You Lead, We Exceed: Labor-Free Video Concept Learning by Jointly Exploiting Web Videos and Images

Chuang Gan[§ *] Ting Yao[†] Kuiyuan Yang[†] Yi Yang[‡] Tao Mei[†]
[§] IIIS, Tsinghua University, Beijing, China
[†] Microsoft Research, Beijing, China
[‡] QCIS, University of Technology Sydney, Australia

Abstract

Video concept learning often requires a large set of training samples. In practice, however, acquiring noise-free training labels with sufficient positive examples is very expensive. A plausible solution for training data collection is by sampling from the vast quantities of images and videos on the Web. Such a solution is motivated by the assumption that the retrieved images or videos are highly correlated with the query. Still, a number of challenges remain. First, Web videos are often untrimmed. Thus, only parts of the videos are relevant to the query. Second, the retrieved Web images are always highly relevant to the issued query. However, thoughtlessly utilizing the images in the video domain may even hurt the performance due to the well-known semantic drift and domain gap problems. As a result, a valid question is how Web images and videos interact for video concept learning. In this paper, we propose a Lead–Exceed Neural Network (LENN), which reinforces the training on Web images and videos in a curriculum manner. Specifically, the training proceeds by inputting frames of Web videos to obtain a network. The Web images are then filtered by the learnt network and the selected images are additionally fed into the network to enhance the architecture and further trim the videos. In addition, Long Short-Term Memory (LSTM) can be applied on the trimmed videos to explore temporal information. Encouraging results are reported on UCF101, TRECVID 2013 and 2014 MEDTest in the context of both action recognition and event detection. Without using human annotated exemplars, our proposed LENN can achieve 74.4% accuracy on UCF101 dataset.

1. Introduction

Motivations. Video concept learning is fundamentally a classification task that predicts whether a video is relevant to a given concept. The significance of the topic is partly reflected in the huge volume of published papers in the area of computer vision in the last decades. For example, support vector machines (SVM) trained on reliable hand-crafted features such as mid-level parts [42, 40], improved dense trajectories [39] and deep neural networks [8, 22, 30, 38] have achieved promising recognition results. A critical step along this process is the acquisition of sufficiently large amounts of quality training data. The acquisition, however, is not a trivial process. For instance, it took long time to construct the ActivityNet [18] and Sport1M [22] datasets, which only contain hundreds of concepts. Such a labor-intensive process will become extremely difficult for the ul-

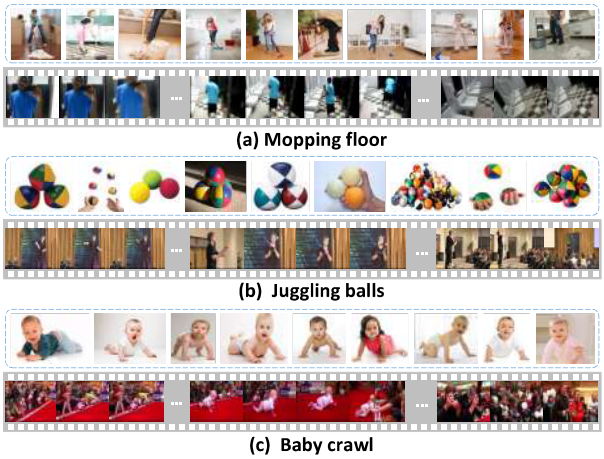
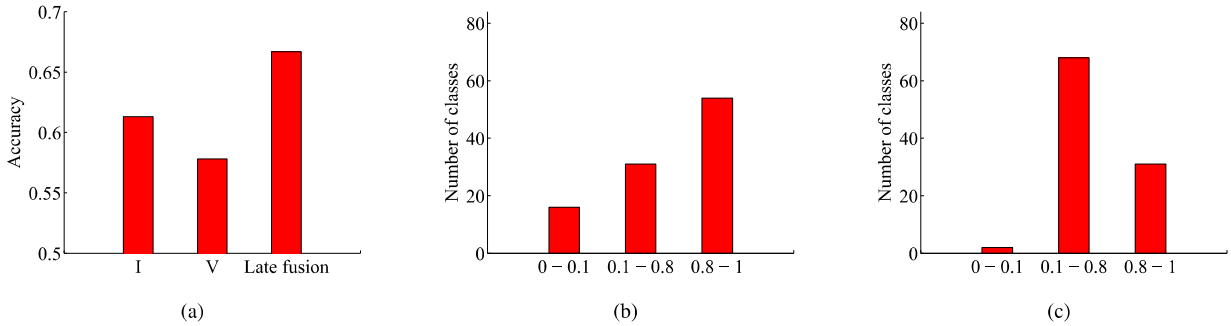


Figure 1. Web image and videos returned by a search engine are usually highly correlated to the query. However, web videos are always untrimmed and contain large portion of irrelevant frames, as indicated by green boxes in this figure. Web images could be noisy due to 1) semantic drift, i.e. the mismatch between query and returned images, for example juggling balls in this figure (b), and 2) domain gap, i.e. the inconsistencies between videos and images, *e.g.* images of baby crawl usually post edited with clean white background.

*This work was done when Chuang Gan was a visiting research student in Microsoft Research Asia.



