Attention-in-Attention Networks for Surveillance Video Understanding in Internet of Things

Ning Xu, An-An Liu, *Member, IEEE*, Wei-Zhi Nie, and Yu-Ting Su

***Abstract*—In this paper, we propose an approach to generate the comprehensive video interpretation for the surveillance video understanding in Internet of Things. The key problem of many visual learning tasks is to adaptively select and fuse diverse and complimentary features for video representation. We design the attention-in-attention (AIA) network to hierarchically explore the attention fusion in an end-to-end manner, and demonstrate the value of this model on the multievent recognition and video captioning challenges. Particularly, it consists of multiple encoder attention modules (EAMs) and a fusion attention module (FAM). Each EAM aims to highlight the space-specific features by selecting the most salient visual features or semantic attributes and averages them into one attentive feature. The FAM can suppress or enhance the activation of multispace attentive features and adaptively co-embed them for comprehensive video representation. Then, one long short-term memory unit decodes the video representations to generate multiple event labels or video captions. This architecture is capable of: 1) adaptively learning the salient space-specific feature representation and 2) co-embedding multispace attentive features into one space for feature fusion. Experiments conducted on the surveillance video dataset (concurrent event dataset) and the popular video captioning datasets (Microsoft Research Video Description Corpus and MSR-Video to Text). It shows that the proposed AIA can achieve competitive performances against the state of the arts.**

***Index Terms*—Attention fusion, multievent recognition, multispace features, video caption.**

I. Introduction

Surveillance video understanding is an important aspect of multimedia content analysis, which consists of the multievent recognition task and the video captioning task. Particularly, the multievent recognition automatically describes multiple discrete events in the surveillance video. It is an important research topic due to the great challenges of large variances of viewpoint, scaling, lighting, cluttered background, etc. Additionally, it has many real-world applications for home security, public security, and law enforcement [1], [2]. On the other hand, the video captioning describes the surveillance video with the continuous natural language. It has brought a profound challenge to both computer vision and natural language processing communities [3]–[8] and also has a variety of practical applications. For example, generating descriptions of videos may help visually impaired people better understand the content of videos and retrieve videos using descriptive texts.

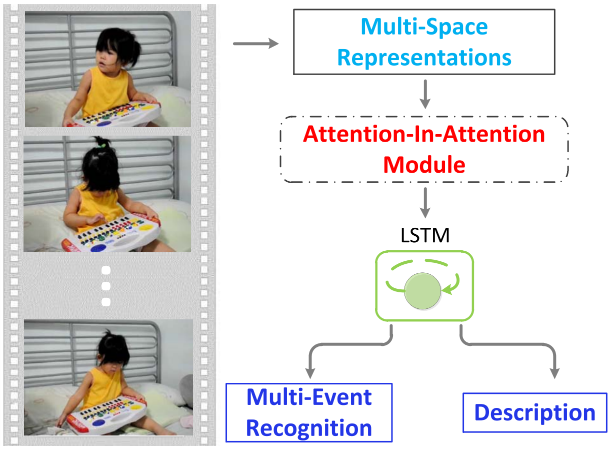

Fig. 1. High-level visualization of our approach. We extract multispace representations of videos, which are incorporated by the proposed AIA module to generate the multiple event labels or describe the main content with natural language for the input video.

Meanwhile, they play the different roles for surveillance understanding, such as: 1) multievent recognition from videos to facilitate search and retrieval of visual information from the Web and 2) video captioning task goes beyond conventional one-versus-all prediction tasks and report the natural language like humans. Both of tasks integrate large-scale visual information and generate outputs presented in the discrete or continuous form, respectively. Furthermore, the common challenge of these visual learning tasks is to select and fuse diverse and complimentary features for video representation [4], [6], [9]. In this paper, we develop a hierarchical network and demonstrate the value of this model on the multievent recognition and video captioning tasks.

Intensive research interests have been paid for these emerging topics. For the multievent recognition, it can be roughly categorized into the simple event recognition [10]–[13] and the complex event recognition [14]–[17]. In the past decades, most study was limited to the simple events recognition on sports videos [10] and news videos [13], which can be characterized by a single shot or a few frames under the constrained environments. For example, Wang *et al.* [11] proposed

Manuscript received June 27, 2017; revised August 23, 2017 and November 5, 2017; accepted November 24, 2017. Date of publication December 4, 2017; date of current version November 14, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61772359 and Grant 61472275 and in part by the Tianjin Research Program of Application Foundation and Advanced Technology under Grant 15JCYBJC16200. *(Corresponding author: An-An Liu.)*

The authors are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: liuanan@tju.edu.cn).

Digital Object Identifier 10.1109/JIOT.2017.2779865

motion relativity and visual relatedness for event recognition. Recently, some works have focused on the complex event recognition, which aims to recognize more generic and complicated events. It includes events with multiple objects (e.g., making a cake), variations in scene (e.g., kitchen, outdoor, etc.), and accompanied by specific motions and sequence within a long video footage. For example, Liu *et al.* [14] have designed a local expert forest model for score fusion from multiple classifiers in complex event recognition.

Existing approaches for video captioning have evolved through three dimensions: 1) template-based methods [18]–[20]; 2) sequence-learning methods [3], [21], [22]; and 3) attention-based methods [4], [7], [8], [23]–[25]. The template-based methods predefine 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. This kind of methods highly depend on hard-coded visual concepts and suffer from the implied limits on the variety of the output. The sequence-learning methods, in contrast, aim to leverage sequence learning models to directly translate video content into sentences, which is mainly inspired by the recent advances of machine translation by deep learning [26]. Specifically, an encoder is first implemented to map a sequence of 2-D/3-D convolutional neural networks (CNNs) features to fixed-length feature vectors in the embedding space and then a decoder is applied to generate a descriptive sentence in the target language. Venugopalan *et al.* [3] leveraged the popular sequence to sequence model to transfer the temporal visual frames to natural language description. Attention-based methods exploit the temporal structure and rich intermediate description of long videos. Specifically, soft attention mechanism [27] is employed to weight each temporal feature vector. Yao *et al.* [4] integrated the soft attention mechanism into an encoder–decoder video captioner to exploit global fine-grained structure.

However, the current methods seldomly explore the hierarchical and adaptive mechanism for attention fusion of multispace features. In this paper, we propose the attention-in-attention (AIA) network that can adaptively learn the compact and salient space-specific feature representation and further co-embed multispace attentive features into one feature space for fusion. AIA utilizes and incorporates multiple attention modules, which have been successfully applied in image captioning [28], video caption [4], machine translation [27], and pose estimation [29], for both multispace salient feature selection and feature fusion. However, only using one-layer attention module is not sufficient to generate rich fine-grained textual descriptions for videos, which is evaluated and discussed in Section IV-G. Therefore, we utilize the hierarchical attention framework to enhance the video representation for captioning. Particularly, we predefines a set of spaces for each kind of feature in this paper. The overall architecture of AIA is illustrated in Fig. 2. During the encoding stage, AIA can leverage multiple off-the-shelf feature extraction methods for space-specific representations. In this paper, we extract features from two modalities.

1) In the visual modality, 2-D CNN features are extracted for visual representation.
2) In the semantic modality, a set of pretrained attribute models are implemented to extract semantic representation for both video frames and entire videos.

With the multispace feature representations, the proposed AIA applies multiple encoder attention modules (EAMs) and a fusion attention module (FAM) for the feature selection and the co-embedding learning, respectively. Each EAM captures the explicit space-specific features by selecting the most salient visual features or semantic attributes and further project them into an attentive feature. Then, the FAM can suppress or enhance the activation of the multispace attentive features and adaptively co-embed them for comprehensive video representation. In the decoding stage, one long short-term memory (LSTM) unit is employed to decode the fusion representation and attentive features to generate multiple event labels or video captions. The hidden state of LSTM, which can record the dynamic of sentence generation and the latent correlation between visual and textual modalities, is further fed to both EAMs and FAM for sentence-induced feature selection and fusion.

