Video Captioning with Transferred Semantic Attributes*

Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei

University of Science and Technology of China, Hefei, China   Microsoft Research, Beijing, China

`panyw.ustc@gmail.com, {tiyao, tmei}@microsoft.com, lihq@ustc.edu.cn`

Abstract

Automatically generating natural language descriptions of videos plays a fundamental challenge for computer vision community. Most recent progress in this problem has been achieved through employing 2-D and/or 3-D Convolutional Neural Networks (CNNs) to encode video content and Recurrent Neural Networks (RNNs) to decode a sentence. In this paper, we present Long Short-Term Memory with Transferred Semantic Attributes (LSTM-TSA)—a novel deep architecture that incorporates the transferred semantic attributes learnt from images and videos into the CNN plus RNN framework, by training them in an end-to-end manner. The design of LSTM-TSA is highly inspired by the facts that 1) semantic attributes play a significant contribution to captioning, and 2) images and videos carry complementary semantics and thus can reinforce each other for captioning. To boost video captioning, we propose a novel transfer unit to model the mutually correlated attributes learnt from images and videos. Extensive experiments are conducted on three public datasets, i.e., MSVD, M-VAD and MPII-MD. Our proposed LSTM-TSA achieves to-date the best published performance in sentence generation on MSVD: 52.8% and 74.0% in terms of BLEU@4 and CIDEr-D. Superior results are also reported on M-VAD and MPII-MD when compared to state-of-the-art methods.

1. Introduction

Video captioning, which is known as describing videos with natural language, has brought a profound challenge to both computer vision and language processing communities. Intensive research interests have been paid for this emerging topic.

Existing approaches to video captioning have evolved through two dimensions: template-based language model [8, 20, 33] and sequence learning method [15, 29, 34, 37]. The former predefines a set of templates for sentence generation following specific grammar rules and aligns each part

---

*This work was performed when Yingwei Pan was visiting Microsoft Research as a research intern.

Input Video:

Attributes from Images:

young, girl, holding, child, little, floor, pair, it, woman, playing

Attributes from Videos:

person, doing, man, room, boy, cleaning, machine, his, someone, riding

Video Caption:

a boy is cleaning the floor

Figure 1. An example of video description generation. The input is a short video clip and the attributes are learnt from images and videos, respectively. The output is a sentence generated by our LSTM-TSA architecture.

of sentence with video content. This category of model, however, highly depends on the pre-defined templates and thus the generated sentences are always with a constant syntactical structure. Sequence learning method, in contrast, is to leverage sequence learning models to directly translate the video content into a sentence, which is mainly inspired from the recent advances by using Recurrent Neural Networks (RNNs) in machine translation [24]. The spirit behind is an encoder-decoder mechanism for translation. More specifically, an encoder 2-D/3-D Convolutional Neural Network (CNN) reads a video and produces a vector of video representations, which in turn is fed into a decoder RNN that generate a natural sentence. While encouraging performances are reported, the CNNs plus RNNs-based sequence learning approaches translate directly from video representations to language, leaving the high-level semantic cues in the video under explored. Moreover, high-level semantic information, i.e., semantic attributes, has shown effective in the vision to language tasks [31] (e.g., image captioning and visual Q&A).

This paper proposes a novel deep architecture, named Long Short-Term Memory with Transferred Semantic Attributes (LSTM-TSA), which takes advantages of incorporating semantic attributes into sequence learning for video captioning. More importantly, take the given video in Figure 1 as an example, the semantic properties learnt from

images often depict static objects and scenes (e.g., "girl," "child" and "floor") while the semantics extracted from videos convey temporal dynamics (e.g., "doing," "cleaning" and "riding"). This has made the attributes mined from images and videos complementary to each other for the generation of sentence (e.g., "a boy is cleaning the floor"). We investigate how the attributes from the two sources can be leveraged for enhancing video captioning. Specifically, given a video, a 2-D/3-D CNN is utilized to extract visual features of selected video frames/clips and the video representations are produced by mean pooling over these visual features. Then, a LSTM network for generating video description is learnt by feeding into both video representations and semantic attributes mined from images and videos. To better leverage the attributes from two sources, a transfer unit is devised to dynamically balance the influence in between given the input word and the hidden state in the LSTM.

The main contribution of this work is the proposal of LSTM-TSA for addressing the issue of exploiting the mutual relationship between video representations and attributes for boosting video captioning. This issue also leads to an elegant view of how complementary attributes from images and videos are jointly exploited for sentence generation, which is a problem not yet fully explored in the literature.

2. Related Work

We briefly group the related works into two categories: video captioning and sequence learning by using attributes. The former draws upon research in automatically generating description to a video, and the later investigates sequence learning for visual content by utilizing the attributes.

Video Captioning. The research in this direction has proceeded along two different dimensions: template-based language methods [8, 11, 20, 33] and sequence learning approaches (e.g., RNNs) [15, 22, 29, 30, 32, 34, 37]. Template-based language methods firstly align each sentence fragments (e.g., subject, verb, object) with detected words from visual content and then generate the sentence with predefined language template. Obviously, most of them highly depend on the templates of sentence and always generate sentence with syntactical structure. [11] is one of the earlier works that builds a concept hierarchy of actions for natural language description of human activities. Rohrbach *et al.* learn a CRF to model the relationships between different components of the input video and generate description for video [20]. Recently, a deep joint video-language embedding model in [33] is designed for video sentence generation. Different from template-based language methods, sequence learning approaches learn the probability distribution in the common space of visual content and textual sentence to generate novel sentences with more flexible syntactical structure. In [30], Venugopalan *et al.* present a LSTM based model to generate video description

tions with the mean pooling representation over all frames. The framework is then extended by inputting both frames and optical flow images into an encoder-decoder LSTM in [29]. Furthermore, Pan *et al.* additionally consider the relevance between sentence semantics and video content as a regularizer in LSTM based architecture [15]. Compared to mean pooling, Yao *et al.* propose to utilize the temporal attention mechanism to exploit temporal structure for video captioning [34].