(a) (b) (c)

Figure 2. Preliminary experiment results: (a) the action recognition performance by using web images only (I), web videos only (V), and Late fuse (Late fuse); (b) the accuracy distribution of different action classes by using web images only; (c) the accuracy distribution of different action classes by using web videos only.

timate goal of labeling thousands of video concepts.

On the other hand, with the success of commercial Web image and video search engines, we can easily crawl sufficient images and videos via Google, Flickr and YouTube, given a concept as a query. Automatic sampling of these Web images and videos for video concept learning thus appears as a natural way of replacing expensive manual labeling. Such a solution sounds promising, though is challenging, particularly for identifying high-quality positive samples. Web images are always well taken, especially for the highlight moments of actions and events. This category of resources, nevertheless, is fragmentally recorded and static. Web videos, in contrast, are untrimmed and with large spatio-temporal variance. Therefore, the videos often contain redundant and irrelevant parts in answering the query. There is no clear mechanism, however, how the Web images and videos could be jointly exploited for video concept learning in a principled way.

Preliminary experiments. To better understand how much Web images and videos could contribute to video concept learning, we conduct a preliminary experiment on the UCF101 action recognition dataset, which contains 101 action categories. First, we collect Web images and videos via the Google image search engine and YouTube, by issuing each action category as a search query. For each category, we crawled around 600 images and 15 videos as positive training examples. To learn video concept detectors, VGGNet [31] is first pre-trained with the ILSVRC-2012 [6] training set of 1.2 million images and then fine-tuned by using Web images and frames of Web videos for action recognition respectively, which is observed to be better than training from scratch [13, 15]. Evaluating the learnt detectors on the test split 2 on UCF101 dataset, Figure 2 (a) shows the accuracies by using Web images, Web videos and their late fusion. There are two observations as shown in the figure: 1) the accuracy by solely using Web images can reach 61.3%, compared with 57.8% using Web videos; 2) with a simple late fusion of the prediction scores of fine-tuned models on Web images and videos, the performance can further be improved to 66.7%. The results essentially indicate that Web images and videos are complementary for learning video concepts.

Figure 2 (b) and (c) further details the performance across different action categories. Overall, different action categories respond quite differently to Web images. Among all the categories, the accuracy surpasses 80% in 54 out of 101 categories. Meanwhile, there are 16 action categories where the accuracy is below 10%. The performance, in contrast, are generally concentrated when exploiting Web videos for each category. There are only 31 categories whose performance is over 80% and 2 categories achieve an accuracy lower than 10%. For instance, the images relevant to the query *mopping floor* (Figure 1 (a)) are all highly related to actions in videos, resulting in good performance by Web images alone. Instead, Web images are found to be quite different in visual appearance from videos due to the domain gap [29] for queries such as *baby crawl* (Figure 1 (c)), and Web videos show better performance. In the extreme case where all Web images are found to be less helpful because of semantic drift [4], as for the query *juggling balls* (Figure 1 (b)), the accuracy of a detector learnt on Web images drops to 0, while the performance can still reach 40% by relying on Web videos. As indicated by our results, allowing an interaction between Web images and videos could lead to better performance for video concept learning. In particular, Web videos should lead the training process, while the learning is enhanced by further involving Web images.

Contributions. By consolidating the idea of jointly exploiting Web images and videos for video concept learning, we present a Lead-Exceed Neural Network (LENN), as shown in Figure 3. Specifically, the training process starts by feeding into all the key frames of Web videos to learn an initial neural network. Then, the network is utilized to

predict on Web images and filter out the noisy ones. The selected Web images further fine-tune the initial network to enhance the whole architecture. The refined architecture is employed to trim Web videos and localize the relevant frames of Web videos to video concept. Finally, Long Short-Term Memory (LSTM) [8] networks are applied on the localized video frames to explore long term temporal information for video concept learning. In summary, this paper makes the following contributions:

- To the best of our knowledge, this is the first in-depth study of utilizing web image and video data, which are arbitrary and noisy, for real world video concept recognition without any human supervision.
- Coupling with the powerful feature learning frameworks Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM), we pave a new Lead–and–Exceed way of video concept learning, which maximizes the instinct strengths of web videos and images while minimizes the side effects caused by semantic drift, domain gap, noises in irrelevant frames, and so forth.
- Experimental results on three large-scale video datasets demonstrate that the proposed system outperforms other webly-supervised approaches and certain few-shot supervised approaches as well.

The remaining sections are organized as follows. Section 2 describes related work on video concept learning and learning from the Web. Section 3 presents our Lead–Exceed Neural Network (LENN) for the video concept learning by jointly exploiting Web images and videos. Section 4 provides empirical evaluations, followed by the discussion and conclusions in Section 5

2. Related Work

Our research involves two research directions, which will be reviewed briefly in this section.

Video Concept Learning. Video concept learning, such as action recognition and event detection, has been widely explored in the community of computer vision and multimedia [47]. A detailed survey can be found in [26]. A considerable portion of these works are about video representation. Improved dense trajectories (IDT) [39] and its variant [24, 41] combined with Fisher vector coding [28] show state-of-the-art performance.

