# Leveraging Weak Semantic Relevance for Complex Video Event Classification

Chao Li, Jiewei Cao, Zi Huang, Lei Zhu
The University of Queensland
Australia
{c.li1, j.cao3}@uq.edu.au, huang@itee.uq.edu.au
leizhu0608@gmail.com

Heng Tao Shen
UESTC
China
shenhengtao@hotmail.com

## Abstract

*Existing video event classification approaches suffer from limited human-labeled semantic annotations. Weak semantic annotations can be harvested from Web-knowledge without involving any human interaction. However such weak annotations are noisy, thus can not be effectively utilized without distinguishing its reliability. In this paper, we propose a novel approach to automatically maximize the utility of weak semantic annotations (formalized as the semantic relevance of video shots to the target event) to facilitate video event classification. A novel attention model is designed to determine the attention scores of video shots, where the weak semantic relevance is considered as attentional guidance. Specifically, our model jointly optimizes two objectives at different levels. The first one is the classification loss corresponding to video-level groundtruth labels, and the second is the shot-level relevance loss corresponding to weak semantic relevance. We use a long short-term memory (LSTM) layer to capture the temporal information carried by the shots of a video. In each timestep, the LSTM employs the attention model to weight the current shot under the guidance of its weak semantic relevance to the event of interest. Thus, we can automatically exploit weak semantic relevance to assist video event classification. Extensive experiments have been conducted on three complex large-scale video event datasets i.e., MEDTest14, ActivityNet and FCVID. Our approach achieves the state-of-the-art classification performance on all three datasets. The significant performance improvement upon the conventional attention model also demonstrates the effectiveness of our model.*

## 1. Introduction

Video event classification is widely applied in many real-world applications, such as security surveillance, human-computer interaction, etc. It has been becoming one of the most significant research problems in computer vision, multimedia, and artificial intelligence communities [18, 45, 44, 14, 41, 2]. The big intra-class variation in visual content is a major challenge for video event classification. On the one hand, most of the events (e.g., "birthday party", "wedding ceremony") have unconstrained content which includes various entities (e.g., objects, people, animals) with diverse interactions. On the other hand, even for the videos illustrating the same event, they may have very different visual appearances. For instance, with regard to "repairing an appliance", the appliance could be a television in black or a washing machine with white paint. The above facts lead to big intra-class variation in visual content for an event. To alleviate this variation, high quality semantic annotations are in demand.

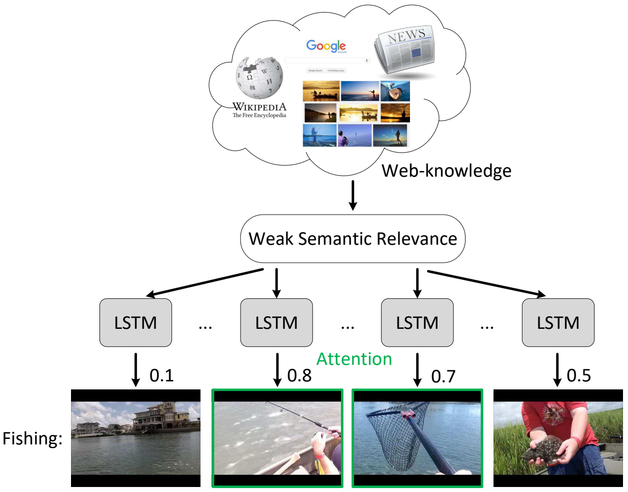

Figure 1: Illustration of the proposed framework. Our framework first harvests weak semantic knowledge from Web-knowledge, then uses it as a weak guidance to the attention model. The LSTM layer then employ the attention model to assign attention score for each shot in a video.

Essentially, a video consists of a sequence of shots. Generally, not all the shots are relevant to the event represented by the video. A natural way to evaluate the importance of a video shot to its belonging event is exploiting its semantic relevance [1, 30, 10] to the event of interest. Specifically, when classifying a video, we wish to pay more attention on the shots with high semantic relevance to the target event, and neglect the ones with low relevance. How to assign semantic annotations to each video shot and how to measure its relevance to the target event are two major research issues to address in video event classification.

In some recent works [1, 30, 10], a small number of event-related semantic concepts (less than 100) are predefined. Concept detectors are then trained on the manually annotated video shots. The response scores of testing video shots w.r.t these detectors are used as semantic relevance to the target event. To prepare the annotated video training set, a large amount of human effort is required. Moreover, when a new event comes, it needs to annotate each video again. In contrast to the prohibitive labour cost on obtaining sufficient semantic annotations for every shot in millions of videos, Li et al. [18] propose to automatically discover latent concepts in a data-driven manner. Furthermore, weak semantic relevance can be conveniently gained from easily accessible Web-knowledge [6, 29]. For instance, in [20, 6], event-related Web images are downloaded from Google and Flickr by directly searching the event names. The authors assume these Web images are in high relevance to their corresponding events and can be used in fine-tuning CNNs for video event classification. In [38], CNNs pre-trained on object and scene classification tasks are respectively applied on videos. The probabilistic outputs of these CNNs are considered as semantic relevance w.r.t object and scene respectively, which are further used as the input features to a fusion network.

Once a reliable semantic relevance has been achieved, a straightforward way to utilize it is to directly combine it with low-level shot features (e.g., SIFT [19], STIP [17], Dense Trajectory [35] ). For example, before aggregating the shot features of a video into a global bag-of-words (BoW) vector, we can weight them by their semantic relevance to the target event. However the weak semantic relevance gained from Web-knowledge is not always reliable yet even noisy due to the domain gap [40, 29] between the Web-knowledge and the videos. Directly employing it without distinguishing its reliability can not maximize its utility. Even worse, it may bring noise to the final representation, resulting in inferior classification performance.