The main contributions of this paper are threefold.

1) We propose an AIA network to hierarchically explores the attention fusion for multispace salient feature selection and fusion.
2) We perform comprehensive evaluations on the multi-event recognition task and the video captioning task. The proposed AIA method can achieve competitive performances against the state of the arts on one surveillance video dataset and two popular video captioning datasets.
3) We further verify the effectiveness of AIA by comparing its multiple variations.

II. Related Works

This paper can be uniquely positioned in the context of two recent research directions in multievent recognition task and video captioning task.

A. Multievent Recognition

Multievent recognition is a challenging task due to its dynamic attributes and semantic richness. It can be roughly categorized into simple event recognition and complex event recognition.

Simple events are unusual events or sports events that last for short time and have small intraclass variations [30]–[33]. Particularly, Adam *et al.* [34] presented an algorithm using multiple local monitors which collect low-level statistics to recognize certain types of unusual events in surveillance videos. Xu *et al.* [10] considered both textual information (i.e., Web-casting text) and visual information (i.e., broadcast video) to recognize events from live sports game. In [35], video sequence is treated as a space-time volume, which features based on optical flow are extracted for event recognition. In [13], a model based on a multiresolution, multisource, and multimodal bootstrapping framework has been developed for events recognition in news videos.

Unlike the simple event recognition, the complex event recognition is much more complicated, occur in much

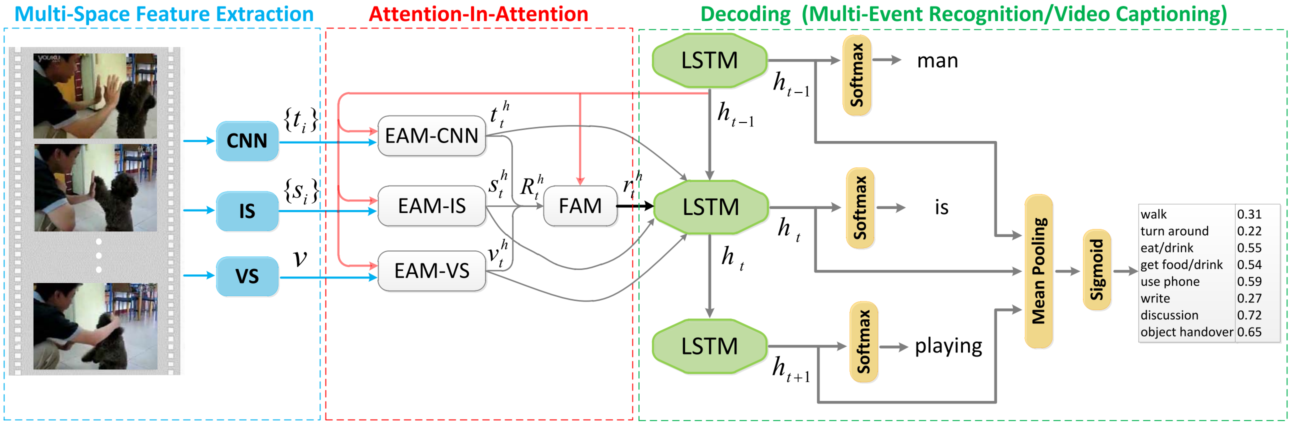

Fig. 2. Illustration of the AIA framework, which consists of three components. First, in the multispace feature extraction component, the space-specific features are obtained by multiple off-the-shelf feature extraction methods. Second, the AIA component is utilized for feature selection and fusion, which includes multiple EAMs and FAM to generate the space-specific attentive features and further project them into a space with the identical dimension. Third, in the decoding component, one LSTM unit is employed to decode the fusion representation and attentive features, where the hidden state is simultaneously fed into both EAMs and FAM for sentence-induced feature selection and fusion. Particularly, for the multievent recognition task, we first mean pool the hidden states from the LSTM unit and then assign a probability distribution by the sigmoid layer. For the video captioning task, the LSTM unit sequentially generates the description by the softmax layer. The proposed framework is learned end-to-end and we can easily add extra branches for additional features. CNN: convolutional neural network; IS: image-wise semantic; VS: video-wise semantic; EAM: encoder attention module; and FAM: fusion attention module.

longer videos and have huge intraclass variations [36]–[39]. Particularly, Yang and Shah [36] designed an approach that discovers data-driven concepts from multimodality signals, which a sparse video representation is learned for event recognition. Natarajan *et al.* [38] proposed to combine multiple features from different modalities to improve multimedia event recognition. Vahdat *et al.* [39] presented a compositional model that leveraged a novel multiple kernel learning algorithm to incorporate structured latent variables.

B. Video Captioning

The literature on visual captioning can be roughly divided into three categories, including template-based methods, sequential learning-based methods, and attention-based methods.

The template-based methods predefine 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 [18], [19]. Guadarrama *et al.* [18] designed semantic hierarchies to choose an appropriate level of the specificity and accuracy of sentence fragments. Rohrbach *et al.* [19] 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 highly rely on hard-coded visual concepts and suffer from the implied limits on the variety of the output.

The sequential learning-based methods have been widely applied to video captioning, 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 [3], [21], [22], [40]. This problem is analogous to translate 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 captioning method [22] extended the image caption methods by simply pooling the features of multiple frames to form a single representation. Venugopalan *et al.* [3] applied the sequence to sequence model to transfer the temporal visual information to natural language description and further extended it by inputting both appearance features and optical flow. However, this strategy can only work for short video clips, which only contain one major event with limited visual variation, and ignore the rich fine-grained information conveyed by the video stream.

Recently, the attention-based methods employ soft attention mechanism [27] to weight each temporal feature vector in order to exploit the temporal structure and rich intermediate description of long videos. For instance, Yao *et al.* [4] proposed to exploit temporal structure based on soft attention mechanism, which allows to go beyond local temporal modeling and learns to select the most relevant temporal segments for video captioning. Ballas *et al.* [24] leveraged convolutional gated recurrent unit-recurrent neural network (GRU-RNN) to extract visual representation and generate sentence based on the LSTM text-generator with soft-attention mechanism. Yu *et al.* [7] further exploited temporal- and spatial-attention mechanisms to selectively focus on visual elements during generation.

However, seldom work has been done to hierarchically learn and integrate the multispace representations from visual/semantic modalities in the data-driven manner. In this paper, we explore the comprehensive video representation under the multispace features condition for video understanding in Internet of Things.

III. AIA Network

In this section, we first give an overview of the proposed framework and the details of the computational pipeline in Section III-A. Then, the three key components, including multispace feature extraction, AIA module, and decoding module, will be detailed in Sections III-B–III-C, respectively.

A. Framework Overview

AIA network can adaptively and jointly perform the procedures of space-specific feature selection and multispace attentive feature fusion. Fig. 2 depicts the AIA framework for video captioning. Given several complementary information (e.g., visual features and semantic attributes), AIA first considers each information as a unique space, and uses a space-specific attention module to encode the salient knowledge with the output of an attentive feature individually. Further, a FAM is stacked to co-embed multispace features for comprehensive video representation. Then, we employed one LSTM unit to decode the fusion representation and attentive features to generate probabilities for each event or sentences word by word.

