# Further Understanding Videos through Adverbs: A New Video Task

**Bo Pang, Kaiwen Zha, Yifan Zhang, Cewu Lu***
Shanghai Jiao Tong University
{pangbo, kevin_zha, zhangyf_sjtu, lucewu}@sjtu.edu.cn

## Abstract

Video understanding is a research hotspot of computer vision and significant progress has been made on video action recognition recently. However, the semantics information contained in actions is not rich enough to build powerful video understanding models. This paper first introduces a new video semantics: the Behavior Adverb (BA), which is a more expressive and difficult one covering subtle and inherent characteristics of human action behavior. To exhaustively decode this semantics, we construct the Videos with Action and Adverb Dataset (VAAD), which is a large-scale dataset with a semantically complete set of BAs. The dataset will be released to the public with this paper. We benchmark several representative video understanding methods (originally for action recognition) on BA and action recognition. The results show that BA recognition task is more challenging than conventional action recognition. Accordingly, we propose the BA Understanding Network (BAUN) to solve this problem and the experiments reveal that our BAUN is more suitable for BA recognition (11% better than I3D). Furthermore, we find these two semantics (action and BA) can propel each other forward to better performance: promoting action recognition results by 3.4% averagely on three standard action recognition datasets (UCF-101, HMDB-51, Kinetics).

## Introduction

Recognizing semantic labels in visual data is an essential computer vision task. It is worth noting that these semantic labels are part of our language system – we can use language to describe these semantics. For example, object detection/recognition (Ren et al. 2015; Redmon and Farhadi 2016) can be considered as exploiting "noun" in visual data. To understand "verb", action recognition (Donahue et al. 2015; Wu et al. 2015; Srivastava, Mansimov, and Salakhudinov 2015; Yue-Hei Ng et al. 2015; Karpathy et al. 2014; Ji et al. 2013) has been extensively studied. Moreover, "adjective" labels (e.g., cool, dark, beautiful) can be explored through attribute learning (Petrosino and Gold 2010). But till now, few works have focused on an important kind of words — "**Adverb**", which can properly express the attributes of action, as well as the attitude and mood of the subject. From the viewpoint of language research (White 1991), these concepts convey more sensitive semantics compared with nouns and verbs, making them a suitable choice for developing video understanding models! Motivated by these, we propose a new task: BA recognition (Fig. 1).


Figure 1: BA examples. As a rich video semantics, BA can describe the mood, attitude and action attributes. Note that each BA presents diverse patterns on different actions.

Note that the BA recognition task is different from fine-grained actions task. Compared to fine-grained action categories, BA categories own different feature space. For example, Latin dance can be finely categorized into Rumba, Chacha and so on, which requires the model to extract subtle dance features, but still action-related features. But for BA, the model needs to distinguish "professionally" and "amateurishly", "happily" and "sadly", which requires the model to understand the video meticulously and capture relatively action-irrelevant features since the same BA may present diverse movement patterns in different actions.

To dissect this new video semantics, we build a new benchmark: Videos with Action and Adverb Dataset (VAAD), containing 150k video clips, covering 2K action-adverb pairs (e.g., "smoking sadly") over 200 action categories described by 51 adverb categories, where three types of BA that can respectively describe the subject's mood, attitude and action attributes (e.g., "quietly" and "easily") are included. We highlight three features of VAAD. Firstly,

*Corresponding author.

there is an average of 11 distinct adverbs modifying each action (no dull actions that can only be labeled by few adverbs). Secondly, our adverb categories are based on semantics rather than words, for example, we do not take "smoke sadly" and "smoke sorrowfully" as different categories. Thirdly, the dataset is multi-labeled, where an action can be labeled with multiple BAs expressing mood, attitude and action attributes simultaneously.