Motivated by the promising results of deep networks (particularly ConvNets) on image analysis tasks [23, 37, 31, 20], there have also been a number of attempts to develop a deep architectures for video recognition. Karpathy *et al.* [22] compared several architectures for action recognition. Tran *et al.* [38] proposed to learn generic spatial-temporal features by using 3D ConvNets for video recognition. Simonyan *et al.* [30] proposed two-stream networks to capture spatial and motion information using frames and stacked optical flows as inputs, respectively. More recently, Recurrent Neural Networks (RNNs), which are well-suited for modeling sequential information have also proven effective on video recognition. Srivastava *et al.* [34] proposed an LSTM encoder-decoder framework to learn video representations in an unsupervised manner [34]. Donahua *et al.* [8] trained a two-layer LSTM network for action classification. Ng *et al.* [27] further demonstrated that a five-layer LSTM network can achieve slightly better results. However, these approaches are all based on the assumption that we have high-quality labeled data that can be used for training. To the best of our knowledge, there are no previous works exploring how to obtain reasonable results using noisy Web data.

Learning from Web Data. As commercial visual search engines became mature, many researchers have pushed hard in the direction of learning visual models using Web data [5, 7, 35, 11, 25]. To combat the problems of noise and data bias, [4, 5, 7] proposed semi-supervised approaches to jointly learn robust visual models and find clean exemplars, hoping the simple examples learned first could detect harder and more complex examples. In the video domain, Duan *et al.* [10] describe a system that uses a large amount of weakly labeled Web videos for visual event recognition by measuring the distance between two videos and a new transfer learning method. Chen *et al.* [36] and Duan *et al.* [9] proposed domain transfer approaches from Web images for action localization and event recognition task. Habibian *et al.* [17] obtain textual descriptions of videos from the Web and learn a multimedia embedding for few-example event recognition. Nevertheless, these approaches all require humans to annotate a few positive videos as seeds. To alleviate the tedious human burdens and achieve labor-free video concept learning, several researchers have attempted to learn video concept detectors by crawling images and videos [45, 16] after querying the event name and potential associated queries. However, the quality of the concepts is low compared with the fully-supervised approach, due to the fact that Web video search engines are a weak form of supervision, providing no spatial or temporal localization. This means that the untrimmed video contains large quantities of unrelated frames, which will confuse the classifier training. In [3, 32], the authors attempt to learn video concepts from Web images. However, the performance is still limited, due to the well-known domain gap problem [29], even though concept pruning and a domain adaptation approach [32] have been proposed to address the domain shift problem. To eliminate these concerns, we propose a novel framework to learn video concept detectors by leveraging image and video web data together.

Our work is also related to zero-shot video retrieval [12, 1, 43, 32, 14, 44]. Given a textual query, state-of-the-art event retrieval system is performed by selecting concepts

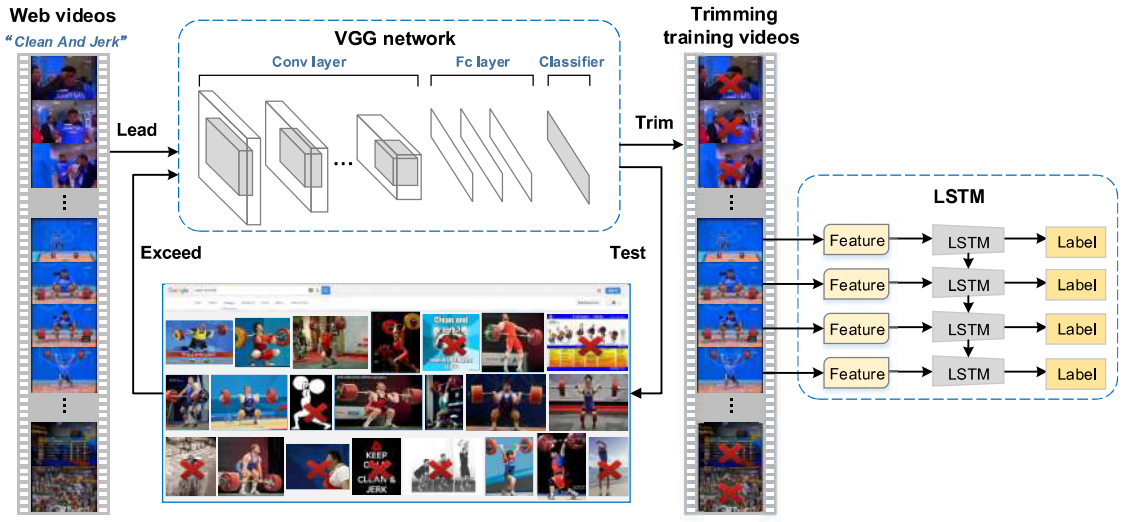


Figure 3. The overview of our approach: Web videos are firstly used to train a Lead Network by fine-tuning the VGGNET [31]. Then the Leading Network is applied to Web images to filter out noisy images due to semantic drift or domain gap. After further fine-tuning the Leading Network by adding related images, we obtain an Exceeding Network which is then used to filter out irrelevant frames. Finally, the remaining related frames are fed into a LSTM network to incorporate temporal information.

linguistically related to the query and fusing the concept responses on unseen videos. The key building block of zero-shot video retrieval is a pre-defined large vocabulary of concepts. Therefore, the output of our framework can serve as the input of zero-shot video retrieval systems.