For each space in the AIA network, we leverage individual off-the-shelf feature extraction methods to obtain input representations. Particularly, for the visual space, we extract the CNN [41] feature on each frame and generate a series of *image-wise features* $\{t_i\}$. For the image semantic, we extract a set of related attributes. They explicitly represent each frame and construct a sequence of *image-wise semantic* $\{s_i\}$. Similarly, we mine global information for videos and generate *video-wise semantic* v. Additionally, the attention module can convert an entire input sequence $(v_1, \ldots, v_n)$ to an attentive vector, which adaptively selects and encodes salient space-specific knowledge. With the multispace feature representations, AIA employs multiple EAMs to capture the explicit space-specific features. We then obtain the attentive inputs to the FAM, which co-embeds them for comprehensive video representation. One LSTM unit is used to decode the fusion representation and attentive features and simultaneously facilitate the multispace fusion via hidden states.

In order to leverage multispace features, we utilize four attention modules with one LSTM unit. This framework is highly extensible since we can easily add extra branches for additional features.

B. Multispace Feature Extraction

In this section, we individually describe the procedures of feature extraction from multiple sources.

1) Visual Feature: We extract one feature vector per frame, leading to a series of *image-wise features*. Additionally, we could also extract other forms of visual features, such as features from an area of a frame, several consecutive frames, etc. In this paper, we only consider *image-wise features*, denoted by $\{t_i\}$, which are commonly used for visual representations.

Particularly, 2-D GoogLeNet [41] CNN is used to extract fixed-length representation (with the help of the popular implementation in Caffe [42]). Features are extracted from the pool5/7x7_s1 layer. We select 28 equally spaced frames out of each video and feed them into the CNN to obtain 1024-D image-wise feature vectors.

2) Semantic Feature: Using attribute-based models as a high-level representation has shown potential in many computer vision tasks such as object recognition, image annotation and image retrieval [43], [44]. The semantic properties captured in images often depict static objects and scenes (e.g., "boy," "dog," and "floor") while the semantic cues extracted from videos convey the temporal dynamics (e.g., "playing," "cleaning," and "riding"). This has made the attributes mined from images and videos complementary to generate the sentence for the video (e.g., "a boy is playing with a dog"). In this paper, we leverage and integrate the attributes from two sources to enhance video captioning.

Particularly, we describe the visual content in terms of a set of attributes. We first construct image semantic vocabulary $V_{\text{att}}^{\text{img}}$ by extracting attributes from MSCOCO image captions. We select the c $(c = 256)$ most frequent words as the high-level semantic attributes, which include object names (nouns), motions (verbs) or properties (adjectives). Given the $V_{\text{att}}^{\text{img}}$, we associated each MSCOCO image with a set of attributes according to its captions. We followed the split strategy [28] of MSCOCO for semantic modeling. The semantic attributes were trained with SVM classifiers [45] using GoogleNet features. Finally, the SVM predictions were aggregated as a 256-way vector and used as *image-wise semantic* representation denoted by $\{s_i\}$.

For video semantics, we implemented the same strategy and selected 256 common words to construct video attribute vocabulary $V_{\text{att}}^{\text{vid}}$ on each video captioning benchmark individually. We utilized the SVM classifiers [45] for semantic modeling. Different from the image-wise domain, we perform mean pooling on the GoogleNet features to generate a single *video-wise semantic* representation v for each video.

3) Attention-in-Attention Module: Given the image-wise features $\{t_i\}$, image-wise semantic $\{s_i\}$, and video-wise semantic v, AIA can adaptively and jointly perform the procedures of space-specific feature selection and multispace attentive feature fusion with the hierarchical attention modules.

In many cases, a description only related to a small region of a video. For example, in Fig. 1, although there are multiple objects in the video: girl, piano, bed, pillow, and wall but the dominated objects for the descriptions only relate to girl and piano. Therefore, only using one-layer attention module to generate the video description could lead to suboptimal results due to the noises introduced from regions that are irrelevant to the video content. Instead, via the FAM progressively, AIA is able to gradually filter out noises and pinpoint the dominated visual or semantic information that are highly relevant to video descriptions.

We adopt the popular soft attention module [27], which allows to weight each input feature vector $V = \{v_1, \ldots, v_n\}$. This approach has been used successfully to exploit spatial or temporal structure [4], [28]. Here, we adopt it to exploit more informative representations of videos.

Due to the variability of the length of videos, it is challenging to directly input all these vectors to the model at every time step. A simple strategy is to compute the average of features V, $y_t = (1/n) \sum_{i=1}^{n} v_i$, and input this average vector to each time step t of the model. However, this strategy collapses all available information into a single vector, neglecting the inherent temporal structure. Instead, the dynamic weighted sum of the feature vectors is taken such that

$$y_t = \sum_{i=1}^{n} \alpha_i^{(t)} v_i \quad (1)$$

where $\sum_{i=1}^{n} \alpha_i^{(t)} = 1$ and $\alpha_i^{(t)}$'s are computed at each time step t inside the LSTM unit (Section III-C). $\alpha_i^{(t)}$ is regarded as the attention weights at time t.

The attention weight $\alpha_i^{(t)}$ reflects the relevance of the ith feature vector in the input video given all the previously generated words, i.e., $w_1, \ldots, w_{t-1}$. Hence, we takes the previous hidden state h_{t-1} as input for both EAMs and FAM, which summarizes all the previously generated words. Specifically, we compute basic attention scores $e_i^{(t)}$ conditioning on the input vector h_{t-1} at each time step t

$$e_i^{(t)} = w^T \tanh(W_a h_{t-1} + U_a v_i + b_a) \tag{2}$$

where w, W_a, U_a, and b_a are the parameters that are estimated together with all the other parameters of the encoder and decoder networks.

Then we feed the relevance scores $e_i^{(t)}$ through a sequential softmax layer to obtain a set of attention weights $\{\alpha_i^{(t)}\}$

$$\alpha_i^{(t)} = \frac{\exp\{e_i^{(t)}\}}{\sum_{j=1}^{n} \exp\left\{e_j^{(t)}\right\}}. \tag{3}$$

For convenience, the output of the attention module at a given time step t is regarded as the attentive feature and denoted by $y_t^h = \text{Attention}(h_{t-1}, \{v_i\})$.

In AIA, we applied multiple attention modules to encode multispace features individually, which is named as EAM. Each EAM aims to highlight the space-specific features by selecting the most salient visual features or semantic attributes. Then, we obtain the corresponding attentive features from image-wise features $\{t_i\}$, image-wise semantic $\{s_i\}$ and video-wise semantic v as follows:

$$\begin{aligned} t_t^h &= \text{EAM} - \text{CNN}\ (h_{t-1}, \{t_i\}) \\ s_t^h &= \text{EAM} - \text{IS}\ (h_{t-1}, \{s_i\}) \\ v_t^h &= \text{EAM} - \text{VS}\ (h_{t-1}, \{v\}). \end{aligned} \tag{4}$$

Particularly, the attentive features are the weighted averages, which selectively enhance the features that are relevant to the individual space. However, for complicated descriptions, only one-layer EAM is not sufficient to associate dominating elements for video descriptions. Therefore, a multispace FAM is stacked to co-embed attentive features for comprehensive video representation and can be formulated as

$$\begin{aligned} R_t^h &= \left[t_t^h, s_t^h, v_t^h\right] \\ r_t^h &= \text{FAM}\ \left(h_{t-1}, R_t^h\right) \end{aligned} \tag{5}$$

where we aggregated the set of attentive features as R_t^h, which is fed into the FAM to form the more fine-grained fused representation r_t^h. FAM focuses more on the visual or semantic elements that corresponds to the descriptions and further project them into a space with the identical dimension.