Motivated by the above facts, we propose a long short-term memory (LSTM) [9] framework with a novel attention model which takes semantic relevance gained from Web-knowledge as weak guidance. Attention model [23] is recently used on image and video captioning tasks [39, 43, 42]. When a caption is being generated for an image, the caption model pays attention to different regions in each step. Inspired by their success, we design a novel attention model to automatically evaluate the weight of current testing video shot based on its weak semantic relevance to the event of interest. As aforementioned, the semantic relevance generated from Web-knowledge is weak and noisy. To maximize its utility, the proposed attention model assigns attention score to the current video shot automatically in each timestep by taking the semantic relevance as a weak guidance rather than simply considering the semantic relevance as the weight of each shot. The score of a testing video to a target event will be then computed based on its weighted shots.

The contributions of our work are summarised as follows:

- To leverage weak semantic relevance for video event classification, our framework jointly optimizes two objectives at two levels. The first one is the classification loss corresponding to the video-level groundtruth label, and the second one is the shot-level relevance loss corresponding to the weak semantic relevance.

- To maximize the utility of weak semantic relevance for video event classification, we propose a novel attention model. Instead of entirely following the weak semantic relevance, the proposed attention model takes it as a weak guidance to automatically weight each testing video shot.

- We conduct extensive experiments on three large-scale video event datasets, i.e., MEDTest14, ActivityNet and FCVID. The experimental results demonstrate the effectiveness of the proposed framework w.r.t leveraging weak semantic relevance for video event classification. We achieve state-of-the-art classification performance on all of these three datasets.

## 2. Related Works

Complex video event classification has attracted wide attention in computer vision, multimedia and artificial intelligence communities. The major challenge to complex video event classification is the high intra-class variation caused by unconstrained content and various visual appearances. To alleviate this issue, methods utilizing semantic information have been proposed [1, 30, 10, 25, 12, 13, 32]. However methods based on human-labelled semantic information [1, 30, 10] require large amount of human effort to create and maintain a semantic information database. Alternatively, in some recent works [6, 29], the methods exploiting Web-knowledge are proposed for zero-shot video event classification. These methods harvest semantic relevance from Web-knowledge, which are utilized by applying heuristic algorithms. Jain et al. [11] use ImageNet objects to encode unseen video classes via semantic embedding. Gan et al. [5] fine-tune a CNN that are pre-trained on ImageNet for video event classification and evidence recounting. In [20, 6], Web images related to events are collected from Google and Flickr by directly searching the event names. The authors assume these Web images are in high relevance to their corresponding events, which can be used in fine-tuning CNNs for video event classification. In [38], CNNs pre-trained on object and scene classification tasks are respectively applied on videos. The probabilistic outputs of these CNNs are considered as semantic relevance w.r.t object and scene respectively, which are further used as the input features of a fusion network. Chang et al. [3] sort the video shots by their semantic relevance, based on which an isotonic regularizer is developed to exploit the ordering information. Different from the above related works, we use semantic relevance generated from Web-knowledge as a weak guidance to our proposed attention model, where an attention score will be assigned to current video shot in each timestep. The whole process is automatic without human interfering.

Video event has plentiful temporal information. For example, the event "birthday party" typically consists of the following activities in sequence: "people singing", "blowing out candles", "applauding", and "cutting cake". Unfortunately, this valuable temporal information is usually neglected by traditional methods (e.g., BoW) for video event classification. In our proposal, we use LSTM [9] to capture the temporal information in complex events. LSTM is a type of the recurrent neural network (RNN) [9], which memorizes useful patterns of preceded observations to provide long range context for the prediction of the current step. There are many applications of LSTM such as sentiment analysis, machine translation, image captioning, etc. [31, 34].

Attention model [23] is recently introduced for image and video captioning tasks [39, 43, 42]. In their models, the current caption word is generated by paying different attentions on different image regions or different video shots in each timestep. The attention models they proposed are only guided by the groundtruth of the captions. Different from these traditional attention models, we design a novel one, which is not only supervised by the video-level groundtruth labels but also takes into account the semantic relevance as a weak guidance to generate attention scores. The proposed attention model aims to maximizes the utility of the weak semantic relevance to assist video event classification.

## 3. The Proposed Approach

In this section, we propose a framework for video event classification, which consists of a novel attention model to generate an attention score for each shot and an LSTM layer to capture the temporal information embedded in video shots. Importantly, the proposed attention model takes the weak semantic relevance as a guidance, where the utility of weak semantic relevance is effectively exploited to serve the video event classification.

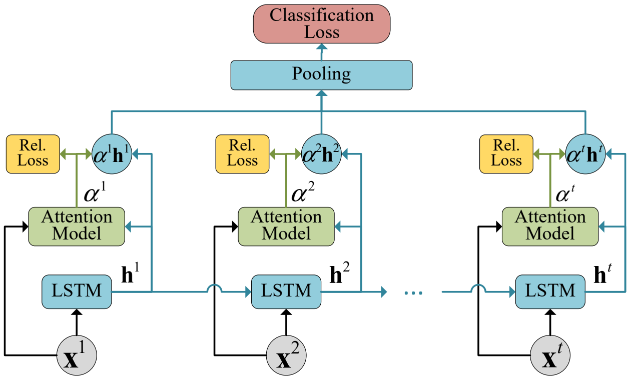

Figure 2: Demonstration of the proposed framework. $\mathbf{x}^t$ represents the feature of shot $t$. $\mathbf{h}^t$ corresponds to the temporal representation returned by LSTM at time $t$. $\alpha^t$ is the attention score for shot $t$, which is evaluated by the proposed attention model. Our framework jointly optimizes two objectives at two different levels respectively. One is the relevance loss at shot-level, and the other one is the classification loss at video-level.