As the first attempt, we propose the BA Understanding Network (BAUN) for this task. Firstly, the mood related BAs are modeled by extracting facial expression features of the target person. Secondly, we adopt a Spatial-Temporal Separated Module (STSM) which extracts the temporal and spatial features separately to understand the deeper-level temporal-spatial information. Thirdly, a group of BAs is hard to recognize such as "inexorably" and "frightenedly", which often needs to be identified in conjunction with other objects in context. To deal with this, we propose the Image Relation Network (IRN) to explore the relations in the video. Finally, we utilize the conditional distribution between action and BA as prior knowledge to further improve the performance of the multi-task recognition (on action and BA).

We conduct comprehensive experiments to show the challenge of BA recognition and evaluate our BAUN. Results reveal that: 1) BA recognition is a challenging task for current video understanding models. 2) BAUN enjoys accuracy gains from the elaborate structure, substantially better than the 3D CNN model. 3) The two semantics, BA and action, can propel each other forward to better performance.

We evaluate our BAUN on both BA and action recognition tasks. For BA recognition, it achieves 11% relative improvements on VAAD over I3D model (Carreira and Zisserman 2017). For action recognition, accuracy increases by 2.5% averagely on VAAD, UCF-101 (Soomro, Zamir, and Shah 2012), HMDB-51 (Kuehne, Jhuang, and others ) and Kinetics (Kay et al. 2017). Furthermore, we transfer the models pre-trained on VAAD with BA semantics to UCF-101, HMDB-51 and Kinetics, and performances improve on action recognition task as expected.

## Related Work

**Action Datasets and Models** Action recognition has made great progress in the past few years due to many excellent datasets, from small simple datasets like KTH (Laptev 2005) and UCF-101 to large-scale, real-world datasets such as YouTube-8M (Abu-El-Haija and others 2016), Sport-1M (Karpathy et al. 2014) and Kinetics (Kay et al. 2017). Some works step to temporal localization. ActivityNet (Caba Heilbron et al. 2015), AVA (Gu, Sun, and others 2017) and Charades (Sigurdsson et al. 2016) contain large numbers of videos with action labels attached to time (AVA also involves spatial localization).

For action recognition tasks, some works extract features from video frames separately and then fuse them together, where RNN is widely used (Donahue et al. 2015; Wu et al. 2015; Srivastava, Mansimov, and Salakhudinov 2015; Pang et al. 2019) and many pooling methods have been developed (Yue-Hei Ng et al. 2015; Karpathy et al. 2014). Another idea is to adopt methods similar to 2D image tasks, including C3D (Ji et al. 2013), I3D (Carreira and Zisserman 2017) and SlowFast Network (Feichtenhofer et al. 2018), which inflate 2D CNNs into 3D, endowing them with an additional temporal dimension equally. Besides, other related works employ methods like optical flow (Simonyan and Zisserman 2014), trajectories (Wang et al. 2011), and pose estimation (Maji, Bourdev, and Malik 2011) to deal with temporal information.

**Video Captioning and Human Expression** BA and VAAD are related to video captioning and expression tasks since video captioning entails predicting adverbs and many common emotional adverbs are included in expression tasks. For video captioning, there are many datasets designed for it, such as MSR-VTT (Xu et al. 2016), YouTube2Text (Guadarrama et al. 2013), and ActivityNet Captions (Krishna et al. 2017). For expression tasks, many models like (Fan et al. 2016) and (Ng et al. 2015) are built on the benchmarks: Emotiw2016 (Dhall et al. 2016), MMI (Valstar and Pantic 2010) and HUMAINE (Douglas-Cowie et al. 2007). The expression recognition mainly focuses on faces, while BA recognition requires recognizing subject's mood and attitude from body language, not merely from facial expressions.

**Relational Reasoning Structures** Relational reasoning is a fundamental property of intelligent species. Relation Network (Santoro et al. 2017) addresses it by computing a function on the feature embedding at all pairs of positions in its input. In (Zhou et al. 2018), the authors focus on exploring the temporal dependencies among video frames. Moreover, relation between objects is also adopted to enhance the object detection performance (Hu et al. 2018). We adopt the relation model to explore the interactions among the target people which are crucial for BA recognition. Unlike the above works, we do not treat it as a bag model, but instead, we add positional information into the representation.

