Generating Video Descriptions with Topic Guidance

Shizhe Chen
Multimedia Computing Lab
School of Information
Renmin University of China
cszhe1@ruc.edu.cn

Jia Chen
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
jiac@cs.cmu.edu

Qin Jin*
Multimedia Computing Lab
School of Information
Renmin University of China
qjin@ruc.edu.cn

Abstract

Generating video descriptions in natural language (a.k.a. video captioning) is a more challenging task than image captioning as the videos are intrinsically more complicated than images in two aspects. First, videos cover a broader range of topics, such as news, music, sports and so on. Second, multiple topics could coexist in the same video. In this paper, we propose a novel caption model, topic-guided model (TGM), to generate topic-oriented descriptions for videos in the wild via exploiting topic information. In addition to predefined topics, i.e., category tags crawled from the web, we also mine topics in a data-driven way based on training captions by an unsupervised topic mining model. We show that data-driven topics reflect a better topic schema than the predefined topics. As for testing video topic prediction, we treat the topic mining model as teacher to train the student, the topic prediction model, by utilizing the full multi-modalities in the video especially the speech modality. We propose a series of caption models to exploit topic guidance, including implicitly using the topics as input features to generate words related to the topic and explicitly modifying the weights in the decoder with topics to function as an ensemble of topic-aware language decoders. Our comprehensive experimental results on the current largest video caption dataset MSR-VTT prove the effectiveness of our topic-guided model, which significantly surpasses the winning performance in the 2016 MSR video to language challenge.

Keywords

Video Captioning; Data-driven Topics; Multi-modalities; Teacher-student Learning

1 Introduction

It is an ultimate goal of video understanding to automatically generate natural language descriptions of video contents. A wide range of applications can benefit from it such as assisting blind people, video editing, indexing, searching or sharing. Drawing on the recent success of image captioning [1–4], where a sentence is generated to describe the image content, more researchers are paying attention to the video captioning task to translate videos to natural language.

However, the open domain videos are quite diverse in topics which makes generating video descriptions more complicated than the image captioning. For various topics ranging from political news to edited movie trailers, the vocabularies and expression styles vary a lot in describing the video contents. For example, for political news videos, words from the political domain occur more frequently. Also political news descriptions are typically in the style of somebody reporting something, which are quite different from descriptions for other topics such as gaming, travel, movie, and animals etc. Besides the topic diversity across videos, even in the same video, its diversity in content and video structure can result in very different video descriptions capturing different topics in the video as shown in Figure 1. Therefore, the topic information is important to guide the caption model to generate better topic-oriented language expression.

Speech Content: 50% of Canada's research Publications have co-authors from other countries this is double the world ...
Groundtruth Captions:
People in white lab coats working on a project.
A person using a laptop.
Canada has double the number co-authors than other countries.
This is a video about technology in Canada.

Speech Content: you can unlock a bus if you really want a driver...
Groundtruth Captions:
A gamer describes how to unlock a feature.
A guy is playing a pc video game.
A lego man with wings is picking a vehicle.
A man is describing a video game as he plays it.

Figure 1: Topic diversity across videos and within one video.

In our previous study [5], we have utilized the predefined topics, the category tags crawled from video meta-data during data collection, to improve the captioning performance. However, the predefined topics are suboptimal for video captioning because: 1) The crawled information contains labelling mistakes which harms the captioning performance; 2) The exclusive topic labels do not capture the topic diversity nature inside the video; 3) The predefined topic schema is not specially designed for the video captioning task, which may not reflect the topic distributions well. In this work, we propose the data-driven topics to deal with the drawbacks of predefined topics. We use the Latent Dirichlet Allocation (LDA) topic mining model to automatically generate topics from the annotated video captions in training set.

In order to use the mined topic information in the captioning task, two questions need to be addressed: 1) how to obtain the topics automatically for testing videos; and 2) how to effectively

*Corresponding author.

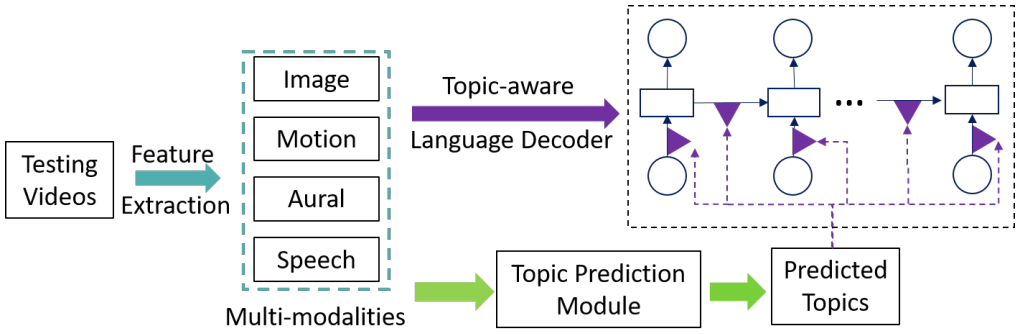


Figure 2: Framework of the proposed topic-guided model (TGM). We use the automatically mined topics as our topic guidance. For testing videos, multimodal features are utilized to predict the mined topics. Then the predicted topics are adopted in the topic-aware language decoder to generate topic-oriented video captions.

employ the topic information to model the difference of sentence descriptions in different video topics.