### 3.1. Weak Semantic Relevance Extraction

In this work we use ImageNet [26] and publicly available NLP corpus such as Wikipedia Dump [37] as our sources of computing weak semantic relevance. The ImageNet dataset has $C = 1000$ categories, each of which comes with an entity description (e.g., laptop computer, german shepherd dog). Assume there are a number of $E$ events in our video dataset. Each event has a text description. We use a Word2Vec embedding [22] that pre-trained on massive natural language corpus to evaluate the semantic relevance between the ImageNet category and the target event based on their text descriptions. In Word2vec embedding, each word is embedded in a continuous vector space. Two words with similar semantic meanings have close Cosine distance in this vector space [24, 22, 21]. Note that, for a description with multiple words, we use the average of these word vectors as its final representation. Now, for each event $e \in [1, E]$ we obtain a $C$-dimensional relevance score vector $\mathbf{S}^e \in \mathbb{R}^C$, in which each element indicates the relevance of the corresponding category to the target event $e$.

For a video $v_i$, we first segment it into a sequence of shots and sample one frame from each as its representation. For each shot $t$, a deep CNN [16, 28] pre-trained on ImageNet is used to output a 1000-way vector $\mathbf{p}_i^t$, which is a probability distribution over 1000 ImageNet categories. In [3], the final semantic relevance score of the $t$-th shot to the target event $e$ is defined as the probabilistic expectation of the relevance scores over all 1000 categories.

$$r_i^{t,e} = \sum_{c=1}^{C} p_{i,c}^t S_c^e \quad (1)$$

where $p_{i,c}^t$ is the $c$-th element in the probability vector of the $t$-th shot in video $v_i$. However, the long tail of this distribution may pollute the final semantic relevance. Inspired by [11], we select the top 50 most responsive elements in $\mathbf{p}_i^t$ and re-normalize them with softmax. The expectation over this new distribution is taken as our final semantic relevance.

This type of semantic relevance is generated from both image domain and natural language domain. Semantic gaps certainly exist among language, image and video domains, resulting in low reliability compared with human-labelled semantic relevance. Hence we call it weak semantic relevance. Note that it is only one method to calculate relevance. Other methods such as heuristic algorithm proposed in [29] can also be applied in our framework.

## 3.2. Problem Formulation

Suppose we have $N$ labelled videos $(v_i, \boldsymbol{l}_i)$ in the training set, where $i \in [1, N]$, $\boldsymbol{l}_i \in \{0,1\}^E$, $l_i^e$ indicates whether $v_i$ belongs to event $e$. The feature of the $t$-th shot from video $v_i$ is represented as $\mathbf{x}_i^t$, where $t \in [1, M_i]$ and $M_i$ is the total number of shots in $v_i$ . Each video $v_i$ is associated with a weak relevance vector $\mathbf{r}_i^e \in \mathbb{R}^{M_i}$, in which each element $r_i^{t,e}$ corresponds to the relevance score of shot $\mathbf{x}_i^t$ to the target event $e$. We denote the set of all videos and labels as $V$ and $L$ respectively, and the set of relevance vectors of all videos as $R$. Under the guidance of the weak semantic relevance, the proposed attention model evaluates the attention score $\alpha_i^{t,e}$ for video shot $\mathbf{x}_i^t$ with regard to event $e$. The proposed framework pays different attentions to different shots when conducting classification.

To effectively leverage weak semantic relevance in our framework, we aim to maximize its utility with the attention model. To this end, we formulate the video event classification task assisted by weak semantic relevance by jointly optimizing the following two losses at two different levels respectively:

$$Loss(V, L, R) = (1 - \lambda_a) L_c(V, L) + \lambda_a L_a(V, R) \quad (2)$$

where $L_c(V, L)$ is the classification loss corresponding to the groundtruth labels $L$, and $L_a(V, R)$ is the relevance loss at shot-level with respect to the guidance from weak semantic relevance $R$ received by the attention model. $\lambda_a$ is the parameter controlling the contribution of the guidance from the weak semantic relevance.

With Equation (2) as the objective function of the overall framework (illustrated in Fig. 2), we develop the specific formulations of $L_c(V, L)$ and $L_a(V, R)$ in the following sections.

## 3.3. Video Event Classification by Paying Attention to Relevant Shots

It is a natural way to focus attention on relevant shots when performing classification on an event video. To achieve that, we use an attention score to measure the relevance of each video shot to its target event. The LSTM layer [9] in our framework is designed to capture the temporal information carried by the shots in a video. In each timestep, the LSTM unit returns the representation for the current shot, which memorizes useful patterns observed in its preceded video shots. We classify a video based on the representation sequence produced by the LSTM layer and the attention score assigned by the proposed attention model. The probability of video $v_i$ being classified to the event $e$ is denoted as $p_i^e$, which is formally defined as:

$$\begin{aligned} p_i^e &= f(\bar{\mathbf{h}}_i^e; \mathbf{w}_f) = \frac{\exp(\mathbf{w}_f^e \cdot \bar{\mathbf{h}}_i^e)}{\sum_{j\in[1,E]} \exp(\mathbf{w}_f^j \cdot \bar{\mathbf{h}}_i^j)} \\ \bar{\mathbf{h}}_i^e &= \frac{1}{Z_i^e} \sum_{t=1}^{M_i} \alpha_i^{t,e} \cdot \mathbf{h}_i^t \\ \mathbf{h}_i^t &= g_l(\mathbf{x}_i^t, \mathbf{h}_i^{t-1}; \mathbf{w}_l) \\ Z_i^e &= \sum_{t=1}^{M_i} \alpha_i^{t,e} \\ &where\ t \in [1, M_i],\ e \in [1, E] \end{aligned} \quad (3)$$