## Videos with Action and Adverb Dataset

**Data Collection** The data is collected from the existing datasets, including the Kinetics, UCF101, and HMDB51.

**Defining the BA-Action Lexicon** We fix the action list first, then make the BA list according to it, and finally group them into BA-action pairs. We adopt language prior knowledge (0.2 million video descriptions) to choose the actions which are frequently accompanied by adverbs, so that we can get rid of "dull" actions. To get a relatively complete list of adverbs, we exploit the Corpus of Contemporary American English, an authoritative corpus of American English (Davies 2008) which provides the frequencies of the commonly used words. Based on the word frequencies, we choose 113 most frequently used action adverbs. After removing the synonyms, there are 51 adverbs left. We build BA-action pairs based on the their frequencies provided by the N-gram data [1]. Every action has about 11 appropriate adverbs and we have 2K adverb-action pairs in total.

**Video Annotating Process** Annotation is implemented in human instance level, because some actions like "kiss" and "hug" involve more than one player, and we need to annotate them respectively. We propose a semi-automatic annotation

[1] https://www.ngrams.info/

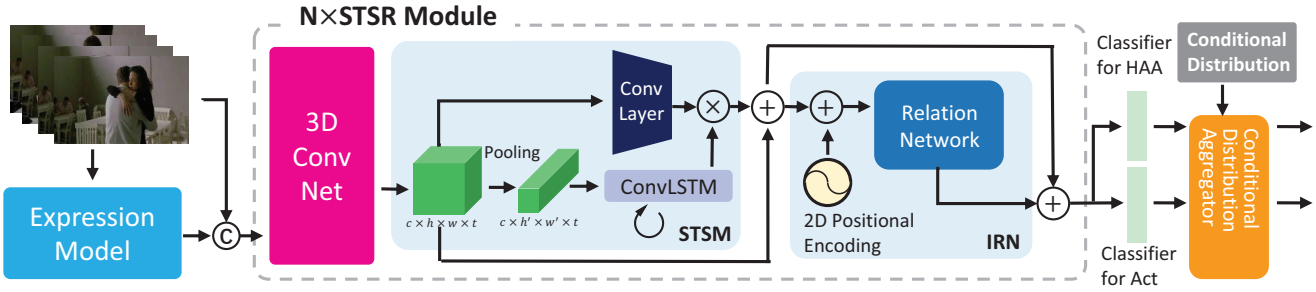


Figure 2: Pipeline of our BAUN. 3D Convolution Network takes the RGB/Flow images concatenated with expression features as inputs and takes charge of extracting short-term temporal-spatial features. Spatial-Temporal Separated Module (STSM) is adopted to split spatial and temporal features and endow the network with a larger temporal receptive field. Image Relation Network (IRN) consisting of a RN and 2D positional encoding has a large spatial receptive field and can effectively exploit the context relations. We stack 3 STSR Modules (N=3) to build the recognition model and the conditional distribution between action and BA is utilized as constraints in the final classification.

framework to effectively localize human instance, where the human bounding box is only labeled at the first frame, and we utilize object tracking model MDNet (Nam and Han 2015), the winner of the VOT-2015 challenge (Kristan et al. 2015) for the following frames. To improve the robustness of annotation, we adopt the Faster-RCNN (Girshick 2015) based human detector to adjust tracking bounding box by averaging the two results. During annotating, the annotator is only required to monitor the tracking results and correct them, which greatly reduces the workload of labeling.

BAs are used to depict concepts that are more subjective than objects and actions. Annotating by only one person may cause ambiguity. Therefore, each video is assigned to three different annotators to annotate the BA. We make sure that those three annotators should possess diverse backgrounds (e.g., nationality, age, gender, education background). After annotation, in order to guarantee the labels consistency, we compare the annotations from the three annotators and relabel the videos suffering large label distance $d$ (>1.6), which is defined based on Manhattan distance:

$$d_v = avg(\sum_{i<j} |a_{v,i} - a_{v,j}|), \quad (1)$$

where $a_{v,i}$ denotes the $i$-th annotation of the video $v$.