Sequence Learning by Using Attributes. Attributes are properties observed in visual content with rich semantic cues and have been widely studied in computer vision for improving the efficacy of visual recognition [17]. Following this elegant recipe, several recent works have attempted to inject attributes into sequence learning for image caption generation. Fang *at al.* [6] leverage Multiple Instance Learning to train attributes detector and then generate sentence through a maximum-entropy language model based on the outputs of attributes detector. Recently, in [31], high-level concepts/attributes are shown to obtain clear improvements on image captioning task when injected into existing state-of-the-art RNN-based model and such visual attributes are also utilized as semantic attention in [36] to enhance image captioning. Most recently, Yao *et al.* [35] feed both image and attributes into RNNs in different ways for enhancing image description generation.

Summary. Our work aims to leverage semantic attributes in video captioning. Different from most of the aforementioned sequence learning models using attributes which mainly focus on sentence generation by solely depending on the attributes learnt in domain, our work contributes by studying not only learning attributes in videos from both image and video domains, but also how the attributes could be better fused by dynamically offering a transfer unit in between for boosting video captioning.

3. Approach

We devise our CNN plus RNN architecture to generate video descriptions under the umbrella of incorporating mined semantic attributes from images and videos. Specifically, we begin this section by presenting the problem formulation and how to learn semantic attributes in videos, followed by our proposed LSTM-TSA video captioning framework. In particular, several variants of our designed transfer unit which is utilized to fuse the attributes learnt from two sources are investigated and discussed.

3.1. Problem Formulation

Suppose we have a video V with N_v sample frames/clips (uniform sampling) to be described by a textual sentence $\mathcal{S}$, where $\mathcal{S} = \{w_1, w_2, ..., w_{N_s}\}$ consisting of N_s words. Let $\mathbf{v} \in \mathbb{R}^{D_v}$ and $\mathbf{w}_t \in \mathbb{R}^{D_w}$ denote the D_v-dimensional video representations of the video V and the D_w-dimensional

textual features of the t-th word in sentence $\mathcal{S}$, respectively. As a sentence consists of a sequence of words, a sentence can be represented by a $D_w \times N_s$ matrix $\mathbf{W} \equiv [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_{N_s}]$, with each word in the sentence as its column vector. Furthermore, we have another two feature vectors $\mathbf{A_i} \in \mathbb{R}^{D_{a_i}}$ and $\mathbf{A_v} \in \mathbb{R}^{D_{a_v}}$ to represent the probability distribution over the high-level attributes for video $\mathcal{V}$ learnt from images and videos, respectively. More details about how we mine and represent the attributes from images and videos will be introduced in Section 3.2.

Inspired by the recent successes of probabilistic sequence models leveraged in statistical machine translation [24] and semantic attributes utilized in image captioning [6, 36], we aim to formulate our video captioning model in an end-to-end fashion based on LSTM [9] which encodes the given video and its learnt attributes from both images and videos into a fixed dimensional vector and then decodes it to the output target sentence. Hence, the video sentence generation problem we exploit here can be formulated by minimizing the following energy loss function as

$$E(\mathbf{v}, \mathbf{A_i}, \mathbf{A_v}, \mathcal{S}) = -\log \Pr(\mathcal{S}|\mathbf{v}, \mathbf{A_i}, \mathbf{A_v}), \quad (1)$$

which is the negative $\log$ probability of the correct textual sentence given the video and detected attributes from both images and videos.

Since the model produces one word in the sentence at each time step, it is natural to apply chain rule to model the joint probability over the sequential words. Thus, the $\log$ probability of the sentence is given by the sum of the $\log$ probabilities over the word and can be expressed as

$$\log \Pr(\mathcal{S}|\mathbf{v}, \mathbf{A_i}, \mathbf{A_v}) = \sum_{t=1}^{N_S} \log \Pr(\mathbf{w}_t|\mathbf{v}, \mathbf{A_i}, \mathbf{A_v}, \mathbf{w}_0, \ldots, \mathbf{w}_{t-1}). \quad (2)$$

By minimizing this loss, the contextual relationship among the words in the sentence can be guaranteed given the video and its learnt attributes from images and videos.

3.2. Semantic Attributes in Video

Attributes Learnt from images. We draw inspiration from recent advances in attribute detection for image captioning [6, 36] and adopt the weakly-supervised approach of Multiple Instance Learning (MIL) on image captioning benchmarks (e.g., COCO [12]) to learn attribute detectors. For an attribute w_a, one image I is regarded as a positive bag of regions (instances) if w_a exists in image I's ground-truth sentences, and negative bag otherwise. By inputting all the bags into a noisy-OR MIL model [38], the probability of the bag b_I which contains attribute w_a is measured on the probabilities of all the regions in the bag as

$$\Pr_I^{w_a} = 1 - \prod_{r_i \in b_I} (1 - p_i^{w_a}), \quad (3)$$
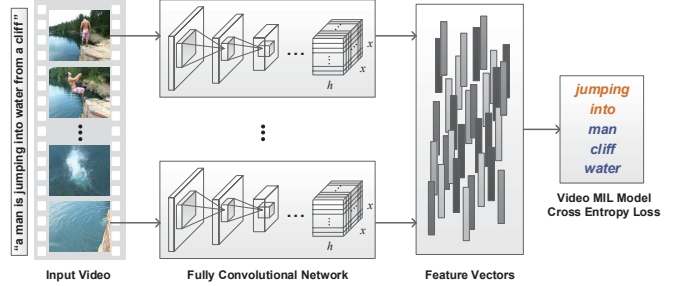