For the first question, we take a teacher-student learning perspective [6] to train the data-driven topic prediction model. The LDA topic mining model is viewed as the teacher to guide the student topic prediction model to learn. A video generally contains multiple modalities including image, motion, aural and speech modalities. Image modality provides rich information for understanding the video's semantic contents such as object and scene. The motion modality presents actions of the objects and the temporal structure of videos. Aural and speech modalities provide additional information for understanding semantic topics from the sound perspective. Hence, we build one general topic prediction model that utilizes all the four modalities and another topic prediction model dedicated to speech modality as its feature representation is different from other modalities.

For the second question, we propose a novel topic-guided model (TGM) to employ the predicted topics, which is based on the encoder-decoder framework [7]. The TGM functions as an ensemble of topic-aware language decoders to learn specific vocabularies and expressions for various video topics. We also compare the TGM with a series of caption models that we propose in this paper to exploit the topic information, including topic concatenation in encoder (TCE), topic concatenation in decoder (TCD), topic embedding addition/multiplication in decoder (TEAD/TEMD). These compared models implicitly use topics to change the input features of the encoder or decoder, while our proposed TGM explicitly modifies the weights in the decoder according to the predicted topics to capture the sentence distributions within the topic more effectively. The framework of the overall system for testing videos is shown in Figure 2. Experimental results on the MSR-VTT dataset demonstrate the effectiveness of our proposed method, which can generate more comprehensive and accurate video descriptions.

In summary, our contributions in this work include: 1) we show that the data-driven topics are more suitable as the topic representation for video captioning than the predefined topics, e.g. the category tags, with respect to the topic accuracy and schema; 2) to the best of our knowledge, we are the first to use the full multi-modalities especially the speech modality to successfully boost the video captioning performance; and 3) the proposed topic-guided model can exploit the topic information more effectively to generate better topic-oriented video descriptions.

The rest of the paper is organized as follows: Section 2 introduces the related work. Section 3 compares the predefined and the data-driven topics. Our proposed topic-guided model is described in Section 4. Section 5 presents experimental results and analysis. Section 6 draws some conclusions.

2 Related Works

There are mainly two directions in previous image/video captioning works. The first is to build rule based systems, which first detect words by object or action recognition and then generate sentences with predefined language constrains. For example, Lebret et al. [1] predict phrases with a bilinear model and generate descriptions using simple syntax statistics. Rohrbach et al. [8] use the Conditional Random Field to learn object and activity labels from the video. Such systems suffer from the expression accuracy and flexibility.

More recently, researches have been focusing on the second direction of encoder-decoder framework [7] which generates sentences based on image/video features in an end-to-end manner. For example, Vinyals et al. [2] utilize the LSTM to generate sentences with CNN features extracted from the image. Venugopalan et al. [9] transfer knowledge from image caption models with the encoder to perform mean pooling over frame CNN features for video captioning. Pan et al. [10] explicitly embed the sentences and the videos into a common space in addition to the video description generation.

There are also works considering to employ semantic concepts in the encoder-decoder framework. For example, Wu et al. [11] directly generate image captions based on the detected semantic concepts. You et al. [4] propose to selectively attend to concept proposals in the decoder. Gan et al. [12] propose the semantic compositional networks (SCN) which works as an ensemble of concept-dependent language decoder. Our topic-guided model is inspired by SCN but with different aims of producing topic-oriented descriptions to address the topic diversity in video captioning. The reasons of using topics rather than semantic concepts are as follows: 1) There are much more objects in a video than in an image but many

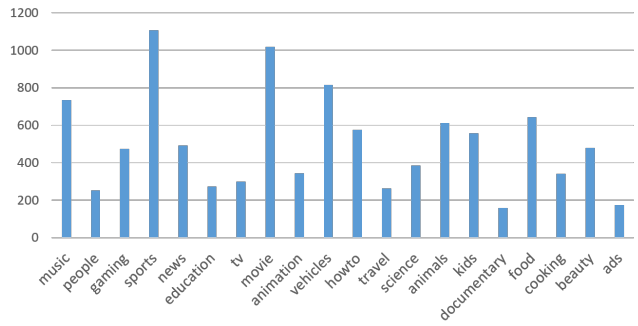


Figure 3: The number of video clips for the 20 predefined video categories of MSR-VTT dataset.

Figure 4: Examples of inaccurate category tags in MSR-VTT dataset. The word in black is the original category tag and the red is the more appropriate tag.

of them might be irrelevant to the video description. 2) The topics contain more additional information than semantic objects such as from the motion, aural and speech modalities; and 3) The prediction accuracy is very important for the model as shown in [12]. The topic classification is much easier than object classification.