C. Decoding Module

In this paper, the decoding component is shared by both multievent recognition task and video captioning task. It is natural to use a RNN as a decoder when the output is sequential words [3], [7], [28]. Among the recent successful applications of RNN, it is noticeable that most of them used LSTM units [46] or their variation, GRUs [47]. In this paper, we leverage the variation of the LSTM units, introduced in [4] and [28], as the decoder. As shown in Fig. 2, we jointly integrate three attentive features (i.e., t_t^h, s_t^h, and v_t^h) and one fusion representation (r_t^h) into the LSTM unit. The multispace coefficients of forget, input, output gates, updated memory content, memory content, and hidden state, respectively, can be formulated by

$$\begin{aligned} f_t &= \sigma\Big(W_f E[w_{t-1}] + U_f h_{t-1} + A_f^t t_t^h + A_f^s s_t^h + A_f^v v_t^h \\ &\qquad + A_f^r r_t^h + b_f\Big) \\ i_t &= \sigma\Big(W_i E[w_{t-1}] + U_i h_{t-1} + A_i^t t_t^h + A_i^s s_t^h + A_i^v v_t^h \\ &\qquad + A_i^r r_t^h + b_i\Big) \\ o_t &= \sigma\Big(W_o E[w_{t-1}] + U_o h_{t-1} + A_o^t t_t^h + A_o^s s_t^h + A_o^v v_t^h \\ &\qquad + A_o^r r_t^h + b_o\Big) \\ \tilde{c}_t &= \tanh\Big(W_c E[w_{t-1}] + U_c h_{t-1} + A_c^t t_t^h + A_c^s s_t^h \\ &\qquad + A_c^v v_t^h + A_c^r r_t^h + b_c\Big) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\ h_t &= o_t \odot c_t \end{aligned} \tag{6}$$

where σ is the element-wise logistic sigmoid function and $\odot$ is an element-wise multiplication. W, U, A, and b are, in order, the weight matrices for the input, the previous hidden state, the weighted average, and the bias. E is a word embedding matrix, and we denote by $E[w_{t-1}]$ an embedding vector of word w_{t-1}.

As shown in Fig. 2, the previous hidden state h_{t-1} of the LSTM is taken as input for the AIA components. For each EAM, h_{t-1} aims to capture the explicit space-specific features by selecting the most salient visual or semantic knowledge. For the FAM, h_{t-1} has influence on the selection mechanism, which attends to certain whether visual or semantic attentive features should be strengthened and further co-embeds them into a space with the identical dimension. Meanwhile, the embedding word is sequentially fed into the LSTM unit, which learns sentence-induced fusion dynamics and generates the descriptions for videos.

For the multievent recognition, we denote a set of event labels as target sequences (usually with different length). During the procedure of training, the AIA method hierarchically explores the attention fusion by minimizing the cross entropy loss in the softmax layer. At the test stage, we remove the softmax layer and take mean pooling over the sequence of probabilities, which generated by the LSTM unit. We assign a distribution by the sigmoid function and each value in the output vector indicates the probability that an event happens in the video. We experimentally set threshold for inference. For the video captioning, we replace event labels as ground truth sentences. Meanwhile, a softmax function is applied to get the probability distribution over the output words in the vocabulary.

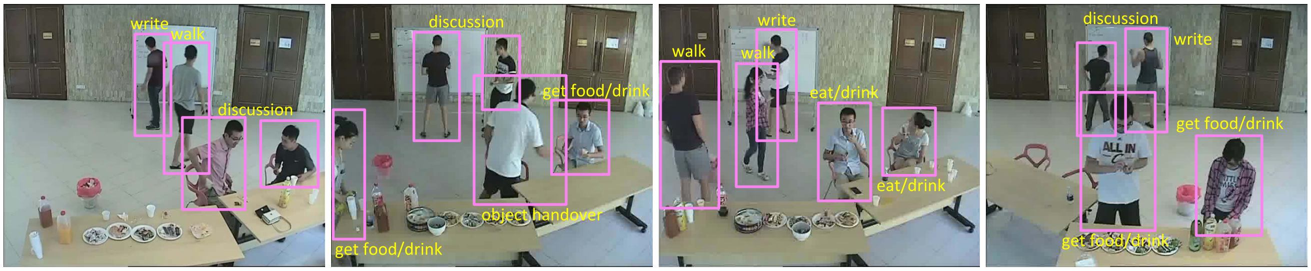

Fig. 3. Exemplary frames from CED. Each sample is overlaid with bounding boxes and corresponding labels.

IV. Experiments

A. Datasets

For the multievent recognition, we evaluate the AIA method on the concurrent event dataset (CED), which is collected by the SeSaMe Research Group, National University of Singapore. The dataset is recorded under the real-world surveillance environment. In total, it recruited nine individuals for seven recording sessions, where each session consists of four to five persons. The duration of each session vary from 16 min to 42 min. During the recording sessions, the individuals are asked to participate in an unconstrained social event with eight unique prescripted events. The recorded sessions are manually labeled based on these events, namely walk (2305), turn around (1992), eat/drink (2527), get food/drink (896), use phone (2921), write (1211), discussion (4756), and object handover (278). The number in the bracket indicates the number of events in CED and exemplary frames are shown in Fig. 3.

For the video captioning, since there no exists natural language description dataset for surveillance settings, we evaluate the AIA network on two popular video captioning datasets.

Microsoft Research Video Description Corpus (MSVD) [48] 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 multilingual descriptions. In this paper, we use only the English descriptions. Following [4], we select 1200 videos as training set, 100 videos as validation set, and 670 videos as test set.

MSR-Video to Text (MSR-VTT) [49] 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 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 paper, we strictly followed the dataset split provided by Xu *et al.* [49].

B. Evaluation Methods

In this paper, we adopt the precision and recall metrics to evaluate the generated multiple event labels. Meanwhile, we adopt the BLEU@N(B@N) [50], METEOR(M) [51], ROUGE-L(R) [52], and CIDEr(C) [53] metrics against all ground truth to evaluate the generated 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@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. ROUGE is primarily recall-based and thus has a tendency to reward long sentences with high recall. CIDEr, a consensus-based evaluation protocol for image descriptions, rewards a sentence for being similar to the majority of human written descriptions. We employ the evaluation source code[1] released by Microsoft COCO Evaluation in this paper.

C. Experimental Setup

For the multievent recognition, we use fourfold cross validation on CED, which are randomly divided into training and test sets with a mutual ratio of 3:1.

For the video captioning, the descriptions of each dataset were preprocessed to lower case, tokenized sentences and removed punctuation. After preprocessing, the numbers of unique words were 13 008 for MSVD and 9730 for MSR-VTT. We represented each word in the sentence as "one-hot" vector (binary index vector in a vocabulary). The word embedding dimensionality is set to 468 and the dimension of hidden layers in LSTM is 3518. We optimized the hyperparameters using random search to maximize the log-probability of the validation set. ADADELTA [54] is employed with the gradient computed by the backpropagation algorithm in terms of learning rate 10^{-4}. To avoid over-fitting, drop-out was utilized at the inputs and outputs of all layers. In testing stage, the beam search strategy [26] is utilized to find the sentence with the highest probability. Our system is implemented using the Theano [55] framework.

From Sections IV-D and IV-E, we intensively discuss the effectiveness of AIA on the multievent recognition task.

