This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2729019, IEEE Transactions on Multimedia

Video Captioning with Attention-based LSTM and Semantic Consistency

Lianli Gao, Zhao Guo, Hanwang Zhang, Xing Xu and Heng Tao Shen, *Senior Member, IEEE*

***Abstract*—Recent progress in using Long Short-Term Memory (LSTM) for image captioning has motivated the exploration of their applications for automatically describing video content with natural language sentences. By taking a video as a sequence of features, LSTM model is trained on video-sentence pairs and learns to associate a video to a sentence. However, most existing methods compress an entire video shot or frame into a static representation, without considering attention mechanism which allows for selecting salient features. Furthermore, existing approaches usually model the translating error, but ignore the correlations between sentence semantics and visual content. To tackle these issues, we propose a novel end-to-end framework named aLSTMs, an attention-based LSTM model with semantic consistency, to transfer videos to natural sentences. This framework integrates attention mechanism with LSTM to capture salient structures of video, and explores the correlation between multi-modal representations (i.e., words and visual content) for generating sentences with rich semantic content. Specifically, we first propose an attention mechanism which uses the dynamic weighted sum of local 2D Convolutional Neural Network (CNN) representations. Then, a LSTM decoder takes these visual features at time t and the word-embedding feature at time t-1 to generate important words. Finally, we uses multi-modal embedding to map the visual and sentence features into a joint space to guarantee the semantic consistence of the sentence description and the video visual content. Experiments on the benchmark datasets demonstrate that our method using single feature can achieve competitive or even better results than the state-of-the-art baselines for video captioning in both BLEU and METEOR.**

***Index Terms*—LSTM, Attention Mechanism, Embedding, Video Captioning.**

I. Introduction

PREVIOUSLY, visual content understanding [1], [2], [3] and Natural Language Processing (NLP) [4] are not correlative with each other. Integrating visual content with language learning to generate descriptions for images, especially for videos, has been regarded as a challenging task and is a critical step towards machine intelligence and many applications in daily scenarios such as image/video retrieval [5], [6], video understanding [7], blind navigation [8] and automatic video subtitling [9].

Thanks to the rapid development of deep Convolutional Neural Network (CNN) [10], [11], recent works have made significant progress for image description generation [8], [12], [13], [14], [15]. However, compared with image captioning, video captioning is more difficult for the diverse sets of objects, scenes, actions, attributes and salient contents. Despite the difficulty of video captioning, there have been a few attempts [16], [17], [18], [19], [20], which are mainly inspired by recent advances in translating with Long Short-Term Memory (LSTM) [21]. The LSTM is proposed to overcome the vanishing gradients problem by enabling the network to learn when to forget previous hidden states and when to update hidden states by integrating memory units. LSTM has been successfully adopted to several tasks, e.g., speech recognition, language translation and image captioning [16]. Thus, we follow this elegant recipe and choose to extend LSTM to generate the video sentence with rich semantic content.

Attention networks are currently a standard part of the deep learning toolkit, contributing to impressive results in neural machine translation [4], visual captioning [13], [18] and question answering [22]. This approach alleviates the bottleneck of compressing a source into a fixed-dimensional vector by equipping a model with a variable-length memory, thereby providing random access into the source is needed. In addition, attention is implemented as a hidden layer which computes a categorical distribution to make a soft-selection over source elements. Thus we incorporate an attention based LSTM model to capture salient temporal structures of videos. Recently, many applications were proposed [16], [17], [18], [19] to directly connect a visual convolution model to deep LSTM networks. For example, Venugopalan *et al.* [16] translate videos to sentences by directly concentrating a deep neural network with a recurrent network. Ideally, video description not only requires modeling and integrating their sequence dynamic temporal attention information into a natural language but also needs to take into account the relationship between sentence semantics and visual content, which to our knowledge has not been simultaneously considered.

Therefore, in this paper we propose a unified framework (see Fig. 1), named aLSTMs, an attention-based LSTM model with semantic consistency. Firstly, to extract more meaningful spatial features, we adopt Inception-v3 neural network [11] which is an extended version of GoogleNet [10]. To exploit temporal information, we introduce one-layer LSTM visual encoder to encode those spatial 2D CNN feature vectors. Then we propose an attention mechanism which takes the dynamic weighted sum of local spatial 2D CNN feature vectors as the input for the LSTM decoder. Finally, we integrate multi-word embedding and cross-view methodology to project the generated words and the visual features into a common space to bridge the semantic gap between videos and the corresponding sentences. It is worthwhile to highlight the following aspects

Lianli Gao, Zhao Guo, Xing Xu and Heng Tao Shen are with the Center of Future Media, School of Computer Science and Engineering, University of Electronic Science and Technology of China, 611731. Hanwang Zhang is with Department of Computer Science, Columbia University, USA. Heng Tao Shen is the correspondence author. Email:shenhengtao@hotmail.com

of the proposed approach:

- Our method incorporates the attention mechanism which allows for salient features by using the dynamic weighted sum of 2D CNN feature representations at frame level.
- We use cross-view model to enforce the consistency between the generated sentence features and visual features. Therefore, aLSTMs can simultaneously explore the learning of LSTM and visual-semantic embedding.
- Experiments on the benchmark datasets demonstrate that our method achieves comparable or even better results than the state-of-the-art methods for video captioning in both BLEU and METEOR.

The rest of the paper is organized as follows. Related work is discussed in Section II. Section III describes the details of the proposed approach. We present the experimental settings and results in Section IV. Finally, Section V concludes the paper.

II. Related Work

A. Image/Video Recognition

Recognition of image and video is a fundamental and challenging problem in computer vision [23], [24], [25]. Dramatic progress has been achieved by supervised convolutional models on image-based action recognition task [26], [27], [28]. Rapid progress has been made in the past few years, especially in image feature learning, and various pre-trained CNN models are proposed. However, such image based deep features cannot be directly applied to process videos due to the lack of dynamic information. Du *et al.* [27] propose to learn spatio-temporal features using deep 3D CNN and shows good performance on various video analysis tasks. To effectively learn the spatial-temporal signals and features, Sun *et al.* [26] propose a new deep architecture, called factorized spatio-temporal convolutional networks, which factorizes the original 3D spatio-temporal convolution kernel learning as a sequential process of learning 2D spatial kernels in the lower network layers. Thanks to the emergence of LSTM [21], it is able to model sequence data and learn patterns with wider range of temporal dependencies. Donahue *et al.* [28] integrate CNN and LSTM to learn spatio-temporal information from videos. It extracts 2D CNN features from video frames and then the 2D CNN features are fed into a LSTM network to encode the videos' temporal information.