**Dataset Statistics and Discussion** In total, there are 51 BAs, 200 actions, 2K BA-action pairs, and 150K videos in VAAD. A target person may be annotated with more than one BA and the average number of BAs per person is 1.81. There are more samples of common adverbs (such as "happily" or "slowly") than specific ones (such as "calmly"). "But this is how it should be! Recognition models need to operate on realistic "long tailed" action distributions rather than being scaffolded using artificially balanced datasets." (Gu, Sun, and others 2017)

## BA Semantics and BAUN

In this section, we will analyze the intrinsic difficulties of BA semantics, introduce our solution: BA Understanding Network (BAUN) and conduct in-depth discussions on why its structures work properly for BA semantics. The BAUN consists of five important modules: Expression Model, 3D Convolutional Network, STSM, IRN and Conditional Distribution Aggregator as illustrated in Fig. 2. To strengthen the feature extraction capability, we integrate the 3D Convolutional Network, STSM and IRN into the Spatial-Temporal Separated Relation Module (STSR Module) and stack 3 of them in the pipeline (N=3).

### Mood Related BAs & Expression Model

Different from action recognition, BA is required to understand the mood of target person, such as "sadly" and "happily', which can be conveyed efficiently through facial expressions. Therefore, we adopt the expression model (Fan et al. 2016) to facilitate indicating the target's mood.

We concatenate the expression features $\mathbf{f}_e \in \mathbb{R}^{C_1 \times T \times H \times W}$ extracted by the expression model with the original RGB or optical flow images $\mathbf{x} \in \mathbb{R}^{C_2 \times T \times H \times W}$ to generate the concatenated input $\mathbf{x}_c \in \mathbb{R}^{(C_1+C_2) \times T \times H \times W}$:

$$\mathbf{f}_e = \mathrm{EprModel}(\mathbf{x}) \quad (2)$$

$$\mathbf{x}_c = \mathrm{Concat}(\mathbf{x}, \mathrm{Deconv}(\mathbf{f}_e)) \quad (3)$$

where $C, T, H, W$ are the channel number, time stamps, the height and width of the feature map, EprModel is the expression model and Concat is the concatenation operation. Note that $\mathbf{f}_e$ is deconvolved to the same size as $\mathbf{x}$.

### 3D Convolutional Network

3D convolutional networks (3DCN) are widely used on video-related tasks (Carreira and Zisserman 2017; Feichtenhofer et al. 2018; Ji et al. 2013). We adopt it as the basics of our model to generate the temporal-spatial features $\mathbf{f}_b = \mathrm{3DCN}(\mathbf{x}_c)$ for it can efficiently process the information in a local manner by its convolutional kernels.

We adopt the Inception (Szegedy et al. 2015) module and similar with the I3D (Carreira and Zisserman 2017), its specific settings in each STSR Module are listed in table 1.

### Deeper Spatial-Temporal Features & STSM

Compared with action, BA encodes deeper spatial and temporal information. Almost all the actions can be identified

Table 1: Structure of the 3DCN in BAUN.

| Module | 3DCN Settings |
|---|---|
| STSR Module 1 | 3D Conv Layer<br>MaxPooling<br>3 × 3D Conv Layer |
| STSR Module 2 | 2 × Inception Module |
| STSR Module 3 | 5 × Inception Module |

by the shape changes of objects, which can be further summarized into a displacement pattern recognition problem. For example, to distinguish "walk" and "run", the model pays attention to the displacement of the legs and arms, and obviously, "run" brings drastic displacements. Whereas for BA, only recognizing displacement is less effective. For example, when punching a sandbag, "heavily" and "lightly" punching share nearly the same displacement pattern. Here, BA models need to pay more attention to the pattern of "speed" and "strength". Compared with "displacement", they are deeper concepts because they are the first and second order derivative of "displacement" with regard to time.