D. Results and Analysis

To evaluate the effectiveness of AIA for the multievent recognition, we comparatively train binary SVM classifiers for each event over four kinds of kernels, including the linear,

[1] [Online]. Available: https://github.com/tylin/coco-caption

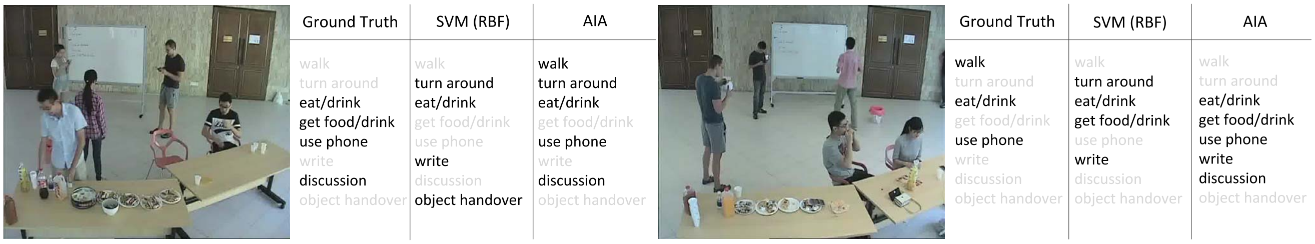

Fig. 4. Samples of predicted results on the CED testing set with SVM and AIA methods. For ease of comparison between different methods, we mark the recognized events by the black color in the label list.

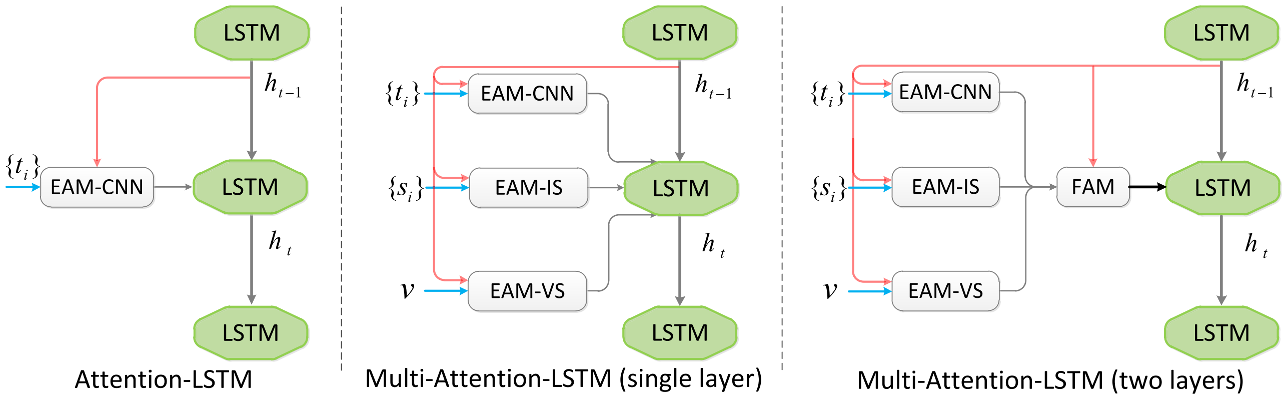

Fig. 5. Frameworks of three variations of the proposed AIA. We explore the effectiveness of multispace features in video captioning, and evaluate the hierarchical soft-attention fusion manner of attentive features.

TABLE I
COMPARISON OF THE AIA METHOD AND SVM CLASSIFERS ON CED (LR: LEARNING RATE)

lr	AIA		Kernels	SVM	
	Precision	**Recall**		**Precision**	**Recall**
1×10^{-3}	66.7%	88.1%	Linear	61.4%	67.9%
5×10^{-4}	68.2%	91.8%	Polynomial	63.6%	62.7%
1×10^{-4}	68.4%	95.2%	RBF	**64.4%**	**68.9%**
5×10^{-5}	**70.0%**	**96.5%**	Sigmoid	61.9%	67.6%

polynomial, RBF, and sigmoid kernels. From Table I, the performance of AIA model tend to increase, as the learning rate is decreasing with the more precise optimization. Meanwhile, the RBF kernel can achieve the best performance than the other SVM kernels in terms of precision and recall metrics, which illustrates the mapped high-dimensional space in the RBF kernel is benefited to the recognition in the surveillance videos. Furthermore, our method is significantly better than SVM classifiers across all kernels. Particularly, AIA can significantly improve the performance in the recall metric while achieving the competitive performance in the precision metric. The best score in precision and recall metrics of AIA is 13.1% and 40.1% better than the best one of SVM, respectively, which demonstrates the effectiveness of the hierarchical attention fusion in the surveillance videos. Fig. 4 shows the predicted results for different events on CED testing set using SVM(RBF) and AIA methods. It is noted that some categories is not recognized by SVM(RBF) but captured by AIA (e.g., "discussion" in the left example; "use phone" in the right example).

TABLE II
COMPARISON OF THREE AIA VARIATIONS ON CED

Model	Precision	Recall
Attention-LSTM	65.7%	85.2%
Multi-Attention-LSTM (single layer)	66.1%	90.7%
Multi-Attention-LSTM (two layers)	68.3%	93.6%
AIA	**70.0%**	**96.5%**

E. Comparison by Architecture Variation

In this section, we comprehensively analyze three variations of the proposed AIA network on CED, as shown in Fig. 5. First, we compare AIA with the method using the single modality and the single attention module, which are denoted as attention-LSTM. Table II shows that AIA is significantly better than attention-LSTM in terms of precision and recall metrics. It demonstrated the advantage of feature selection and fusion in the AIA network for multievent recognition in the surveillance videos.

To explore the influence of FAM, we train two variational networks of AIA, namely multiattention-LSTM (single layer) and multiattention-LSTM (two layers) as shown in Fig. 5. Particularly, we only use multiple EAMs by removing the stacked FAM with other components of AIA unchanged, which is named as multiattention-LSTM (single layer). Meanwhile, we use the LSTM unit to only decode the fusion representation from FAM without the attentive features, which is named as multiattention-LSTM (two layers). From the Table II, we can observe that "two layers" improved performances than "single layer," which illustrates that FAM is benefit to the multievent recognition due to the fusion of visual- and semantic-space attentive features from EAMs. Furthermore, both *single layer* and *two layers* are significantly better than attention-LSTM in terms of precision and recall metrics, which shows that multievent recognition can benefit from the fusion of multispace features.

In order to enhance the robustness of the model, we leverage the attentive features and the fusion representation into the LSTM unit to form the AIA network. From Table II, the performance of AIA is better than the multiattention-LSTM (two layers).

From Sections IV-F–IV-H, we intensively explore the effectiveness of AIA on the video captioning task.

F. Results and Analysis

1) State-of-the-Art Methods: On MSVD, we compare to nine methods as follows.

FGM [56] uses a two step approach to first obtain confidences on subject, verb, object, and scene and then generates a sentence based on a template.

Mean Pool [21] pools image-wise features to create a fixed-length video representation, which is decoded by an LSTM.

S2VT [3] incorporates both RGB and optical flow inputs, and the encoding and decoding of the inputs and word representations are learned jointly in a parallel manner.

Glove-Deep [57] integrates both a neural language model and distributional semantics trained on large text corpora into the LSTM-based architecture for video captioning.

LSTM-E [6] explores the visual-semantic embedding based on the LSTM for video captioning.

TA [4] exploited a weighted attention mechanism to dynamically attend to specific temporal regions of the video while generating sentence.

GRU-RCN [24] leverages convolutional GRU-RNN to extract visual representation and generate sentence based on the LSTM with soft-attention mechanism.

HRNE [25] encodes the frame sequence with hierarchical RNN and decodes the sentence with attention mechanism.

Boundary-aware encoder [58] proposes a LSTM-based cell which can identify discontinuity points between frames or segments and modify the temporal connections of the encoding layer accordingly.

TDDF [8] integrates complementary features from multiple channels to linearly fuse heterogenous dynamics according to model status.

h-RNN [7] exploits both spatial and temporal attention mechanisms for paragraph generation.

On MSR-VTT, there are relatively less work and we compare to seven methods.

v2t nagvigator [59], Aalto [61], and VideoLAB are the top three results in the 2016 MSR-VTT video to language grand challenge.[2]

TA [49] and TDDF and TDDF17 are the same methods as above which are evaluated in MSR-VTT.