B. Image/Video Captioning

To further bridge the gap between video/image understanding and natural language processing, generating description for image or video becomes a hot research topic. It aims to generate a sentence to describe the image/video content. Due to the development of Recurrent Neural Network (RNN) [29] and LSTM, researchers have striven to automatically describe an image/video with a correct and novel natural language sentence [30], [31], [20], [32], [33], [34], [35], [36], [37], [38].

Inspired by the advantages in multi-modal learning and machine translation, Ryan *et al.* [32] construct a joint multi-modal embedding approach to project image features extracted by a deep CNN model and text features encoded by a LSTM network to a common space. Then, a decoder is applied to decode image content into visual sentences using structure-content neural language model. In [39], they develop a so-called correlation component manifold space learning (CCMSL) to learn a common feature space by capturing the correlations between the heterogeneous databases. Karpathy *et al.* [30] leverage dataset of images and their sentence descriptions to learn about the inter-modal correspondences between language and visual data. It firstly exploits two modalities through a multi-modal embedding to align regions of image and snippets from corresponding sentence. Next, it uses image-sentence and region-snippet pairs to train a multi-modal recurrent neural network to generate novel descriptions of regions and images. In [12], a multi-modal Recurrent Neural Networks (m-RNN) model is proposed for image captioning, which directly models the probability of generating a word given previous words and images. Vinyals *et al.* [8] propose an end-to-end neural network system to generate sentences for images via integrating LSTM and GoogleNet [10]. In [20], an attribute with high-level concepts is incorporated into a CNN-RNN network as the external input. This work provides a fully trainable attribute-based deep neural network, which yields significantly good performance. Justin *et al.* [31] introduce a dense captioning approach, which not only detects object region proposals but also generalizes phrases and sentences to describe image region and full image content respectively. It simultaneously takes the object detection and description task into account, and proposes a Fully Convolutional Localization Network (FCLN) architecture that processes an image with a single efficient forward pass, requiring no external regions proposals. This network can be trained end-to-end with a single round of optimization.

Following image captioning, there are several researches [16], [17], [40], [41], [19] focusing on video captioning. In [16], it firstly proposes an end-to-end LSTM-based model for video-to-text generation. This work only leverages the local 2D CNN feature from frame-level, then performs mean pooling over 2D features across each video to form a fixed-dimensional video-level feature. Compared with image content, video has both spatial and temporal structure. In order to efficiently translate video to language, approaches should take both temporal and spatial information into account. Inspired by this, an end-to-end sequence-to-sequence model [17] is proposed to generate captions for videos. It incorporates a stacked LSTM which firstly reads the sequence of CNN outputs and then generates a sequence of words. Pan *et al.* [41] propose a novel approach, namely Hierarchical Recurrent Neural Encoder (HRNE), which exploits multiple time-scale abstraction of the temporal information with two-layer LSTMs network. Pan *et al.* [40] propose a framework which explores the learning of LSTM, and aims to locally maximize the probability of the next word given previous words and visual content features. The visual features are generated by a VGGNet and C3D network. In addition, it proposes a visual-semantic embedding, which enforces the relationship between the entire sentence semantics and the visual content. To obtain the most representative and high-quality description for a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2729019, IEEE Transactions on Multimedia

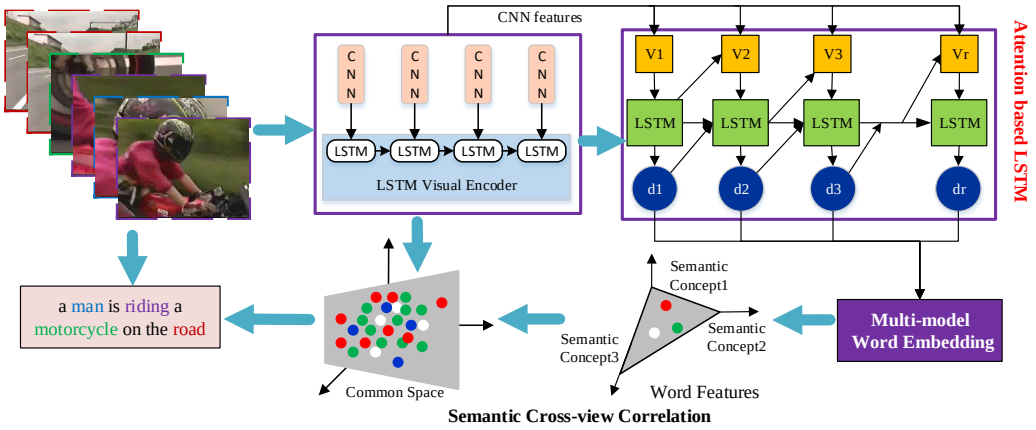

Fig. 1. The framework of our proposed method aLSTMs which consists of LSTM visual encoder, attention-based LSTM and semantic cross-view correlation. To illustrate the effectiveness of our aLSTMs framework, the generated words within a sentence are marked with different colors, which correspond to the frames marked with the same color.

target video, Li *et al.* [19] propose a summarization-based video captioning method, which constructs an adjacency graph on sentence sequences, then adopts this graph to re-rank the generated candidate sentences.

Attention networks have proven to be an effective approach for embedding categorical inference within a deep neural network. This mechanism, which learns to automatically select the most relevant source data to generate output data, has made a great success in machine translation [4], visual question answering [22] and video/image captioning [13], [18], [42]. Since not all source words in a sentence are equally salient for machine translation, and also the generated word is usually relevant to a subset of source words, it is important for a model to identify the importances or weights of source words for translation. In [4], two types of attention-based model are proposed for machine translation: a global mechanism in which all source words are attended, and a local one whereby only a subset of source words are considered at a time. Yang *et al.* [22] introduce a multiple-layer stack attention neural network to answer questions according to an image. In [13], two attention-based LSTM models are proposed. They are capable of well aligning the most relevant visual content to the next word of the sentence. In [42], a set of visual concepts corresponding to each image are firstly obtained by running a set of attribute detectors. Next it learns to selectively attend to semantic concept proposals and feds them into hidden states of recurrent neural networks. Since not all frames in a video are equally salient for a short description generation and an event may last in multiple frames, it is important for a model to identify which frames are more salient. In [43], they propose a content similarity based fast reference frame selection algorithm for reducing the computational complexity of the multiple reference frames based inter-frame prediction. In [18], they propose a temporal attention based LSTM model which combines local temporal modeling to automatically select the most relevant temporal segments to generate the next word.

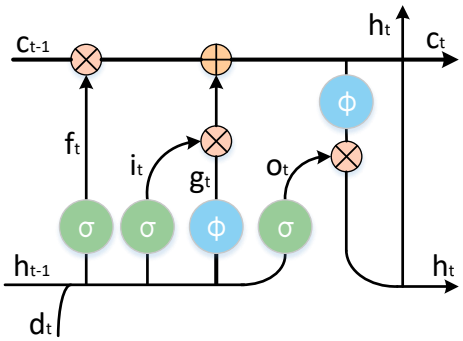

Fig. 2. An illustration of the LSTM unit.