where $f(\cdot\ ;\ \mathbf{w}_f)$ is the softmax scoring function, parameterized by $\mathbf{w}_f$. $[\mathbf{h}_i^1, \mathbf{h}_i^2, ..., \mathbf{h}_i^{M_i}]$ is the representation sequence produced by the LSTM layer, where $\mathbf{h}_i^t$ is the representation returned by the LSTM layer in timestep $t$. It is further weighted by the attention score sequence $[\alpha_i^{1,e}, \alpha_i^{2,e}, ..., \alpha_i^{M_i,e}]$ evaluated by the proposed attention model. The weighted average of this representation sequence, i.e., $\bar{\mathbf{h}}_i^e$, is taken as the input by the softmax function $f$. $g_l(\cdot\ ,\ \cdot\ ;\ \mathbf{w}_l)$ is the updating function within each LSTM unit and $\mathbf{w}_l$ is the corresponding parameters. The attention score $\alpha_i^{t,e}$ for shot $\mathbf{x}_i^t$ with regard to event $e$ is calculated in each timestep by the attention model.

Accordingly, we define the video-level classification loss as the following categorical cross-entropy loss:

$$L_c(V, L) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{e=1}^{E} l_i^e \log(p_i^e) \quad (4)$$

Note that, for a video with multiple labels, we normalize its label vector $\boldsymbol{l}_i$ by $L_1$ norm to get a probability vector.

## 3.4. The Proposed Attention Model

The attention model [23] is recently incorporated in LSTM framework for sequence generation task, such as image captioning [23, 43, 39] and video captioning [42]. Its basic idea is that, when generating a caption for an image or video, in each timestep, the attention model computes the weight, i.e., attention score, for every individual visual region (e.g., image regions, video shots). Based on the combination of the weighted visual regions, the LSTM layer generates a word for the current timestep.

However the above attention models are only supervised by the ground-truth labels, i.e., the captions of images or videos. To effectively leverage weak semantic relevance in video classification, we design a novel attention model which is not only supervised by the groundtruth event labels, but also guided by weak semantic relevance.

For a video $v_i$, in timestep $t$, we define the attention score vector $\boldsymbol{\alpha}_i^t$ for shot $\mathbf{x}_i^t$ by the following equations:

$$\begin{aligned} &\boldsymbol{\alpha}_i^t = g_a(\mathbf{h}_i^t, \mathbf{x}_i^t; \mathbf{w}_a) \\ &where\ t \in [1, M_i] \end{aligned} \tag{5}$$

where $g_a(\cdot\ ,\ \cdot\ ;\ \mathbf{w}_a)$ is an attention network with softmax output and being parameterized by $\mathbf{w}_a$. Each element $\alpha_i^{t,e}$ in $\boldsymbol{\alpha}_i^t$ is the attention score of shot $\mathbf{x}_i^t$ with respect to event $e$. We use a multi-layer perceptron as our attention network conditioned on shot feature $\mathbf{x}_i^t$ and its corresponding representation $\mathbf{h}_i^t$ produced by the LSTM layer.

Note that, most existing attention models are designed for captioning, i.e., word sequence generation, where strong relations between neighbouring words exist. Basically, these models compute the attention score for current timestep $t$ purely based on the previous representation $\mathbf{h}_i^{t-1}$ [39, 42]. In video event classification task, we focus on the discriminative power of the final video representations. Therefore, our attention network is conditioned on $\mathbf{h}_i^t$ and $\mathbf{x}_i^t$. More specifically, we feed the concatenated vector $[\mathbf{h}_i^t, \mathbf{x}_i^t]$ to our attention network, where $\mathbf{h}_i$ captures the temporal information of the observed video shots and $\mathbf{x}_i$ preserves the inherent visual appearance of the current shot.

The weak semantic relevance can be hardly used as the attention score directly to weight video shots, because it is noisy and not reliable enough. Instead of completely rely on it, we utilise it in our attention model as a weak guidance. The attention loss $L_a(V, R)$ is correspondingly formulated as:

$$L_a(V,R) = \frac{1}{N}\sum_{i=1}^{N} ||\boldsymbol{\alpha}_i^t - \mathbf{r}_i^t||^2 \tag{6}$$

where $\boldsymbol{\alpha}_i^t$ is the attention score vector of video $v_i$, calculated by Equation (5). This loss function implies that $\boldsymbol{\alpha}_i^t$ follows a Gaussian distribution with mean $\mathbf{r}_i^t$. By this way the proposed attention model takes the weak relevance as a priori when computing the attention score for the current shot.

The overall objective function, i.e., Equation (2), is optimized using stochastic gradient descent. By minimizing this objective function, our model exploits weak semantic relevance by the proposed attention model to facilitate video classification.