For video captioning, various video topics result in quite diverse expressions compared with image captioning. Previous works have explored generating descriptions for narrow-domain videos such as YouCook [13] and TACoS [14], whose vocabularies and expressions are similar in the dataset. However, for open-domain videos with various topics such as the MSR-VTT dataset [15], Jin et al. [5] exploit the predefined video categories in the encoder and significantly improve the captioning performance, which results in their winning of the MSR video to language challenge [16]. In our work, we further analyze the qualities of the predefined categories and propose to mine topics in a data-driven approach that leads to better accuracy and topic schema.

Multi-modality nature is also emphasized in video captioning. For the motion modality in videos, Yao et al. [17] explore the temporal structure with local C3D features and global temporal attention mechanism. Venugopalan et al. [18] propose the sequence to sequence structure which utilizes the LSTM as encoder to capture the temporal dynamics of videos. Pan et al. [19] further propose the hierarchical RNN encoder as well as the temporal-spatial attention. Aural modality has also been explored for video captioning. Jin et al. [5, 20] and Ramanishka et al. [21] integrate the visual and aural features in the encoder by early fusion and show that the multi-modal fusion was beneficial to improve captioning performance. In our work, we consider more modalities in videos especially for the use of speech content modality.

3 Predefined vs. Data-driven Topics

Our previous study [5] has shown that using predefined topics such as category tags can significantly boost video captioning performance. In this section, we analyze the qualities of these predefined topics, and propose a data-driven approach to develop better topics for the captioning task.

3.1 Predefined Topics: Category Tags

Each video clip in the MSR-VTT dataset contains a predefined category tag derived from the meta-data of the video. The distribution of the category tags is shown in Figure 3. The predefined category tags reflect the variety of the video topics, but there are mainly three disadvantages of them:

(1) Inaccurate category labels: The predefined category labels contain a certain amount of labelling mistakes as shown in the examples in Figure 4, which greatly harm the captioning performance.

(2) Exclusive topic distributions: The users can only assign one of the category labels. Such one-hot topic representation cannot reflect the topic diversity inside the video.

(3) Suboptimal topic schema: a) Ambiguous category definition. For example, the 'people' category is too general to classify. b) Overlap between different categories. For example, the 'food' and 'cooking' categories cover almost similar videos. c) Large mixed categories. Some categories contain much more videos than others and are mixed with many subclasses. d) Indirect connection with captioning task. The category tags are defined to organize videos in the wild, they are not specifically defined for video captioning.

Therefore, although the predefined category tags have benefited video captioning a lot, there is still much room for improvement by defining a better topic schema to represent the diversity of videos for the video captioning task.

3.2 Data-driven Topics

In order to overcome the drawbacks of the predefined category tags, we propose a data-driven way to generate a more suitable set of video topics. The human generated groundtruth captions provide us with rich and accurate annotations about the videos, which also reflect a more task-related topic distributions of the videos. Thus, we propose to mine topics from the groundtruth video captions in the training set. We note that it requires no additional labelling effort on the dataset.

We observe that the multiple human generated groundtruth captions sometimes do not agree with each other even for the same video. For example, Figure 1 shows an example video and its groundtruth captions which describe the video from different aspects including detailed frame contents, speech contents and general video contents. Such example indicates that a video usually

Table 1: Examples of some data-driven topics with their representative words and their co-occurrence with predefined category tags.

topic id	#videos	co-occurrence with predefined categories	representative words
1	182	music:43%, people:21%	people dancing group girls dance women music video dances
3	281	music:50%	stage man playing singing band performing song guitar music
13	439	food:63%, cooking:29%	food cooking kitchen dish person bowl ingredients pan preparing
8	864	news:15%, edu:14%, sci:13%	man talking talks guy speaking person sitting giving camera

contains several topics, which are reflected in its multiple diverse groundtruth captions. The above observation aligns with the generation process of Latent Dirichlet Allocation (LDA) model [22]:
1. the model first draws a topic index $z_{di} \sim Multinomial(\theta_d)$ from the video, where $\theta_d \sim Dirichlet(\alpha), d = 1, ..., D$.
2. the model draws the observed word $w_{ij} \sim Multinomial(\beta_{z_{di}})$ from the selected topic, where $\beta_k \sim Dirichlet(\eta), k = 1...K$.
Here, we group the multiple groundtruth captions of a video into one document to mine latent topics from the training data. Stopwords are removed and the the bag-of-words representation is used as our document feature.

3.3 Relation between Predefined Topics and Data-driven Topics

We study the relation between the predefined topics and data-driven topics based on their co-occurrence in videos. For each video in the training set, we have its corresponding predefined category tag and the data-driven topic distribution calculated from the LDA model. To simplify the calculation of co-occurrence, we assign each video with the most likely topic. Table 1 shows co-occurrence between some predefined categories and data-driven topics. We can see that some predefined categories are split into different topics. For example, the music category is mainly separated into topic 1 of dancing and topic 3 of singing. Some content similar categories are combined together as one topic, i.e. topic 13 consisting of food and cooking categories. And categories that are different but express similar content are also merged. For example, news, educations and science categories are merged into one as most of the descriptions under these categories are "somebody is talking about something". In summary, it shows that the data-driven topics are quite promising and reflect video content distributions better than predefined categories.