Figure 2. Video MIL framework.

where $p_i^{w_a}$ is the probability of the attribute w_a predicted by region r_i and can be calculated through a sigmoid layer after the last convolutional layer in the CNN architecture [6]. Here the adopted CNN architecture is a fully convolutional network extended from recent popular CNN [23] that shows superior performance for video representation learning [7, 14]. Specifically, the dimension of convolutional activations from the last convolutional layer is $x \times x \times h$ and h represents the representation dimension of each region, resulting in $x \times x$ response map which preserves the spatial dependency of the image. Then, a cross entropy loss is calculated based on the probabilities of all the attributes at the top of the whole architecture to optimize image MIL model. With the learnt image MIL model on image captioning dataset, we compute the probability distribution on all the attributes for each sampled frame and perform mean pooling over distributions of all the sampled frames to obtain the final representations $\mathbf{A_i}$ of attributes learnt from images.

Attributes Learnt from videos. To detect attributes from videos, one natural way is to directly train image MIL model on video frames. However, as a video is a sequence of frames with large variations, simply assigning video-level description to each sampled frame will lead to the issue of semantics shift and thus involve noise in the process of attribute learning. To solve this problem, a video MIL model is particularly devised to learn attributes from videos, as shown in Figure 2.

Given an attribute w_a, we treat the spatial regions of all the N_V sampled frames in video V as one bag, which is considered as positive if w_a exists in video V's descriptions and negative otherwise. By feeding all the bags into the fully convolutional network with the same architecture in image MIL model, we calculate the probability of bag b_V which contains attribute w_a on the probabilities of all the regions in the bag as

$$\Pr_V^{w_a} = 1 - \prod_{j \in [1, N_V]} \prod_{r_{ij} \in b_V^{(j)}} (1 - p_{ij}^{w_a}), \quad (4)$$

where $p_{ij}^{w_a}$ is the probability of the attribute w_a predicted by the i-th region in the j-th frame and $b_V^{(j)}$ denotes the set of
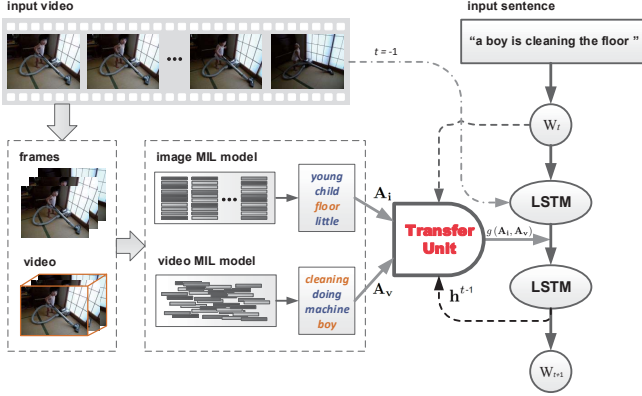
Figure 3. The overview of our LSTM-TSA for video captioning (better viewed in color). The video representation is produced by mean pooling over the visual features of sampled frames/clips extracted by a 2-D/3-D CNN, which is injected into LSTM only at the initial time. Image and video MIL models are used to mine semantic attributes from images and videos respectively, which are additionally incorporated into LSTM for boosting video captioning. To better leverage the mined attributes from two sources, a transfer unit is devised to dynamically fuse them into LSTM.

all the regions in the j-th frame. Specifically, in our training, all the N_V sampled frames from one video are taken as a batch and each frame is fed into the same fully convolutional network followed by a sigmoid layer, resulting in $x \times x$ response map whose element represents the probability $p_{ij}^{w_a}$ of attribute w_a detected in region r_{ij}. Similar to image MIL model, a cross entropy loss layer is designed at the top of the whole architecture to optimize our video MIL model. As such, the proposed video MIL model is trained holistically among all the frames in the video and the probability distribution calculated by Eq.(4) are employed as representations $\mathbf{A_v}$ of attributes learnt from videos.

3.3. Video Captioning with Semantic Attributes

With the detected high-level semantic attributes learnt from images and videos, we propose a Long Short-Term Memory with Transferred Semantic Attributes from Images and Videos (LSTM-TSA$_{IV}$) model for video captioning. The basic idea of LSTM-TSA$_{IV}$ is to translate the video representation from a 2-D and 3-D CNN to the desired output sentence through LSTM-type RNN model by additionally injecting the high-level semantic attributes learnt from both images and videos. Specifically, a transfer unit is designed to dynamically control the impacts of semantic attributes from the two sources on sentence generation.