III. The Proposed Approach

Our task is to generate language sentences for videos. In this section, we first define the terms and notations. Next, we introduce our aLSTMs approach. An objective function is built by integrating two loss functions which simultaneously consider video translation and semantic consistency. Specifically, one loss function aims to guarantee the translation from videos to words, while another loss function tries to bridge the semantic gap with semantic cross-view correlations. The detailed information about solution is given as well.

A. Terms and Notations

Suppose we have a video $\mathbf{V}$ to be described by a textual sentence $D = \{d_1, \cdots, d_{N_d}\}$ consisting of N_d words. Let $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_{N_x}\} \in \mathbb{R}^{M \times N_x}$ and $\mathbf{D}' = \left\{\mathbf{d}'_1, \cdots, \mathbf{d}'_{N_d}\right\} \in \mathbb{R}^{L \times N_d}$ denote the visual and the textual features, where $\mathbf{d}'_i = \mathbf{E}d_i$ is the word representation of a single word d_i and N_x is the total number of feature vectors. Let M and L denote the dimension of visual feature and textual feature. $\mathbf{X}$ is extracted using deep neural networks, which will be described in the experiment.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2729019, IEEE Transactions on Multimedia

B. Attention-based Long Short-Term Memory Decoder

To date, modeling sequence data with recurrent neural network has been proven successful in the process of machine translation, speech recognition, image/video captioning [16], [17] etc. However, it is still difficult to train a standard RNN due to the vanishing gradient problem [44]. LSTM, an updated version of standard RNN, solved this issue by learning patterns with wider range of temporal dependencies. As videos and natural sentences are both sequential data, LSTM is applied as the basic component for our aLSTMs.

The main idea of Attention-based Long Short-Term Memory is to integrate attention mechanism into the LSTM. A basic LSTM unit (see Fig. 2) consists of a single memory cell, an input activation function, and three gates (input $\mathbf{i}_t$, forget $\mathbf{f}_t$ and output $\mathbf{o}_t$). $\mathbf{i}_t$ allows incoming signal to alter the state of the memory cell or block it. $\mathbf{f}_t$ controls what to be remembered and what to be forgotten by the cell and somehow can avoid the gradient from vanishing or exploding when back propagating through time. Finally, $\mathbf{o}_t$ allows the state of the memory cell to have an effect on other neurons or prevent it. Basically, the memory cell and gates in a LSTM block are defined as follows:

$$\begin{aligned}\begin{pmatrix}\mathbf{i}_t\\ \mathbf{f}_t\\ \mathbf{o}_t\\ \mathbf{g}_t\end{pmatrix} &= \begin{pmatrix}\sigma\\ \sigma\\ \sigma\\ \phi\end{pmatrix} Z_{L+r,r}\begin{pmatrix}\mathbf{E}d_{t-1}\\ \mathbf{h}_{t-1}\end{pmatrix}\\ \mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t\\ \mathbf{h}_t &= \mathbf{o}_t \odot \phi(\mathbf{c}_t)\end{aligned} \tag{1}$$

where $\mathbf{E}$ denotes an embedding matrix, σ represents the logistic sigmoid non-linear activation function mapping real numbers to $(0, 1)$ and can be thought as knobs that LSTM learns to selectively forget its memory or accept current input, ϕ denotes the hyperbolic tangent function tanh, $\odot$ is the element-wise product with the gate value, $Z_{L+r,r}$ denotes the parameters of the LSTM. Let L and r denote the embedding and LSTM dimensionality respectively.

Compared with images, videos contain more complex temporal information which should be aligned to language data. Thus we extend the attention mechanism (see Fig. 3) introduced by [13] to support video captioning. The new form of LSTM is defined as:

$$\begin{pmatrix}\mathbf{i}_t\\ \mathbf{f}_t\\ \mathbf{o}_t\\ \mathbf{g}_t\end{pmatrix} = \begin{pmatrix}\sigma\\ \sigma\\ \sigma\\ \phi\end{pmatrix} Z_{L+r+M,r}\begin{pmatrix}\mathbf{E}d_{t-1}\\ \mathbf{h}_{t-1}\\ \mathbf{v}_t\end{pmatrix} \tag{2}$$

$$\begin{aligned}\mathbf{v}_t &= \sum_{i=1}^{N_x}\beta_i^t\mathbf{x}_i, \quad s_i^t = W_s\phi\left(W_h\mathbf{h}_{t-1} + W_x\mathbf{x}_i + b_s\right)\\ \beta_i^t &= \frac{\exp\left(s_k^t\right)}{\sum_{k=1}^{N_x}\exp\left(s_k^t\right)}, \text{s.t,} \quad \sum_{i=1}^{N}\beta_i^t = 1\end{aligned} \tag{3}$$

where $\mathbf{v}_t$ represents context vector which is a dynamic representation of the relevant representation of the video input at time t. M is the dimension of $\mathbf{v}_t$. In addition, β_i^t is the attention weights at time t describing the relevance of the i-th feature in the input video. Given the previous hidden state $\mathbf{h}_{t-1}$ of the LSTM decoder and the i-th video feature, it returns the unnormalized relevance score s_i^t. Once the relevance scores for all the features $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_{N_x}\}$ are computed, the LSTM is able to obtain β_i^t at each time step t. The W_s, W_h, W_x and b_s are the parameters to be estimated.

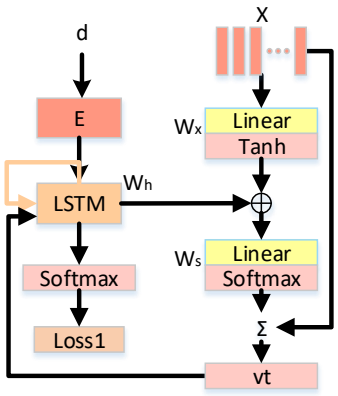

Fig. 3. An illusion of our temporal attention mechanism in LSTM decoder process.

In addition, to capture rich temporal information, we introduce an one-layer LSTM visual encoder, called LSTM Visual Encoder. The LSTM recently has made a great success in the process of action recognition. Inspired by this, in our framework we propose to integrate the updated GoogleNet with one-layer LSTM visual encoder to encode video temporal information. Specifically, the last output of the LSTM Visual Encoder is used to initialize the first LSTM unit of our attention based LSTM network to facilitate video captioning.

