Computer Vision and Image Understanding 163 (2017) 113–125

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Hierarchical & multimodal video captioning: Discovering and transferring multimodal knowledge for vision to language

An-An Liu[a,*], Ning Xu[a], Yongkang Wong[b], Junnan Li[c], Yu-Ting Su[a], Mohan Kankanhalli[d]

[a] School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China
[b] Smart Systems Institute, National University of Singapore, Singapore
[c] NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore
[d] School of Computing, National University of Singapore, Singapore

ARTICLE INFO

Article history:
Received 14 September 2016
Revised 11 April 2017
Accepted 27 April 2017
Available online 8 May 2017

Keywords:
Video to text
Semantic discovery
Multi-modal fusion
Deep learning

ABSTRACT

Recently, video captioning has achieved significant progress through the advances of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Given a video, deep learning approach is applied to encode the visual information and generate the corresponding caption. However, this direct visual to textual translation ignores the rich intermediate description, such as objects, scenes, actions, etc. In this paper, we proposed to discover and integrate the rich and primeval external knowledge (i.e., frame-based image caption) to benefit the video caption task. We propose a Hierarchical & Multimodal Video Caption (HMVC) model to jointly learn the dynamics within both visual and textual modalities for video caption task, which infers an arbitrary length sentence according to the input video with arbitrary number of frames. Specifically, we argue that the module for latent semantic discovery transfers external knowledge to generate complex and helpful complementary cues. We comprehensively evaluate the HMVC model on the Microsoft Video Description Corpus (MSVD), the MPII Movie Description Dataset (MPII-MD), and the novel dataset for 2016 MSR Video to Text challenge (MSR-VTT), and have attained a competitive performance. In addition, we evaluate the generalization properties of the proposed model by fine-tuning and evaluating the model on different datasets. To the best of our knowledge, this is the first time such analysis has been applied for the video caption task.

1. Introduction

Object, event, and attribute recognition (Liu et al., 2015; 2017; 2016, Yan et al. 2013, 2016) and retrieval (Cheng and Shen, 2016; He et al., 2016; Nie et al., 2013; Zhang et al., 2016; 2013; Zhu et al., 2017) have been widely investigated. Recently, several works (Song et al., 2016; Venugopalan et al., 2015a; Xu et al., 2015b; Yao et al., 2015; Yu et al., 2016) have tackled vision to language task, which aims to describe the visual content with natural language. It presents a particular challenge in the fields of computer vision and machine learning. On one hand, it requires developing sophisticated algorithms to explore cross-modality semantic mapping. On the other hand, human language is designed specifically so as to communicate information between humans, whereas even the most carefully composed image is the culmination of a complex set of physical processes over which humans have little control (Wu et al., 2016).

1.1. Motivation

Although there exist significant differences between visual and textual modalities, the state-of-the-art methods inspired by machine translation have been surprisingly successful. Especially with the rapid development of deep learning approaches in multiple areas, many sequential learning methods, such as Chen and Zitnick (2015), Donahue et al. (2015), Karpathy and Li (2015), Kiros et al. (2014), Kuznetsova et al. (2014), Mao et al. (2015), Rohrbach et al. (2013), Vinyals et al. (2015) and Pan et al. (2016a), have been proposed for visual caption to deal with the problems induced by the previous rule-based methods (Yang et al., 2011; Zha et al., 2007). The current methods usually leverage Convolutional Neural Network (CNN) features as visual *encoder* to explicitly produce fixed-length vector representations (Venugopalan et al., 2015b) or utilize the Long Short Term Memory (LSTM) architecture to implicitly generate visual representation (Venugopalan et al., 2015a) and then feed them into the RNN-based *decoder* to generate captions. However, this direct translation from visual content to natural language might ignore rich visual knowledge embedded within the video. For example, it might not capture the description of

* Corresponding author.
E-mail address: anan0422@gmail.com (A.-A. Liu).

http://dx.doi.org/10.1016/j.cviu.2017.04.013
1077-3142/

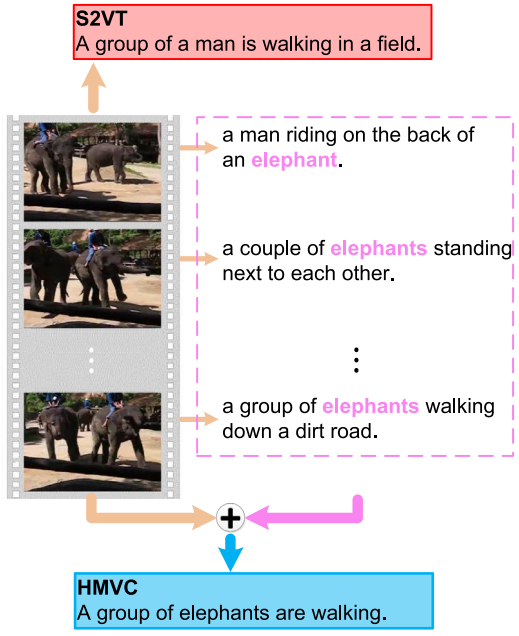

Fig. 1. S2VT model (Venugopalan et al., 2015a) (marked as red) generates text description directly from the visual source. In contrast, the proposed HMVC model (marked as blue) integrates the rich intermediate knowledge by leveraging the latent semantic knowledge in each frame (marked as pink) to compose descriptions of the given video. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

diverse objects, scenes, or actions, which can be reliably detected with state-of-the-art algorithms. Moreover, it is hard to determine the salient content and describe the dominant event appropriately. Fig. 1 shows the comparison of video captions generated by the popular sequence-to-video-text (S2VS) model (Venugopalan et al., 2015a) and our proposed method. Since a video may contain several events which involve multiple people and objects, the visual semantic regions, which do not occur frequently, can be easily forgotten due to the complicated relationships and visual dynamics. Therefore, the S2VS model failed to generate "elephant". Comparatively, the proposed model can generate "elephant" for the corresponding snippet and take advantage of both visual features and the generated semantic concepts for video caption. Therefore, it can succeed in remembering this latent and temporary visual and textual knowledge for inference.

1.2. Contribution

In this paper, we focus on discovering and integrating rich visual and textual knowledge to benefit video captioning. Specifically, we propose a Hierarchical & Multimodal Video Caption (HMVC) model to jointly learn the dynamics within both visual and textual modalities to infer an arbitrary length sentence based on the input video with arbitrary number of frames. For the encoding stage, we design the HMVC architecture with three-layer LSTMs to learn the multimodal and dynamic representation. For the decoding stage, we leverage the LSTM-based framework for word generation one by one. The proposed HMVC model is evaluated on three popular video caption datasets, namely the Microsoft Video Description Corpus (MSVD) (Chen and Dolan, 2011), the MPII Movie Description Dataset (MPII-MD) (Rohrbach et al., 2015b), and the novel dataset for 2016 MSR Video to Text challenge (MSR-VTT) (Xu et al., 2016). The extensive comparison experiments demonstrate that the proposed method can achieve competitive performance on all datasets. Furthermore, the discovered latent knowledge corresponding to the fine-grain video units can benefit model learning. The main contributions are three-fold:

- Discovery of internal knowledge: We design the hierarchical model to jointly learn the temporal dynamics within both visual and textual modalities to enrich the multi-view representation of one video.
- Transfer of external knowledge: The module for latent semantic discovery with respect to individual video units can transfer external knowledge to generate the complex interactions of humans, objects and events as one important complementary cue for model learning. Furthermore, it can also be regarded as leveraging the large-scale image caption dataset to benefit video caption modeling with limited video caption datasets.
- Comprehensive evaluation: We conduct a detailed evaluation on various types of visual encoders for the video caption model and demonstrate that both internal and external knowledge are complementary to each other for model learning. We have also study the generalization properties of the proposed video caption model by fine-tuning and evaluating the model on different dataset.

The rest of the paper is organized as follows. Section 2 reviews related work on image and video caption. Section 3 delineates the details of the proposed HMVC model. The experiment configuration is described in Section 4, whereas the experimental results and discussion are presented in Section 5. We conclude the paper in Section 6.

2. Related work

The literature on visual caption task can be broadly divided into two categories. The first category is the template-based methods, which predefines the specific grammar rules and splits sentences into several terms (e.g., subject, verb, object, etc.). With such sentence fragments, each term is aligned with visual content and then the sentence is generated (Guadarrama et al., 2013; Kulkarni et al., 2013; Rohrbach et al., 2013). The other category depends on learning sequential models (e.g., RNNs) in the co-embedding space of visual and textual modalities. Several works have explored these kind of methods with topic models (Barnard et al., 2003; Jia et al., 2011) and deep learning approaches (Donahue et al., 2015; Kiros et al., 2014; Mao et al., 2014; Venugopalan et al., 2015a; 2015b; Vinyals et al., 2015; Yao et al., 2015) to generate flexible natural language description. In the following section, we will detail image and video caption, respectively.