3. Approach

In our framework, all supervision information for both images and videos are crawled from Web, and the information is gradually reinforced during the learning process. Web videos and images are presumed to be complementary to each other. Due to the heterogeneity of videos and images, even a video may have irrelevant frames given a query, it is unlikely that images which are visually similar to the irrelevant frames will be also retrieved by a search engine. For example, no images similar to the last three frames of the first video (*mopping floor*) in Figure 1 (a) will be retrieved using the query *mopping floor*. Similarly, given a query, the retrieved noisy images are very unlikely to appear in a video clip, *e.g.* the clean background images of *baby crawling* in Figure 1 (c). To leverage such complementary information from Web videos and images, videos are firstly used to train a Lead Network to model the appearances of related frames and unrelated frames. Then the Lead Network is used to filter out noisy images. By further refining the Lead Network on the remaining images, we obtain an Exceed Network, which is then used to filter out unrelated frames. After pruning videos, only related frames are fed into a LSTM network to further incorporate temporal information. The whole video concept learning framework is summarized in Figure 3, which consists of four major components: data gathering, Lead Network training using Web videos, Exceed Network training using Web images and a LSTM network to model temporal information. Each component will be detailed in following subsections.

3.1. Data Gathering

The Web is the richest source for training data gathering. In our framework, all supervised information is gathered from Web-based search engines.

For the image domain, we use category names with minor changes (*e.g. doing balance beam* for the class *balance beam*) and *photo* filter to query Google image search and download the retrieved images. The *photo* filter removes artificial images that rarely appear in videos. To comply with the query format of Google image search engine, all occurrences of *without*, *non-* and *not* are replaced with the minus sign. With this procedure, about 600 images on average are gathered for each query.

For the video domain, we download the queried videos at the best quality available from YouTube. In order to control both storage and computational cost, we limit the retrieved video to be less than 15 minutes in length. In practice, 90% of videos have a duration between 5 and 10 minutes. Around 60% of the videos are in resolution 1280× 720, while the majority have a frame rate of 30 FPS (frames per second). In the paper, we crawl about 15 videos on average for each query.

3.2. Lead Network

Web videos directly describe the visual appearance of video concepts with less domain gap. So we start by training a Lead Network by using Web videos. For this training, each video is decomposed into a set of frames. Using all video frames would be computationally expensive and is not necessary, as there is lots of redundancy between frames. Thus, we only use the key frames. To extract these, we start with detecting shot boundaries by calculating color histograms for all frames. For each frame, we then calculate the L_1 distance between the previous color histogram and the current one. If the distance is larger than a certain threshold, this frame is marked as a shot boundary. After detecting the shot, frames within a shot are similar, so we use the frame in the middle to represent the shot, defining it as the key frame. By using this algorithm, we extract around 200 key frames for a 5 minute video.

Encouraged by the state-of-art performance achieved by CNNs in several action recognition task [30], we also choose CNNs as a building block of our framework. Training a CNN starting from a random initialization is time-consuming and also requires large quantities of annotated training data, while CNNs pre-trained from ImageNet have been proven to generalize well to other vision tasks with domain-specific fine-tuning. Thus we choose a pre-trained CNN for a warm start. Specifically, we choose the VGGNET networks [31] released by Oxford to conduct experiments, which contains sixteen convolutional layers and three fully connected layers. The output of the last fully-connected layer is fed into a 1000-way softmax layer with multinomial logistic regression used to define the loss function, which is equivalent to defining a probability distribution over the 1000 classes. To fine-tune the VGGNET, we set the output number of the last fully-connected layer and the softmax layer as the number of video concepts, and initialize the network with pre-trained weights, except that the weights for the last fully-connected layer are randomly initialized.

3.3. Exceed Network

Though with less domain gap, the Lead Network trained on videos suffers from unfocused problem. To suppress the effect of unrelated frames for the Lead Network, we resort to using supervised information from the image domain. For a video concept, the related images with distinctive action scenes will be helpful to keep related frames in videos with implicit supervised information derived from the image capture process. While promising, the Web images are noisy and some exhibit semantic drift, as e.g. the example of *juggling ball*. The top returned images are all about the ball itself, not juggling.

To remove useless Web images and keep related ones, we use the Lead Network to perform filtering. The Lead Network is trained on both related frames and unrelated frames, and favors related Web images since unrelated frames rarely appear as single images, as they are not informative enough to capture.

Formally, suppose we have M crawled images from C video concepts. Each data sample is the form of (I_m, y_m), where $y_m \in \{1, 2, ..., \mathsf{C}\}$ is the category label of the m-th image. Each image I_m is fed into the Lead Network in a feed-forward pass, and yields a probability distribution $p_m \in \mathbb{R}^{\mathsf{C}}$ over the C video concepts. We use $p_m(c)$ to denote the probability of image m being in the c^{th} category. We keep images whose $p_m(y_m)$ is above a threshold η_I as related images labeled by the Lead Network. Empirically, η_I is set as 0.5 in our experiments which is good enough to filter unrelated images.

The cleaned Web images are used to further fine-tune the Lead Network and obtain the Exceed Network. The Exceed Network is more focused on video concept related appearance enhanced by related web images. The Exceed Network is further taken back to trim Web videos to keep related frames. Suppose a video V_i from video concept y_i contains a set of key frames $V_i = \{v_{i1}, v_{i2}, \cdots, v_{in_i}\}$, where n_i denotes the total number of key frames in V_i. We feed each key frame into the Exceed Network, and obtain its probability score on y_i. The key frames with scores above threshold η_V will be selected to train the temporal model. η_V is set as 0.5, the same as η_I.