C. Loss 1: Translation from Videos to Words

In the LSTM decoding phase, the LSTM computes context vector $\mathbf{v}_t$ given an input sequence $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_{N_x}\}$ and the hidden state $\mathbf{h}_{t-1}$. Inspired by the principle of translating images, we treat the activation value indexed by a training word d_t in the softmax layer of our sentence generator as the likelihood of generating that word:

$$P(d_t|\mathbf{v}_t; d_1, d_2, \cdots, d_{t-1}; \mathbf{E}) \tag{4}$$

The cost of generating that training word is then defined as the negative logarithm of the likelihood. We further define the cost of generating the words as:

$$Loss_1 = -\sum_{t=1}^{N_d} log(P(d_t|\mathbf{v}_t; d_1, d_2, \cdots, d_{t-1}; \mathbf{E})) \tag{5}$$

where N_d denotes the total number of words in sentence and d_i denotes the i-th word in sentence D. By minimizing $Loss_1$, the contextual relationship among the words in the sentence can be guaranteed, making the sentence coherent and smooth.

D. Loss 2: Bridging the Semantic Gap with Semantic Cross-view Correlation

Given a set of video visual features $\mathbf{X}$, we perform one-layer LSTM process over all the feature vectors across each video

to generate a single M-dimensional visual feature $\mathbf{X}_e$ with spatial and temporal information. Simultaneously, we perform "mean pooling" process over all the embedding vectors $\mathbf{D}^{'}$ for constructing a single L-dimensional sentence feature $\mathbf{D}^{'}{}_{mean}$ for each description corresponding to the video (usually $M \neq L$).

Then, the visual and sentence features are mapped into a common high level abstract space by a linear projection, which is the simplest isomorphic function:

$$R_I : R^M \rightarrow A^{\mathrm{C}} \qquad R_D : R^L \rightarrow A^{\mathrm{C}} \tag{6}$$

where $R_I \in R^{M \times C}$ and $R_D \in R^{L \times C}$ are linear projection matrix, R^M and R^L are video feature space and description feature space respectively. In order to construct a cross-correlation between two modalities, we require visual and sentence features of the same instance to be equal in A^{C}:

$$R_I\left(\mathbf{X}_e\right) = R_D\left(\mathbf{D}^{'}{}_{mean}\right) \tag{7}$$

Here, we let $M = C$, then the Eq.7 can be rewritten by left multiplication inverse R_I:

$$X_e = {R_I}^{-1} R_D\left(\mathbf{D}^{'}{}_{mean}\right) = R\mathbf{D}^{'}{}_{mean}, \forall i \tag{8}$$

where $R = R_I^{-1} R_D$ is the linear projection matrix. Therefore, we can approximate the above formulate by optimizing the cross-correlation:

$$Loss_2 = \left\| X_e - R\mathbf{D}^{'}{}_{mean} \right\|_F^2 \tag{9}$$

By minimizing the semantic correlation score, the semantic consistence between the generated words and the video visual context can be guaranteed, making the sentence with rich semantic context.

E. Attention based LSTM with Semantic Cross-view Correlation

In our video captioning problem, on one hand, the generated descriptive words must be able to depict the main contents of a video precisely, and on the other hand, the generated words should be organized coherently in language. Therefore, we formulate the video captioning problem by minimizing the following energy loss function:

$$E(\mathbf{X}, D) = (1 - \lambda)\, Loss_2 + \lambda Loss_1 \tag{10}$$

where λ is the trade-off parameter between our two losses.

Sentence generation: Sampling and BeamSearch are two main approaches that used to generate a sentence given a video [13]. In this paper, we choose BeamSearch method which iteratively considers the set of the k best sentences up to time t as candidates to generate sentence of time $t + 1$, and keeps only best k results of them. Finally, we approximates $D = argmax_{D'} Pr(D^{'}|X)$ as our best generated description, where Pr indicates the probability. In our entire experiment, we set the beam size of BeamSearch as 5.

IV. Experiments

We evaluate our algorithm on the task of video captioning. We first study the influence of parameters on aLSTMs, and then compare our results with the state-of-the-art methods.

A. Datasets

We consider two publicly available datasets that have been widely used in previous work.

The Microsoft Video Description Corpus (MSVD). This video corpus consists of 1,970 short video clips, which is well suited for training and evaluating an automatic video description generation model. In total, this video dataset has approximately 80,000 description pairs and about 16,000 vocabulary words, which are provided by Microsoft Research [45]. Following [18], [17], we split the data set into a training, a validation and a testing set with 1,200, 100 and 670 videos, respectively.

MSR Video to Text (MSR-VTT). In 2016, Xu *et al.* [46] proposed a new large-scale video benchmark for video understanding and especially video captioning tasks. Specifically, this dataset contains 10,000 web video clips, and each clip is annotated with approximately 20 natural language sentences, which are generated by 1,327 AMT workers. In addition, it covers the most comprehensive categorizes (i.e., 20 categories) and a wide variety of visual content, and contains 200,000 clip-sentence pairs. To date, it represents the largest dataset in terms of vocabulary and sentence. Following [46], this dataset is split according to 65%:30%:5%, resulting in 6,513, 2,990 and 497 clips in the training, testing and validation sets, respectively.

B. Implementation Details

Preprocessing. For MSVD dataset, we firstly convert all descriptions to lower cases, and then use wordpunct_tokenizer method from NLTK [1] toolbox to tokenize sentences and remove punctuations. Therefore, it yields a vocabulary of 15,903 in size for the training split. For MSR-VTT dataset, captions have been tokenized, thus we directly split descriptions using blank space, thus it yields a vocabulary of 23,662 in size for training split.

Inspired by [18], we pre-process each video clip by firstly selecting equally-spaced 28 frames out of the first 360 frames and then feeding them into a CNN network proposed in [18]. Thus, for each selected frame we obtain a 2048-dimensional feature vector, which are extracted from the $pool3$ layer.

Training details. In the training phase, in order to deal with sentences with varying lengths, we add a begin-of-sentence tag <START> to start each sentence and an end-of-sentence tag <END> to end each sentence. In the testing phase, we input <START> tag into our attention-based LSTM to trigger video description generation process. For each word generation, we choose the word with the maximum probability and stop until we reach <END>.

In addition, all the LSTM unit sizes are set as 512 and the word embedding size is set as 512, empirically. Our objective function 10 is optimized over the whole training video-sentence pairs with mini-batch 64 in size of MSVD and 256 in size of MSR-VTT. We adopt adadelta [47] which is an adaptive learning rate approach to optimize our loss function. We utilize dropout regularization with the rate of 0.5 in all layers and clip gradients element wise at 10. we stop training

[1] http:/s/www.nltk.org/index.html

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2729019, IEEE Transactions on Multimedia

TABLE I
THE EFFECT OF GOOGLENETS. THE PERFORMAMANCE IS REPORTED OVE THE MSVD DATASET.

Evaluation Metrics	GoogleNet based aLSTMs	Inception-v3 based aLSTMs
B@1	78.9	**81.8**
B@2	66.1	**70.8**
B@3	55.9	**61.1**
B@4	45.3	**50.8**
METEOR	32.3	**33.3**
CIDEr	68.7	**74.8**

our model until 500 epoches are reached or until the evaluation metric does not improve on the validation set at the patience of 20.