The problem of image annotation with natural language at the scene level has long been studied in the field of computer vision. For the template-based method, objects are first detected and recognized in the images, and then the sentences are generated with syntactic and semantic constraints (Farhadi et al., 2010; Kulkarni et al., 2013; Li et al., 2011; Sun et al., 2015; Yang et al., 2011; Zha et al., 2007). For instance, Farhadi et al. (2010) took advantage of the detected objects to infer a triplet of scene elements, which can be converted to sentence by the predefined language template. Li et al. (2011) considered the relationships among the detected objects and composed image descriptions using web-scale *N*-grams. A more sophisticated *CRF*-based method with attribute detection was proposed by Kulkarni et al. (2013). Recently, more and more researchers are engaging into leveraging sequential learning for description generation (Chen and Zitnick, 2015; Fang et al., 2015; Karpathy and Li, 2015; Kiros et al., 2014; Lebret et al., 2015; Mao et al., 2015; Vinyals et al., 2015; Xu et al., 2015b). Kiros et al. (2014) proposed to use a log-bilinear model with bias features derived from the image to discover the co-embedding space be-

tween visual and textual modalities. Fang et al. (2015) proposed a three-step approach for image caption, which consists of word detection by multiple instance learning, sentence generation by the language model, and sentence re-ranking by deep embedding. Recently, several researchers proposed a popular deep learning architecture by connecting CNN with RNN to map vision to language directly. Mao et al. (2015) proposed a multimodal RNN (m-RNN) to estimate the probability distribution of the next word given previous words and the deep CNN feature of an image at each time step. Chen and Zitnick (2015) learn a bi-directional mapping between images and their sentence-based descriptions, which allows to reconstruct visual features given an image description. Johnson et al. (2016) developed the Fully Convolutional Localization Network architecture based on CNN-RNN models, which contains a differentiable localization layer to predict a set of descriptions across image regions. Xu et al. (2015b) proposed an attention-driven image caption model by discovering the semantic-related image regions to benefit sentence generation. Jia et al. (2015) applied additional retrieved sentences to guide the LSTM in generating captions. The LSTM-in-LSTM architecture (Song et al., 2016) consists of an inner LSTM and an outer LSTM, which respectively encodes the spatially concurrent visual objects and captures the correspondence of sentences and images.

In the video domain, similar approaches have been proposed for video description generation. The early research directly applied video representation to template-based or statistical model-based machine translations (Barbu et al., 2012; Guadarrama et al., 2013; Rohrbach et al., 2013). Guadarrama et al. (2013) designed semantic hierarchies to choose an appropriate level of the specificity and accuracy of sentence fragments. Barbu et al. (2012) and Rohrbach et al. (2013) generated sentences by mapping semantic sentence representation, modeled with a CRF model, to high-level concepts such as actors, actions and objects. Specifically, Rohrbach et al. (2013) learned to model the relationships between different components of the input video for descriptions. The advantage of template-based methods is that the resulting captions are more likely to be grammatically correct. However, they still rely on hard-coded visual concepts and suffer from the implied limits on the variety of the output. Recently, sequence learning is widely applied to video caption, where an encoder maps a sequence of video frames to fixed-length feature vectors in the embedding space and a decoder then generates a translated sentence in the target language (Donahue et al., 2015; Venugopalan et al., 2015a; 2015b; Yao et al., 2015). This problem is analogous to translating a sequence of words in the input language to a sequence of words in the output language in the area of machine translation. The early video caption method (Donahue et al., 2015) extended the image caption methods by simply pooling the features of multiple frames to form a single representation. However, this strategy can only work for short video clips where there is only one major event with limited visual variation. To avoid this issue, more sophisticated visual feature encoding methods were proposed in recent works, using either a recurrent encoder (Donahue et al., 2015; Pan et al., 2016c; Venugopalan et al., 2015a; Xu et al., 2015a) or an attention model (Long et al., 2016; Yao et al., 2015). Yao et al. (2015) proposed to utilize the temporal attention mechanism to exploit temporal structure as well as a spatio-temporal convolutional neural network (Ji et al., 2013) to obtain local action features. Based on the popular sequence-to-sequence (S2S) framework (Cho et al., 2014; Sutskever et al., 2014), Venugopalan et al. (2015a) transfer the temporal visual information to natural language description and further extended it by inputting both appearance features and optical flow. However, this straightforward extension of the S2S framework not only ignores the hierarchical structure of video stream, such as frame, shot, and scene, but also ignores the multi-modal information conveyed by the video stream to infer the natural language description (Farhadi et al., 2010; Kulkarni et al., 2013; Li et al., 2011; Yang et al., 2011).

In contrast to the previous video caption works, we argue that the latent intermediate knowledge (e.g., semantic description) is the important complementary information for video caption. Moreover, semantics discovery benefits from the transfer and external knowledge learning by leveraging a large amount of image caption datasets (Yan et al., 2014). Therefore, the goal of this work is to propose an effective video caption method by jointly learning the dynamics within both visual and textual modalities, where we must deal with the problem of limited video caption datasets. As a result, our algorithm can learn a video caption model to infer an arbitrary length sentence matching the input video with arbitrary number of frames.

3. Approach

Inspired by Sutskever et al. (2014) and Venugopalan et al. (2015a), we propose a Hierarchical & Multimodal Video Caption (HMVC) model to leverage the hierarchical and multimodal information embedded in video. Differing from prior works (Venugopalan et al., 2015a; Yao et al., 2015), our proposed model transfers the latent intermediate knowledge from an external data source to enhance the video caption quality. We leverage the large-scale image knowledge on a trained image caption model, which is utilized to transfer frame-level images into textual descriptions. Briefly, the HMVC model consists of three layers of Long Short Term Memory (LSTM) cell that encodes both visual and textual information and decodes them into a natural language description. The top layer (visual model) encodes the visual feature stream. The middle layer (textual model) encodes the latent textual feature stream generated by a given image caption model. The bottom layer (language model) encodes the given video caption groundtruth and the hidden representation of the caption stream to generate the output sequential words. Formally, let $\boldsymbol{V} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n)$ be a set of visual features from n frames video, and $\boldsymbol{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_m)$ be its groundtruth caption. Previous works directly estimate the conditional probability $p(\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_m \mid \boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n)$ (Venugopalan et al., 2015a). In contrast, the proposed HMVC model aims to discover the latent frame-wise semantic knowledge, $\boldsymbol{S} = (\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n)$, from the corresponding frames, and furthermore optimize $p(\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_m \mid (\boldsymbol{v}_1, \boldsymbol{s}_1), (\boldsymbol{v}_2, \boldsymbol{s}_2), \ldots, (\boldsymbol{v}_n, \boldsymbol{s}_n))$. A conceptual diagram is shown in Fig. 2.

In the following section, we first revisit the LSTM networks and discuss the extension with image caption information. Then, we detail the proposed HMVC model, followed by an image caption model with CNN-RNN framework. Lastly, we provide details of the visual and textual feature representation.

3.1. Long Short Term Memory (LSTM)

The proposed HMVC model employs LSTM architecture (Hochreiter and Schmidhuber, 1997) to encode knowledge at every time step for what inputs have been observed up to this step. Fig. 3 illustrates the structure of a LSTM cell. It consists of one memory cell and three controllable gates, including the input gate i, the output gate o, and the forget gate f. Given the input $\boldsymbol{x}_t$ at time step t, the LSTM unit is updated by following:

$$
\begin{aligned}
\boldsymbol{i}_t &= \sigma(\boldsymbol{R}_{xi}\boldsymbol{x}_t + \boldsymbol{R}_{hi}\boldsymbol{h}_{t-1} + \boldsymbol{R}_{ci}\boldsymbol{c}_{t-1} + \boldsymbol{b}_i),\\
\boldsymbol{f}_t &= \sigma(\boldsymbol{R}_{xf}\boldsymbol{x}_t + \boldsymbol{R}_{hf}\boldsymbol{h}_{t-1} + \boldsymbol{R}_{cf}\boldsymbol{c}_{t-1} + \boldsymbol{b}_f),\\
\boldsymbol{c}_t &= \boldsymbol{f}_t\boldsymbol{c}_{t-1} + \boldsymbol{i}_t \tanh(\boldsymbol{R}_{xc}x_t + \boldsymbol{R}_{hc}\boldsymbol{h}_{t-1} + \boldsymbol{b}_c),\\
\boldsymbol{o}_t &= \sigma(\boldsymbol{R}_{xo}\boldsymbol{x}_t + \boldsymbol{R}_{ho}\boldsymbol{h}_{t-1} + \boldsymbol{R}_{co}\boldsymbol{c}_t + \boldsymbol{b}_o),\\
\boldsymbol{h}_t &= \boldsymbol{o}_t \tanh(\boldsymbol{c}_t)
\end{aligned}
\tag{1}
$$

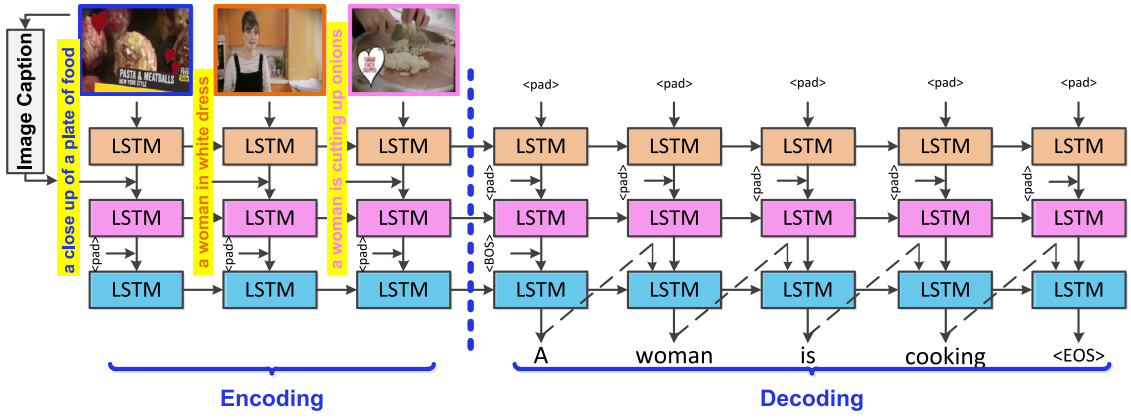

Fig. 2. The architecture of the proposed Hierarchical & Multimodal Video Caption (HMVC) model. It consists of three layers of LSTM that encodes both visual and textual information and decodes them into a natural language description. The top layer (visual model) encodes the visual feature stream. The middle layer (textual model) encodes the latent textual feature stream generated by a given image caption model. The bottom layer (language model) encodes the given text input and the hidden representation of the caption stream. ⟨*BOS*⟩ and ⟨*EOS*⟩ indicates begin-of-sentence and end-of-sentence tag, respectively. A zero vector is used as ⟨*pad*⟩ when there is null input at corresponding time step.