Implementation details. Each key frame is resized with the shorter side to be 256 pixels which is compatible with the input requirement of VGGNET. During Leed Network training, all key frames are randomly shuffled, and organized as mini-batches with size of 128 for VGGNET fine-tuning by using stochastic gradient descend. The learning rate starts from 10^{-3} and decreases to 10^{-4} after 20K iterations, then to 10^{-5} after 40K iterations. The training is stopped after 60K iterations. During Exceed Network training, we take the selected web images inputs to further enhance the initial trained Lead Network. The learning rates starts from 10^{-3} and decreases to 10^{-4} after 30K iterations. The training will be stopped after 60K iterations.

3.4. Long Short-Term Memory Machines

Besides appearance information in each related frame, temporal information also contains discriminative signals for video concept learning. Thus, after related frames are selected for Web videos, we further utilize Long Short-term Memory (LSTM) to capture such temporal information. Long Short-term Memory (LSTM) [19] is a type of recurrent neural network (RNN) that solves the vanishing and exploding gradients problem of conventional RNN architectures when trained using back-propagation. Standard LSTM architecture includes an input layer, a recurrent LSTM layer and an output layer. The recurrent LSTM layer

has a set of memory cells, forget gates, input gates and output gates, which allow it to maintain long-term memory and reset its memory, respectively.

Denote an input sequence $\mathbf{X}$ as $\{x_1, x_1, \cdots, x_T\}$, where each x_t is a feature vector of a video frame at time t. Through the LSTM, the input sequence is mapped to an output sequence $\mathbf{Y} = \{y_1, y_2, \cdots, y_T\}$ as follows:

$$i_t = \sigma(W_{it}x_t + W_{ir}r_{t-1} + W_{ic}c_t + b_i), \quad (1)$$
$$f_t = \sigma(W_{fx}x_t + W_{rf}r_{t-1} + W_{cf}c_{t-1} + b_f), \quad (2)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot g(W_{cx}x_t + W_{cr}r_{t-1} + b_c), \quad (3)$$
$$o_t = \sigma(W_{ox}x_t + W_{or}r_{t-1} + W_{oc}c_t + b_o), \quad (4)$$
$$m_t = o_t \odot h(c_t), \quad (5)$$
$$r_t = W_{rm}m_t, \quad (6)$$
$$y_t = W_{yr}r_t + b_y. \quad (7)$$

Here $W's$ and $b's$ are the weight matrices and biases, respectively. $\odot$ denotes element-wise multiplication and c is activation of a memory cell. i, f, o are activations of the input gate, forget gate and output gate respectively. m and r are the recurrent activation before and after projection. σ is the sigmoid function. g and h are the $tanh$ function.

We take the frames selected by Exceed Network to train a LSTM network. The top layer is a softmax classifier. We use the LSTM implemented by Caffe [21], and set the rolling time k as 25 and the number of hidden state as 256. The LSTM weights are learnt by using the BPTT algorithm with a mini-batch size of 10. And the learning rate starts from 10^{-3} and decreases to 10^{-4} after 50K iterations. The training is stopped after 100K iterations.

4. Experiment

We empirically verify the merit of our video concept learning framework in two aspects: 1) how Web videos and images complement each other and 2) comparisons with state-of-the-art zero/one shot learning methods. To achieve this goal, two sets of experiments were conducted on video action recognition and event detection, respectively.

4.1. Dataset

We validate our framework on three large-scale video recognition datasets. One is for action recognition, and the other two are for video event detection.

UCF101 [33]. This is a large video dataset collected from YouTube for action recognition, which contains 101 action classes, 13K clips and 27 hours of video data. The task is considered challenging since lots of videos are captured under poor lighting, cluttered background, or severe camera motion. As our framework doesn't require a training set, we only use the three provided test-splits with around 3,800 videos each for evaluation. Performance is measured in terms of classification accuracy.

TRECVID MED 2013[1] and 2014 dataset[2]. These are two largest publicly available video corpora in the literature for video event detection. They have been introduced by NIST for all participants in the TRECVID competition and research community to conduct experiments. MEDTest 13 contains 20 events E006 – E015 and E021 – E030, while MEDTest 14 has 20 events E021 – E040, where E021 – E030 are shared by both datasets. Each dataset contains three different partitions, i.e., *Background*, *100EX* and *MEDTest*. *Background* contains about 5000 background videos not belonging to any of the target events; *100EX* contains 100 positive videos for each event, are used as the training set; *MEDTest* contains around 25,000 videos (over 960 hours of videos), with per-video ground truth annotations for 20 event categories. Since we focus on utilizing Web data to train event detectors, we just use the 5000 videos in the *Background* set (not using any positive videos from *100EX*) during training. To evaluate the performance, we apply the official metric: average precision (AP) per event, and mean Average Precision (mAP) by averaging AP on all events.

Implementation details. For testing on UCF101 dataset, we uniformly sample 25 frames per video on the testing videos and then utilize a spatial network or a LSTM network to do predictions. To arrive at a video-level classification score, we rely on simply late fusion. Testing the spatial model is achieved by averaging the classification score on key frames. For the testing on TRECVID MED dataset using LSTM model, we produce 25 key frame long clips with a 12-frame overlap between two consecutive clips and the classification score of a video is the average of the scores of all clips. Similarly, the average of scores predicted on all key frames by spatial model is taken as the classification result of a video.