MMVD [61] proposed a video captioner based on the S2VT [3] model and extended it to fuse multimodal information.

TABLE III
PERFORMANCE COMPARISON ON MSVD

Model	**B@4**	**M**	**R**	**C**
FGM [56]	13.7	23.9	-	-
Mean Pool [21]	33.3	29.1	-	-
S2VT [3]	-	29.8	-	-
Glove-Deep [57]	42.1	31.4	-	-
LSTM-E [6]	45.3	31.0	-	-
TA [4]	41.9	29.6	-	51.7
GRU-RCN [24]	43.3	31.6	-	68.0
HRNE [25]	43.8	33.1	-	-
Boundary-Aware Encoder [58]	42.5	32.4	-	-
TDDF [8]	45.8	**33.3**	**69.7**	**73.0**
h-RNN [7]	**49.9**	32.6	-	65.8
AIA	49.5	32.7	67.6	67.0

Multimodal fusion [62] designed a multimodal fusion encoder and use the RNN decoder to integrate information such as image, motion, aural, speech, and so on.

2) Results on MSVD: We report the results on MSVD in Table III. AIA achieves the competitive scores in terms of BLEU@4 and METEOR among all the methods. Particularly, the BLEU@4 of AIA can achieve 49.5% on MSVD dataset, making the relative improvement over the state-of-the-art techniques including nonattention approaches (FGM, mean pool, glove-deep, and LSTM-E) and attention-based approaches (TA, GRU-RCN, HRNE, TDDF, and boundary-aware encoder). The results indicate the advantage of feature selection and co-embedding fusion in the AIA framework, which generates better video representation suitable for different sentences that share a similar context.

TA [4] applied the attention module on the concatenated feature channels, which uses single modality with single attention module for temporal structure fusion. However, our method hierarchically exploits multiple modalities into multiple attention modules for attention fusion. Therefore, AIA outperforms it by 18.1% in terms of BLEU@4, which shows that the fusion strategy in the AIA manner is better. Compared to GRU-RCN [24] and HRNE [25], our method achieves better results in BLEU@4, which confirms the effectiveness of AIA. Boundary-aware encoder [58] encodes the content and temporal structure of input clips by a boundary detection module while AIA further improve the performance by 16.5% in terms of BLEU@4. It shows the robustness of the fusion representation learned from the hierarchical attention framework. Meanwhile, AIA is over 8.1% better than TDDF in terms of BLEU@4, which uses the pure linear dynamic fusion method to boost the video captioning. It further demonstrates the attention-based feature selection and fusion is reasonable. However, the performance of AIA is a little worse than TDDF under some metrics, since TDDF simultaneously integrates VGG and 3-D CNN features to train captioners. Additionally, we notice that h-RNN [7] outperforms the others in terms of BLEU@4. h-RNN proposed a better language model which can generate multiple descriptions for a video. Comparatively, this paper focuses on fusing a comprehensive video representation to improve the encoder part but not the language decoder part.

It is reasonable to expect the performance of the proposed AIA can be further improved by utilizing more complimentary

[2] [Online]. Available: http://ms-multimedia-challenge.com/leaderboard

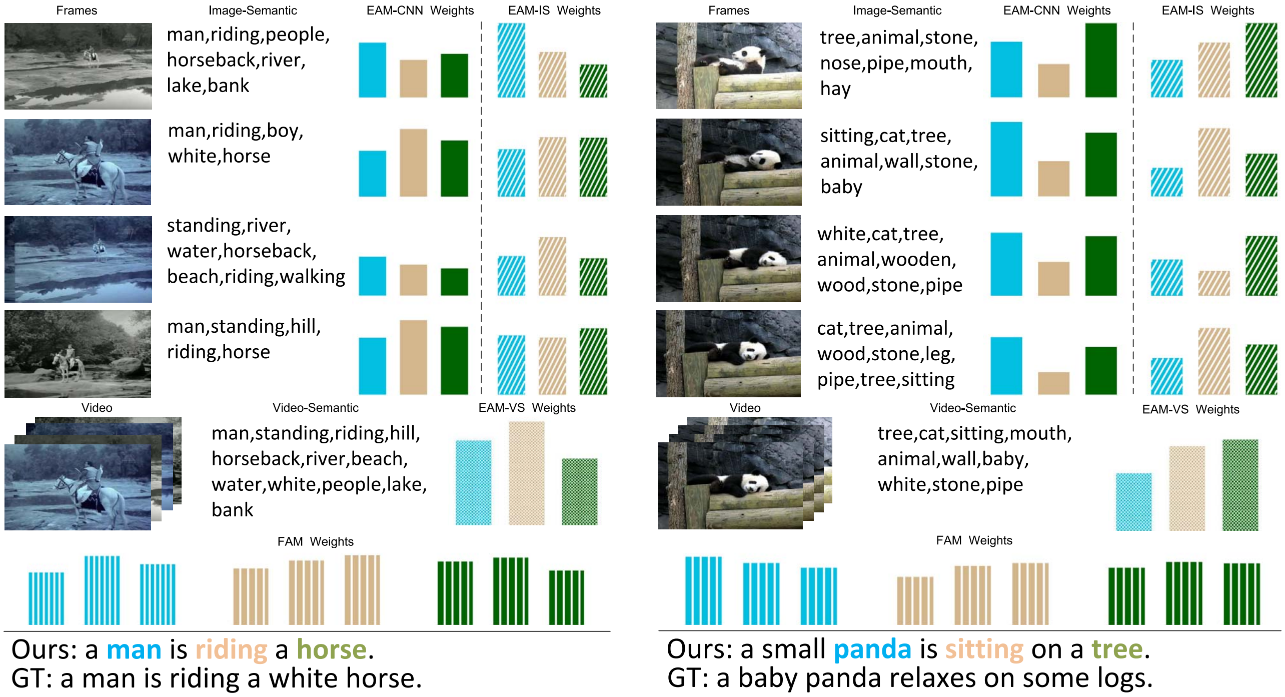

Fig. 6. Qualitative cases from the test part of MSVD. First, the image-semantic and video-semantic attributes are presented on the right of each frame or video, respectively. Then, we visualize the corresponding weights of EAM-CNN and EAM-IS, which are shown by the bar plots of two columns. The weights of EAM-VS are shown on the right of video-semantic attributes. The weights of FAM are visualized in the bottom of each panel, according to the order of attentive image, attentive image-semantic and attentive video-semantic spaces for each word. Particularly, we use the same color to associate the various weights with the corresponding words. The generated sentences of AIA are presented and GT represents randomly selected ground truth sentences.

TABLE IV
PERFORMANCE COMPARISON ON MSR-VTT

Model	B@4	M	R	C
Rank:1, v2t navigator [59]	40.8	28.2	60.9	44.8
Rank:2, Aalto [60]	39.8	26.9	59.8	45.7
Rank:3, VideoLAB	39.1	27.7	60.6	44.1
TA [49]	40.5	**29.9**	-	-
TDDF [8]	37.3	27.8	59.2	43.8
MMVD [61]	40.7	28.6	61.0	**46.5**
Multi-modal Fusion [62]	**43.7**	29.0	**61.4**	45.7
AIA	41.2	29.1	60.3	45.8

TABLE V
COMPARISON OF THREE AIA VARIATIONS ON MSVD

Model	B@4	M	R	C
Attention-LSTM [4]	41.9	29.6	-	51.7
Multi-Attention-LSTM (single layer)	46.1	32.1	64.5	66.6
Multi-Attention-LSTM (two layers)	48.1	32.7	63.2	66.9
AIA	**49.5**	**32.7**	**67.6**	**67.0**

information, such as more appearance, motion, and attribute features.