Evaluation metrics. To evaluate the performance, we employ three different standard evaluation metrics: BLUE [48], METEOR [49] and CIDEr [50]. More specifically, for the MSVD dataset, the prior results of most of the three metrics have been reported in previous studies. For MSR-VTT, previous approaches [16], [18], [46] only report performance in terms of BLUE and METEOR.

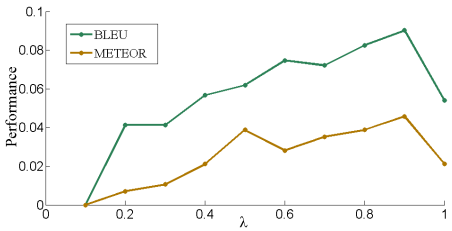

Fig. 4. The effect of the λ on the MSVD dataset.

C. The Study of Influence Factors

The effect of λ. In this sub-experiment, we aim to study the effect of λ. Since λ is the trade-off between the loss of translation from video to words and the loss of semantic gap between video and sentence, it affects the performance of our algorithm. In this subsection, we study the performance variance with different λ values. Specifically, in order to make the performance curves falling into a comparable scale, we normalize BLEU and METEOR scores with the following function:

$$P^{'} = \frac{P - min(P)}{min(P)} \tag{11}$$

where P and $P^{'}$ represent the original and normalized performance values.

We tune λ from 0.1 to 1 on the MSVD dataset and the results are shown in Fig.4. When λ is relatively small, e.g., 0.1, the worst performance is achieved. That is to say, if the impact of semantic cross-view correlation loss is dominant in constructing the final loss, the performance is unsatisfactory. With the increase of λ, the performance is improved until reaching the peak at $\lambda = 0.9$. Further raising λ will result in a slight drop. In the following experiments, we set $\lambda = 0.9$.

The effect of GoogleNets. Inception-v3 neural network [11] is a upgrade version of GoogleNet [10]. In this sub-experiment, we study the influence of above two different versions of GoogleNet in our framework on the MSVD dataset and the results are show in Tab. I. From Tab. I, we have following observations:

- Inception-v3 based aLSTMs performs better than GoogleNet based aLSTMs in terms of all evaluation metrics.
- Compared with GoogleNet based aLSTMs, the performance of Inception-v3 based aLSTMs is improved by average 4.57% in terms of BLUE, 1.0% in terms of METEOR and 6.1% in terms of CIDEr. This indicates that Inception-v3 is able to extract more informative visual features than GoogleNet for our aLSTMs.

The effect of LSTM visual encoder and semantic cross-view correlation. In this sub-experiment, our task is to evaluate the effect of LSTM visual encoder as well as semantic cross-view correlation on our aLSTMs. Specifically, the 2D-CNN has been proved to contain rich spatial information. However, a video not only contains spatial but also temporal structure information. Thus, in this study we proposed a LSTM visual encoder to encode the video temporal information to support video captioning. As mentioned in Section III, our approach integrates multi-word embedding with a cross-view methodology to project the generated words and the visual feature into a common space which can enhance video captioning using semantic cross-view correlation. Therefore, this sub-experiment is conducted on both datasets and the results are shown in Tab. II. We can make the following observations:

- For both datasets, our aLSTMs outperforms the aLSTMs without one-layer LSTM visual encoder. For aLSTMs without visual encoder, the aLSTMs takes the average mean pooling value of 28 frames feature representations to initialize the first LSTM unit of the attention network. Specifically, taking advantage of visual LSTM unit to encode temporal information can perform better with 3.0% (B@4) and 0.6% (METEOR) increases in MSVD dataset and 0.5% (B@4) and 0.1% (METEOR) increases in MSR-VTT dataset. The results indicate that LSTM visual encoder can provide more accurate video temporal information.
- Our aLSMs achieves better results than aLSTMs without considering the semantic cross-view correlation. Specifically, semantic cross-view correlation based method can achieve better results (with 3.5% (B@4) and 0.5% (METEOR) increases in MSVD dataset and 1.3% (B@4) and 0.4% (METEOR) increases in MSR-VTT dataset). It is possibly due to that semantic cross-view correlation loss guarantees the consistency between visual and sentence information.

D. Comparison to the State-of-the-Art Methods

Baselines. To evaluate the performance of our proposed algorithm, we compare it with the following baselines.

- Factor Graph Model (FGM) [51]: FGM combines a compositional semantics language model in the dependency-tree structure with a deep video model, which is followed by two-layer neural network. Finally, it introduces a

TABLE II
THE EFFECT OF LSTM VISUAL ENCODER AND SEMANTIC CROSS-VIEW CORRELATION. THE PERFORMANCE IS REPORTED ON THE MSVD AND MSR-VTT DATASETS.

	MSVD						MSR-VTT		
Model	B@1	B@2	B@3	B@4	METEOR	CIDEr	B@4	METEOR	CIDEr
aLSTMs without LSTM Visual Encoder	80.1	68.1	58.1	47.8	32.7	72.4	37.5	26.0	39.2
aLSTMs without Semantic Cross-view Correlation	81.2	68.9	58.4	47.3	32.8	74.0	36.7	25.7	38.7
aLSTMs	**81.8**	**70.8**	**61.1**	**50.8**	**33.3**	**74.8**	**38.0**	**26.1**	**43.2**

jointly embedding model to deal with two tasks: video-to-text and text-to-video.

- MP-LSTM [16]: MP-LSTM attempts to directly translate video pixels to natural language sentences by utilizing both deep CNN network and two-layer LSTMs neural network. To generate video representation, the proposed method applies mean pooling over the features of all frames, and the temporal information is ignored.
- Summarization-based Video Captioning (S-VC) [19]: To obtain the most representative and high-quality descriptions for a target video, S-VC introduces a summarization-based video captioning which firstly constructs an adjacency graph on the sentence sequences, and then applied one-layer LSTM neural network to generate multiple sentences. Finally the generated adjacency graph is applied to re-rank the sentences.
- Soft-Attention (SA) [18]: To exploit the local structure, SA makes use of two types of features: frame-level representations extracted from the GoogleNet [10]; and video level features which are extracted by a 3-D ConvNet, which takes a concatenation vector as the input. More specifically, this vector is concatenated by a set of descriptors including, Histograms of Oriented Gradients (HOG), Histograms of Optical Flow (HOF), and Motion Boundary Histogram (MBH). Furthermore, a weighted attention mechanism is proposed to learn weights of specific temporal regions to enhance sentence generation.
- Sequence to Sequence - Video to Text (S2VT) [17]: An end-to-end sequence-to-sequence model for generating caption of videos is proposed in [17], named S2VT. It incorporates a stacked LSTM, which firstly takes the RGB and optical flow descriptions as the input, and then produces a sequence of CNN outputs, which are used to generate a sequence of words to describe the video.
- LSTM-E [40]: This approach explores the learning of LSTM, which aims to locally maximize the probability of the next word given previous words and visual content features. The content features are generated by a VGGNet and C3D network. In addition, LSTM-E introduces a visual-semantic embedding, which enforces the relationship between the entire sentence semantics and the visual content.
- p-RNN [52]: This approach introduces a hierarchical-RNN framework for describing a long video with a paragraph consisting of multiple sentences. This framework consists of two generators: 1) a sentence generator which produces single short sentences that describe specific time intervals and video regions, and 2) a paragraph generator which takes the sentential embedding as input and uses another recurrent layer to output the paragraph state. Such state is then used to initialize the sentence generator. In addition, both sentence and paragraph generators adopt recurrent layers for language modeling.
- HRNE [41]: In order to make use of temporal information of videos, this approach introduces a hierarchical recurrent neural encoder (HRNE) by utilizing a soft attention mechanism to generate captions for videos.