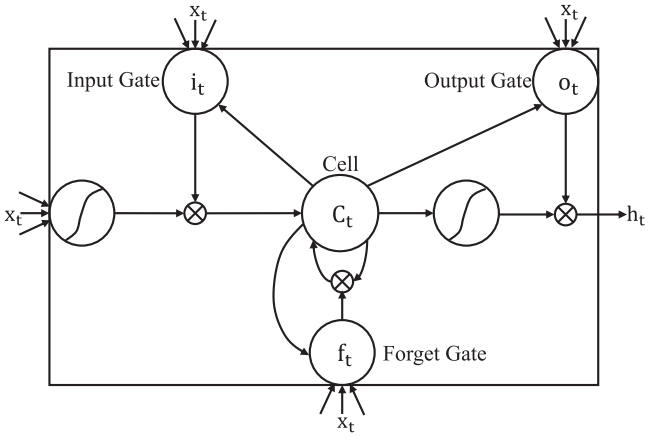

Fig. 3. A Long Short-Term Memory cell.

where $\boldsymbol{h}_t$ is the hidden state at time step t, and the $\boldsymbol{R}$ matrices are the weight matrices between two units.

In the encoding phase, given input sequences $\boldsymbol{V} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n)$, the proposed deep network will predict the latent image caption information $\boldsymbol{S} = (\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n)$, a sequence of hidden states $\boldsymbol{H} = (\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_n)$, and moreover define a distribution over the output sequence $\boldsymbol{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_m)$ as:

$$p(\boldsymbol{W}|\boldsymbol{V}, \boldsymbol{S}) = \prod_{t=1}^{m} p(\boldsymbol{w}_t \mid \boldsymbol{h}_n, \{\boldsymbol{w}_i\}_1^{t-1}, \{\boldsymbol{s}_i\}_1^t) \tag{2}$$

The decoding stage will apply a softmax function on the LSTM's hidden state. For a word in the vocabulary ($\boldsymbol{w} \in \boldsymbol{V}$), the generated word at the tth step can be inferred by:

$$p(\boldsymbol{w}_t|\boldsymbol{h}_n, \{\boldsymbol{w}_i\}_1^{t-1}) = \text{softmax}(\boldsymbol{R}_v \boldsymbol{h}_n + \boldsymbol{b}_v) \tag{3}$$

where $\boldsymbol{R}_v$ and $\boldsymbol{b}_v$ are the trained embedding matrices and bias vector, respectively, for word generation.

Given a training dataset with J samples, $\{< \boldsymbol{v}^j, \boldsymbol{w}^j >\}_{j=1}^J$, the objective function of HMVC can be formulated as:

$$\sum_{j=1}^{J} \sum_{t=1}^{m^j} \log p(\boldsymbol{w}_t^j \mid \boldsymbol{h}_n, \{\boldsymbol{w}_i^j\}_1^{t-1}, \{\boldsymbol{s}_i^j\}_1^t) \tag{4}$$

3.2. *Hierarchical* & *Multimodal Video Caption (HMVC) Model*

HMVC consists of three layers of LSTMs with 1000 hidden units of each LSTM (see Fig. 2). The hidden representation $\boldsymbol{h}_t$ from the upper LSTM is provided as the input $\boldsymbol{x}_t$ to the lower LSTM. The top layer and middle layer in both models are used to model the visual and textual knowledge respectively and the bottom layer is used to model the output word sequence. Language modeling and generation rely on a single layer of LSTM for both encoding and decoding stages, which allows parameter sharing between the encoding and decoding stages (Fig. 4).

In the encoding stage, the top LSTM receives and encodes the sequential visual frames and outputs the hidden representation, $\boldsymbol{h}_t^{\text{top}}$, to the corresponding LSTMs in the middle layer. The input visual frames are also fed into the module of image caption to generate the corresponding text descriptions. The textual information is not provided together with the video data and consequently it is latent information to be discovered to enrich the knowledge for model learning. The discovered image captions are fed into the corresponding LSTM and integrated with the corresponding $\boldsymbol{h}_t^{\text{top}}$ from the top layer as the input of the LSTM in the middle layers. Then, it generates the hidden representation, $\boldsymbol{h}_t^{\text{middle}}$, and outputs them to the bottom layer. The LSTMs of the bottom layer receive $\boldsymbol{h}_t^{\text{middle}}$ from the middle layer and integrate it with null padded input words (zeros). There is no loss during this encoding stage. After all the frames in the video clip are encoded, the bottom LSTM layer is fed with the begin-of-sentence (⟨*BOS*⟩) tag, which can initialize the decoding of the last hidden representation during the encoding stage into a sequence of words. During model training, the model maximizes the log-likelihood of the predicted output sentence given the hidden representation of the visual, the latent caption knowledge, and the previous words it has seen. Given a training dataset with J paired samples, the optimal parameters, θ^*, can be achieved by:

$$\theta^* = \arg\max_{\theta} \sum_{j=1}^{J} \sum_{t=1}^{m^j} \log p(\boldsymbol{w}_t^j|\boldsymbol{h}_n, \{\boldsymbol{w}_i^j\}_1^{t-1}, \{\boldsymbol{s}_i^j\}_1^t; \theta) \tag{5}$$

This log-likelihood can be optimized over the entire training dataset using stochastic gradient descent. The loss is computed only in the decoding stage. Since the loss is propagated back in time, the LSTM learns to generate an appropriate hidden state representation $\boldsymbol{H}$ for all input sequences. The output $\boldsymbol{h}_t^{bottom}$ of the bottom LSTM layer is used to obtain the emitted word $\boldsymbol{w}_t$. We apply a softmax function to get the probability distribution over the words

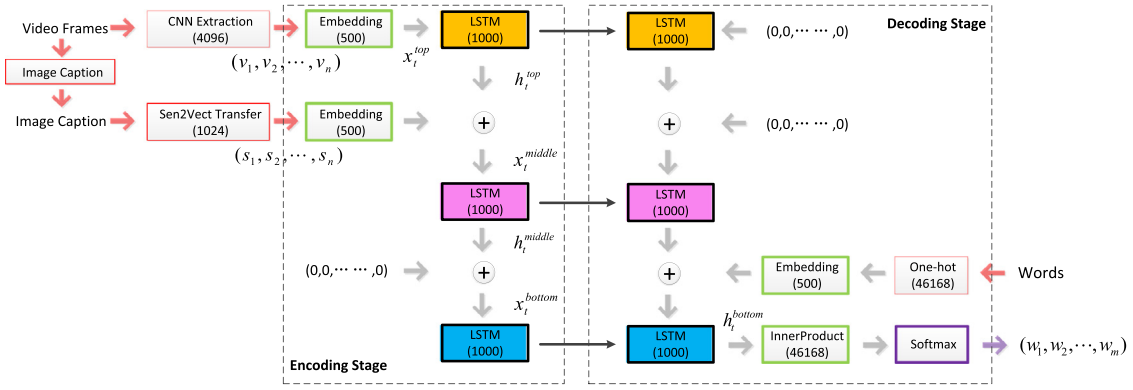

Fig. 4. Illustration of our HMVC model.

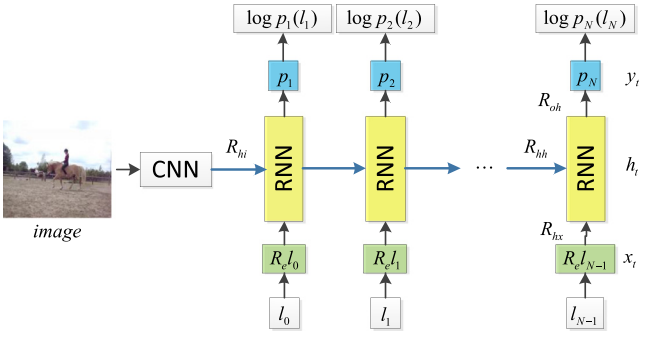

Fig. 5. RNN generative model combined with a CNN image embedder and word embeddings. The RNN takes a word, the context from previous time steps and defines a distribution over the next word in the sentence. The RNN is conditioned on the image information at the first time step. All RNNs share the same parameters. $\boldsymbol{I}_0$ and $\boldsymbol{I}_N$ are special tokens.

$\boldsymbol{w}'_t$ in the vocabulary V. During the decoding phase, the visual and caption representation for LSTMs are simply a vector of zeros that acts as padding input. We require an explicit end-of-sentence tag ($\langle EOS \rangle$) to terminate each sentence since this enables the model to define a distribution over sequences of varying lengths. During the decoding stage for caption generation, we choose the word $\boldsymbol{w}_t$ with the maximum probability based on Eq. (3) until the video caption model emits the $\langle EOS \rangle$ token.

3.3. Visual and text representation

RGB Frame: In this work, we extracted the CNN features by using the output of the fc7 layer (after applying the ReLU nonlinearity) of the VGG-16 (Simonyan and Zisserman, 2014). This CNN model was pretrained on the 1.2M images from the ILSVRC-2012 object classification subset (Deng et al., 2009), and is made publicly available via the Caffe Model Zoo.[1] Specifically, we sampled video frames at the frame rate 1/20 and extracted a single 4096 dimension vector for each frame. Then we learned a linear embedding of the CNN features to a 500 dimensional space, which is the input to the LSTMs in the top layer. The weights of the embedding are learned jointly with the LSTM layers during training (Fig. 5).

Image Caption: Le and Mikolov (2014) proposed an unsupervised algorithm, Sent2Vec, that learns fixed-length feature representations from variable-length pieces of texts. The algorithm represents each sentence by a continuous distributed vector, which gives the potential to capture the ordering and semantics of the words while it is sense about the semantics of the text and more formally the distances between the words. In this paper, we utilized Sent2Vec to project the extracted individual image caption into a common low-dimensional space (1024 dimension in our work). This sentence feature will be further embedded into a 500 dimensional vector and input to the LSTMs in the middle layer.