4.2. Experiment Result on Action Recognition

We first validate the performance of our models that capture the appearance information, then examining whether the better appearance information could improve the temporal model and the final concept detection results.

Comparison with baselines. To the best of our knowledge, this is the first attempt to use Web data to conduct action recognition on UCF101 dataset. To demonstrate the effectiveness of our proposed framework, we compare our framework against other baseline systems:

- **Image:** Directly using Web images to fine-tune the VGGNET.
- **Video:** Directly using Web video key frames to fine-tune the VGGNET.
- **Image + Video:** Using Web images to fine-tune the VGGNet first, then using the fine-tuned model to select

[1] http://nist.gov/itl/iad/mig/med13.cfm
[2] http://nist.gov/itl/iad/mig/med14.cfm

Method	Acc (%)
Image	62.4
Video	58.5
Image + Video	63.2
Noise Mixing	64.6
Late fusion	67.8
Mixing	68.9
Lead-Exceed (Ours)	**74.4**
Lead-Exceed + LSTM (Ours)	**76.3**

Table 1. Comparisons with other approaches on UCF101 dataset.

Method	Acc (%)
Image	58.5
Video	53.4
Image + Video	59.5
Mixing	61.4
Lead-Exceed (Ours)	**65.7**

Table 2. Comparison LSTMs performance on UCF101 dataset when using different appearance models to select relevant frames.

key frames from videos for further fine-tuning.

- **Noise Mixing**; Directly mixing the Web image and video key frames together to fine-tune the VGGNET.
- **Mixing:** Mixing the selected Web image and video key frames in our framework together to fine-tune the VGGNET.
- **Late Fusion:** Using the selected Web images and videos in our framwork separately to fine-tune two VGGNETs and then average their scores as final prediction.

The comparison results are shown in Table 1. To further examine whether the improved appearance model could yield a better temporal model, we use the above models to select 25 key frame for each Web video and put them into an LSTM classifier to train a temporal model. The result of LSTM classifiers and are shown in Table 2.

Result Analysis. From Table 1, we can observe three key findings: 1) Performance can be significantly improved by taking advantage of both Web images and videos. Particularly, comparing with using Web images only and Web videos only, our Lead-Exceed Network can improve the relative performance by 20% and 27%, respectively, which validates the direction of jointly using Web images and videos for video concept learning; 2) Our proposed Lead-Exceed Network performs significantly better than the other two baselines (noise mixing, mixing and late fusion) that use both images and videos, which validate that our method is effective in learning discriminative information by taking full advantage of both images and videos; 3) However, Image + Video performs worst in methods using both videos and images. This is not surprising, since images for a video concept may have semantic drift that will lead the fine-tuning to the wrong direction, and in turn, the fine-tuned model has a hard time selecting the right frames, thus even hurting performance on these categories. From Table 2, we observe that better appearance models can also help better trimming for unconstrained Web videos, achieving highest performance compared with other trimming approaches when using a LSTM to model temporal information. When comparing Table 1 with Table 2, we find that the LSTM model is not as good as the appearance model. We speculate the drop may be cased by the fact that there are only 15 video samples in each class for training. We believe that adding more video data into the training set would further improve the results.

Method	mAP (%)
Concept Discovery [3]	2.3
Bi-concept [16]	6.0
Composite Concept [16]	6.4
EventNet [45]	8.9
Selecting [32]	11.8
Lead-Exceed (Ours)	**16.3**
Lead-Exceed + LSTM (Ours)	**16.7**

Table 3. Comparisons with other state-of-the-art zero-shot event detection systems on MEDtest13.

Method	MEDtest13	MEDtest14
IDTFV	12.4	8.9
VGG	13.8	11.8
Lead-Exceed (Ours)	**16.3**	**14.7**
Lead-Exceed + LSTM (Ours)	**16.7**	**15.8**

Table 4. Comparisons with other stat-of-the-art few-shot event detection approaches.

4.3. Experiment Results on Event Detection

Comparison with Previous Zero-shot Approach. In order to have a better understanding of our approach, we also apply our framework on the large-scale TRECVID MED 2013 and 2014 datasets. We first compare our approach with recent state-of-the-art zero-shot systems that also use Web data to learn event detectors, including (1) Concept Discovery [3], (2) Bi-Concept [16], (3) Composite Concepts [16], (4) EventNet [45], and (5) Selected Concepts [32]. Approach (1) directly uses Web images to train event detectors, while approaches (2) – (4) directly use Web videos to train event detectors, and approach (5) first uses Web images to pre-train a concept detector, and then uses top returned testing videos to re-train an event detector. For a fair comparison, we report our results on MEDtest13 and directly compare with state-of-the-art results quoted from original papers. The results in Table 3 show that our framework beats other zero-shot systems by a large margin. We observe that our proposed algorithm significantly outper-

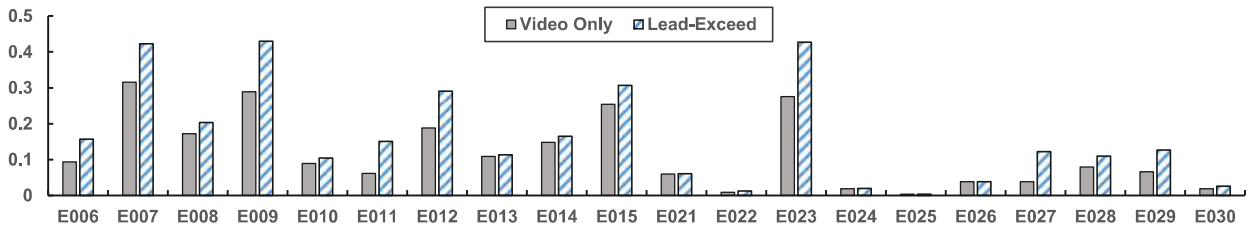


Figure 4. Per-event detection result compared with using videos only on MEDTest 13 dataset.