4 Topic Guidance Model

In this section, we provide our solutions for the following two problems: 1) how to automatically predict topics for testing videos with multi-modalities; 2) how to maximize the effects of the topic information for caption generation.

4.1 Multimodal Features

We extract features from image, motion, aural and speech modalities to fully represent the content of videos.

Image modality: The image modality reflects the static content of videos. We extract activations from the penultimate layers of the inception-resnet [23] pre-trained on the ImageNet as image object features, and the penultimate layers of the resnet [24] pre-trained on the places365 [25] as image scene features, the dimensionality of which are 1536 and 2048 respectively.

Motion modality: The motion modality captures the local temporal motion. We extract features from the C3D model [26] pre-trained on the Sports-1M dataset. We extract activations from the last 3D convolution layer and max-pooling them along the spatial dimension (width and height) to obtain video features with dimensionality of 512. We then applied l_2-norm on the C3D features.

Aural modality: Aural modality is complementary to visual modalities, especially for distinguishing scenes or events. We extract the Mel-Frequency Cepstral Coefficients (MFCCs) [27] as the basic low-level descriptors. Two encoding strategies, Bag-of-Audio-Words [28] and Fisher Vector [29], are used to aggregate MFCC frames into one video-level feature vector, with dimensionality of 1024 and 624 respectively.

Speech modality: Speech modality provides semantic topics and content details of the video. We use the IBM Watson API [30] for speech recognition. Since the backgrounds of the videos are noisy, we clean the speech transcriptions by removing transcriptions with less than 10 words and out-of-vocabulary words based on the training caption vocabulary. Only about half portion of the videos contain speech transcriptions afterwards and a certain amount of transcription errors still exist in the transcriptions. We use the bag-of-words representation as the speech modality feature.

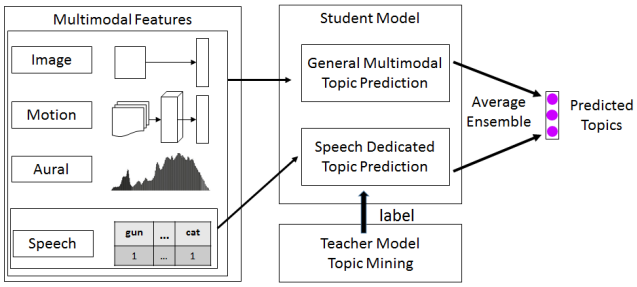


Figure 5: The framework for topic prediction. We treat the problem from a teacher-student perspective. The teacher topic mining model is used to guide the two student topic prediction models to learn based on general multi-modalilties and speech modality respectively.

4.2 Topic Prediction

For predefined topics, i.e. category tag, we train a standard one hidden layer neural network with cross-entropy loss to predict. The inputs of this neural network are the multimodal features as described in the section 4.1.

For data-driven topics, as there is no direct topic class label, we leverage the topic distribution generated from the topic mining LDA model in section 3.2. We take a teacher-student learning perspective [6] to train the data-driven topic prediction model. To be specific, the topic mining LDA model is viewed as the teacher and the topic prediction model is viewed as the student. The teacher, topic mining model, is trained in unsupervised style and it generates label, i.e. topic, to guide the student, topic prediction model, to learn. First, we design two topic prediction models: one is general for all multimodal features and the other one is dedicated to speech modality features. Then, we ensemble predictions from these two models by averaging to get the final prediction.

General Multimodal Topic Prediction Model: This model is designed to predict topic from the video content using all the multimodal features. In the teacher-student perspective, the student model usually learns the output distribution on labels, i.e. dark knowledge [31], rather than output label from the teacher model. Following this way, we choose to use KL-divergence as our loss function. The formulation of KL-divergence is as follows:

$$D_{KL}(P||Q) = \sum_{k=1}^{K} P_k log \frac{P_k}{Q_k} \tag{1}$$

where P, Q are the probability distributions of the mined topics and predicted topics respectively. The $D_{KL}(P||Q)$ is differentiable and thus can be optimized via back-propagation.

Speech Dedicated Topic Prediction Model: As speech modality is very informative but not always available in videos, we build a dedicate topic prediction model using speech features. Different from other modality features, the speech text feature is of high dimensionality with noises and is sparse. Instead of using the same architecture of the general multimodal topic prediction model, we design a very different architecture for the speech dedicated topic prediction model. We make a simple choice of the architecture: reusing the topic mining model for topic prediction on speech text features. As the representation of speech text features is the same as that of features used in the topic mining model, we don't need to reinvent the wheel.