**Our Solution** Simultaneously improving the spatial and temporal (S&T) processing ability leads to higher computing overhead, and from another perspective, the parameters in kernel are hard to be co-adaptive to both S&T features (Pang et al. 2019). Therefore, we go down to consider processing the S&T information separately and propose the Spatial-Temporal Separated Module (STSM). There are two flows in STSM, one temporal flow for extracting long-term temporal features with less spatial interventions, one spatial flow for maintaining spatial features. In this way, the temporal direction possesses longer receptive fields and it can reduce the learning burdens of the spatial direction as well.

With the basic features $\mathbf{f}_b \in \mathbb{R}^{C\times T\times H\times W}$ generated by the 3D Convolutional Network, the temporal flow first extracts the temporal features with less spatial interventions:

$$\bar{\mathbf{f}}_b^{(t,h,w)} = \mathbf{f}_b^{(t,h,w)}\mathbf{w}_1 + \mathbf{b}_1 \tag{4}$$

$$\mathbf{f}_t = \mathrm{AveragePooling}(\bar{\mathbf{f}}_b) \tag{5}$$

where $\bar{\mathbf{f}}_b$ is the linear element-wise transformation of the original features $\mathbf{f}_b$ and $\bar{\mathbf{f}}_b^{(t,h,w)}$ denotes the feature vector of the pixel at position $(t, h, w)$ in the feature maps. In practice, Eq. 4 can be implemented as convolutions with kernel size (1,1,1). $\mathbf{f}_t$ is the temporal features generated by reducing the spatial resolution with an average pooling operation.

With the temporal features, we generate the long-term temporal features on them, which can be expressed as:

$$\mathbf{f}_{lt} = \mathrm{TempNet}(\mathbf{f}_t) \tag{6}$$

where $\mathrm{TempNet}$ is the network that combines the temporal features among all the time stamps, and we utilize the bi-directional ConvLSTM (Xingjian et al. 2015) to implement it. The bi-directional setting allows the model to take both the previous and subsequent time stamps into account. With the pre-long-term features $\mathbf{f}_{lt}$, we adopt a sigmoid function to obtain the final long-term features $\bar{\mathbf{f}}_{lt} = \mathrm{Sigmoid}(\mathbf{f}_{lt})$.

The spatial flow is a shallow 3D convolution network and we merge the spatial features generated by it with the above long-term temporal features to produce a combined one $\mathbf{f}_c$:

$$\mathbf{f}_c = \mathrm{Conv}(\mathbf{f}_b) \times \mathrm{Deconv}(\bar{\mathbf{f}}_{lt}) + \mathbf{f}_b \tag{7}$$

where $\times$ denotes element-wise production and $\bar{\mathbf{f}}_{lt}$ is deconvolved to the same size as $\mathbf{f}_b$. With the sigmoid function, the temporal flow can be regarded as a control gate. Here, "$+\mathbf{f}_b$" denotes a residual connection.

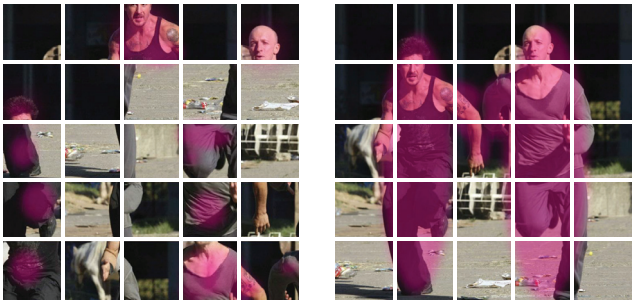

Figure 3: Effects of 2D positional encoding. **Left:** The sample input to IRN without 2D positional encoding, whose location information is lost. **Right:** The raw sample image with correct location distribution, which can be recovered with 2D positional encoding on the left image. Here IRN is designed to explore the relation between the two people as the red masks show in the right image, and we adopt 2D positional encoding to maintain the relative position cues.