3.3.1 Attributes-based LSTM-type Video Captioning

Inspired by the best-performing architecture (factored, two-layer LSTM) in LRCN [5], we devise our attributes-based LSTM-type video captioning model by injecting both video

representation and its detected semantic attributes learnt from images and videos into LSTM, as illustrated in Figure 3. In particular, our LSTM-TSA$_{IV}$ model firstly encodes video representation $\mathbf{v}$ at the initial step and then feeds attributes representations from images and videos as the additional inputs to the second-layer LSTM unit at each time step to emphasize the semantic information more frequently. The LSTM updating procedure in LSTM-TSA$_{IV}$ is as

$$\mathbf{x}^{-1} = f_1\left(\mathbf{T}_v \mathbf{v}\right) + g\left(\mathbf{A_i}, \mathbf{A_v}\right), \tag{5}$$

$$\mathbf{x}^t = f_1\left(\mathbf{T}_s \mathbf{w}_t\right) + g\left(\mathbf{A_i}, \mathbf{A_v}\right), t \in \{0, \ldots, N_s - 1\}, \tag{6}$$

$$\mathbf{h}^t = f_2\left(\mathbf{x}^t\right), t \in \{0, \ldots, N_s - 1\}, \tag{7}$$

where D_e is the dimension of LSTM input, $\mathbf{T}_v \in \mathbb{R}^{D_e \times D_v}$ and $\mathbf{T}_s \in \mathbb{R}^{D_e \times D_w}$ are the transformation matrices for video representation and textual features of word, $\mathbf{x}^t$ and $\mathbf{h}^t$ are the inputs and cell output of the second-layer LSTM unit, f_1 and f_2 are the updating functions within the first/second-layer LSTM units, and g is the transformation function to transfer both $\mathbf{A_i}$ and $\mathbf{A_v}$ into the second-layer LSTM unit.
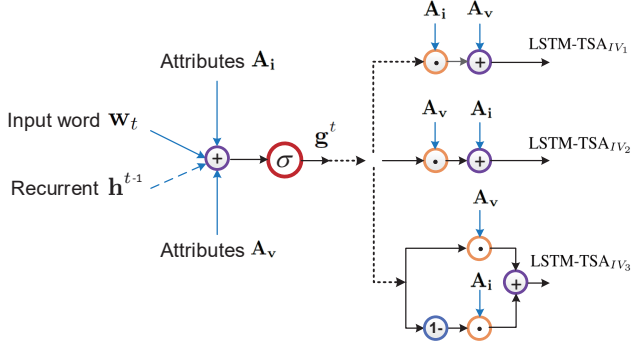
3.3.2 Transfer Unit

To contextually transfer the information of semantic attributes from multiple sources into LSTM, we devise a novel transfer unit, which is treated as the core unit in our proposed LSTM-TSA$_{IV}$ model.

Transfer Gate. A novel gate architecture, named as transfer gate, is especially designed to control the impact of semantic attributes by taking contextual information into account, which is the left part of transfer unit as shown in Figure 4. At the t-th time step, the transfer gate encapsulates both the static information (attributes learnt from images and videos) and dynamic (contextual) information (current input word and previous LSTM hidden state) to select valuable knowledge from attributes, which is applied with feature transformation, to produce a fix-length weight vector and followed by a sigmoid function to squash the real-valued weight vector to a range of $[0, 1]$. Such output weight vector $\mathbf{g}^t$ for transfer gate is computed as

$$\mathbf{g}^t = \sigma(\mathbf{G}_s \mathbf{w}_t + \mathbf{G}_h \mathbf{h}^{t-1} + \mathbf{G}_i \mathbf{A_i} + \mathbf{G}_v \mathbf{A_v}), \tag{8}$$

where D_h is the dimension of LSTM cell output, $\mathbf{G}_s \in \mathbb{R}^{D_e \times D_w}$, $\mathbf{G}_h \in \mathbb{R}^{D_e \times D_h}$, $\mathbf{G}_i \in \mathbb{R}^{D_e \times D_{a_i}}$ and $\mathbf{G}_v \in \mathbb{R}^{D_e \times D_{a_v}}$ are the transformation matrices for textual features of word, cell output of LSTM, representation of attributes learnt from images and videos, respectively, and sigmoid σ is element-wise non-linear activation function.

Transfer Unit

Legend

┈┈▶	Unweighted connections	⊙	Function of dot product
┈─▶	Weighted connections	⊕	Function of sum
σ	Gate activation function	1-	Function of constant subtraction

Figure 4. Three different architectures of transfer unit with transfer gate (left side) in our LSTM-TSA$_{IV}$ framework.

LSTM with Transfer Unit. Then, we formulate our video captioning with semantic attributes learnt from two sources as a multi-source sequence learning problem and modify the architectures of transfer unit which is treated as the additional input to LSTM for our purpose. The core issue for the modification is about whether the transfer gate in our transfer unit should *individually* or *simultaneously* impact the semantic attributes learnt from different sources. Individual impact means that the transfer gate only critically control the information transferred from attributes in one specific source, while directly leverages the attributes from other source unconditionally. Simultaneous impact decouples the influence of transfer gate such that attributes learnt from different sources can be simultaneously guided with transfer gate.

Our preliminary design LSTM-TSA$_{IV_0}$ is the deep fusion without transfer gate by directly utilizing the multimodal layer. Specifically, the additional input to LSTM is calculated as

$$\text{LSTM-TSA}_{IV_0}: g\left(\mathbf{A_i}, \mathbf{A_v}\right) = \mathbf{T_{A_i}}\mathbf{A_i} + \mathbf{T_{A_v}}\mathbf{A_v}, \qquad (9)$$

where $\mathbf{T_{A_i}} \in \mathbb{R}^{D_e \times D_{a_i}}$ and $\mathbf{T_{A_v}} \in \mathbb{R}^{D_e \times D_{a_v}}$ are the transformation matrices for representation of attributes learnt from images and videos, respectively. Please also note that if only semantic attributes learnt from one single source (images/videos) are available, the additional input $g\left(\mathbf{A_i}, \mathbf{A_v}\right)$ to LSTM in LSTM-TSA will be degraded into $g\left(\mathbf{A_i}\right) = \mathbf{T_{A_i}}\mathbf{A_i}$ or $g\left(\mathbf{A_v}\right) = \mathbf{T_{A_v}}\mathbf{A_v}$ and we name these two variants as LSTM-TSA$_I$ and LSTM-TSA$_V$.