As to the problem of missing modalities, we only use videos that contain the corresponding speech modality in the speech dedicated topic prediction model. In general multimodal topic prediction model, the features of missing modalities are padded as zeros. In ensembling, predictions of the missing modalities are not considered in the average process.

4.3 Caption Models with Topic Guidance

Suppose we have multiple video-sentence pairs (V, y) in video captioning dataset, where $\mathrm{y} = \{w_1, \ldots, w_{N_w}\}$ is the sentence with N_w words. Assume the multimodal features of the video are $m_1, \ldots, m_{N_m}$, where N_m is the number of modalities, the multimodal encoder is a neural network that fuses the multimodal features into a dense video representation x as follows:

$$\mathrm{x} = W_e[m_1; \ldots; m_{N_m}] + b_e \tag{2}$$

where W_e, b_e are the parameters in the encoder and $[\cdot]$ denotes the feature concatenation. Since the captioning output is the sequential words, we utilize the LSTM [32] recurrent neural networks as our language decoder:

$$h_t = f(h_{t-1}, w_{t-1}; \theta_d) \text{ for } t = 1, \ldots, N_w \tag{3}$$

where f is the LSTM update function, h_t is the state of LSTM and θ_d is the parameter in LSTM. We initialize h_0 as x to condition on the video representation and w_0 as the sentence start symbol. Then the probability of the correct word conditioned on the video content and previous words can be expressed as:

$$\Pr(w_t|\mathrm{x}, w_0, \ldots, w_{t-1}) = \mathrm{Softmax}(W_d h_t + b_d) \tag{4}$$

where W_d, b_d are parameters. The objective function is to maximize the log likelihood of the correct description:

$$\log \Pr(\mathrm{y}|\mathrm{x}) = \sum_{t=1}^{N_w} \log \Pr(w_t|\mathrm{x}, w_0, \ldots, w_{t-1}) \tag{5}$$

We denote the predicted topics in Section 4.2 as $\mathrm{z} \in \mathbb{R}^K$, where K is the number of topics. In order to effectively exploit the topics z for video captioning, we propose the novel topic-guided model (TGM), and a series of simple yet strong caption models with topic guidance for comparison, called TCE, TCD and TEAD/TEMD for short. Detailed descriptions of these models are as follows.

Topic Concatenation in Encoder (TCE): We fuse the topic distribution z together with multimodal features in the encoder as the topic-aware video representation x:

$$\mathrm{x} = W_e([m_1; \ldots; m_{N_m}; \mathrm{z}]) + b_e \tag{6}$$

The LSTM decoder then generates video description conditioning on the new topic-aware representation x.

Topic Concatenation in Decoder (TCD): In the TCE model, the topic information only occurs at the first step in the decoder, which could easily make the topic guidance "drift away". To enhance the topic guidance, we concatenate the topic distribution z with the word embedding w_{t-1} as the input to the LSTM every step, which is similar to the gLSTM proposed in [33]:

$$h_t = f(h_{t-1}, [w_{t-1}; \mathrm{z}]; \theta_d) \tag{7}$$

The extra input of topic z can guide the language decoder to generate words related to the topic in every step.

Topic Embedding Addition/Multiplication in Decoder (TEAD/TEMD): To generate a more comparable representation for the topic representation z compared to the word embedding, we embed each topic into a latent vector space with the same dimensionality as the word embedding:

$$\mathrm{z}_e = W_z \mathrm{z} + b_z \tag{8}$$

where W_z, b_z are topic embedding parameters. We perform addition or multiplication on the topic embedding and word embedding to generate the topic-aware input feature for the language decoder every step, which are expressed as:

$$\text{TEAD}: \ h_t = f(h_{t-1}, w_{t-1} \oplus \mathrm{z}_e; \theta_d) \tag{9}$$

$$\text{TEMD}: \ h_t = f(h_{t-1}, w_{t-1} \odot \mathrm{z}_e; \theta_d) \tag{10}$$

where $\oplus, \odot$ are element-wise addition and multiplication respectively.

Topic-Guided Model (TGM): The above TCD and TEAD/TEMD models only implicitly using the topic information as the global guidance, which modify the inputs to the language decoder in every step and cannot take into account of the overall expressions

within the topic. Inspired by Gan et al. [12], therefore, we further propose the topic-guided model (TGM) that explicitly functions as an ensemble of topic-aware language decoders to capture different sentence distributions for each topic. The structure of the TGM is shown in the right side in Figure 2, which can automatically modify the weight matrices in LSTM according to the topic distribution z.