## Context BAs & Image Relation Network

Some BAs need to be identified in conjunction with other person in context, and this is not necessary in conventional action recognition tasks. For example, to understand a target is running "in fear", the model needs to notice that there is someone chasing after him/her, and to understand someone is behaving "pitilessly", the model needs to detect that he/she is ignoring someone's begging. This can be seen as a visual reasoning task which cannot be successfully solved as a conventional pattern recognition task and this kind of reasoning is widely required in psychologically relevant and relatively subjective BAs. To deal with this, we adopt a simple but powerful reasoning model (Santoro et al. 2017), which achieves the state-of-the-art on CLEVR (Johnson et al. 2017), combine it with a 2D positional encoding method and propose the Image Relation Network (IRN).

**Network Structure** The IRN works in the image level. With the video level feature $\mathbf{f}_c$ from STSM, the IRN treats it as a group of image level features $[\mathbf{f}_s^{(0)}, \mathbf{f}_s^{(1)}, ..., \mathbf{f}_s^{(T)}]$. Every pixel in the image level feature maps is treated as an elementary unit and the relation is calculated among these pixels. For a image level feature map $\mathbf{f}_s \in \mathbb{R}^{C\times H\times W}$, the network flattens it into $\mathbf{f}_s \in \mathbb{R}^{C\times(H*W)}$. Following (Santoro et al. 2017), the relation features $\mathbf{f}_r$ can be calculated as:

$$\mathbf{e}^{(i)} = \mathrm{ReLU}(\mathbf{f}_s^{(i)}\mathbf{w}_e + \mathbf{b}_e) \tag{8}$$

$$\mathbf{r}^{(i,j)} = g_\theta([\mathbf{e}^{(i)}, \mathbf{e}^{(j)}]) \tag{9}$$

$$\mathbf{f}_r = f_\phi(\sum_{i<j} \mathbf{r}^{(i,j)}) \tag{10}$$

where $\mathbf{f}_s^{(i)}$ is the feature vector of pixel $i$ in the image level feature maps, $\mathbf{e}$ is the elements needed by the relation network, $\mathbf{r}^{(i,j)}$ is the relation vector of element $\mathbf{e}^{(i)}$ and $\mathbf{e}^{(j)}$, and $\mathbf{f}_r$ summarizes the relations, which has a receptive field as large as the feature map. We implement $g_\theta$ and $f_\phi$ as MLPs with ReLU activation function. The final video-level features after IRN can be described as:

$$\mathbf{f}_f = [\mathbf{f}_r^{(0)}, \mathbf{f}_r^{(1)}, ..., \mathbf{f}_r^{(T)}] + \mathbf{f}_c \tag{11}$$

**2D Positional Encoding** The above Image Relation Network is a bag model and the location information of each pixel is easily lost, as shown in Fig. 3. In order to better capture location information, similar with the 1D positional encoding of Transformer (Vaswani et al. 2017), we propose a 2D positional encoding method to keep pixel's location information by adding the encoding with the feature maps: $\mathbf{f}_s \leftarrow \mathbf{f}_s + \mathcal{E}$, and the encoding $\mathcal{E} \in \mathbb{R}^{H_F \times W_F \times C_E}$ follows:

$$\zeta(p,c) = \begin{cases} \frac{p}{224^{c/C_E}}, c = 2k \\ \frac{p}{224^{(c-1)/C_E}}, c = 2k+1 \end{cases} \tag{12}$$

$$\mathcal{E}((p_x,p_y),c) = \begin{cases} \sin(\zeta(p_x,c)) + \cos(\zeta(p_y,c)), c = 2k \\ \cos(\zeta(p_x,c)) - \sin(\zeta(p_y,c)), c = 2k+1 \end{cases} \tag{13}$$