As emphasised before, the proposed attention model is supervised not only by video-level groundtruth event label, but also under the weak guidance of the shot-level semantic relevance. We can examine this by investigating the propagation path of gradient w.r.t attention scores: according to Equations (2), (3), (4), (5) and (6), the gradient w.r.t the attention model is:

$$\frac{\partial Loss(V,L,R)}{\partial \boldsymbol{\alpha}_i} = (1-\lambda_a)\frac{\partial L_c(V,L)}{\partial \bar{\mathbf{h}}_i}\cdot\frac{\partial \bar{\mathbf{h}}_i}{\partial \boldsymbol{\alpha}_i} + \lambda_a \frac{\partial L_a(V,R)}{\partial \boldsymbol{\alpha}_i} \tag{7}$$

Similarly, the LSTM layer is also supervised by these two level losses. The gradient w.r.t the parameters of LSTM layer, i.e., $\mathbf{w}_l$ is:

$$\begin{aligned} \frac{\partial Loss(V,L,R)}{\partial \mathbf{w}_l} &= (1-\lambda_a)\frac{\partial L_c(V,L)}{\partial \bar{\mathbf{h}}_i}\cdot\frac{\partial \bar{\mathbf{h}}_i}{\partial \mathbf{w}_l} \\ &+ \lambda_a \frac{\partial L_a(V,R)}{\partial \boldsymbol{\alpha}_i}\cdot\frac{\partial \boldsymbol{\alpha}_i}{\partial \bar{\mathbf{h}}_i}\cdot\frac{\partial \bar{\mathbf{h}}_i}{\partial \mathbf{w}_l} \end{aligned} \tag{8}$$

The above equations clearly illustrates how the proposed framework learns from two different knowledge sources, i.e., event videos and Web-collected weak semantic relevance.

# 4. Experiments

In this section, we conduct extensive experiments to evaluate the effectiveness of our framework and the ability of the proposed attention model with leveraging weak semantic relevance.

## 4.1. Experiment Setup

**Dataset.** The performance study is conducted on three large-scale benchmark video event datasets, i.e., MEDTest14 [33], ActivityNet [8] and FCVID [15].

MEDTest14 [33] is a commonly-used benchmark dataset covering 20 events for complex video event classification. Each event has 100 positive training examples, and all events share about 5,000 negative training examples. The test set has approximately 23,000 videos.

ActivityNet [8] is recently released for complex human activity recognition. It comprises 28K videos of 203 activity categories collected from YouTube. The lengths of the videos range from several minutes to half an hour. The total length of the whole dataset is 849 hours. Many of the videos in this dataset are shot by amateurs in uncontrolled environments, where the variances within the same activity category are often large. ActivityNet provides trimmed and untrimmed videos for evaluation. Following the settings in [38], we adopt a more challenging untrimmed setting for our experiments. ActivityNet consists of training, validation, and test splits. The test split is not publicly available, as the authors are reserving the test data for a potential future competition. Hence we use validation set as our test set as well as [38] does.

FCVID [15] consists of 91,223 Web videos annotated manually according to 239 categories. The total duration of all videos is 4,232 hours and the average duration per video is 167 seconds. The categories in FCVID cover a wide range of topics like social events (e.g., "tailgate party"), procedural events (e.g., "making cake"), objects (e.g., "panda"), scenes (e.g., "beach"), etc. We use its standard split of 45,611 videos for training and 45,612 videos for testing.

**Implementation Details.** Due to the computational limitation of our experimental environment, we construct a moderate sized network by segmenting each video into 30 shots. The color histogram difference between consecutive frames is considered as the indicator of shot boundaries. Other segmentation algorithms can also be employed in our framework. For videos that have more than 30 shots, an agglomerative clustering alike method is applied to repeatedly merge two shortest shots into one in each round until the number of all shots is reduced to 30. For videos with less than 30 shots, we simply pad them with zeros at the tail. The middle frame of each shot is selected as its representative, whose feature is extracted by applying a very deep CNN architecture (from fc6 layer of VGG-19 [28]). We also use its probability output to compute the weak semantic relevance for each frame as explained in Sec. 3.1. Since unidirectional LSTM can only capture the previously observed temporal patterns (related to the current timestep) in a video, we adopt bidirectional LSTM [27, 7] to capture the intact temporal context (previous and post). Stochastic gradient descent algorithm with momentum is used to train our model. The batch size, momentum, and dropout rate (applied on both LSTM layer and fully connected layer) are set to be 64, 0.9 and 0.1 respectively. The learning rate is set to be 0.01 initially and divided by 10 after every 10K iterations. Finally, we employ mean average precision (mAP) to evaluate the overall performance on all three datasets.

**Compared methods.** The proposed approach are compared with the following alternative methods including two baseline methods and four state-of-the-art methods that also utilize weak semantic relevance generated from Web:

1. SVM-WA. The weak semantic relevance is directly used to weight video shot features without considering its reliability. The weighted shot features in a video are then average-pooled into a global feature vector, on which SVM is applied for classification.

2. LSTM-NR. It is a variant of the proposed method without utilizing weak semantic relevance. It is equivalent to LSTM with a conventional attention model.

3. Nearly-Isotonic SVM (NISVM) [3]. This state-of-the-art method sorts the video shots by their semantic relevance. An isotonic regularizer is introduced to impose larger weights on the shots with higher semantic relevance.

4. Ma et al. [20]. The authors download 393K event-related Web images from Google and Flickr by directly searching the event names. These Web images are assumed in high relevance to their corresponding events and are further used in fine-tuning CNNs.

5. Jiang et al. [15]. It combines multiple state-of-the-art handcrafted visual features (e.g., improved dense trajectories) and deep features. The authors use a regularized deep neural network to exploit feature and class relationships.

6. OSF[38]. In this paper, the CNNs pre-trained on object and scene classification tasks are respectively applied on videos. The probabilistic outputs of these CNNs are considered as the semantic relevance w.r.t object and scene respectively and are used as the input features of a fusion network.

| | ActivityNet | FCVID | MEDTest14 |
|---|---|---|---|
| SVM-WA | 50.8% | 69.9% | 28.1% |
| LSTM-NR | 55.1% | 73.2% | 29.1% |
| Ours | **61.6%** | **77.8%** | **36.3%** |

Table 1: Comparisons with baseline methods on ActivityNet, FCVID and MEDTest14 datasets

| | ActivityNet | FCVID |
|---|---|---|
| Ma et al. [20] | 53.8% | - |
| Heilbron et al. [8] | 42.5% | - |
| Jiang et al. [15] | - | 73.0% |
| OSF [38] | 56.8% | 76.5% |
| Ours | **61.6%** | **77.8%** |

Table 2: Comparisons with state-of-the-art methods on ActivityNet and FCVID datasets. Our method achieve best classification performance on both of the two datasets.