TABLE III
EXPERIMENT RESULTS ON THE MSVD DATASET. WE COMPARE OUR METHOD WITH THE BASELINES USING STATIC FRAME-LEVEL FEATURES ONLY IN THIS TABLE.

Model	B@1	B@2	B@3	B@4	METEOR	CIDEr
FGM	-	-	-	-	23.9	-
S2VT	-	-	-	-	29.2	-
MP-LSTM	-	-	-	33.3	29.1	-
S-VC	-	-	-	35.1	29.3	-
SA	-	-	-	40.3	29.0	51.7
LSTM-E	74.9	60.9	50.6	40.2	29.5	-
p-RNN	77.3	64.5	54.6	44.3	31.1	-
HRNE	78.4	66.1	55.1	43.6	32.1	-
HRNE-SA	79.2	66.3	55.1	43.8	33.1	-
aLSTMs	**81.8**	**70.8**	**61.1**	**50.8**	**33.3**	**74.8**

Results on MSVD dataset. In this subsection, we show the comparison of our approach with the baselines on the MSVD dataset. In addition, some of the above baselines only utilize video features generated by a single deep network, while others (i.e., S2VT, SA, LSTM-E and p-RNN) make uses of both single network and multiple network generated features. Therefore, in this subsection, we first compare our method with approaches using static frame-level features extracted by a single network, and then we compare our method with methods utilizing different deep features or their combinations.

Firstly, we compare our method with approaches on the MSVD dataset and the results are shown in Tab. III. All the compared approaches use static frame-level features only. More specifically, baselines including MP-LSTM, S-VC and LSTM-E ignore temporal dependencies along video sequences. The averages of frame-level features are adopted to represent videos. From Tab. III, we have the following observations:

- Compared with the best counterpart (i.e., p-RNN) which only takes spatial information, our method has 6.5% improvement on B@4 and 2.1% on METEOR.
- Hierarchy structure in HRNE reduces the length of input flow and composites multiple consecutive input at a higher level, which increases the learning capability and enables the model encode richer temporal information of multiple granularities. Our approach (50.8% B@4, 33.3% METEOR) also performs better than HRNE (43.6%

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2729019, IEEE Transactions on Multimedia

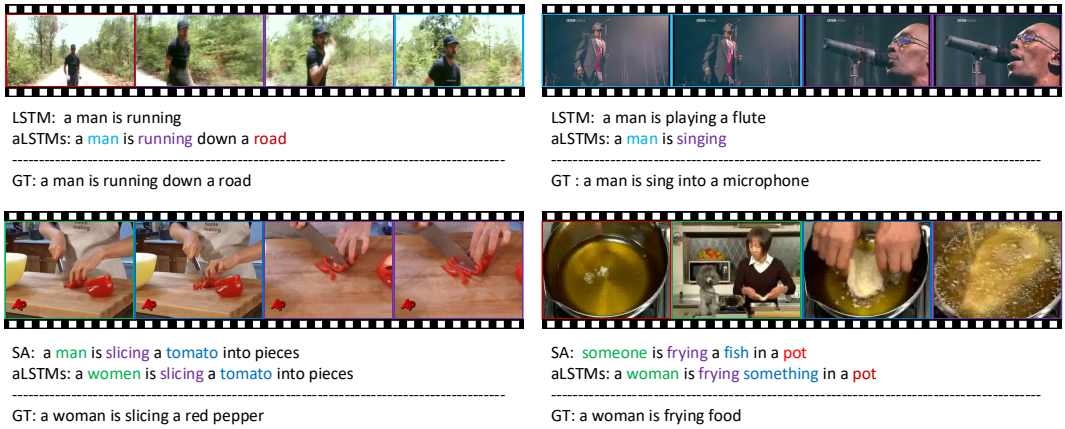

Fig. 5. Examples of video captioning results. The results of LSTM and SA are reported in [18]. We mark words and frames with different colors to show the importance of visual and semantic relationships.

TABLE IV
EXPERIMENT RESULTS ON THE MSVD DATASET. WE COMPARE OUR METHOD WITH THE BASELINES WITH MULTIPLE FEATURES.

Model	B@1	B@2	B@3	B@4	METEOR
S2VT(VGGNet+Optical flow)	-	-	-	-	29.8
SA(GoogleNet+3D-CNN)	-	-	-	41.9	29.6
LSTM-E(VGGNet)	74.9	60.9	50.6	40.2	29.5
LSTM-E(C3D)	75.7	62.3	52.0	41.7	29.9
LSTM-E(VGGNet+C3D)	78.8	66.0	55.4	45.3	31.0
p-RNN(VGGNet)	77.3	64.5	54.6	44.3	31.1
p-RNN(C3D)	79.7	67.9	57.9	47.4	30.3
p-RNN(C3D+VGGNet)	81.5	70.4	60.4	49.9	32.6
aLSTMs(Inception-v3)	**81.8**	**70.8**	**61.1**	**50.8**	**33.3**

B@4, 32.1% METEOR) and HRNE-SA (43.8% B@4, 33.1% METEOR)in video captioning task. This proves the effectiveness of our semantic gap factor.

- S2VT uses one-layer LSTM as video encoder to explore videos' temporal information. Our aLSTMs achieves better result than S2VT (33.3% vs 29.2% on METERO).
- SA uses attention-based LSTM model to explore the dynamic representation from videos. Our aLSTMs achieves better result than SA (74.8% vs 51.7% on CIDEr, respectively). This indicates that the semantic gap factor is beneficial to the effectiveness of video captioning.