Then based on the above core design issue, we derive three different architectures of transfer unit as depicted in Figure 4, respectively named as LSTM-TSA$_{IV_1}$ to LSTM-TSA$_{IV_3}$. The first design (LSTM-TSA$_{IV_1}$) individually as-

signs the attributes learnt from images with the weight vector of transfer gate to dynamically select the favorable information which will be fused as the additional input to L-STM. The second design (LSTM-TSA$_{IV_2}$) is similar except that the calculated weight vector of transfer gate is only allocated to the attributes learnt from videos. Both designs are relatively straightforward to implement by multiplying the transformed representation of attributes from one specific source with the weight vector of transfer gate through dot product. The last design (LSTM-TSA$_{IV_3}$) is a compromise version between the former two architectures, by simultaneously controlling the two attributes learnt from different sources with decoupled weight vectors from transfer gate, which is also treated as a linear combination between the attributes learnt from images and videos.

Specifically, given the output weight vector $\mathbf{g}^t$ of transfer gate in the time step t, the three variants of our transfer unit are designed as

$$\text{LSTM-TSA}_{IV_1}: g\left(\mathbf{A_i}, \mathbf{A_v}\right) = \mathbf{T_{A_i}}\mathbf{A_i} \odot \mathbf{g}^t + \mathbf{T_{A_v}}\mathbf{A_v}, \qquad (10)$$

$$\text{LSTM-TSA}_{IV_2}: g\left(\mathbf{A_i}, \mathbf{A_v}\right) = \mathbf{T_{A_i}}\mathbf{A_i} + \mathbf{T_{A_v}}\mathbf{A_v} \odot \mathbf{g}^t, \qquad (11)$$

$$\text{LSTM-TSA}_{IV_3}: g\left(\mathbf{A_i}, \mathbf{A_v}\right) = \mathbf{T_{A_i}}\mathbf{A_i} \odot \left(1 - \mathbf{g}^t\right) + \mathbf{T_{A_v}}\mathbf{A_v} \odot \mathbf{g}^t, \quad (12)$$

where $\odot$ denotes the element-wise dot product function.

4. Experiments

We evaluate and compare our proposed LSTM-TSA with state-of-the-art approaches by conducting video captioning task on three video captioning benchmarks, i.e., Microsoft Research Video Description Corpus (MSVD) [3], Montreal Video Annotation Dataset (M-VAD) [26] and MPII Movie Description Corpus (MPII-MD) [19]. The first is the most popular video captioning benchmark of YouTube videos and the other two are both recently released large-scale movie description datasets.

4.1. Datasets and Settings

MSVD. MSVD contains 1,970 video snippets collected from YouTube. There are roughly 40 available English descriptions per video. In our experiments, we follow the setting used in prior works [8, 15], taking 1,200 videos for training, 100 for validation and 670 for testing.

M-VAD. M-VAD is a recent collection of large-scale movie description dataset. It is composed of about 49,000 DVD movie snippets, which are extracted from 92 DVD movies. Each movie clip is accompanied with single sentence from semi-automatically transcribed descriptive video service (DVS) narrations.

MPII-MD. MPII-MD is another recent collection of movie description dataset, similar to M-VAD. It contains around 68,000 movie snippets from 94 Hollywood movies and each snippet is equipped with a single sentence from movie scripts and DVS.

Settings. We uniform sample 25 frames/clips for each video and each word in the sentence is represented as "one-hot" vector (binary index vector in a vocabulary). For video representations, we take the output of 4096-way fc6 layer from the 19-layer VGG [23] pre-trained on Imagenet ILSVRC12 dataset [21] and 4096-way fc6 layer from C3D [27] pre-trained on Sports-1M video dataset [10] as frame/clip representation respectively, and concatenate the features from VGG and C3D as the input video representation. For representation of attributes learnt from images, we select the 1,000 most common words on COCO [12] as the high-level semantic attributes in the image domain and train the attribute detectors with image MIL model [6] purely on the COCO training data, resulting in the final 1,000-way vector of probabilities. For the representation of attributes learnt from videos, 1,000 most common words on each video captioning benchmark are selected individually as semantic attributes in each specific video domain and the corresponding attribute detectors are trained with proposed video MIL model. The dimension of the input and hidden layers in LSTM are both set to 1,024. In testing stage, we adopt the beam search strategy and set the beam size to 4.

For quantitative evaluation of our proposed models, we adopt three common metrics in image/video captioning tasks: BLEU@N [16], METEOR [2], and CIDEr-D [28]. All the metrics are computed by using the codes[1] released by Microsoft COCO Evaluation Server [4].

4.2. Compared Approaches

To empirically verify the merit of our LSTM-TSA models, we compared the following state-of-the-art methods.

(1) LSTM [30]: LSTM attempts to directly translate from video pixels to natural language with a CNN plus RNN framework. The video representation is generated by performing mean pooling over the frame features across the entire video.

(2) Sequence to Sequence–Video to Text (S2VT) [29]: S2VT incorporates both RGB and optical flow inputs, and the encoding and decoding of the inputs and word representations are learnt jointly in a parallel manner.

(3) Temporal Attention (TA) [34]: TA combines the frame representation from GoogleNet [25] and video clip representation based on a 3-D CNN trained on hand-crafted descriptors. Furthermore, a weighted attention mechanism is exploited to dynamically attend to specific temporal regions of the video while generating sentence.

(4) Long Shot-Term Memory with visual-semantic Embedding (LSTM-E) [15]: LSTM-E utilizes both 2-D CNN and 3-D CNN to learn video representation, and simultaneously explores the learning of LSTM and visual-semantic embedding for video captioning.