Although there are other video classification methods, they are either based on feature ensemble or fusion of snippet scores [36] but not utilizing semantic information, hence do not apply in our comparable experiments.

## 4.2. Comparison with Baseline Methods

To examine the extent to which the weak semantic relevance harvested from Web-knowledge can facilitate video classification, we compare our method with two baseline models SVM-MA and LSTM-NR. Table 1 shows the video

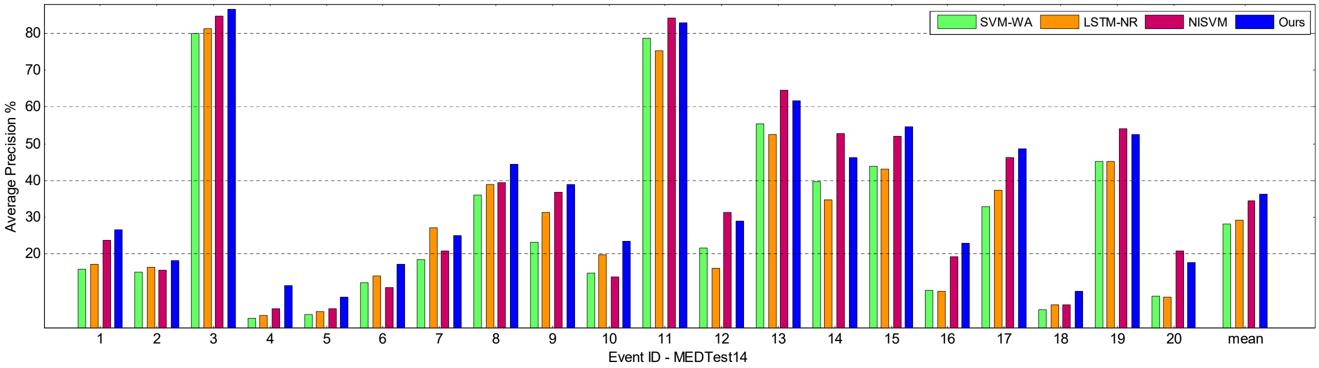

Figure 3: Results on MEDTest14 dataset. The mean APs of SVM-WA, LSTM-NR, NISVM and our full model are 28.1%, 29.1%, 34.4% and 36.3% respectively

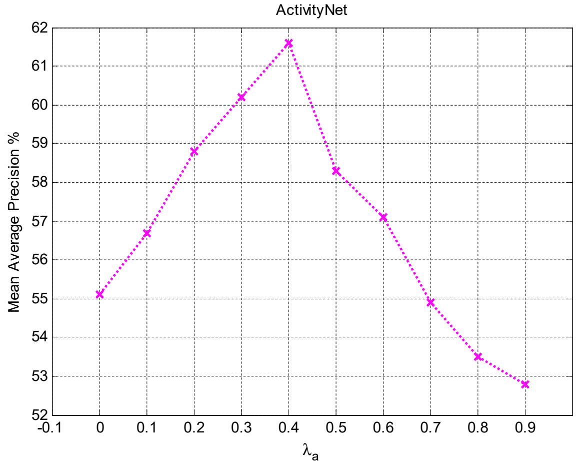

Figure 4: The effect of the relevance loss, i.e., trade-off parameter $\lambda_a$, on ActivityNet dataset

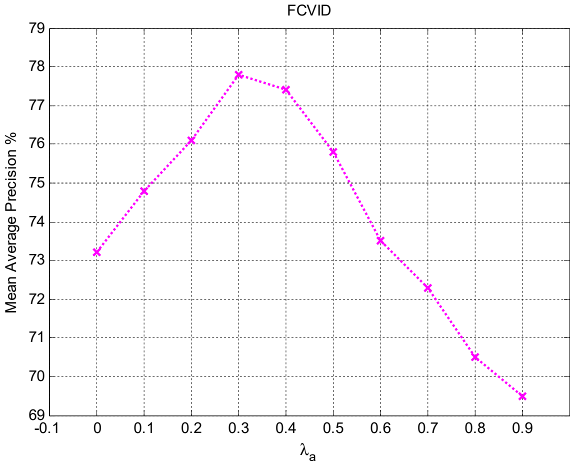

Figure 5: The effect of the relevance loss, i.e., trade-off parameter $\lambda_a$, on FCVID dataset

classification performance of the evaluated baseline methods and the proposed approach. The proposed method outperforms SVM-MA by large margins i.e., 10.8%, 7.9%, and 8.2% on ActivityNet, FCVID, and MEDTest14 respectively. It apparently evidences our assumption discussed in Sec. 3.1 that utilising the weak semantic relevance without distinguish its reliability may result in inferior classification performance. The automatic learning process in our proposal effectively distinguishes useful information from noisy weak semantic relevance.

Our method is also compared with its variant LSTM-NR. The main difference between these two methods lies on the attention model training process, where the conventional attention model used in LSTM-NR is supervised by the groundtruth event label while our novel attention model also takes weak semantic relevance as a weak guidance. We get this variant by setting the parameter $\lambda_a$ in Eq. (2) as 0. As shown in Table 1, the proposed model outperforms its variant over all three datasets. It indicates that our attention model which leverages semantic relevance as weak guidance is superior than the conventional one. The weak semantic relevance makes significant contribution to achieving promising classification performance.