Secondly, we compare our aLSTMs to other video captioning approaches including S2VT, SA, LSTM-E and p-RNN and the results are shown in Tab. IV. In this experiment, our aLSTMs only uses one Inception-v3 feature as the input but the compared approaches combine multiple deep CNN features. From Tab. IV, we have the following observations:

- Utilizing both spatial and temporal video information can enhance the video captioning performance. VGGNet and GoogleNet are used to generate spatial information, while optical flow and C3D are used for capturing temporal information. For example, compared with LSTM-E(VGGNet) and LSTM-E (C3D), LSTM-E(VGGNet+C3D) achieves higher performance (45.3% B@4 and 31.0% METEOR). In addition, for p-RNN, p-RNN(C3D+VGGNet) (49.9% B@4 and 32.6% METEOR) performs better than both p-RNN(VGGNet) (44.3% B@4 and 31.3% METEOR) and p-RNN(C3D) (47.4% B@4 and 30.3% METEOR).
- Our aLSTMs achieves the best results (50.8% B@4 and 33.3% METEOR). For S2VT(VGGNet+Optical flow), SA(GoogleNet+3D-CNN), LSTM-E(VGGNet+C3D) and p-RNN(C3D+VGGNet), they use two networks VGGNet/GoogleNet and optical flow/C3D to capture video's spatial and temporal information, respectively. Compared with them, our approach firstly utilizes Inception-V3, an updated version of GoogleNet, to capture frame-level features, and then integrates with a LSTM visual encoder to obtain video temporal information.

In Fig. 5, we show some examples of video captioning results. Four videos are used for demonstration and four frames are extracted from each video. The top two examples demonstrate the influence of temporal information for caption generation. It indicates that LSTM encoder can generate good sentences compared with the basic LSTM approach. In addition, the bottom two examples prove that considering the semantic cross-view correlation improves the generation of sentences with precise nouns. Such nouns enable the description with rich context. Furthermore, our aLSTMs is possible to generate more precise nouns for video captioning. In Fig.5, the relationship between words and the visual content is marked with different colors and the results show that our generated words has stronger correlation with the corresponding visual content. For instance, in the fourth video, "women" is connected with the second frame and "pot" is connected with the third frame.

In addition, in order to further analyze our approach, we show some descriptions generated on sample Youtube clips from MSVD (see Fig. 6). All the descriptions are generated by our aLSTMs model. In Fig. 6, three categories of captions are selected: 1) Correct descriptions; 2) Incorrect nouns in description; 3) Incorrect verbs in description. Each category is presented in one column. Also, for each category, four video clips are selected. For each video, two frames are used for demonstration.

The left column of Fig .6 shows that the sentences generated by our aLSTMs correctly describe the corresponding videos. In the middle column, the generated sentence contains some

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2729019, IEEE Transactions on Multimedia

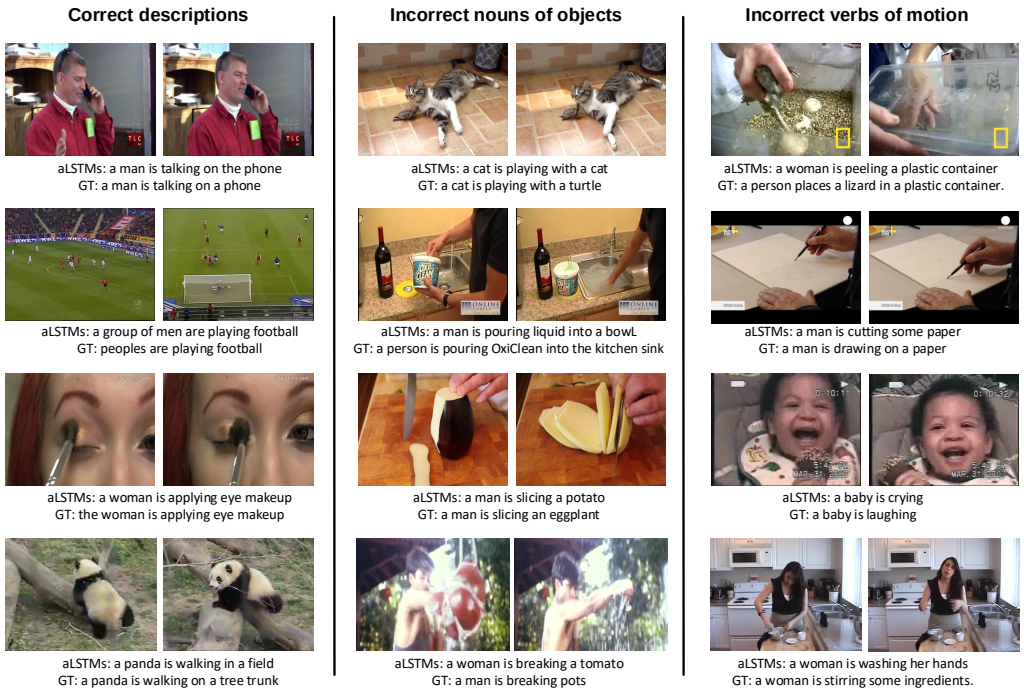

Fig. 6. Examples of video captioning results.

incorrect objects. For example, for the top video, the generated caption is "a cat is playing with a cat" and the ground truth is "a cat is playing with a turtle". By carefully observing the video frames, we find that: 1) the "turtle" is quite smaller than the cat; 2) more importantly the color of turtle is quite similar with the dark color of the cat; and 3) the turtle is connected with the cat. In this case, the updated GoogleNet is impossible to identify the turtle. In the third video, the generated caption is "a man is slicing a potato". The incorrect noun is "potato" and it should be "eggplant". This is because the peeled "eggplant" is quite similar with "potato" not only in shape but also in color.

The right column provides four videos with incorrect verbs. For instance, for the second video, "a man is cutting some paper" contains the error verb and it should be "A man is drawing on a paper". By checking the video frames, we can see that the "drawing" action has high similarity with "cutting". In addition, for the third video, the correct caption should be "a baby is laughing", while our model generates a wrong verb "crying". This is due to the reason that our training dataset does not provide training samples to distinguish face expressions. This indicates that to date, existing video captioning training datasets are still incomplete and require further refinement and expansion.

Results on MSR-VTT dataset. We compare our model with the state-of-the-art methods on the MSR-VTT dataset. Tab. V shows the comparison and depicts the BLEU, METEOR for each method. Tab. V shows that our model performs the best (38.0% @B4 and 26.1% METEOR).

TABLE V
COMPARISON OF STATE-OF-THE-ART ALGORITHMS WITH OUR PROPOSED METHOD ON THE MSR-VTT DATASET.

Model	B@4	METEOR
MP-LSTM (GoogleNet)	34.6	24.6
MP-LSTM (VGGNet)	34.8	24.8
MP-LSTM (C3D+VGGNet)	35.8	25.3
SA (GoogleNet)	35.2	25.2
SA (VGGNet)	35.6	25.4
SA (C3D)	36.1	25.7
SA (C3D+VGGNet)	36.6	25.9
aLSTMs	**38.0**	**26.1**

V. CONCLUSION

In this paper, we have proposed a framework aLSTMs, which is implemented by simultaneously minimizing the relevance loss and semantic cross-view loss. On two popular video description datasets, the results of our experiments demonstrate the success of our approach, which achieves comparable or even superior performance compared with the current state-of-the-art models. In the future, we will modify our model to work on domain-specific datasets, e.g., movies.