Let us take one of weight matrices in the LSTM as an example, and other parameters in LSTM cell are alike. We define weight $W_\tau \in \mathbb{R}^{n_h \times n_w \times K}$, where n_h is the number of hidden units and n_w is the dimension of word embedding. The W_τ can be viewed as the ensemble of K topic-specific LSTM weight matrices. The topic-related weight matrix $W(\mathrm{z}) \in \mathbb{R}^{n_h \times n_w}$ can be specified as

$$W(\mathrm{z}) = \sum_{k=1}^{K} \mathrm{z}_k W_\tau[k] \tag{11}$$

where z_k is the k-th topic in z; $W_\tau[k]$ denote the k-th matrix of W_τ. In this way, the video topic z can automatically generate its corresponding LSTM decoders to produce the topic-oriented video descriptions. However, the parameters are increasing with K which may result in over-fitting easily. So the ideas in [34] are used to share parameters by factorizing $W(\mathrm{z})$ as follows:

$$W(\mathrm{z}) = W_a \cdot diag(W_b \mathrm{z}) \cdot W_c \tag{12}$$

where $W_a \in \mathbb{R}^{n_h \times n_f}$, $W_b \in \mathbb{R}^{n_f \times K}$ and $W_c \in \mathbb{R}^{n_f \times n_w}$. n_f is the number of factors. W_a and W_c are shared among all topics, while W_b can be viewed as the latent topic embedding.

5 Experiments

5.1 Experimental Setup

Dataset: The MSR-VTT corpus [15] is currently the largest video to language dataset with a wide variety of video contents. It consists of 10,000 video clips with 20 human generated captions per clip. Each video also contains a predefined category tag, which is one of the 20 popular video categories in web videos. Following the standard data split, we use 6,513 videos for training, 497 videos for validation and the remained 2,990 for testing.

Data Preprocessing: We convert all descriptions to lower case and remove all the punctuations. We add begin-of-sentence tag <BOS> and end-of-sentence tag <EOS> to our vocabulary. Words which appear more than twice are selected, resulting in a vocabulary of size 10,868. The maximum length of a generated caption is set to be 30.

Training Settings: We empirically set the feed forward neural networks for topic prediction to have one hidden layer with 512 units. The dimension of 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 [35] with learning rate of 10^{-4}. Beam search with beam width of 5 is used to generate sentences during testing process. The baseline system is the vanilla encoder-decoder framework with multimodal features (we call it the multimodal baseline).

In our experiments, we find that the dimensionality of the features to the encoder should not be too high in order to avoid over-fitting, so only the image object, video motion, and aural features are used as input to the encoder.

Table 2: Captioning performance of the predefined categories and data-driven topics using the best caption model with topic guidance. The acronyms B, M, R, C denote BLEU@4, METEOR, ROUGE-L and CIDEr respectively.

	B	M	R	C
multimodal	0.4211	0.2872	0.6170	0.4608
pred category TGM	0.4276	0.2885	0.6167	0.4754
pred topic TGM	**0.4397**	**0.2921**	0.6234	**0.4970**
category TCE	0.4343	0.2921	**0.6249**	0.4890

Evaluation Metrics: We evaluate the caption results comprehensively on all major metrics, including BLEU [36], METEOR [37], ROUGE-L [38] and CIDEr [39].

5.2 Evaluation

Table 2 presents captioning performances on testing set with predefined category tags and the data-driven topics using their corresponding best caption model with topic guidance. We can see that the different topic guidances (the second to forth rows) all greatly improve the performance of the multimodal baseline (the first row). Since the data-driven topics on testing set are predicted as shown in Figure 5, we also use the predicted category tags for a fair comparison. The guidance from predicted data-driven topics outperforms that from the predicted category tags on all four evaluation metrics, and the Student's t-test shows the improvement is significant with p-value$<$ 0.002. Even compared with the category tags assigned by video uploaders, the predicted data-driven topics also slightly boost the captioning performance on multiple metrics with the Student's t-test p-value$<$ 0.01 on BLEU@4 and CIDEr metrics, which shows the performance gain is robust. These results suggest that the data-driven topics are more suitable as the topic representation than predefined topics for video captioning.

To demonstrate the effectiveness of the proposed topic-guided model (TGM), we further compare the TGM with other caption models with topic guidance. As shown in Table 3, the TGM achieves the best performance among all the caption models on all four metrics especially for the CIDEr score. It suggests that modifying the weights of the decoder according to the topic distributions can employ the topic guidance more effectively to generate better topic-oriented descriptions. Our proposed TGM model also achieves better performance than the winning performance in 2016 MSR video to language challenge [5], where we use multimodal features and select best models by the predefined categories.

Figure 6 presents some examples in the testing set. In addition to more accurate video descriptions, the TGM can also generate more novel concepts such as the llama. Our statistics on the generated sentence show that the number of unique caption words generated by TGM is 397, while it is 360 by the multimodal baseline model.

5.3 Ablation Experiments

Interactive Caption with Manually Annotated Topic: Our model offers the flexibility of manually assigned topics to the video. It means that we could interactively annotate the topics for testing videos based on relevance between video contents and the representative words in topics to generate better captions. Results in Table 4

Table 3: Captioning performance comparison among caption models with topic guidance (using the predicted data-driven topics). The acronyms B, M, R, C denote BLEU@4, METEOR, ROUGE-L and CIDEr respectively.

