M&M TGM model, we present some video examples with the video description from Vanilla model and M&M TGM on testing set of MSR-VTT.

In Figure 6, we can see that M&M TGM can generate more accurate video descriptions than Vanilla model even though they utilize the same multimodal features. In Figure 7, though the descriptions from Vanilla and M&M TGM are both correct, M&M TGM model can produce more detailed information about the video contents. We also observe that M&M TGM employs more unique words with 493 compared to the 391 unique words in Vanilla model.

The reason behind these quality improvements could be that M&M TGM can narrow down the sentence generation space according to the predicted latent topics. This enables the model to focus on subtler differences between similar concepts such as the soccer and rugby sports in the upper row of Figure 6, and cover more detailed descriptions under the topic such as making airplane vs. folding paper in the upper row of Figure 7 with more specialized words.

5.7 Human Interaction in Captioning

Besides the improvement on the caption performance, our topic-guided model can provide an interface for users to be involved in the caption generation. For example, users could simply assign a category tag to the uploaded videos with minimum costs to refine the automatic generated video descriptions. To evaluate the

Vanilla Baseline: a person is folding a piece of paper.
M&M TGM : **a person is making a paper airplane.**
Random GT: a man is showing how to fold a paper airplane.

Vanilla Baseline: a girl is playing with a doll in a room.
M&M TGM : **a girl in a pink dress is playing with her doll.**
Random GT: little girl in pink dress says cute and playing with toys.

Figure 7: Examples on MSR-VTT testing set. M&M TGM can generate more detailed descriptions than Vanilla model.

Table 6: The influence of human interactions on Youtube2Text dataset with noisy and clean expert-defined topics assignment.

	BLEU4	METEOR	ROUGE$_l$	CIDEr
noisy topic	46.47	32.99	69.20	74.24
clean topic	**49.35**	**35.10**	**71.14**	**82.79**

performance boosts with the human enhancement to our topic-guided model, we manually re-annotate the expert-defined topics for Youtub2Text dataset as the clean version and compare its caption performance with the noisy assigned topics. Results are shown in Table 6 which demonstrates that a huge gain can be achieved with the manual correction of the topics in low cost.

6 Conclusions

In this paper, we propose a novel topic-guided caption model to address the topic diversity challenge for open-domain video captioning task. The proposed model can predict the latent topics of videos and then generate topic-oriented video descriptions with the topic guidance jointly in an end-to-end manner. We utilize the multimodal topic mining approach to construct video topics automatically and take a teacher-student learning perspective to predict the latent topics purely from video multimodal contents. The topic-aware decoder can exploit the predicted topics to adjust its weights to fit the topic-dependent sentence distributions. Our experimental results on two public video caption benchmark datasets show that the proposed model can generate more accurate and detailed descriptions within different topics and improves the performance consistently on all metrics on both datasets. Furthermore, we show that our model has very good generalization ability across datasets. The proposed topic-guided caption model can be considered as a generic framework, which could be integrated with other techniques such as temporal attention or hierarchical video encoder. We will study such integration in the future.