VI. ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Project 61502080, Project 61632007, and the Fundamental Research Funds for the Central Universities under Project ZYGX2016J085, Project ZYGX2014Z007.

REFERENCES

[1] J. Song, L. Gao, F. Nie, H. T. Shen, Y. Yan, and N. Sebe, "Optimized graph learning using partial tags and multiple features for image and video annotation," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 4999–5011, 2016.

[2] X. Zhu, Z. Huang, J. Cui, and H. T. Shen, "Video-to-shot tag propagation by graph sparse group lasso," *IEEE Transactions on Multimedia*, vol. 15, no. 3, pp. 633–646, 2013.

[3] Z. Pan, Y. Zhang, and S. Kwong, "Efficient motion and disparity estimation optimization for low complexity multiview video coding," *IEEE Transactions on Broadcasting*, vol. 61, no. 2, pp. 166–176, 2015.

[4] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *EMNLP*, 2015, pp. 1412–1421.

[5] F. Shen, C. Shen, Q. Shi, A. Van Den Hengel, and Z. Tang, "Inductive hashing on manifolds," in *CVPR*, 2013, pp. 1562–1569.

[6] J. Song, L. Gao, L. Liu, X. Zhu, and N. Sebe, "Quantization-based hashing: a general framework for scalable image and video retrieval," *Pattern Recognition*, 2017.

[7] C. Gan, T. Yao, K. Yang, Y. Yang, and T. Mei, "You lead, we exceed: Labor-free video concept learning by jointly exploiting web videos and images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 923–932.

[8] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *CVPR*, 2015, pp. 3156–3164.

[9] Z. Guo, L. Gao, J. Song, X. Xu, J. Shao, and H. T. Shen, "Attention-based LSTM with semantic consistency for videos captioning," in *ACM Multimedia*, 2016, pp. 357–361.

[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015, pp. 1–9.

[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, June 2016.

[12] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Explain images with multimodal recurrent neural networks," *arXiv preprint arXiv:1410.1090*, 2014.

[13] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015, pp. 2048–2057.

[14] A. Karpathy, A. Joulin, and F. F. F. Li, "Deep fragment embeddings for bidirectional image sentence mapping," in *Advances in neural information processing systems*, 2014, pp. 1889–1897.

[15] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the long-short term memory model for image caption generation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2407–2415.

[16] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. J. Mooney, and K. Saenko, "Translating videos to natural language using deep recurrent neural networks," in *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, 2015, pp. 1494–1504.

[17] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence-video to text," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4534–4542.

[18] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, "Describing videos by exploiting temporal structure," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4507–4515.

[19] G. Li, S. Ma, and Y. Han, "Summarization-based video caption via deep neural networks," in *ACM Multimedia*, 2015, pp. 1191–1194.

[20] Q. Wu, C. Shen, A. v. d. Hengel, L. Liu, and A. Dick, "Image captioning with an intermediate attributes layer," *arXiv preprint arXiv:1506.01144*, 2015.

[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[22] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *CVPR*, 2016.

[23] L. Gao, J. Song, F. Nie, F. Zou, N. Sebe, and H. T. Shen, "Graph-without-cut: An ideal graph learning for image segmentation," in *AAAI*, 2016, pp. 1188–1194.

[24] J. Song, L. Gao, M. M. Puscas, F. Nie, F. Shen, and N. Sebe, "Joint graph learning and video segmentation via multiple cues and topology calibration," in *ACM Multimedia*, 2016, pp. 831–840.

[25] J. Song, Y. Yang, Z. Huang, H. T. Shen, and J. Luo, "Effective multiple feature hashing for large-scale near-duplicate video retrieval," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1997–2008, 2013.

[26] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4597–4605.

[27] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2015, pp. 4489–4497.

[28] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *CVPR*, 2015, pp. 2625–2634.

[29] M. Auli, M. Galley, C. Quirk, and G. Zweig, "Joint language and translation modeling with recurrent neural networks," in *ACL*, 2013, pp. 1044–1054.

[30] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015, pp. 3128–3137.

[31] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," *arXiv preprint arXiv:1511.07571*, 2015.

[32] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014.

[33] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng, "Semantic compositional networks for visual captioning," *arXiv preprint arXiv:1611.08002*, 2016.

[34] X. Long, C. Gan, and G. de Melo, "Video captioning with multi-faceted attention," *arXiv preprint arXiv:1612.00234*, 2016.

[35] A. Rohrbach, M. Rohrbach, and B. Schiele, "The long-short story of movie description," in *German Conference on Pattern Recognition*. Springer, 2015, pp. 209–221.

[36] K. Tran, X. He, L. Zhang, J. Sun, C. Carapcea, C. Thrasher, C. Buehler, and C. Sienkiewicz, "Rich image captioning in the wild," in *CVPR Workshops*, 2016, pp. 49–56.

[37] Y. Pan, T. Yao, H. Li, and T. Mei, "Video captioning with transferred semantic attributes," *arXiv preprint arXiv:1611.07675*, 2016.

[38] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," *arXiv preprint arXiv:1611.05594*, 2016.

[39] Q. Tian and S. Chen, "Cross-heterogeneous-database age estimation through correlation representation learning," *Neurocomputing*, vol. 238, pp. 286–295, 2017.

[40] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, "Jointly modeling embedding and translation to bridge video and language," in *CVPR*, 2016, pp. 4594–4602.

[41] P. Pan, Z. Xu, Y. Yang, F. Wu, and Y. Zhuang, "Hierarchical recurrent neural encoder for video representation with application to captioning," in *CVPR*, 2016, pp. 1029–1038.

[42] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *CVPR*, June 2016.

[43] Z. Pan, P. Jin, J. Lei, Y. Zhang, X. Sun, and S. Kwong, "Fast reference frame selection based on content similarity for low complexity HEVC encoder," *J. Visual Communication and Image Representation*, vol. 40, pp. 516–524, 2016.

[44] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[45] D. L. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 190–200.

[46] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *CVPR*, 2016.

[47] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[48] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "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, 2002, pp. 311–318.

[49] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, vol. 29, 2005, pp. 65–72.

[50] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *CVPR*, 2015, pp. 4566–4575.

[51] R. Xu, C. Xiong, W. Chen, and J. J. Corso, "Jointly modeling deep video and compositional text to bridge vision and language in a unified framework." in *AAAI*. Citeseer, 2015, pp. 2346–2352.

[52] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu, "Video paragraph captioning using hierarchical recurrent neural networks," in *CVPR*, 2016.