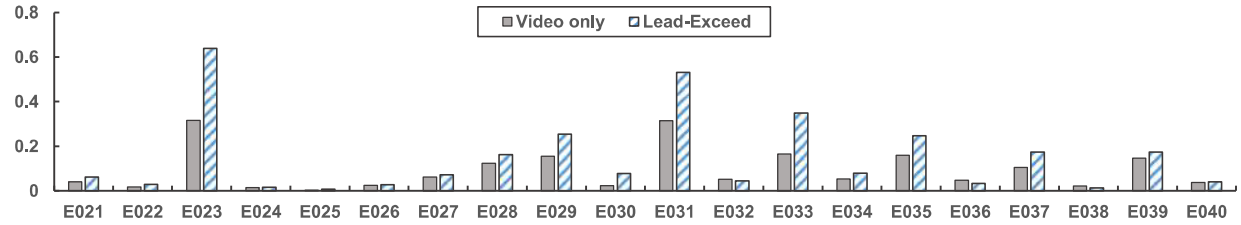


Figure 5. Per-event detection result compared with using videos only on MEDTest 14 dataset.

forms the previous approaches. For additional analysis, we also provide event class-specific results in Figure 4 and 5. For these two we report the number that using video only and our Lead-Exceed Network. We observe that for 19 out of 20 classes in MEDtest13 and 17 of 20 classes in MED14 dataset, confirming that our proposed Lead-Exceed Network can better leverage the complementary strengths of images and videos. The failure cases are due to the Lead Network failing to model an event (with the average precision below 0.05). When the Lead Network is seriously bad, our Exceed Network is unable to enhance the results.

Comparison with state-of-the-art Few-shot Approaches. We also compare our approach with state-of the art approaches using 5 positive exemplars by using the best hand-crafted features: Improved Dense Trajectory with Fisher Vector and best high-level VGG CNN features. Trajectory features have proven to be the most reliable hand-crafted features for action recognition and event recognition, consisting of five different descriptors (trajectories, HOG, HOF, MBHX and MBHY) to capture the shape and temporal motion information of videos. We adopt the improved trajectories proposed by [39] to extract local features for each video in the UCF101 dataset. We use the default parameters, which results in 426 dimensions in total. Then the PCA operations are performed separately on each of the 5 descriptor types to keep half of the dimensions. After PCA, the local features reduce to 213 dimensions. Finally, each video is encoded in a Fisher Vector [28] based on a GMM of 256 Gaussians, producing a 109,056-dimensional vector. For VGG CNN features, we take the key frames of videos as input to forward pass the VGGNET and extract the fc6 activation. To arrive at video-level representations, we rely on simply average pooling. To train the event detector, we use LIBSVM [2], with fixed parameter $C = 1$, as recommended in [46]. In the results shown in Table 4, our webly supervised approach remarkably can achieve better results than when 5 human-annotated exemples are fed into traditional supervised learning approaches.

5. Conclusion

In this paper, we present a simple but effective labor-free video concept learning framework by jointly utilizing noisy Web videos and images. Our approach can leverage the complementary nature of the two media, by drawing on the novel idea of a Lead–Exceed Neural Network (LENN). Experimental results on three large video recognition datasets confirm that our framework can learn high-quality video concept detectors without annotating any positive exemplars. We believe this paper opens up avenues for exploitation of Web data to achieve next cycle performance gains in the video learning task.

Acknowledgement. This work was supported in part by the National Basic Research Program of China Grant 2011CBA00300, 2011CBA00301, the National Natural Science Foundation of China Grant 61033001, 61361136003, partially supported by the Data to Decisions Cooperative Research Centre (www.d2dcrc.com), and partially supported by the ARC DECRA and DP.

References

[1] M. ain, J. C. van Gemert, T. Mensink, and C. G. Snoek. Objects2action: Classifying and localizing actions without any video example. *ICCV*, 2015. 3

[2] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2011. 8

[3] J. Chen, Y. Cui, G. Ye, D. Liu, and S. Chang. Event-driven semantic concept discovery by exploiting weakly tagged internet images. In *ICMR*, 2014. 3, 7

[4] X. Chen and A. Gupta. Webly supervised learning of convolutional networks. *ICCV*, 2015. 2, 3

[5] X. Chen, A. Shrivastava, and A. Gupta. NEIL: Extracting visual knowledge from web data. In *ICCV*, pages 1409–1416, 2013. 3

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 2

[7] S. K. Divvala, A. Farhadi, and C. Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *CVPR*, pages 3270–3277, 2014. 3