Video Caption: The output sequence of words are represented using one-hot vector encoding (1-of-D coding, where D is the size of the vocabulary). Similar to the visual and textual features, the one-hot feature of individual word is also embedded to a lower 500 dimensional space by applying a linear transformation. The projection matrix was learned via back propagation. The embedded word vector was concatenated with the hidden states of the middle-layer LSTMs to form the input to the bottom-layer LSTMs. Softmax was applied to infer the output word at time step t based on Eq. (3).

4. Experimental setup

In this section, we first provide an overview of the datasets, followed by the objective evaluation metrics and details of model training.

4.1. Video description datasets

- **Microsoft Research Video Description Corpus (MSVD)** (Chen and Dolan, 2011) is a collection of 1970 manually selected YouTube snippets, which cover a wide range of daily activities. On average, each video consists of 40 manually annotated sentence descriptions. The duration of each clip is between 10 and 25 s and the original corpus has multi-lingual descriptions. In this work, we use only the English descriptions. Following Xu et al. (2015c), We randomly select 1200 videos as training set, 100 videos as validation set and 670 videos as test set.
- **MPII Movie Description Dataset (MPII-MD)** (Rohrbach et al., 2015b) contains 68,337 video clips extracted from 94 Hollywood movies. Each clip is accompanied with a single sentence description which is selected from movie scripts and audio description data. Although the movie snippets are manually aligned to the descriptions, this dataset is very challenging due to the high diversity of visual and textual content, as well as the fact that most snippets have only a single reference sentence. The average duration of each video is 94 frames, which is shorter than the samples from MSVD. We strictly followed the dataset splits provided by Rohrbach et al. (2015b).
- **MSR-Video to Text (MSR-VTT)** (Xu et al., 2016) is a new large-scale video benchmark for video understanding, especially for the task to translate video to text. The dataset consists of top 150 video search results for each of the top 257 representative queries collected from a commercial video search engine. Each

[1] https://github.com/BVLC/caffe/wiki/Model-Zoo.

video clip is manually annotated with 20 sentences using Amazon Mechanical Turn workers. In total, MSR-VTT consists of 10K video clips with 41.2 h and 200K clip-sentence pairs, covering a comprehensive list of 20 categories and a wide variety of video content. In this work, We strictly followed the dataset split provided by Xu et al. (2016).

4.2. Evaluation metrics

To evaluate the generated sentences, we adopt the BLEU@N (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014), ROUGE-L (Lin, 2004) and CIDEr (Vedantam et al., 2015) metrics against all ground truth sentences. These metrics are widely used in machine translation and have already shown to be well correlated with human judgment. Specifically, BLEU@N measures the fraction of *N*-gram that are in common between a hypothesis and a reference or set of references. The unigram scores (BLEU@1) account for the adequacy of the information retained by the translation, while *N*-gram scores (BLEU@2, BLEU@3, and BLEU@4) account for the fluency. METEOR computes unigram precision and recall, extending exact word matches to include similar words based on WordNet synonyms and stemmed tokens. This has been used for image description evaluation with promising results (Elliott and Keller, 2014). ROUGE is primarily recall-based and thus has a tendency to reward long sentences with high recall. Recently, Vedantam et al. proposed CIDEr, a consensus-based evaluation protocol for image descriptions, which rewards a sentence for being similar to the majority of human written descriptions. We employ the evaluation source code[2] released by Microsoft COCO Evaluation (Chen et al., 2015) in this work. In addition, the 2016 Microsoft Research video to language grand challenge carried out human evaluation on the submitted results on MSR-VTT dataset. Humans were asked to rank the generated sentences from 1 to 5 (lower-better) with respect to coherence, relevance, and whether it is helpful for blind.

4.3. Model training

4.3.1. Proposed HMVC model

In all evaluations, we unrolled the LSTM to a fixed 80 time steps during training, which is an empirical trade-off between memory consumption and available information for LSTM learning. This setting allowed us to fit multiple videos in single mini-batches, which passed through the network sequentially. For videos requiring fewer than 80 time steps of both words and frames encoding, the remaining inputs were padded with an all-zero vector. For longer videos, we truncated the input frames to ensure that the sum of the number of frames and words is within 80 time steps. For model test, we did not constrain the length of the video and our model was fed with all sampled frames. We use Beam Search (Sutskever et al., 2014), which iteratively considers the set of c best sentences up to time t as candidates to generate sentences at time $t+1$, and only keeps the best c results. We set the c as 1. To avoid over-fitting, drop-out was utilized at the inputs and outputs of each LSTM. We used an initial learning rate of 0.01 which was exponentially decreased at each 500 snapshots by a factor of 0.9. The momentum was set to 0.9 while the max iteration was 25,000. Additionally, we pre-processed all text to lower case, tokenized the sentences and removed punctuation. Further, we pre-trained the models and then fine-tuned them on respective datasets to improve performance.

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

4.3.2. Image caption model

Ideally, both image caption model and video caption model are trained using end-to-end approach. However, existing video caption datasets have insufficient training data for both the video caption and image caption modeling. Although TACoS-MultiLevel dataset (Yu et al., 2016) contains frame-wise captions, it is a close-domain dataset and consists of different actors, fine-grained activities and small interacting objects in daily cooking scenarios. Its vocabulary only contains 2864 unique lexical entries. Therefore, the semantics frame-level information modeled with this dataset could not described the large-scale open-domain web videos, such as MSR-VTT. Hence, it motivated us to leverage large-scale image caption dataset (e.g., MSCOCO, etc.) to train a powerful image caption model for transferring external semantic knowledge to benefit both multi-modal video representation and video caption inference. In our work, we learned the image caption model with the MSCOCO (Lin et al., 2014) dataset, which contains 123,000 images and each image was annotated with 5 sentences. Meanwhile, we constructed a vocabulary with filtered words that occurred at least 5 times. We used mini-batches of 100 image-sentence pairs and momentum of 0.9 to optimize the model. We cross-validated the learning rate and the weight decay, while dropout regularization was applied in all layers except the recurrent layers (Pham et al., 2014).

5. Experimental results

In this section, we first evaluate the proposed HMVC model against the state-of-the-art methods on MSVD, MPII-MD, MSR-VTT dataset. Then, we provide comprehensive analysis of the proposed model with variations in different modalities and network architectures. The study the generalization properties of HMVC model, we pre-trained and fine-tuned our model with various combinations of dataset, and evaluate on different target dataset. Finally, we provide qualitative evaluation on the generated video captions.

5.1. Comparison against the state of the arts

In this section, we evaluate the performance of the proposed HMVC model against the state-of-the-art methods on MSVD and MPII-MD. The quantitative results is shown in Tables 1 and 2. Overall, HMVC achieves competitive or better performances on both dataset, which shows a clear benefit from the discover and combine of both visual information and complementary textual knowledge.

As shown in Table 1, HMVC achieves competitive performance comparing with all the competing methods on MSVD. First, we compare our model against the template-based Factor Graph Model (FGM) (Thomason et al., 2014) methods. It uses a two step approach to obtain confidences on subject, verb, object and scene elements, followed by combining these elements to generate a sentence. FGM ignores the temporal information in the video sequence and suffers from the constraint on the variety of the output due to the fixed template. The Mean Pooling model (Venugopalan et al., 2015b) first pools AlexNet's fc7 activations across all frames to create a fixed-length vector representation of the video, and uses a LSTM model to decode the video descriptor into a sequence of words. Specifically, the model is pre-trained on image caption datasets (i.e., Flickr30k (Young et al., 2014) and MSCOCO (Lin et al., 2014)) and fine-tuned on MSVD. HMVC model shows noticeable performance improvement over such methods as it models the temporal sequential information during the encoding stage.

Pan et al. (2016b) redefine the special rule for video representation by mean pooling over the visual features of frames/clips while using single LSTM layer to generate descriptions for video content. Specifically, the LSTM-E model explores the AlexNet, VGG-19 and

Table 1
Performance comparison with BLEU@N (B), METEOR (M), ROUGE-L (R), and CIDEr (C) metrics on MSVD.

Model	B@1	B@2	B@3	B@4	M	R	C
Template							
- FGM (Thomason et al., 2014)	–	–	–	–	23.9	–	–
Mean pooling							
- AlexNet (Venugopalan et al., 2015b)	–	–	–	31.2	26.9	–	–
- AlexNet (COCO) (Venugopalan et al., 2015b)	–	–	–	33.3	29.1	–	–
- GoogleNet (Yao et al., 2015)	–	–	–	38.7	28.7	–	44.8
- AlexNet (LSTM-E) (Pan et al., 2016b)	74.5	59.8	49.3	38.9	28.3	–	–
- VGG (LSTM-E) (Pan et al., 2016b)	74.9	60.9	50.6	40.2	29.5	–	–
- C3D (LSTM-E) (Pan et al., 2016b)	75.7	62.3	52.0	41.7	29.9	–	–
- VGG+C3D (LSTM-E) (Pan et al., 2016b)	78.8	66.0	55.4	45.3	31.0	–	–
Soft Attention							
- GoogleNet+C3D (Yao et al., 2015)	–	–	–	41.9	29.6	–	51.7
- TDDF (Zhang et al., 2017)	–	–	–	45.8	**33.3**	**69.7**	**73.0**
Paragraph Captioning							
- Hierarchical-RNN (Yu et al., 2016)	**81.5**	**70.4**	**60.4**	**49.9**	32.6	–	65.8
S2VT							
- AlexNet (Flow) (Venugopalan et al., 2015a)	–	–	–	–	24.3	–	–
- AlexNet (RGB) (Venugopalan et al., 2015a)	–	–	–	–	27.9	–	–
- VGG (RGB random order) (Venugopalan et al., 2015a)	–	–	–	–	28.2	–	–
- VGG (RGB) (Venugopalan et al., 2015a)	–	–	–	–	29.2	–	–
- VGG (RGB)+AlexNet (Flow) (Venugopalan et al., 2015a)	–	–	–	–	29.8	–	–
- Glove-Deep (Venugopalan et al., 2016)	–	–	–	42.1	31.4	–	–
HMVC	78.0	65.2	54.9	44.3	32.1	68.9	68.4

Table 2
Performance comparison with BLEU@N (B), METEOR (M), ROUGE-L (R), and CIDEr (C) metrics on MPII-MD.