### 4.3. Comparison with State-of-the-art Methods

In this section, we compare our method with four state-of-the-art methods: NISVM [3], Ma et al. [20], Jiang et al. [15], and OSF[38]. In Fig. 3 and Table 2, we report the results of the performance study over all three datasets.

NISVM [3] is similar to our method. The same points we share are that, we both aim to assign larger weights to video shots with higher semantic relevance and the same sources to obtain weak semantic relevance are used. For a fair comparison, we adopt same settings with [3]. On MEDTest14 dataset, we use Eq. (1) to compute the semantic relevance as in [3], without selecting top 50 most responsive elements. We quote their best results to compare with ours. In Fig. 3, the mean APs of NISVM and our model are 34.4% and 36.3% respectively. Our method outperforms NISVM on 14 events out of 20 events. NISVM sorts the video shots by semantic relevance and only considers the ordering information among video shots. As discussed before, our method employs both the semantic relevance as a weak guidance to the proposed attention model and a bidirectional LSTM layer to capture the long-term temporal context among video shots. Hence, our model can exploit more valuable information from both of the semantic relevance and the temporal patterns in video shots.

For event categories 11, 12, 13 and 14, corresponding to “bee keeping”, “wedding shower”, “non-motorized vehicle repair”, and “fixing musical instrument”, our method performs not as good as NISVM. After carefully investigating the videos for these four events, we find out that most of these videos are comprised of static scenes, such as “farm”, “church”, etc. In result, the temporal information is overwhelmed by the strong static visual appearance and the LSTM layer in our model is overfitted. The fact that LSTM-NR performs even worse than SVM-WA on these four events also proves this observation.

In [20], Ma et al. evaluate several recent proposed very deep CNN architectures such as VGG-16, VGG-19 [28] and M2048 [4] etc., for fine-tuning. We quote their best result on ActivityNet dataset from their original paper. Observed from Table 2, the proposed method outperforms their method by a clear margin of 6.7% on ActivityNet. The possible reasons are as follows. Firstly, the compared method does not explicitly distinguish the reliability of the event-related images, which may brings noise to the CNNs and be used for fine-tuning. It is not clear how robust the CNNs are to the noises. Secondly, an LSTM layer is used in our model to capture the temporal information in videos, while the CNNs used in [20] for fine-tuning can only capture the spatial visual appearance of images.

We have also quoted the best results of [38] on ActivityNet and FCVID datasets from their original paper in Table 2. This demonstrates the superior effectiveness of our model with regard to utilizing weak semantic relevance. Note that, their method leverages semantic relevance from three aspects i.e., object, scene, and low-level CNN feature, each of which corresponds to a different source domain. In this paper, our method only utilizes one source of semantic relevance. However, it can be naturally extended to combine heterogeneous semantic relevance sources and is expected to achieve an even better performance.

Jiang et al. [15] combines multiple state-of-the-art hand-crafted visual features (e.g., improved dense trajectories) and deep features for video event classification. They use a regularized deep neural network to exploit feature and class relationships. As clearly shown in Table 2, our model with the consideration of semantic relevance is more effective. In the meanwhile, our method is expected to be further improved by considering motion features for video shot representation, while we only use static CNN feature for model simplicity in this work.

### 4.4. Experimental Study of The Contribution of Weak Semantic Relevance

In this section, we conduct empirical analysis on the contribution of the weak semantic relevance. In Fig. 4 and Fig. 5 we depict the performance on ActivityNet and FCVID respectively of the proposed method against different values of parameter $\lambda_a$ in Eq. (2). A larger value of $\lambda_a$ means a larger weight on the weak semantic relevance. On ActivityNet dataset our model achieves the best classification performance when $\lambda_a$ = 0.4, and on Fcvid dataset it works best when $\lambda_a$ = 0.3. On both of these two datasets, when $\lambda_a$ increases larger than 0.4, the classification performance drops dramatically. It implies that, when $\lambda_a$ gets larger than 0.4 our model starts to be dominated by the weak semantic relevance. This phenomenon can be understood as follows: the semantic relevance we extract from Web-knowledge is not reliable enough to contribute more than “40%” (corresponds to $\lambda_a$=0.4) to the classification task. If more reliable semantic relevance can be obtained, a larger value should be imposed to $\lambda_a$, i.e., let semantic relevance contributes more for better classification performance.

## 5. Conclusion

In this paper, we propose a framework with a novel attention model to automatically utilize weak semantic relevance to assist video classification task. This framework jointly optimizes two objectives at video-level and shot-level separately, which explicitly affect video classification from both global-level (i.e., video-level labels) and local-level (i.e., shot-level attention scores). To alleviate the effect of the noises carried by the weak semantic relevance, we use weak semantic relevance as a weak guidance in the proposed attention model, instead of considering it as the attention score directly. This process significantly improves the effectiveness of our proposed model.

Comprehensive performance studies have been conducted by comparing our method with six other methods over three large-scale benchmark datasets. The effectiveness of our method is evidenced by its superior performances compared with others.

Our framework can also be smoothly extended and improved by generating weak semantic relevance from heterogenous information sources or combining multiple advanced visual features for video shot representation.

**Acknowledgments.** This work is supported by ARC FT130101530, FT120100718, DP170103954 and National Natural Science Foundation of China No. 61632007.

## References

[1] S. Bhattacharya, M. M. Kalayeh, R. Sukthankar, and M. Shah. Recognition of complex events: Exploiting temporal dynamics between underlying concepts. In *CVPR*, pages 2243–2250, June 2014.