[8] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. *CVPR*, 2015. 1, 3

[9] L. Duan, D. Xu, and S.-F. Chang. Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach. In *CVPR*, pages 1338–1345, 2012. 3

[10] L. Duan, D. Xu, I.-H. Tsang, and J. Luo. Visual event recognition in videos by learning from web data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1667–1680, 2012. 3

[11] J. Fu, Y. Wu, T. Mei, J. Wang, H. Lu, and Y. Rui. Relaxing from vocabulary: Robust weakly-supervised deep learning for vocabulary-free image tagging. In *ICCV*, pages 1985–1993, 2015. 3

[12] C. Gan, M. Lin, Y. Yang, Y. Zhuang, and A. G. Hauptmann. Exploring semantic inter-class relationships (SIR) for zero-shot action recognition. In *AAAI*, 2015. 3

[13] C. Gan, N. Wang, Y. Yang, D.-Y. Yeung, and A. G. Hauptmann. DevNet: A deep event network for multimedia event detection and evidence recounting. In *CVPR*, pages 2568–2577, 2015. 2

[14] C. Gan, Y. Yang, L. Zhu, D. Zhao, and Y. Zhuang. Recognizing an action using its name: A knowledge-based approach. *International Journal of Computer Vision*, pages 1–17, 2016. 3

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CVPR*, pages 580–587, 2014. 2

[16] A. Habibian, T. Mensink, and C. G. Snoek. Composite concept discovery for zero-shot video event detection. In *ICMR*, page 17, 2014. 3, 7

[17] A. Habibian, T. Mensink, and C. G. Snoek. Videostory: A new multimedia embedding for few-example recognition and translation of events. In *ACM Multimedia*, pages 17–26, 2014. 3

[18] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles. Activitynet: A large-scale video benchmark for human activity understanding. 1

[19] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 5

[20] W. Huang, D. Zhao, F. Sun, H. Liu, and E. Chang. Scalable gaussian process regression using deep neural networks. In *IJCAI*, pages 3576–3582, 2015. 3

[21] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM Multimedia*, volume 2, page 4, 2014. 6

[22] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 1, 3

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 3

[24] Z. Lan, M. Lin, X. Li, A. G. Hauptmann, and B. Raj. Beyond gaussian pyramid: Multi-skip feature stacking for action recognition. *CVPR*, 2015. 3

[25] X. Liang, S. Liu, Y. Wei, L. Liu, L. Lin, and S. Yan. Towards computational baby learning: A weakly-supervised approach for object detection. In *ICCV*, pages 999–1007, 2015. 3

[26] T. Mei, Y. Rui, S. Li, and Q. Tian. Multimedia search reranking: A literature survey. *ACM Computing Surveys (CSUR)*, 46(3):38, 2014. 3

[27] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. *CVPR*, 2015. 3

[28] D. Oneata, J. Verbeek, C. Schmid, et al. Action and event recognition with fisher vectors on a compact feature set. In *ICCV*, 2013. 3, 8

[29] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, pages 213–226. 2010. 2, 3

[30] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014. 1, 3, 5

[31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015. 2, 3, 4, 5

[32] B. Singh, X. Han, Z. Wu, V. I. Morariu, and L. S. Davis. Selecting relevant web trained concepts for automated event retrieval. *ICCV*, 2015. 3, 7

[33] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 6

[34] N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised learning of video representations using lstms. *ICML*, 2015. 3

[35] C. Sun, C. Gan, and R. Nevatia. Automatic concept discovery from parallel text and visual corpora. In *ICCV*, pages 2596–2604, 2015. 3

[36] C. Sun, S. Shetty, R. Sukthankar, and R. Nevatia. Temporal localization of fine-grained actions in videos by domain transfer from web images. *ACM Multimedia*. 3

[37] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *CVPR*, 2015. 3

[38] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. C3D: Generic features for video analysis. *ICCV*, 2015. 1, 3

[39] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, 2013. 1, 3, 8

[40] L. Wang, Y. Qiao, and X. Tang. Motionlets: Mid-level 3d parts for human motion recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2674–2681, 2013. 1

[41] L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. *CVPR*, 2015. 3

[42] L. Wang, Y. Qiao, and X. Tang. Mofap: A multi-level representation for action recognition. *International Journal of Computer Vision*, pages 1–18, 2015. 1

[43] S. Wu, S. Bondugula, F. Luisier, X. Zhuang, and P. Natarajan. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In *CVPR*, pages 2665–2672, 2014. 3

[44] T. Yao, T. Mei, C.-W. Ngo, and S. Li. Annotation for free: Video tagging by mining user search behavior. In *ACM Multimedia*, pages 977–986, 2013. 3

[45] G. Ye, Y. Li, H. Xu, D. Liu, and S.-F. Chang. Eventnet: A large scale structured concept library for complex event detection in video. In *ACM Multimedia*, pages 471–480, 2015. 3, 7

[46] S.-I. Yu, L. Jiang, Z. Mao, X. Chang, X. Du, C. Gan, Z. Lan, Z. Xu, X. Li, Y. Cai, et al. Informedia@ TRECVID 2014 MED and MER. 8

[47] Z.-J. Zha, T. Mei, Z. Wang, and X.-S. Hua. Building a comprehensive ontology to refine video concept detection. In *multimedia information retrieval*, pages 227–236, 2007. 3