Model	B@1	B@2	B@3	B@4	M	R	C
Template							
- SMT (best variant) (Rohrbach et al., 2015b)	–	–	–	0.5	5.6	13.2	8.1
- Visual-Labels (Rohrbach et al., 2015a)	–	–	–	**0.8**	7.0	16.0	**10.0**
Mean pooling							
- VGG (Venugopalan et al., 2015a)	–	–	–	–	6.7	–	–
S2VT							
- VGG (RGB) (Venugopalan et al., 2015a)	–	–	–	–	**7.1**	–	–
- Glove-Deep (Venugopalan et al., 2016)	–	–	–	–	6.8	–	–
HMVC	**16.9**	**5.4**	**1.6**	0.6	**7.1**	**17.0**	8.9

C3D visual-semantic embedding and the performances of BLEU@2 & BLUE@4 by concatenating VGG and C3D features is slightly better than HMVC due to the robust visual representations. The soft attention model in Yao et al. (2015) enables the decoder to exploit the temporal structure, which is a combination of weighted attention over a fixed set of video frames with input features from GoogleNet and a 3D-ConvNet trained on HoG, HoF and MBH features from an activity classification model. In compare with soft attention model (GoogleNet+C3D (Yao et al., 2015)), HMVC reports almost 2% of improvement in the BLEU-4 and METEOR scores, which shows the advance by adding intermediate image caption knowledge. We notice that TDDF (Zhang et al., 2017) outperforms HMVC. While TDDF proposed a task-driven dynamic fusion method which can adaptively choose different fusion patterns according to model status, our work, however, is focusing on jointly learn the dynamics within both visual and textual knowledge, not the dynamics fusion part. Yu et al. (2016) present an approach that employs hierarchical-RNN to generate multiple sentences for caption task. Comparatively, the proposed HMVC focuses on how to integrate multiple modalities for video captioner, which generates only one sentence to describe a video. Specifically, hierarchical-RNN (Ji et al., 2013) contains a sentence generator and a paragraph generator, which jointly leverage both temporal- and spatial-attention mechanisms. The HMVC model discovers the internal knowledge within both visual and textual modalities, as well as transfer external knowledge by leveraging the large-scale image dataset. However, it ignores to exploit spatial attention and performs lower than hierarchical-RNN (Ji et al., 2013).

For the more challenging MPII-MD dataset, HMVC improve the performance over SMT (Rohrbach et al., 2015b) method and mean pooling model (Venugopalan et al., 2015a) by 1.5% and 0.4%, respectively, in terms of METEOR. The METEOR score by HMVC is slightly better than Visual-Labels (Rohrbach et al., 2015a), where the latter is a LSTM-based approach that uses diverse visual concepts, including object detectors, activity and scene classifiers. HMVC performs better on METEOR comparing with the methods of S2VT (Venugopalan et al., 2016; 2015a), where the two-layer LSTM is employed. Specifically, the Glove-Deep (Venugopalan et al., 2016) fine-tuned the pre-trained linguistic network on the paired video-text corpus, which performs worse than HMVC. It shows that our three-layer architecture is more robust and competitive in the field of knowledge transfer learning. Although the performance on MPII-MD is improved slightly, video caption task on this dataset is still extremely challenging due to the limited frame-wise image caption references.

The comparison results of 2016 MSR-VTT video to language grand challenge[3] are shown in Fig. 6. HMVC achieves 37.1% BLEU@4 and 26.7% METEOR on the test set and ranks the 6th on the objective evaluation and 7th on the subjective human evaluation. The performance of HMVC is slightly lower than the best model, i.e., decrease by 3.7% in BLEU@4 and 1.5% in METEOR,[4] due to following reasons:

[3] http://ms-multimedia-challenge.com/challenge#video.

[4] The approaches proposed by the other teams have not been released when this paper is submitted.

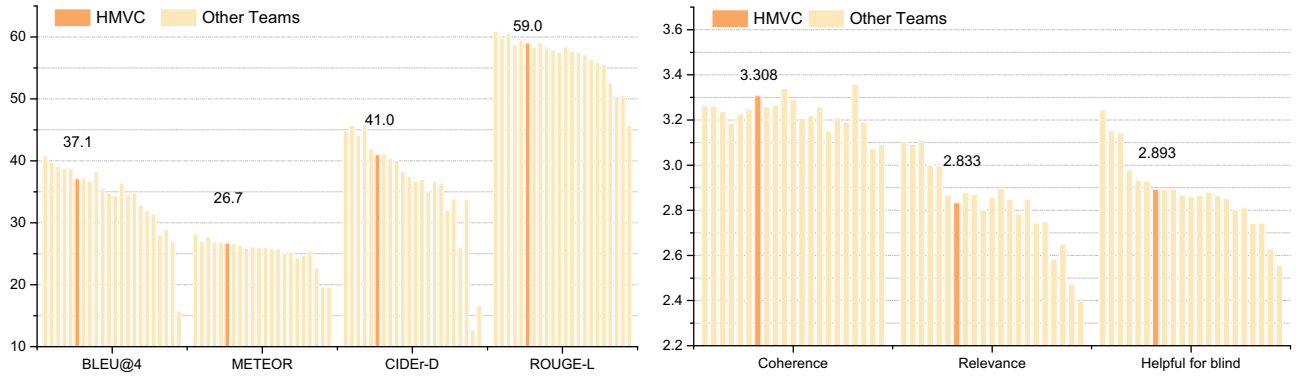

Fig. 6. Performance of proposed MSVD model on the test set of 2016 MSR-VTT Challenge. Overall, the proposed MSVD model ranked 6th and 7th on the objective and subjective evaluation, respectively.

Table 3
Performance evaluation of single-modality model (i.e., Visual-SE and Textual-SE) and proposed multi-modality HMVC model on MSVD.

Model	B@1	B@2	B@3	B@4	M	R	C
Visual-SE							
- AlexNet (Flow) (Venugopalan et al., 2015a)	–	–	–	–	24.3	–	–
- AlexNet (RGB) (Venugopalan et al., 2015a)	–	–	–	–	27.9	–	–
- VGG (RGB random order) (Venugopalan et al., 2015a)	–	–	–	–	28.2	–	–
- VGG (RGB) (Venugopalan et al., 2015a)	–	–	–	–	29.2	–	–
- VGG (RGB)+AlexNet (Flow) (Venugopalan et al., 2015a)	–	–	–	–	29.8	–	–
Textual-SE	74.0	60.3	49.4	36.9	28.0	65.5	40.9
HMVC	**78.0**	**65.2**	**54.9**	**44.3**	**32.1**	**68.9**	**68.4**

- The proposed HMVC model infers video captions with beam size of 1. This directly limited the diversity of the probably generated captions by the current language model (Karpathy and Li, 2015). Therefore, it may yield uninformative and even boring description. The more reasonable approach it to utilize multiple detectors to identify all entities, their mutual interactions and wider context to generate various descriptions for long videos.
- We utilized the popular CNN-RNN framework to generate image level description of video frames. This image caption model was learned on the MSCOCO dataset, which only consists of 8791 words (Karpathy and Li, 2015). The limited vocabulary is insufficient to generate accurate image captions for videos in this challenge as it contains a wide variety of video content.

Additionally, in terms of METEOR, we find that HMVC is slightly worse than Multi-modal Fusion (Jin et al., 2016) and TDDF (Zhang et al., 2017) on MSR-VTT, which benefit from integrating Aural & Meta modality features and dynamics fusion mechanism, respectively.

5.2. Performance comparison with respect to different modalities

In this section, we evaluate the influence of the visual and textual modalities on the proposed HMVC model. Specifically, we consider a standard S2VT model (Venugopalan et al., 2015a) with two stacked LSTM (denoted as Visual-SE), which only inputs sequential video frames and then generates language description. We also consider a textual based mode, namely Textual-SE, where the video frames at each time step is replaced by the sequential frame-wise image captions. In different from Visual-SE and Textual-SE, the proposed HMVC incorporates both modalities with the designed three stacked LSTMs. The illustration of the two architectures are shown in Fig. 7.

Table 3 shows the comparison results on MSVD dataset. Overall, HMVC achieves 32.1% METEOR and significantly outperforms Visual-SE models and Textual-SE model. This observation demonstrates that HMVC can benefit from exploiting multi-modal knowledge for model learning. The Textual-SE model achieves the slightly better performance compared with the AlexNet (Flow)-based model and the AlexNet (RGB)-based model. This is because sequential caption knowledge can capture the global stability and salient semantics within the whole video sequence. The Visual-SE model outperforms the Textual-SE model with VGG features, which demonstrates that the model can benefit from utilizing more discriminative feature representation. Fusion of both RGB and Flow feature can further augment the performance.

5.3. Performance comparison with respect to different network architectures

In this section, we comprehensively analyze three variants of the proposed HMVC. HMVC-2 is the counterpart of HMVC by alternating the inputs of the top and middle LSTM layers and both encode the multimodal information simultaneously. Differing from HMVC/HMVC-2, HMVC-DE/HMVC-DE-2 separate sequential learning for both visual and text modalities into two separate encoding stages to avoid potential negative influence between both modalities due to asynchronous temporal dynamics. The illustration of the three variants are shown in Fig. 7.