(5) Convolutional Gated-Recurrent-Unit Recurrent Networks (GRU-RCN) [1]: GRU-RCN leverages convolutional GRU-RNN to extract visual representation and generate sentence based on the LSTM text-generator with soft-attention mechanism [34].

(6) hierarchical Recurrent Neural Networks (h-RNN) [37]: Proposed most recently, h-RNN exploits both spatial and temporal attention mechanisms for video captioning.

(7) Hierarchical Recurrent Neural Encoder (HRNE) [13]: HRNE encodes the frame sequence with hierarchical RNN and decodes the sentence with attention mechanism.

(8) Long Short-Term with Transferred Semantic Attributes (LSTM-TSA): We design three runs for our proposed framework, i.e., LSTM-TSA$_I$, LSTM-TSA$_V$, and LSTM-TSA$_{IV}$. The input semantic attributes of the first two runs LSTM-TSA$_I$ and LSTM-TSA$_V$ are purely mined from images and videos, respectively. The last run LSTM-TSA$_{IV}$ is to fuse semantic attributes from both images and videos. Note that LSTM-TSA$_{IV_3}$ is particularly exploited as LSTM-TSA$_{IV}$ here. The comparisons between four variants of LSTM-TSA$_{IV}$ w or w/o transfer gate will be discussed in Section 4.4.

4.3. Performance Comparison

Quantitative Analysis. Table 1 shows the performances of different models on MSVD dataset. Overall, the results across six evaluation metrics consistently indicate that our proposed LSTM-TSA$_{IV}$ achieves better performance than all the state-of-the-art techniques including non-attention models (LSTM, S2VT, LSTM-E) and attention-based approaches (TA, GRU-RCN, h-RNN, HRNE). In particular, the CIDEr-D of our LSTM-TSA$_{IV}$ can achieve 74.0% which is to-date the highest performance reported on MSVD dataset, making the relative improvement over TA, GRU-RCN, h-RNN by 43.1%, 8.8%, and 12.5%, respectively. By additionally incorporating attributes to LSTM model, LSTM-TSA$_I$ and LSTM-TSA$_V$ lead to a performance boost, indicating that visual representations are augmented with high-level semantic attributes and thus do benefit the learning of video sentence generation. As expected, LSTM-TSA$_V$ whose attributes are trained in domain outperforms LSTM-TSA$_I$ which predicts the attributes learnt on image domain. LSTM-TSA$_{IV}$ utilizing attributes learnt from images and videos significantly improves LSTM-TSA$_V$. The result indicates the advantage of leveraging the learnt attributes jointly from two domains which are complementary for boosting video captioning.

The performance comparisons in terms of METEOR on two movie datasets M-VAD and MPII-MD are summarized in Table 2. The METEOR scores on the two datasets are much lower than those on MSVD, due to the high diversity of visual and textual content in movies. Our LSTM-TSA$_{IV}$ consistently outperforms other baselines in two datasets.

Table 1. METEOR, CIDEr-D, and BLEU@N scores of our LSTM-TSA and other state-of-the-art methods on MSVD dataset. All values are reported as percentage (%).

Model	METEOR	CIDEr-D	BLEU@1	BLEU@2	BLEU@3	BLEU@4
LSTM [30]	29.1	-	-	-	-	33.3
S2VT [29]	29.8	-	-	-	-	-
TA [34]	29.6	51.7	80.0	64.7	52.6	41.9
LSTM-E [15]	31.0	-	78.8	66.0	55.4	45.3
GRU-RCN [1]	31.6	68.0	-	-	-	43.3
h-RNN [37]	32.6	65.8	81.5	70.4	60.4	49.9
HRNE [13]	33.1	-	79.2	66.3	55.1	43.8
LSTM-TSA$_I$	32.4	71.5	81.0	69.6	60.2	50.2
LSTM-TSA$_V$	32.6	71.7	82.1	70.7	61.1	50.5
LSTM-TSA$_{IV}$	**33.5**	**74.0**	**82.8**	**72.0**	**62.8**	**52.8**

Table 2. METEOR (M) scores (%) of our LSTM-TSA and other state-of-the-art methods on (a) M-VAD and (b) MPII-MD datasets.

(a) M-VAD dataset.

Model	M
TA [34]	4.3
LSTM [30]	6.1
Visual-Labels [18]	6.4
S2VT [29]	6.7
LSTM-E [15]	6.7
HRNE [13]	6.8
LSTM-TSA$_I$	6.4
LSTM-TSA$_V$	6.9
LSTM-TSA$_{IV}$	**7.2**

(b) MPII-MD dataset.

Model	M
SMT [19]	5.6
LSTM [30]	6.7
Visual-Labels [18]	7.0
S2VT [29]	7.1
LSTM-E [15]	7.3
LSTM-TSA$_I$	7.4
LSTM-TSA$_V$	7.6
LSTM-TSA$_{IV}$	**8.0**

The METEOR of LSTM-TSA$_{IV}$ can reach 7.2% and 8.0%, which makes the relative improvement over the best competitor HRNE in M-VAD and LSTM-E in MPII-MD by 5.9% and 9.6%, respectively. Similar to the observations on MSVD, LSTM-TSA$_I$ and LSTM-TSA$_V$ exhibit better performance than LSTM by further taking attributes into account for video captioning. In addition, LSTM-TSA$_V$ performs better than LSTM-TSA$_I$ and larger degree of improvement is attained when exploiting attributes from both images and videos by LSTM-TSA$_{IV}$.