7 Acknowledgments

References

[1] Jimmy Ba and Rich Caruana. 2014. Do deep nets really need to be deep?. In *NIPS*. 2654–2662.
[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).
[3] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.
[4] Shizhe Chen, Jia Chen, and Qin Jin. 2017. Generating Video Descriptions with Topic Guidance. In *ICMR*. 5–13.
[5] Steven Davis and Paul Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing* 28, 4 (1980), 357–366.
[6] Inderjit S Dhillon, Yuqiang Guan, and Brian Kulis. 2004. Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 551–556.
[7] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei Efros. 2012. What makes paris look like paris? *ACM Transactions on Graphics* 31, 4 (2012).
[8] Yinpeng Dong, Hang Su, Jun Zhu, and Bo Zhang. 2017. Improving Interpretability of Deep Neural Networks with Semantic Information. *arXiv preprint arXiv:1703.04096* (2017).
[9] Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. 2016. Semantic Compositional Networks for Visual Captioning. *arXiv preprint arXiv:1611.08002* (2016).
[10] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep Sparse Rectifier Neural Networks.. In *Aistats*, Vol. 15. 275.
[11] 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*. 2712–2719.
[12] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[13] Qin Jin, Jia Chen, Shizhe Chen, Yifan Xiong, and Alexander Hauptmann. 2016. Describing Videos using Multi-modal Fusion. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 1087–1091.
[14] Qin Jin and Junwei Liang. 2016. Video description generation using audio and visual cues. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM, 239–242.
[15] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[16] Alex Krizhevsky, Geoffrey E Hinton, and others. 2010. Factored 3-way restricted boltzmann machines for modeling natural images. In *International conference on artificial intelligence and statistics*. 621–628.
[17] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2011. Baby talk: Understanding and generating image descriptions. In *Proceedings of the 24th CVPR*. Citeseer.
[18] Michael Denkowski Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. *ACL* (2014), 376.
[19] Rémi Lebret, Pedro H. O. Pinheiro, and Ronan Collobert. 2015. Phrase-based Image Captioning. In *ICML*. 2085–2094.
[20] Stefan Lee, Senthil Purushwalkam Shiva Prakash, Michael Cogswell, Viresh Ranjan, David Crandall, and Dhruv Batra. 2016. Stochastic Multiple Choice Learning for Training Diverse Deep Ensembles. In *NIPS*. 2119–2127.
[21] Hongzhi Li, Joseph G Ellis, Heng Ji, and Shih-Fu Chang. 2016. Event specific multimodal pattern mining for knowledge base construction. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 821–830.
[22] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. 2016. Tgif: A new dataset and benchmark on animated gif description. In *CVPR*. 4641–4650.
[23] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, Vol. 8. Barcelona, Spain.
[24] R Memisevic and G Hinton. 2007. Unsupervised Learning of Image Transformations. In *CVPR*. 1–8.
[25] 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*. 1029–1038.
[26] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. 2016. Jointly Modeling Embedding and Translation to Bridge Video and Language. In *CVPR*. 4594–4602.
[27] Stephanie Pancoast and Murat Akbacak. 2014. Softening quantization in bag-of-audio-words. In *ICASSP*. IEEE, 1370–1374.
[28] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 311–318.

[29] Vasili Ramanishka, Abir Das, Dong Huk Park, Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, and Kate Saenko. 2016. Multimodal Video Description. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 1092–1096.

[30] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2016. Self-critical Sequence Training for Image Captioning. *arXiv preprint arXiv:1612.00563* (2016).

[31] Microsoft Research. 2016. ACM Multimedia MSR video to language challenge. *http://www.acmmm.org/2016/wp-content/uploads/2016/04/ACMMM16_GC_MSR_Video_to_Language_Updated.pdf* (2016).

[32] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. 2013. Image classification with the fisher vector: Theory and practice. *International journal of computer vision* 105, 3 (2013), 222–245.

[33] Zhiqiang Shen, Jianguo Li, Zhou Su, Minjun Li, Yurong Chen, Yu-Gang Jiang, and Xiangyang Xue. 2017. Weakly Supervised Dense Video Captioning. In *CVPR*.

[34] Rakshith Shetty and Jorma Laaksonen. 2016. Frame-and segment-level features and candidate pool evaluation for video caption generation. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 1073–1076.

[35] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*. 3104–3112.

[36] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. 2016. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. (2016).

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

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

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

[40] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. 2014. Translating Videos to Natural Language Using Deep Recurrent Neural Networks. *Computer Science* (2014).

[41] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*. 3156–3164.

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

[43] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *arXiv:1502.03044* (2015).

[44] Zhilin Yang, Ye Yuan, Yuexin Wu, William W Cohen, and Ruslan R Salakhutdinov. 2016. Review networks for caption generation. In *NIPS*. 2361–2369.

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

[46] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. *arXiv:1603.03925* (2016).

[47] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. 2016. Video Paragraph Captioning Using Hierarchical Recurrent Neural Networks. In *CVPR*. 4584–4593.