As shown in Table 4, HMVC can consistently outperform the other three architectures on BLUE@4, METEOR and CIDEr metrics, and achieve competitive performance on ROUGE metric. Generally speaking, the synchronized multimodal learning approach (i.e., HMVC and HMVC-2) works better than the asynchronous multimodal learning counterpart (i.e., HMVC-DE and HMVC-DE-2). This is because the top and middle layers of HMVC model can learn to fuse multi-model features at each time step, and consequently obtain a compact and discriminative feature to induce video caption generation. In comparison, HMVC-DE model learns the visual

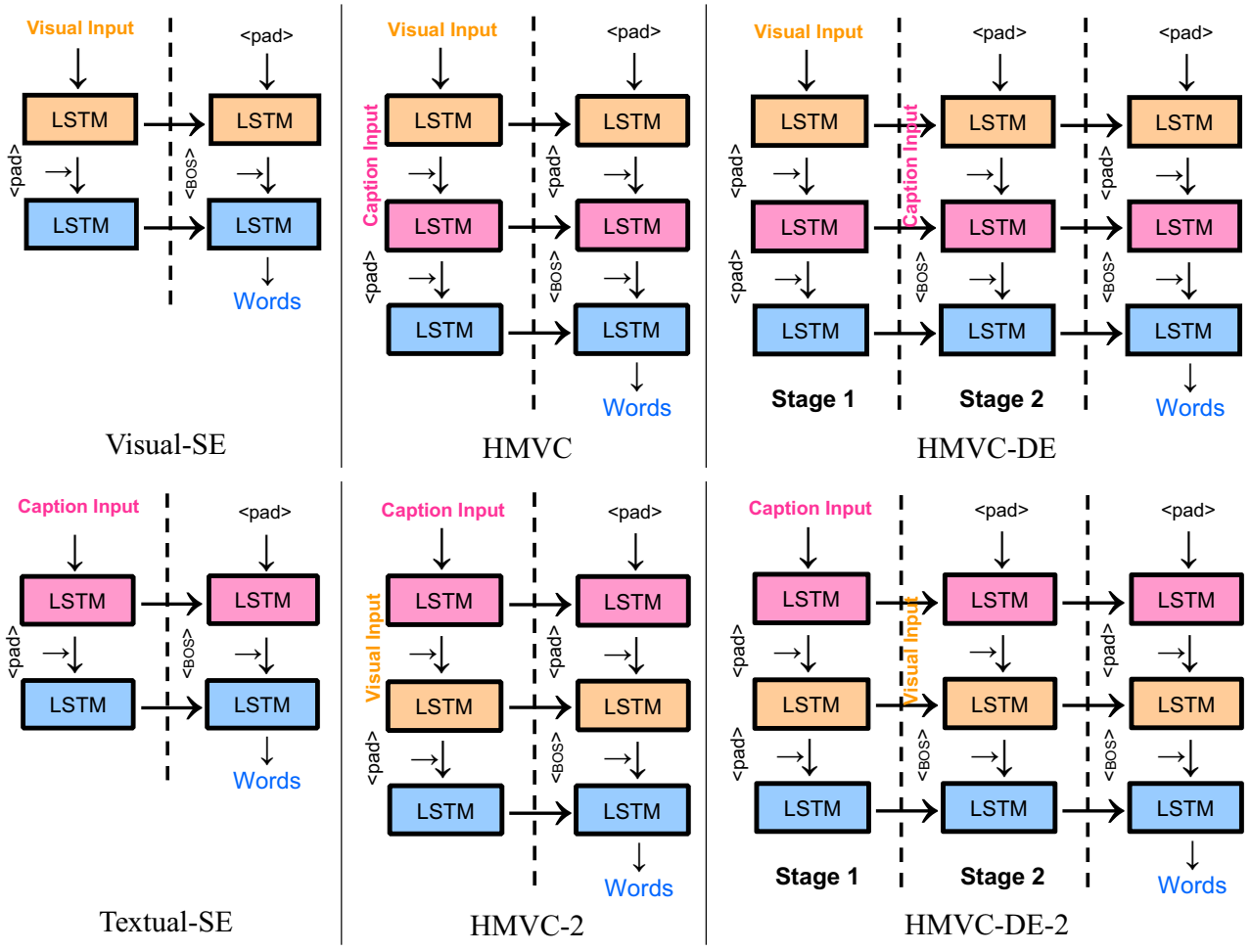

Fig. 7. Illustration of various network configurations for video caption. The left column illustrates a two stacked LSTM that encode single modality, where the top model (Visual-SE) and bottom model (Textual-SE) encode the visual features and textual features, respectively. The middle column is the proposed HMVC model with variations in encoding sequence for the top-layer and middle-layer LSTM. The right column is a dual-stage encoding architectural (HMVC-DE) where each stage only encode one feature type.

Table 4
Comparison of multiple variant architectures on MSVD.

Criteria	B@4	M	R	C
HMVC-2	**44.3**	31.7	**69.2**	66.7
HMVC-DE	41.8	30.2	67.4	56.1
HMVC-DE-2	43.7	31.5	68.5	66.7
HMVC	**44.3**	**32.1**	68.9	**68.4**

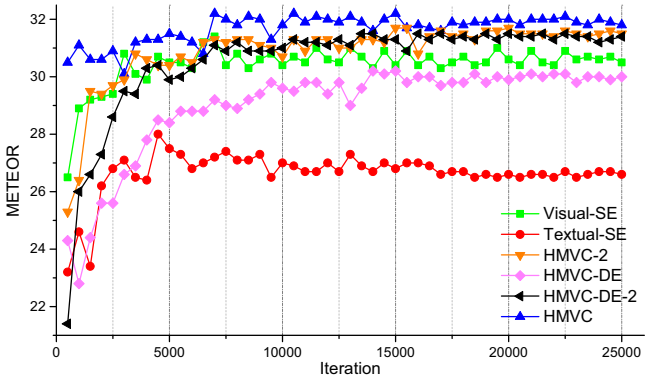

Fig. 8. Ensembles of networks with different snapshots. All results are reported on the test set of MSVD.

and textual sequential dynamics separately. Moreover, HMVC-DE model double the sequence length during the encoding stages and directly increase the difficulty and uncertainty for model learning.

To compare the robustness of different architectures, we visualize the METEOR scores with respective to the iteration numbers by different methods. As shown in Fig. 8, it is obvious that HMVC is more robust than the competing methods by properly fusing multi-modal information since it can achieve the highest Meteor score with smallest fluctuation after the lest iterations. It also shows that HMVC model can consistently outperform HMVC-DE model across all iterations. Furthermore, HMVC-DE, which is equal to HMVC-2 initialized with visual information, can significantly outperform HMVC-2. Similarly, HMVC-DE-2, which is equal to HMVC initialized with textual information, can also outperform HMVC. These observations further demonstrate the superiority of the proposed method on multi-modal fusion.

5.4. Exploring the generalization properties of HMVC

The MSVD dataset contains 1970 clips where each video has 30 or more sentence descriptions. On the other hand, the MPII-MD dataset contains more clips (54,076 in total) but each video

Table 5
Scores of performance comparison for evaluating the generalization ability of HMVC.

Pre-trained Dataset	Fine-tuned Dataset	Evaluation Dataset	B@1	B@2	B@3	B@4	M	R	C
MPII_MD	MSVD	MSVD	77.3	64.4	54.4	43.8	31.6	68.2	64.0
MSVD	MPII_MD	MSVD	56.0	33.2	17.1	7.7	18.7	45.5	12.7
MSVD	MPII_MD	MPII_MD	13.8	4.8	1.9	0.8	6.6	16.1	10.5
MPII_MD	MSVD	MPII_MD	5.1	1.0	0.3	0.1	2.9	6.0	2.9

has only one description. Therefore, both datasets do not contain enough video samples and the associated sentence descriptions to learn the complicated video caption model. This motivates us to explore the generalization ability of HMVC, which to the best of our knowledge, has yet to be discussed in the video caption task. Obviously, it is not reasonable to train the model with multiple datasets together since they may contain the samples from different domains and have different vocabularies. In order to integrate large-scale video samples and sentence descriptions of both datasets, we selected one of them as the target dataset and the other as the auxiliary dataset. Specifically, we designed to pre-train the model on the auxiliary dataset and then fine-tune the model on the target dataset. In this evaluation, we followed the dataset splits introduced in Section 4.1 and performed the comparison experiments on (1) fine-tuning the pre-trained MPII-MD model by MSVD; and (2) fine-tuning the pre-trained MSVD model by MPII-MD. We evaluated the fine-tuned models on both datasets respectively. The results are shown in Table 5.

Surprisingly, the fine-tuned HMVC modeled on the target domain (MSVD/MPII-MD), which has been pre-trained on the auxiliary domain (MPII-MD/MSVD), performs a little worse than the HMVC model that directly trained on the target domain (MSVD/MPII-MD) without pre-training. Moreover, if we fine-tuned the HMVC model on the auxiliary domain (MSVD/MPII-MD), which has been pretrained on the target domain (MPII-MD/MSVD), the model works rather bad compared to the HMVC model, which is directly trained on the target domain (MPII-MD/MSVD) without pre-training. These observations suggest that although the proposed deep learning architecture for video caption is a general framework towards open-domain video understanding without defining specific rules and leveraging heuristics, the model learning is still constrained by the domain embedded in each datasets. For future work, one specific direction is to develop sophisticated cross-domain learning algorithms to leverage multiple small-scale video caption dataset to augment the generalization of the trained model. This can help avoid the severe burden of preparing large-scale video caption groundtruth.

5.5. Qualitative evaluation of HMVC

In this section, we provide some representative video caption examples from the test set of MSVD dataset (Chen and Dolan, 2011). As shown in Fig. 9, for each video, we randomly extract four frame-based images with generated captions shown at the time step in the left panel, and the corresponding descriptions of the video, both generated and reference, are positioned with various colored shadows in the right panel. To better exploit the effect of textual knowledge on our model, we divide the video examples into three categories based on the quality of the video and image captions.