[2] X. Chang, Y. Yang, G. Long, C. Zhang, and A. G. Hauptmann. Dynamic concept composition for zero-example event detection. In *AAAI*, pages 3464–3470, 2016.

[3] X. Chang, Y. Yang, E. P. Xing, and Y. Yu. Complex event detection using semantic saliency and nearly-isotonicSVM. In *ICML*, 2015.

[4] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference, BMVC 2014, Nottingham, UK, September 1-5, 2014*, 2014.

[5] C. Gan, N. Wang, Y. Yang, D. Yeung, and A. G. Hauptmann. Devnet: A deep event network for multimedia event detection and evidence recounting. In *CVPR*, pages 2568–2577, 2015.

[6] 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*, June 2016.

[7] A. Graves, N. Jaitly, and A. Mohamed. Hybrid speech recognition with deep bidirectional LSTM. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, December 8-12, 2013*, pages 273–278, 2013.

[8] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015.

[9] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[10] H. Izadinia and M. Shah. Recognizing complex events using large margin joint low-level event model. In *12th European Conference on Computer Vision*, pages 430–444, 2012.

[11] M. Jain, J. C. van Gemert, T. Mensink, and C. G. M. Snoek. Objects2action: Classifying and localizing actions without any video example. In *ICCV*, pages 4588–4596, 2015.

[12] L. Jiang, A. G. Hauptmann, and G. Xiang. Leveraging high-level and low-level features for multimedia event detection. In *ACM International Conference on Multimedia*, pages 449–458, 2012.

[13] L. Jiang, S. Yu, D. Meng, Y. Yang, T. Mitamura, and A. G. Hauptmann. Fast and accurate content-based semantic search in 100m internet videos. In *ACM International Conference on Multimedia*, pages 49–58, 2015.

[14] Y. Jiang, S. Bhattacharya, S. Chang, and M. Shah. High-level event recognition in unconstrained videos. *IJMIR*, 2(2):73–101, 2013.

[15] Y. Jiang, Z. Wu, J. Wang, X. Xue, and S. Chang. Exploiting feature and class relationships in video categorization with regularized deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, to appear.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1106–1114, Dec. 2012.

[17] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

[18] C. Li, Z. Huang, Y. Yang, J. Cao, X. Sun, and H. T. Shen. Hierarchical latent concept discovery for video event detection. *IEEE Trans. Image Processing*, 26(5):2149–2162, 2017.

[19] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[20] S. Ma, S. A. Bargal, J. Zhang, L. Sigal, and S. Sclaroff. Do less and achieve more: Training cnns for action recognition utilizing action images from the web. *Pattern Recognition*, 2017.

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[22] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.

[23] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu. Recurrent models of visual attention. In *NIPS*, pages 2204–2212, 2014.

[24] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

[25] V. Ramanathan, K. D. Tang, G. Mori, and L. Fei-Fei. Learning temporal embeddings for complex video analysis. In *ICCV*, 2015.

[26] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[27] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE Trans. Signal Processing*, 45(11):2673–2681, 1997.

[28] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[29] B. Singh, X. Han, Z. Wu, V. I. Morariu, and L. S. Davis. Selecting relevant web trained concepts for automated event retrieval. In *ICCV*, pages 4561–4569, 2015.

[30] C. Sun and R. Nevatia. ACTIVE: activity concept transitions in video event classification. In *ICCV*, pages 913–920, Dec. 2013.

[31] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112, 2014.

[32] L. Torresani, M. Szummer, and A. W. Fitzgibbon. Efficient object category recognition using classemes. In *11th European Conference on Computer Vision*, pages 776–789, Sept. 2010.

[33] TRECVID-MED. https://www.nist.gov/itl/iad/mig/med-2014-evaluation. 2014.

[34] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, 2015.

[35] H. Wang, A. Kläser, C. Schmid, and C. Liu. Action recognition by dense trajectories. In *CVPR*, pages 3169–3176, June 2011.

[36] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool. Temporal segment networks: Towards good practices for deep action recognition. In *14th European Conference on Computer Vision*, pages 20–36, 2016.

[37] WikiDump. https://dumps.wikimedia.org/enwiki/latest/.

[38] Z. Wu, Y. Fu, Y.-G. Jiang, and L. Sigal. Harnessing object and scene semantics for large-scale video understanding. In *CVPR*, June 2016.

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

[40] Y. Yang, Y. Luo, W. Chen, F. Shen, J. Shao, and H. T. Shen. Zero-shot hashing via transferring supervised knowledge. In *Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, The Netherlands, October 15-19, 2016*, pages 1286–1295, 2016.

[41] Y. Yang, Z. Zha, Y. Gao, X. Zhu, and T. Chua. Exploiting web images for semantic video indexing via robust sample-specific loss. *IEEE Trans. Multimedia*, 16(6):1677–1689, 2014.

[42] L. Yao, A. Torabi, K. Cho, N. Ballas, C. J. Pal, H. Larochelle, and A. C. Courville. Describing videos by exploiting temporal structure. In *ICCV*, pages 4507–4515, 2015.

[43] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *CVPR*, pages 4651–4659, 2016.

[44] L. Yu, Z. Huang, J. Cao, and H. T. Shen. Scalable video event retrieval by visual state binary embedding. *IEEE Trans. Multimedia*, 18(8):1590–1603, 2016.

[45] L. Yu, Y. Yang, Z. Huang, P. Wang, J. Song, and H. T. Shen. Web video event recognition by semantic analysis from ubiquitous documents. *IEEE Trans. Image Processing*, 25(12):5689–5701, 2016.