3) Results on MSR-VTT: We report the results on MSR-VTT in Table IV. It can be observed that AIA consistently outperforms the top three teams of the MSR-VTT challenge in terms of BLEU@4 and METEOR. However, we find that the METEOR of our approach is slightly worse than only using visual attention [49], which benefits from integrating frame representation from GoogLeNet and video clip representation based on a 3-D CNN trained on hand-crafted descriptors. It illustrates that higher-level visual representations are helpful in improving the quality of models. Furthermore, AIA outperforms the TDDF by 10.5% relatively in terms of BLEU@4 and 4.7% in terms of METEOR. It demonstrates the advantages of attention fusion manner in AIA than the pure linear fusion manner in TDDF. As to MMVD [61] and multimodal fusion [62], both of them has a higher performance than our framework in terms of BLEU@4, ROUGE, and CIDEr. They utilize extra data (e.g., aural data) to preprocess the videos and encode multimodal features to build the more robust video representation. However, both of them are absent of the attention module for multimodal fusion while AIA only use less amounts of data but achieve the competitive performance in some metrics such as METEOR, ROUGE, and CIDEr. It demonstrates the advantages of hierarchical framework in AIA.

G. Comparison by Architecture Variation

Similar to Section IV-E, we analyze three variations of the AIA network on MSVD, as shown in Fig. 5, and obtained the consistent results on CED. Particularly, attention-LSTM model has been exploited by Yao *et al.* [4] for video captioning. Table V shows that AIA network performs better than attention-LSTM, which shows the advantage of the hierarchical feature selection and fusion. Meanwhile, multiattention-LSTM (two layers) improved the results than multiattention-LSTM (single layer), because FAM in the second layer can adaptively co-embeds attentive features for comprehensive video representation. It shows that using *two layers* attention module is more robust than *single layer* for video captioning. Additionally, *single layer*, *two layers*, and AIA networks are significantly better than attention-LSTM in terms of all metrics, which illustrates that the fusion of multispace features can boost the video captioning.

H. Qualitative Analysis

Fig. 6 shows a few video clips and their generated and ground-truth descriptions from the MSVD dataset. We can clearly see that the generated sentences can well describe the video clips. Meanwhile, we, respectively, visualize various weights of the AIA network, including EAM-CNN, EAM-IS, EAM-VS, and FAM.

We first explained the left case. For the EAM-CNN and EAM-IS weights, when the model is about to generate the word "horse," it focuses mostly on the second and fourth frames where the appearance of *horse* is more discriminative than others and the attribute *horse* is detected in the corresponding image-semantic features, respectively. For the EAM-VS weights, the "man" and "riding" are higher than *horse* because the video-semantic feature fails to capture the attribute *horse*. Then, the attentive features from three EAMs are leveraged by FAM. For the FAM weights, the *man* and *riding* are learned with the lower image weights than image/video-semantic weights. It mainly because the attributes *man* and *riding* are always appear in each image/video-semantic feature while the appearance of *man* and *riding* is nin-significant in the second and fourth frames. Meanwhile, for the word *horse*, the video-semantic weight is lower than others due to the absence of the attribute *horse* in video semantics.

For the right case, the same conclusions can be observed. Particularly, for the EAM-CNN weights, it can be observed that the "panda" and "tree" are always higher than "sitting" due to the more discriminative appearance. For the EAM-IS weights, the *panda* always keep a low weight for each frame due to the absence of the attribute *panda* in image semantics. Meanwhile, we find that the *sitting* will increase when the corresponding image semantic successes to capture the attribute *sitting*. The same reason is also explained for the *sitting* in the EAM-VS weights. For the FAM weights, when the model is about to generate the word *panda*, the image weight is higher than the other two semantic weights since the image/video semantic detectors fail to capture the attribute *panda*.

V. Conclusion

In this paper, we present the AIA network that can adaptively learn the compact and salient space-specific attentive features and further co-embed them into one feature space for surveillance video understanding. Particularly, AIA applies multiple EAMs and an FAM. Each EAM aims to highlight the explicit space-specific features by selecting the most salient visual features or semantic attributes. The FAM aims to suppress or enhance the activation of the attentive features and project them into a space for comprehensive video representation. Finally, one LSTM unit is employed to decode the video representations and simultaneously induces the procedures of AIA fusion. The proposed AIA achieves competitive performance on one surveillance video dataset and two widely evaluated video captioning datasets. Meanwhile, we explore the effectiveness of the AIA network by comparing three variations of architectures. In the future, we will explore how to use multiple stacked attention modules for multispace feature fusion, which will generate more robust video representations to boost the multievent recognition and the video captioning.

References

[1] Y. Liu, L. Nie, L. Han, L. Zhang, and D. S. Rosenblum, "Action2activity: Recognizing complex activities from sensor data," in *Proc. IJCAI*, Buenos Aires, Argentina, 2015, pp. 1617–1623.

[2] A. Farhadi and M. K. Tabrizi, "Learning to recognize activities from the wrong view point," in *Proc. ECCV*, Marseille, France, 2008, pp. 154–166.

[3] S. Venugopalan *et al.*, "Sequence to sequence—Video to text," in *Proc. ICCV*, 2015, pp. 4534–4542.

[4] L. Yao *et al.*, "Describing videos by exploiting temporal structure," in *Proc. ICCV*, Santiago, Chile, 2015, pp. 4507–4515.

[5] H. Zhang, X. Shang, H. Luan, M. Wang, and T. Chua, "Learning from collective intelligence: Feature learning using social images and tags," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 13, no. 1, 2016, Art. no. 1.

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

[7] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu, "Video paragraph captioning using hierarchical recurrent neural networks," in *Proc. CVPR*, Las Vegas, NV, USA, 2016, pp. 4584–4593.

[8] X. Zhang *et al.*, "Task-driven dynamic fusion: Reducing ambiguity in video description," in *Proc. CVPR*, Honolulu, HI, USA, 2017, pp. 6250–6258.

[9] L. Liu, L. Shao, X. Li, and K. Lu, "Learning spatio-temporal representations for action recognition: A genetic programming approach," *IEEE Trans. Cybern.*, vol. 46, no. 1, pp. 158–170, Jan. 2016.

[10] C. Xu, J. Wang, K. Wan, Y. Li, and L. Duan, "Live sports event detection based on broadcast video and Web-casting text," in *Proc. ACM MM*, Santa Barbara, CA, USA, 2006, pp. 221–230.

[11] F. Wang, Y.-G. Jiang, and C.-W. Ngo, "Video event detection using motion relativity and visual relatedness," in *Proc. ACM MM*, Vancouver, BC, Canada, 2008, pp. 239–248.

[12] L. Nie, S. Yan, M. Wang, R. Hong, and T.-S. Chua, "Harvesting visual concepts for image search with complex queries," in *Proc. ACM MM*, Nara, Japan, 2012, pp. 59–68.

[13] G. Wang, T.-S. Chua, and M. Zhao, "Exploring knowledge of subdomain in a multi-resolution bootstrapping framework for concept detection in news video," in *Proc. ACM MM*, Vancouver, BC, Canada, 2008, pp. 249–258.

[14] J. Liu, S. McCloskey, and Y. Liu, "Local expert forest of score fusion for video event classification," in *Proc. ECCV*, Florence, Italy, 2012, pp. 397–410.

[15] S. Oh *et al.*, "Multimedia event detection with multimodal feature fusion and temporal concept localization," *Mach. Vis. Appl.*, vol. 25, no. 1, pp. 49–69, 2014.

[16] H. Ma and W. Liu, "Progressive search paradigm for Internet of Things," *IEEE MultiMedia*, to be published.

[17] Y. Yan *et al.*, "Event oriented dictionary learning for complex event detection," *IEEE Trans. Image Process.*, vol. 24, no. 6, pp. 1867–1878, Jun. 2015.