- **Correct video caption and image caption**: This category contains cases (see Fig. 9: a–c) where image captions involve salient knowledge, e.g., "elephants", "riding a wave", and "preparing food". HMVC generates the correct descriptions involving relevant objects and actions. Interestingly, in the case (a), both Visual-SE and Textual-SE fail to recognize "elephant" but HMVC succeeded on generating this word. In the case (b), Visual-SE summarizes the phrase "is swimming in a pool", Textual-SE captures the object "a man" and HMVC combines them together for video caption. In the case (c), Visual-SE captures the action "cooking" and Textual-SE summarizes some detailed procedure for cooking. Furthermore, HMVC generates the summary caption for video. In conclusion, the single modality is not enough to support the integration of the global knowledge. The complementing connection between top and middle LSTM layers of HMVC as shown in Fig. 2 can boost the fusion on the both one-sided modalities for stronger and conclusive video description.
- **Correct video caption with incorrect image caption**: This category contains cases (see Fig. 9: d–f) where insufficient information is captured in textual knowledge, e.g., the missing "playing the piano", "applying makeup" and "ribs". But HMVC still succeed in generating novel words depending on sufficient visual knowledge. Specifically, in the case (d), the textual knowledge only captures the phrase "cat sitting" from the single image but fails to obtain the action "playing" and the object "piano". The high-frequency visual concepts "piano" and "play" in the image sequence rescue this incomplete case. Similarly, in the case (e) and (f), the visual concepts fill a gap in the textual knowledge, which presents the advantages of HMVC in the caption task.
- **Incorrect video caption and image caption**: This category contains cases (see Fig. 9: g–i) where HMVC fails to reappear the losing key concepts, e.g., the missing "toy piano", "tiger" and "cheetah". Specifically, in the case (g), the image caption model fails to recognize "toy piano", meanwhile, this concept doesn't appear in the video training set. Therefore, both HMVC and other video caption model cannot output "toy piano" while one acceptable sentence " A little girl is playing." is generated reasonably. Therefore, the transfer learning method should be designed to address the task of generating descriptions of novel objects which are not presented in paired visual-sentence dataset. In the case (h–i), both of image caption models fail to capture key concepts, namely, "tiger" and "cheetah". Meanwhile, there are 122 occurrences of "tiger" and 24 occurrences of "cheetah" in the video training set but 439 occurrences of "animal" and 2027 occurrences of "cat", which are far more than the number of "tiger" and "cheetah", leading to overfit in the caption task. So, HMVC generated descriptions based on "animal" (acceptable) and "cat" (false) respectively. Therefore, it is a significant problem how to improve the performance of classification of caption model.

6. Conclusions

In this paper, we propose a novel deep learning based video caption model, namely Hierarchical & Multimodal Video Caption (HMVC), consists of a three-layer LSTMs where each layer are designed for visual feature, textual feature, and caption's word vector, respectively. The HMVC model discovers the internal knowledge by jointly learn the dynamics within both visual and textual knowledge conveyed in the fine-grain video units, as well as transfer external knowledge by leveraging the large-scale image knowledge

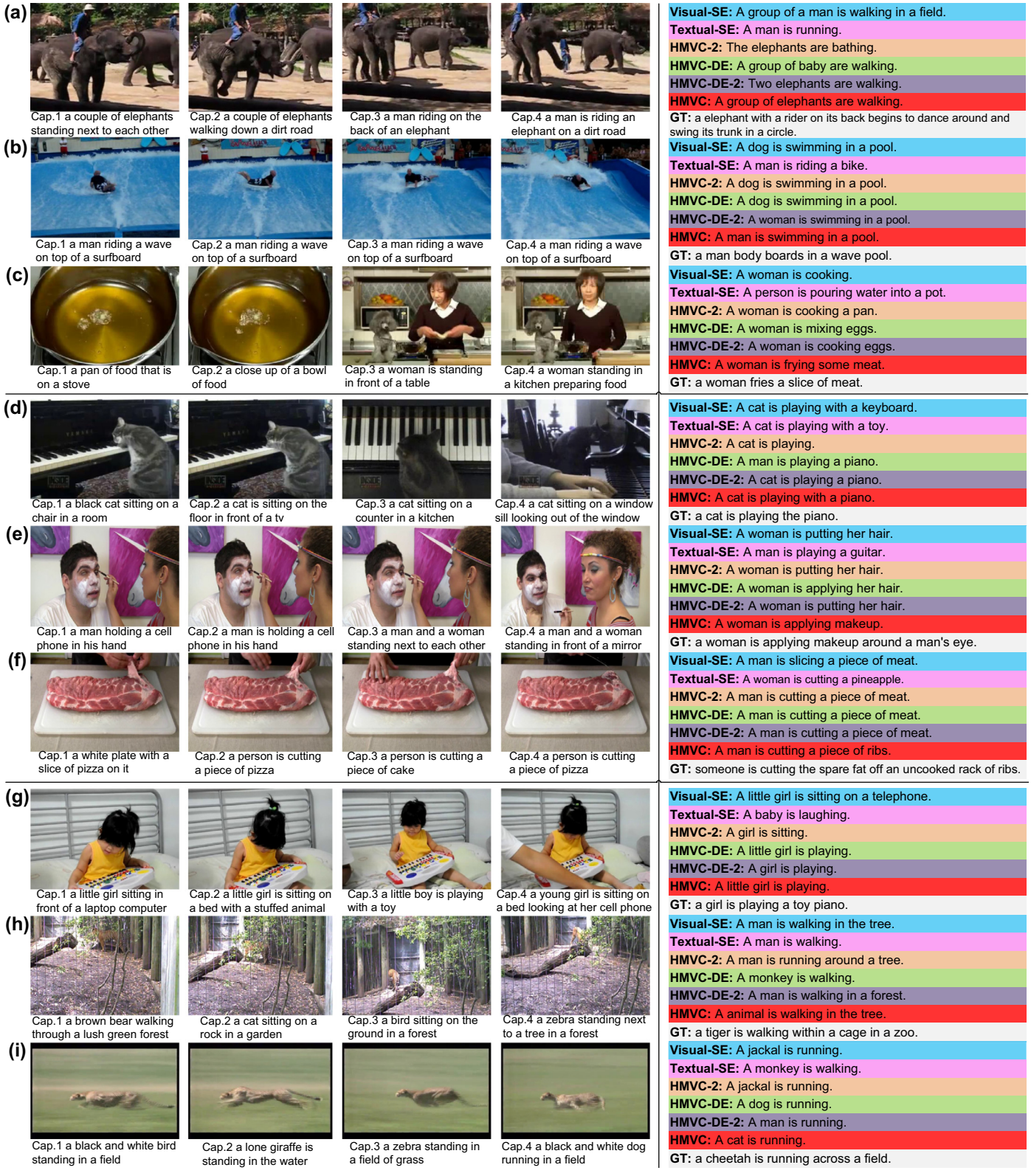

Fig. 9. Qualitative cases from the test part of MSVD dataset (Chen and Dolan, 2011). The left panel presents four randomly extracted frames and the generated captions, while the right panel reports the generated sentences from models as shown in Fig. 7. GT represents randomly selected ground truth sentence. The discussions of the caption quality is shown in Section 5.5.

on a trained frame-level image caption model. Comprehensive evaluation on MSVD (Chen and Dolan, 2011), MPII-MD (Rohrbach et al., 2015b) and MSR-VTT (Xu et al., 2016) datasets demonstrate the effectiveness of HMVC model in four standard evaluation metrics. Furthermore, we compare HMVC with its three variants networks, which illustrates the synchronized learning manner works better than the asynchronous learning manner due to the fusion of multi-model features. In addition, we also explore the generalize properties of video caption model by using different datasets for initialized model training, fine-tuning, and evaluation. To the best of our knowledge, this is the first time such properties is discussed in video caption task. In future, we intend to fully explore the intermediate components (e.g., objects, attributes, relationships, etc.) for tasks such as image captioning, video captioning and question answering.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (61472275, 61572356, 61502337, 61100124), the Tianjin Research Program of Application Foundation and Advanced Technology (15JCYBJC16200), the grant of China Scholarship Council (201506255073), the grant of Elite Scholar Program of Tianjin University (2014XRG-0046). This research is partly supported by the National Research Foundation, Prime Ministers Office, Singapore under its International Research Centre in Singapore Funding Initiative.

References

Barbu, A., Bridge, A., Burchill, Z., Coroian, D., Dickinson, S.J., Fidler, S., Michaux, A., Mussman, S., Narayanaswamy, S., Salvi, D., Schmidt, L., Shangguan, J., Siskind, J.M., Waggoner, J.W., Wang, S., Wei, J., Yin, Y., Zhang, Z., 2012. Video in sentences out. In: Conference on Uncertainty Uncertainty in Artificial Intelligence, pp. 102–112.

Barnard, K., Duygulu, P., Forsyth, D.A., de Freitas, N., Blei, D.M., Jordan, M.I., 2003. Matching words and pictures. JMLR 3, 1107–1135.

Chen, D., Dolan, W.B., 2011. Collecting highly parallel data for paraphrase evaluation. In: Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 190–200.

Chen, X., Fang, H., Lin, T., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L., 2015. Microsoft COCO captions: data collection and evaluation server. CoRR abs/1504.00325.

Chen, X., Zitnick, C.L., 2015. Mind's eye: a recurrent visual representation for image caption generation. In: CVPR, pp. 2422–2431.

Cheng, Z., Shen, J., 2016. On very large scale test collection for landmark image search benchmarking. Signal Process. 124, 13–26.

Cho, K., van Merrienboer, B., Bahdanau, D., Bengio, Y., 2014. On the properties of neural machine translation: encoder-decoder approaches. In: Workshop on Syntax, Semantics and Structure in Statistical Translation, pp. 103–111.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F., 2009. Imagenet: a large-scale hierarchical image database. In: CVPR, pp. 248–255.

Denkowski, M., Lavie, A., 2014. Meteor universal: language specific translation evaluation for any target language. In: Workshop on Statistical Machine Translation, pp. 376–380.

Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Darrell, T., Saenko, K., 2015. Long-term recurrent convolutional networks for visual recognition and description. In: CVPR, pp. 2625–2634.

Elliott, D., Keller, F., 2014. Comparing automatic evaluation measures for image description. In: ACL, pp. 452–457.

Fang, H., Gupta, S., Iandola, F.N., Srivastava, R.K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J.C., Zitnick, C.L., Zweig, G., 2015. From captions to visual concepts and back. In: CVPR, pp. 1473–1482.