Qualitative Analysis. Figure 5 shows a few video examples with the detected semantic attributes from images and videos respectively, human-annotated ground truth sentences and sentences generated by two approaches, i.e., LSTM and our LSTM-TSA$_{IV}$. From these exemplar results, it is easy to see that the two automatic methods can generate somewhat relevant and logically correct sentences, while our model LSTM-TSA$_{IV}$ can predict more accurate words by jointly exploiting video representations and semantic attributes learnt from images and videos for enhancing video captioning. For instance, compared to subject term "a man" and verb term "cutting" in the sentence generated by LSTM for the first video, "a woman" and "lying" in our LSTM-TSA$_{IV}$ are more relevant to the video content,

since the word "woman" and "lying" predicted as one attribute from images and videos respectively are directly fed into LSTM to guide the sentence generation. Similarly, verb term "cleaning" detected as an attribute from videos and object term "floor" learnt from images present the third image more exactly. Moreover, our LSTM-TSA$_{IV}$ can generate more descriptive sentence by enriching the semantics with attributes. For instance, with the detected term "forest," the generated sentence "a bear is walking in the forest" of the fifth video depicts the video content more comprehensive. This confirms that video captioning is benefited by leveraging complementary attributes learnt from images and videos.

4.4. Experimental Analysis

We further verify the effectiveness of our proposed video MIL framework for attribute learning and compare the different variants of our designed transfer unit.

Evaluation of Video MIL Framework. There are generally two directions for attribute learning on videos. One is to perform image MIL model on individual video frame and the other is our proposed video MIL model to jointly utilize all the sampled frames from one video, as shown in Figure 2. Table 3 compares the sentence generation performances of the LSTM-TSA$_V$ model with semantic attributes only learnt from videos by these two different MIL models on MSVD dataset. The results across different metrics consistently indicate that LSTM-TSA$_V$ with semantic attributes learnt by video MIL model leads to a better performance, demonstrating the advantage of exploring semantic information among all the sampled frames from one video holistically, as opposed to locally based on individual frame.

Evaluation of Transfer Unit. Next, we turn to evaluate different variants of our designed transfer unit towards sentence generation. The performances on MSVD dataset of our LSTM-TSA$_{IV}$ are shown in Table 4, by combining attributes learnt from images and videos with different variants of transfer unit. LSTM-TSA$_{IV_0}$ directly calculates
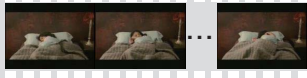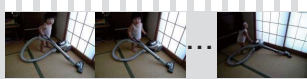
	Attributes from videos: lying: 0.578 person: 0.519 young: 0.369 girl: 0.323 three: 0.296 little: 0.276 boy: 0.254 man: 0.216 trying: 0.215 doing: 0.198	Attributes from images: bed: 0.854 laying: 0.579 man: 0.550 person: 0.290 sleeping: 0.262 white: 0.222 lying: 0.216 young: 0.177 woman: 0.168 two: 0.164	**GT:** a little girl is laying in bed **LSTM:** a man is cutting a piece of paper **LSTM-TSA**$_{IV}$**:** a woman is lying on a bed
	Attributes from videos: flying: 0.998 man: 0.998 flight: 0.941 air: 0.885 sky: 0.845 person: 0.753 takes: 0.657 someone: 0.583 jet: 0.568 something: 0.525	Attributes from images: plane: 0.562 airplane: 0.445 air: 0.271 airport: 0.268 jet: 0.262 runway: 0.230 white: 0.222 sitting: 0.199 it: 0.177 large: 0.134	**GT:** a plane is running on a run way **LSTM:** a car is landing **LSTM-TSA**$_{IV}$**:** a plane is flying
	Attributes from videos: person: 0.962 doing: 0.732 man: 0.675 room: 0.633 boy: 0.564 cleaning: 0.398 machine: 0.382 his: 0.368 someone: 0.333 riding: 0.258	Attributes from images: young: 0.420 girl: 0.319 holding: 0.308 child: 0.210 little: 0.200 floor: 0.186 pair: 0.185 it: 0.176 woman: 0.168 playing: 0.166	**GT:** a baby is cleaning **LSTM:** a boy is playing with a toy **LSTM-TSA**$_{IV}$**:** a boy is cleaning the floor
	Attributes from videos: riding: 0.710 man: 0.707 two: 0.503 each: 0.455 other: 0.453 together: 0.445 going: 0.404 bike: 0.401 talk: 0.400 motor: 0.399	Attributes from images: man: 0.543 woman: 0.409 sitting: 0.391 two: 0.342 wearing: 0.341 riding: 0.311 smiling: 0.281 young: 0.233 people: 0.210 motorcycle: 0.202	**GT:** a man and woman is riding a motorcycle **LSTM:** a woman is riding a horse **LSTM-TSA**$_{IV}$**:** a man and woman are riding a motorcycle
	Attributes from videos: animals: 0.806 ground: 0.756 something: 0.743 black: 0.636 man: 0.611 animal: 0.603 baby: 0.506 forest: 0.453 searching: 0.434 walking: 0.416	Attributes from images: bear: 0.521 forest: 0.460 walking: 0.369 woods: 0.362 some: 0.335 area: 0.242 standing: 0.220 two: 0.212 grass: 0.188 rocks: 0.186	**GT:** bear eats dirt **LSTM:** a badger is walking **LSTM-TSA**$_{IV}$**:** a bear is walking in the forest

Figure 5. Attributes and sentences generation results on MSVD dataset. The attributes from videos and images are predicted by our video MIL model and image MIL model in [6], respectively, and the output sentences are generated by 1) Ground Truth (GT): One selected ground truth sentence, 2) LSTM, and 3) our LSTM-TSA$_{IV}$.