	B	M	R	C
pred topic TCE	0.4353	0.2916	0.6226	0.4783
pred topic TCD	0.4246	0.2861	0.6171	0.4714
pred topic TEAD	0.4333	0.2893	0.6221	0.4765
pred topic TEMD	0.4304	0.2884	0.6207	0.4694
pred topic TGM	**0.4397**	**0.2921**	**0.6234**	**0.4970**
v2t_navigator [5]	0.4080	0.2820	0.6090	0.4480

multimodal: a group of people are dancing.
category TCE: a group of people are playing basketball.
pred topic TGM: a group of people are wrestling.

multimodal: a group of people are sitting on the grass.
category TCE : a group of people are playing with a horse.
pred topic TGM: a group of llamas are running in a field.

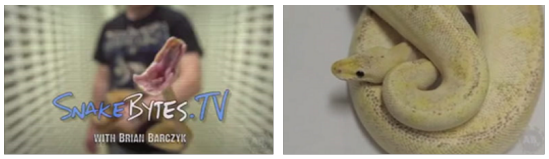
Figure 6: Examples on testing set to demonstrate the effectiveness of the TGM.

Table 4: Captioning performance of predicted topics and the annotated topics. The acronyms B, M, R, C denote BLEU@4, METEOR, ROUGE-L and CIDEr respectively.

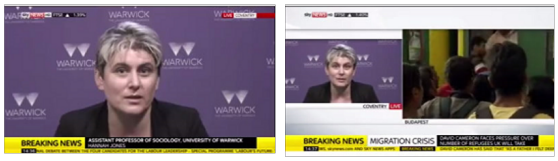
	B	M	R	C
pred topic TGM	0.4397	0.2921	0.6234	0.4970
anno topic TGM	**0.4548**	**0.3009**	**0.6339**	**0.5418**

and examples in Figure 7 presents the captioning performance with the annotated topics. Both show that with more accurate topic information the captioning performance can be further improved.

Influence of Multi-Modalities: As an implicit evaluation, we can evaluate the topic prediction performance with the annotated topics on testing set. The prediction accuracies with different modalities are shown in Figure 8. The performance of the aural and speech modality are evaluated only on videos containing the corresponding modality. Though the aural modality alone do not perform well, the ensemble of aural with image and motion modalities improves the prediction significantly with an absolute 6% boost. By further



Pred topic: a man is showing how to make a dish.
Anno topic: a man is talking about a snake.
GT: a man is talking about pigments showing a white snake.



Pred topic: there is a man in a suit is talking to a woman.
Anno topic: a man in black shirt is presenting the latest news.
GT: a person is showing information on the screen.

Figure 7: Examples on testing set. Pred topic and anno topic denote using the predicted and annotated topics in the topic-guided model. GT is a random groundtruth caption sentence.

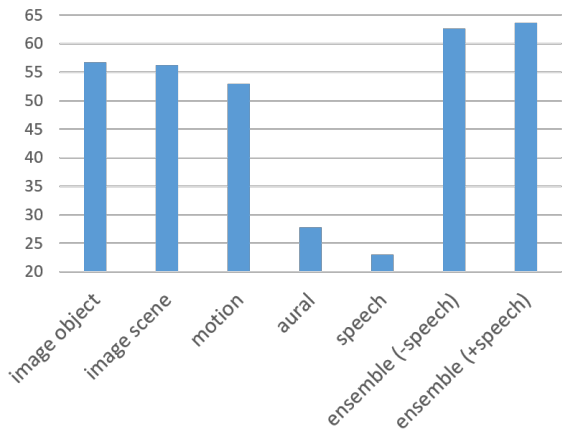


Figure 8: Accuracy of topic prediction with different modalities and ensembles. (-speech) and (+speech) means ensemble without and with speech modality.

using the speech modality predictions, the accuracy is improved from 62.61% to 63.65% on testing set.

To explicitly explore the usefulness of speech modality in video captioning, we use two kinds of predicted topics which are obtained by the fusion with or without the predictions from speech modality. Results are presented in Table 5. We can see that the captioning performance achieves large gain in all metrics although the speech modality gets only 1.04% absolute prediction accuracy improvement as mentioned above. It mainly results from the similar topic probability distributions using the shared topic mining model for speech modality. So the speech modality is quite useful to generate topic proposals and thus boost the captioning performance.

Table 5: Captioning performance with or without using speech modality for topic prediction. The acronyms B, M, R, C denote BLEU@4, METEOR, ROUGE-L and CIDEr respectively.

	B	M	R	C
w/o speech	0.4266	0.2910	0.6192	0.4874
with speech	**0.4397**	**0.2921**	**0.6234**	**0.4970**

Table 6: Captioning performance with different numbers of topics. The acronyms B, M, R, C denote BLEU@4, METEOR, ROUGE-L and CIDEr respectively.