Farhadi, A., Hejrati, S.M.M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D.A., 2010. Every picture tells a story: generating sentences from images. Lect. Notes Comput. Sci. 6314, 15–29.

Guadarrama, S., Krishnamoorthy, N., Malkarnenkar, G., Venugopalan, S., Mooney, R.J., Darrell, T., Saenko, K., 2013. YouTube2Text: recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In: ICCV, pp. 2712–2719.

He, X., Zhang, H., Kan, M., Chua, T., 2016. Fast matrix factorization for online recommendation with implicit feedback. In: ACM SIGIR, pp. 549–558.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Ji, S., Xu, W., Yang, M., Yu, K., 2013. 3D convolutional neural networks for human action recognition. TPAMI 35 (1), 221–231.

Jia, X., Gavves, E., Fernando, B., Tuytelaars, T., 2015. Guiding the long-short term memory model for image caption generation. In: ICCV, pp. 2407–2415.

Jia, Y., Salzmann, M., Darrell, T., 2011. Learning cross-modality similarity for multinomial data. In: ICCV, pp. 2407–2414.

Jin, Q., Chen, J., Chen, S., Xiong, Y., Hauptmann, A.G., 2016. Describing videos using multi-modal fusion. In: ACMMM, pp. 1087–1091.

Johnson, J., Karpathy, A., Fei-Fei, L., 2016. Densecap: fully convolutional localization networks for dense captioning. In: CVPR, pp. 4565–4574.

Karpathy, A., Li, F., 2015. Deep visual-semantic alignments for generating image descriptions. In: CVPR, pp. 3128–3137.

Kiros, R., Salakhutdinov, R., Zemel, R.S., 2014. Multimodal neural language models. In: ICML, pp. 595–603.

Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., Berg, T.L., 2013. Babytalk: understanding and generating simple image descriptions. TPAMI 35 (12), 2891–2903.

Kuznetsova, P., Ordonez, V., Berg, T.L., Choi, Y., 2014. TREETALK: composition and compression of trees for image descriptions. TACL 2, 351–362.

Le, Q.V., Mikolov, T., 2014. Distributed representations of sentences and documents. In: ICML, pp. 1188–1196.

Lebret, R., Pinheiro, P.H.O., Collobert, R., 2015. Phrase-based image captioning. In: ICML, pp. 2085–2094.

Li, S., Kulkarni, G., Berg, T.L., Berg, A.C., Choi, Y., 2011. Composing simple image descriptions using web-scale n-grams. In: CoNLL, pp. 220–228.

Lin, C.-Y., 2004. ROUGE: a package for automatic evaluation of summaries. In: ACL Workshop, pp. 74–81.

Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: common objects in context. Lect. Notes Comput. Sci. 8693, 740–755.

Liu, A., Su, Y., Jia, P., Gao, Z., Hao, T., Yang, Z., 2015. Multipe/single-view human action recognition via part-induced multitask structural learning. IEEE Trans. Cybern. 45 (6), 1194–1208.

Liu, A., Su, Y., Nie, W., Kankanhalli, M.S., 2017. Hierarchical clustering multi–task learning for joint human action grouping and recognition. TPAMI 39 (1), 102–114.

Liu, A.A., Xu, N., Nie, W.Z., Su, Y.T., Wong, Y., Kankanhalli, M., 2016. Benchmarking a multimodal and multiview and interactive dataset for human action recognition. IEEE Trans. Cybern. PP (99), 1–14.

Long, X., Gan, C., de Melo, G., 2016. Video captioning with multi-faceted attention. CoRR abs/1612.00234.

Mao, J., Xu, W., Yang, Y., Wang, J., Yuille, A.L., 2014. Explain images with multimodal recurrent neural networks. NIPS.

Mao, J., Xu, W., Yang, Y., Wang, J., Yuille, A.L., 2015. Deep captioning with multimodal recurrent neural networks (m-RNN). ICLR.

Nie, L., Wang, M., Gao, Y., Zha, Z., Chua, T., 2013. Beyond text QA: multimedia answer generation by harvesting web information. IEEE Trans. Multimedia 15 (2), 426–441.

Pan, P., Xu, Z., Yang, Y., Wu, F., Zhuang, Y., 2016. Hierarchical recurrent neural encoder for video representation with application to captioning. In: CVPR, pp. 1029–1038.

Pan, Y., Mei, T., Yao, T., Li, H., Rui, Y., 2016. Jointly modeling embedding and translation to bridge video and language. In: CVPR, pp. 4594–4602.

Pan, Y., Yao, T., Li, H., Mei, T., 2016. Video captioning with transferred semantic attributes. CoRR abs/1611.07675.

Papineni, K., Roukos, S., Ward, T., Zhu, W., 2002. BLUE: a method for automatic evaluation of machine translation. In: Annual Meeting of the Association for Computational Linguistics, pp. 311–318.

Pham, V., Bluche, T., Kermorvant, C., Louradour, J., 2014. Dropout improves recurrent neural networks for handwriting recognition. In: International Conference on Frontiers in Handwriting Recognition, pp. 285–290.

Rohrbach, A., Rohrbach, M., Schiele, B., 2015. The long-short story of movie description. In: GCPR, pp. 209–221.

Rohrbach, A., Rohrbach, M., Tandon, N., Schiele, B., 2015. A dataset for movie description. In: CVPR, pp. 3202–3212.

Rohrbach, M., Qiu, W., Titov, I., Thater, S., Pinkal, M., Schiele, B., 2013. Translating video content to natural language descriptions. In: ICCV, pp. 433–440.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. ICLR.

Song, J., Tang, S., Xiao, J., Wu, F., Zhang, Z.M., 2016. Lstm-in-lstm for generating long descriptions of images. Comput. Visual Media 2 (4), 379–388.

Sun, C., Gan, C., Nevatia, R., 2015. Automatic concept discovery from parallel text and visual corpora. In: ICCV, pp. 2596–2604.

Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. In: NIPS, pp. 3104–3112.

Thomason, J., Venugopalan, S., Guadarrama, S., Saenko, K., Mooney, R.J., 2014. Integrating language and vision to generate natural language descriptions of videos in the wild. In: COLING, pp. 1218–1227.

Vedantam, R., Zitnick, C.L., Parikh, D., 2015. CIDEr: consensus-based image description evaluation. In: CVPR, pp. 4566–4575.

Venugopalan, S., Hendricks, L.A., Mooney, R.J., Saenko, K., 2016. Improving lstm-based video description with linguistic knowledge mined from text. In: EMNLP, pp. 1961–1966.

Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R.J., Darrell, T., Saenko, K., 2015. Sequence to sequence - video to text. In: ICCV, pp. 4534–4542.

Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R.J., Saenko, K., 2015. Translating videos to natural language using deep recurrent neural networks. In: NAACL HLT, pp. 1494–1504.

Vinyals, O., Toshev, A., Bengio, S., Erhan, D., 2015. Show and tell: a neural image caption generator. In: CVPR, pp. 3156–3164.

Wu, Q., Wang, P., Shen, C., Dick, A.R., van den Hengel, A., 2016. Ask me anything: free-form visual question answering based on knowledge from external sources. In: CVPR, pp. 4622–4630.
Xu, H., Venugopalan, S., Ramanishka, V., Rohrbach, M., Saenko, K., 2015. A multi-scale multiple instance video description network. CoRR abs/1505.05914.
Xu, J., Mei, T., Yao, T., Rui, Y., 2016. MSR-VTT: a large video description dataset for bridging video and language. In: CVPR, pp. 5288–5296.
Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A.C., Salakhutdinov, R., Zemel, R.S., Bengio, Y., 2015. Show, attend and tell: neural image caption generation with visual attention. In: ICML, pp. 2048–2057.
Xu, R., Xiong, C., Chen, W., Corso, J.J., 2015. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In: AAAI, pp. 2346–2352.
Yan, Y., Ricci, E., Subramanian, R., Lanz, O., Sebe, N., 2013. No matter where you are: Flexible graph-guided multi-task learning for multi-view head pose classification under target motion. International Conference on Computer Vision (ICCV).
Yan, Y., Ricci, E., Subramanian, R., Liu, G., Sebe, N., 2014. Multi-task linear discriminant analysis for multi-view action recognition. IEEE Transactions on Image Processing (TIP) 23 (12), 5599–5611.
Yan, Y., Ricci, E., Subramanian, R., Liu, G., Lanz, O., Sebe, N., 2016. A multi-task learning framework for head pose estimation under target motion. IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI) 38 (6), 1070–1083.
Yang, Y., Teo, C.L., Daumé III, H., Aloimonos, Y., 2011. Corpus-guided sentence generation of natural images. In: EMNLP, pp. 444–454.
Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C.J., Larochelle, H., Courville, A.C., 2015. Describing videos by exploiting temporal structure. In: ICCV, pp. 4507–4515.
Young, P., Lai, A., Hodosh, M., Hockenmaier, J., 2014. From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. TACL 2, 67–78.
Yu, H., Wang, J., Huang, Z., Yang, Y., Xu, W., 2016. Video paragraph captioning using hierarchical recurrent neural networks. In: CVPR, pp. 4584–4593.
Zha, Z., Mei, T., Wang, Z., Hua, X., 2007. Building a comprehensive ontology to refine video concept detection. In: SIGMM, pp. 227–236.
Zhang, H., Shang, X., Luan, H., Wang, M., Chua, T., 2016. Learning from collective intelligence: feature learning using social images and tags. TOMCCAP 13 (1), 1:1–1:23.
Zhang, H., Zha, Z.-J., Yang, Y., Yan, S., Gao, Y., Chua, T.-S., 2013. Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval. In: ACMMM, pp. 33–42.
Zhang, X., Gao, K., Zhang, Y., Zhang, D., Tian, Q., 2017. Task-driven dynamic fusion: reducing ambiguity in video description. CVPR.
Zhu, L., Shen, J., Xie, L., Cheng, Z., 2017. Unsupervised visual hashing with semantic assistant for content-based image retrieval. IEEE Trans. Knowl. Data Eng. 29 (2), 472–486.