Table 3. METEOR, CIDEr-D, and BLEU@4 scores of our proposed model LSTM-TSA$_V$ with semantic attributes only learnt from videos by two different MIL models on MSVD dataset. One is to perform image MIL model on individual video frame and the other is our proposed video MIL model as shown in Figure 2. All values are reported as percentage (%).

Model	METEOR	CIDEr-D	BLEU@4
Image MIL model	32.0	70.6	48.8
Video MIL model	**32.6**	**71.7**	**50.5**

Table 4. METEOR, CIDEr-D, and BLEU@4 scores of our proposed model LSTM-TSA$_{IV}$ with semantic attributes learnt from both images and videos on MSVD dataset. Results are shown utilizing the different input architectures of LSTM w/o transfer gate.

Model	METEOR	CIDEr-D	BLEU@4
LSTM-TSA$_{IV_0}$	32.7	71.7	50.3
LSTM-TSA$_{IV_1}$	32.9	71.5	51.2
LSTM-TSA$_{IV_2}$	33.0	72.3	50.5
LSTM-TSA$_{IV_3}$	**33.5**	**74.0**	**52.8**

an element-wise sum of the feature mappings of attributes from images and videos as a combination, which is fed into LSTM as an additional input. Thus, this additional input is shared and fixed at each time step in LSTM. In contrast, LSTM-TSA$_{IV_1}$, LSTM-TSA$_{IV_2}$ and LSTM-TSA$_{IV_3}$ fuses the two attributes with a transfer gate that dynamically computes a distinct weight based on the two attributes, the current input word and the previous hidden state in LSTM, and then computes the additional inputs to LSTM by applying the weight to attributes from images, videos and both, respectively. As such, the weight offers a more precise control of impacts from semantic attributes by integrating context information and is different at each time step. As indicated by our results, utilizing transfer gate which dynamically balances the influence between attributes learnt from images and videos can constantly lead to better performance than LSTM-TSA$_{IV_0}$. A larger performance gain is attained when applying the weight on attributes from both.

5. Discussions and Conclusions

We have presented Long Short-Term Memory with Transferred Semantic Attributes (LSTM-TSA) architecture which explores both video representations and semantic attributes for video captioning. Particularly, we study the problems of how to mine attributes from images and videos and how to fuse them in an elegant manner for enhancing sentence generation. To verify our claim, we have presented video MIL framework to holistically explore semantic information in a video and a transfer unit to contextually control the impacts of attributes learnt from images and videos. Experiments conducted on three widely adopted video captioning datasets validate our proposal and analysis. Performance improvements are clearly observed when comparing to other captioning techniques.

Our future works are as follows. First, attention mechanism will further be incorporated into our LSTM-TSA architecture for further boosting video captioning. Second, we will investigate how to leverage semantic attributes for multiple sentence or paragraph generation for videos.

References

[1] N. Ballas, L. Yao, C. Pal, and A. Courville. Delving deeper into convolutional networks for learning video representations. In *ICLR*, 2016.

[2] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL workshop*, 2005.

[3] D. L. Chen and W. B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *ACL*, 2011.

[4] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[5] J. Donahue, L. A. 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.

[6] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, et al. From captions to visual concepts and back. In *CVPR*, 2015.

[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 *CVPR*, 2016.

[8] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *ICCV*, 2013.

[9] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 1997.

[10] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[11] A. Kojima, T. Tamura, and K. Fukunaga. Natural language description of human activities from video images based on concept hierarchy of actions. *IJCV*, 2002.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[13] P. Pan, Z. Xu, Y. Yang, F. Wu, and Y. Zhuang. Hierarchical recurrent neural encoder for video representation with application to captioning. *arXiv preprint arXiv:1511.03476*, 2015.

[14] Y. Pan, Y. Li, T. Yao, T. Mei, H. Li, and Y. Rui. Learning deep intrinsic video representation by exploring temporal coherence and graph structure. In *IJCAI*, 2016.

[15] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui. Jointly modeling embedding and translation to bridge video and language. In *CVPR*, 2016.

[16] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

[17] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, 2011.

[18] A. Rohrbach, M. Rohrbach, and B. Schiele. The long-short story of movie description. In *GCPR*, 2015.

[19] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele. A dataset for movie description. In *CVPR*, 2015.

[20] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele. Translating video content to natural language descriptions. In *ICCV*, 2013.

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015.

[22] R. Shetty and J. Laaksonen. Video captioning with recurrent networks based on frame-and video-level features and visual content classification. In *ICCV workshop*, 2015.

[23] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[24] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014.

[25] 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.

[26] A. Torabi, C. Pal, H. Larochelle, and A. Courville. Using descriptive video services to create a large data source for video annotation research. *arXiv preprint arXiv:1503.01070*, 2015.

[27] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015.

[28] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015.

[29] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. Sequence to sequence - video to text. In *ICCV*, 2015.

[30] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko. Translating videos to natural language using deep recurrent neural networks. In *NAACL HLT*, 2015.

[31] Q. Wu, C. Shen, L. Liu, A. Dick, and A. v. d. Hengel. What value do explicit high level concepts have in vision to language problems? In *CVPR*, 2016.

[32] J. Xu, T. Mei, T. Yao, and Y. Rui. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, 2016.

[33] 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*, 2015.

[34] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. In *ICCV*, 2015.

[35] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei. Boosting image captioning with attributes. *arXiv preprint arXiv:1611.01646*, 2016.

[36] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *CVPR*, 2016.

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

[38] C. Zhang, J. C. Platt, and P. A. Viola. Multiple instance boosting for object detection. In *NIPS*, 2005.