#topics	B	M	R	C
10	0.4258	0.2889	0.6196	0.4789
20	**0.4397**	**0.2921**	**0.6234**	**0.4970**
30	0.4229	0.2901	0.6169	0.4799

Influence of the Number of Topics: We also explore the performance of TGM with different numbers of topics. As shown in Table 6, the number of topics 20 achieves the best performance, which is the balance between the topic prediction performance and the topic guidance performance. When there is fewer number of topics, the accuracy of topic prediction is higher but it provides less guidance to generate video descriptions. When there is more number of topics, though the topic guidance becomes strong, the captioning performance suffers from the low topic prediction accuracy. Since the more the topics are the more the topics resemble semantic concepts, it suggests that using a small number of topics is enough for the video captioning task and are superior to the large number of detected concepts.

6 Conclusions

Descriptions of videos with diverse topics vary a lot in vocabularies and expression styles. In this paper, we propose a novel topic-guided model to deal with the topic diversity nature of videos. It can generate better topic related descriptions for videos in various topics. Our experimental results show that the topic information is very useful to guide the caption model for more topic appropriate description generation and topics automatically mined in data-driven way are superior to the predefined topics as the topic guidance. Multimodal features especially the speech modality features are vital to predict topics for testing videos. Our proposed topic-guided model which functions as an ensemble of topic-aware language decoders can utilize the topic information more effectively than other caption models. It significantly improves the multimodal baseline performance on the current largest video caption dataset MSR-VTT, outperforming the winning performance in the 2016 MSR video to language challenge. In the future work, we will continue to improve the topic prediction performance and jointly learn the topic representation and caption generation end-to-end.

7 Acknowledgments

This work is supported by National Key Research and Development Plan under Grant No. 2016YFB1001202.

References

[1] Rémi Lebret, Pedro H. O. Pinheiro, and Ronan Collobert. Phrase-based image captioning. In *ICML*, pages 2085–2094, 2015.
[2] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, 2015.
[3] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv:1502.03044*, 2(3):5, 2015.
[4] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. *arXiv:1603.03925*, 2016.
[5] Qin Jin, Jia Chen, Shizhe Chen, Yifan Xiong, and Alexander Hauptmann. Describing videos using multi-modal fusion. In *ACM*, pages 1087–1091, 2016.
[6] Lei Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? *NIPS*, pages 2654–2662, 2013.
[7] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473*, 2014.
[8] Marcus Rohrbach, Wei Qiu, Ivan Titov, Stefan Thater, Manfred Pinkal, and Bernt Schiele. Translating video content to natural language descriptions. In *ICCV*, pages 433–440, 2013.
[9] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. Translating videos to natural language using deep recurrent neural networks. *Computer Science*, 2014.
[10] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly modeling embedding and translation to bridge video and language. In *CVPR*, pages 4594–4602, 2016.
[11] Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton van den Hengel. What value do explicit high level concepts have in vision to language problems? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 203–212, 2016.
[12] Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. Semantic compositional networks for visual captioning. In *CVPR*, 2017.
[13] Pradipto Das, Chenliang Xu, Richard F. Doell, and Jason J. Corso. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *CVPR*, pages 2634–2641, 2013.
[14] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *TACL*, 1:25–36, 2013.
[15] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016.
[16] Msr video to language challenge. *http://www.acmmm.org/2016/wp-content/uploads/2016/04/ACMMM16_GC_MSR_Video_to_Language_Updated.pdf*.
[17] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher J. Pal, Hugo Larochelle, and Aaron C. Courville. Describing videos by exploiting temporal structure. In *ICCV*, pages 4507–4515, 2015.
[18] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *ICCV*, pages 4534–4542, 2015.
[19] Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, and Yueting Zhuang. Hierarchical recurrent neural encoder for video representation with application to captioning. *arXiv:1511.03476*, 2015.
[20] Qin Jin and Junwei Liang. Video description generation using audio and visual cues. In *ICMR*, pages 239–242. ACM, 2016.
[21] Vasili Ramanishka, Abir Das, Dong Huk Park, Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, and Kate Saenko. Multimodal video description. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 1092–1096. ACM, 2016.
[22] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
[23] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. 2016.
[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.
[25] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Antonio Torralba, and Aude Oliva. Places: An image database for deep scene understanding. *arXiv:1610.02055*, 2016.
[26] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497. IEEE, 2015.
[27] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.
[28] Stephanie Pancoast and Murat Akbacak. Softening quantization in bag-of-audio-words. In *ICASSP*, pages 1370–1374. IEEE, 2014.
[29] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image classification with the fisher vector: Theory and practice. *International journal of computer vision*, 105(3):222–245, 2013.

[30] Ibm watson speech to text api. http://www.ibm.com/watson/developercloud/speech-to-text.html.

[31] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *Computer Science*, 14(7):38–39, 2015.

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

[33] Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. Guiding the long-short term memory model for image caption generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2407–2415, 2015.

[34] R Memisevic and G Hinton. Unsupervised learning of image transformations. In *CVPR*, pages 1–8, 2007.

[35] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[36] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[37] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *In Proceedings of the Ninth Workshop on Statistical Machine Translation*. Citeseer, 2014.

[38] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8. Barcelona, Spain, 2004.

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