[18] S. Guadarrama *et al.*, "Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition," in *Proc. ICCV*, Sydney, NSW, Australia, 2013, pp. 2712–2719.

[19] M. Rohrbach *et al.*, "Translating video content to natural language descriptions," in *Proc. ICCV*, Sydney, NSW, Australia, 2013, pp. 433–440.

[20] 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 *Proc. AAAI*, 2015, pp. 2346–2352.

[21] S. Venugopalan *et al.*, "Translating videos to natural language using deep recurrent neural networks," in *Proc. NAACL HLT*, 2015, pp. 1494–1504.

[22] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. CVPR*, Boston, MA, USA, 2015, pp. 2625–2634.

[23] K. Cho, A. C. Courville, and Y. Bengio, "Describing multimedia content using attention-based encoder–decoder networks," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1875–1886, Nov. 2015.

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

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

[26] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. NIPS*, 2014, pp. 3104–3112.

[27] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. ICLR*, 2015.

[28] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. ICML*, Lille, France, 2015, pp. 2048–2057.

[29] X. Chu *et al.*, "Multi-context attention for human pose estimation," in *Proc. CVPR*, 2017, pp. 1831–1840.

[30] Z. Gao, H. Zhang, G. P. Xu, Y. B. Xue, and A. G. Hauptmann, "Multi-view discriminative and structured dictionary learning with group sparsity for human action recognition," *Signal Process.*, vol. 112, pp. 83–97, Jul. 2015.

[31] A.-A. Liu, Y.-T. Su, W.-Z. Nie, and M. S. Kankanhalli, "Hierarchical clustering multi-task learning for joint human action grouping and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 102–114, Jan. 2017.

[32] W. Liu, X. Liu, H. Ma, and P. Cheng, "Beyond human-level license plate super-resolution with progressive vehicle search and domain priori GAN," in *Proc. ACM MM*, Mountain View, CA, USA, 2017, pp. 1618–1626.

[33] Z. Liu *et al.*, "Fusion of magnetic and visual sensors for indoor localization: Infrastructure-free and more effective," *IEEE Trans. Multimedia*, vol. 19, no. 4, pp. 874–888, Apr. 2017.

[34] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 555–560, Mar. 2008.

[35] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient visual event detection using volumetric features," in *Proc. ICCV*, vol. 1. Beijing, China, 2005, pp. 166–173.

[36] Y. Yang and M. Shah, "Complex events detection using data-driven concepts," in *Proc. ECCV*, Florence, Italy, 2012, pp. 722–735.

[37] W. Liu, T. Mei, Y. Zhang, C. Che, and J. Luo, "Multi-task deep visual-semantic embedding for video thumbnail selection," in *Proc. CVPR*, Boston, MA, USA, 2015, pp. 3707–3715.

[38] P. Natarajan *et al.*, "Multimodal feature fusion for robust event detection in Web videos," in *Proc. CVPR*, Providence, RI, USA, 2012, pp. 1298–1305.

[39] A. Vahdat, K. J. Cannons, G. Mori, S. Oh, and I. Kim, "Compositional models for video event detection: A multiple kernel learning latent variable approach," in *Proc. ICCV*, Sydney, NSW, Australia, 2013, pp. 1185–1192.

[40] A.-A. Liu *et al.*, "Hierarchical & multimodal video captioning: Discovering and transferring multimodal knowledge for vision to language," *Comput. Vis. Image Understand.*, vol. 163, pp. 113–125, Oct. 2017.

[41] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. CVPR*, Boston, MA, USA, 2015, pp. 1–9.

[42] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM MM*, Orlando, FL, USA, 2014, pp. 675–678.

[43] Y. Su and F. Jurie, "Improving image classification using semantic attributes," *Int. J. Comput. Vis.*, vol. 100, no. 1, pp. 59–77, 2012.

[44] Q. Wu, P. Wang, C. Shen, A. R. Dick, and A. van den Hengel, "Ask me anything: Free-form visual question answering based on knowledge from external sources," in *Proc. CVPR*, Las Vegas, NV, USA, 2016, pp. 4622–4630.

[45] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.

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

[47] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. EMNLP*, Doha, Qatar, 2014, pp. 1724–1734.

[48] D. L. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proc. ACL*, Portland, OR, USA, 2011, pp. 190–200.

[49] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in *Proc. CVPR*, Las Vegas, NV, USA, 2016, pp. 5288–5296.

[50] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. ACL*, Philadelphia, PA, USA, 2002, pp. 311–318.

[51] M. J. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proc. WMT@ACL*, Baltimore, MD, USA, 2014, pp. 376–380.

[52] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proc. ACL Workshop*, 2004, pp. 74–81.

[53] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proc. CVPR*, Boston, MA, USA, 2015, pp. 4566–4575.

[54] M. D. Zeiler, "ADADELTA: An adaptive learning rate method," *arXiv e-prints*, vol. abs/1212.5701, 2012.

[55] Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," *arXiv e-prints*, vol. abs/1605.02688, 2016.

[56] J. Thomason, S. Venugopalan, S. Guadarrama, K. Saenko, and R. J. Mooney, "Integrating language and vision to generate natural language descriptions of videos in the wild," in *Proc. COLING*, 2014, pp. 1218–1227.

[57] S. Venugopalan, L. A. Hendricks, R. J. Mooney, and K. Saenko, "Improving lstm-based video description with linguistic knowledge mined from text," in *Proc. EMNLP*, Austin, TX, USA, 2016, pp. 1961–1966.

[58] L. Baraldi, C. Grana, and R. Cucchiara, "Hierarchical boundary-aware neural encoder for video captioning," in *Proc. CVPR*, Honolulu, HI, USA, 2017, pp. 1657–1666.

[59] J. Dong, X. Li, W. Lan, Y. Huo, and C. G. M. Snoek, "Early embedding and late reranking for video captioning," in *Proc. ACM MM*, 2016, pp. 1082–1086.

[60] R. Shetty and J. Laaksonen, "Frame-and segment-level features and candidate pool evaluation for video caption generation," in *Proc. ACM MM*, Amsterdam, The Netherlands, 2016, pp. 1073–1076.

[61] V. Ramanishka *et al.*, "Multimodal video description," in *Proc. ACM MM*, Amsterdam, The Netherlands, 2016, pp. 1092–1096.

[62] Q. Jin, J. Chen, S. Chen, Y. Xiong, and A. G. Hauptmann, "Describing videos using multi-modal fusion," in *Proc. ACM MM*, Amsterdam, The Netherlands, 2016, pp. 1087–1091.

Ning Xu is currently pursuing the doctoral's degree at the School of Electrical and Information Engineering, Tianjin University, Tianjin, China.

His current research interests include computer vision and machine learning.

An-An Liu (M'10) received the Ph.D. degree from the School of Electronic Engineering, Tianjin University, Tianjin, China.

He is currently an Associate Professor with the School of Electrical and Information Engineering, Tianjin University. He was a Visiting Scholar with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA, and a Visiting Professor with the SeSaMe Centre, National University of Singapore, Singapore. His current research interests include computer vision and machine learning.

Wei-Zhi Nie received the Ph.D. degree from the School of Electronic Engineering, Tianjin University, Tianjin, China.

He is an Assistant Professor with the School of Electrical and Information Engineering, Tianjin University. He was a Visiting Scholar with the National University of Singapore, Singapore. His current research interests include computer vision and machine learning.

Yu-Ting Su received the Ph.D. degree from the School of Electronic Engineering, Tianjin University, Tianjin, China.

He is a Professor with the School of Electrical and Information Engineering, Tianjin University. His current research interests include computer vision and machine learning.