where $(p_x, p_y)$ is the position of the pixel in the feature map, $c$ is the index of the channel, $k$ is an integer constant, $C_E$ is the total number of the encoding channels and $H_F, W_H$ are the height and width of the input feature maps. The wavelengths of the encoding channels range from $2\pi$ to $224 \cdot 2\pi$. This 2D encoding method utilizes the same number of channels as the original 1D method but can explicitly encode the 2D position without confusion. Specifically, it would allow the model to easily learn to attend by relative 2D positions, since for any $(m, n)$, $\mathcal{E}((p_{x+m}, p_{y+n}))$ can be represented as a linear function of $\mathcal{E}((p_x, p_y))$ or $\mathcal{E}((p_x, p_{y+1}))$.

## Conditional Distribution of Action and BA

In theory, we can set up a task to recognize the action-adverb pairs, where the conditional distributions between action and BA can serve as prior knowledge. However, in practice, there are 2K categories of action-adverb pairs, thus it is difficult to do this multi-class classification task. Therefore, we adopt an indirect way to utilize the prior knowledge. We assume that the conditional distributions between action and BA, $\mathcal{P}(act|adv)$ and $\mathcal{P}(adv|act)$, are consistent between training and validation set and we employ the statistic value of training set $\hat{\mathcal{P}}_{train}(\cdot|\cdot)$ as $\mathcal{P}(\cdot|\cdot)$. In multi-task training scheme (see the Multi-Task section), the classifier at the end of the model outputs the predicted softmax values for action and BA: $\mathbf{p}_{act}$ and $\mathbf{p}_{adv}$, and by incorporating the conditional distributions as constraints into them, the final results $\overline{\mathbf{p}}_{act}$ and $\overline{\mathbf{p}}_{adv}$ can be expressed as:

$$\overline{\mathbf{p}}_{act}^{(i)} = \mathbf{p}_{act}^{(i)} \sum_j (\mathcal{P}(adv^{(j)}|act^{(i)}) \mathbf{p}_{adv}^{(j)} \tag{14}$$

$$\overline{\mathbf{p}}_{adv}^{(i)} = \mathbf{p}_{adv}^{(i)} \sum_j (\mathcal{P}(act^{(j)}|adv^{(i)}) \mathbf{p}_{act}^{(j)} \tag{15}$$

where $i$ and $j$ are the index of action and adverb categories.

# Experiments

## Experimental Setup

**Dataset** We conduct BA recognition on our VAAD which is the only dataset with BA semantics, and on top of VAAD, we also evaluate our BAUN on action datasets: UCF-101 (Soomro, Zamir, and Shah 2012), HMDB-51 (Kuehne, Jhuang, and others ) and Kinetics (Kay et al. 2017).

**Target Attention Mechanism** In VAAD dataset, we label the boundingbox and it can be used for BA detection problem as well. As we mainly focus on recognition problem, the boundingboxes therefore need to be loaded onto the model so we adopt a target attention mechanism to achieve it. In the scheme, we lower the brightness outside the boundingboxes. Let $\sigma, a, B, c$ denote the brightness decay value, the raw value of the point, the attention area, and the center of attention area respectively, the brightness value of point $p$ after decaying can be written as: $\max(0, a - |p-c| \amalg (p \notin B) \times \sigma)$, where $\amalg$ is the indicator function.

**Implementation Details** For all approaches, we follow each convolutional layer with a batch normalization layer (Ioffe and others 2015) and a ReLU activation function. When training on VAAD, 64 clips are fed into the network in each iteration. For ConvNet-LSTM and Two-Stream models, we use ResNet-50 pre-trained on ImageNet as the backbone and Adam (Kingma and Ba 2014) optimizer with the learning rate initialized as $10^{-4}$ and decreased to $10^{-5}$ after 8 epochs, while for I3D and BAUN, we use the backbone pre-trained on ImageNet and SGD with the learning rate initialized as $10^{-2}$ and decreased to $10^{-3}$ after 6 epochs.

For our BAUN, we pre-train the 3D convolution network on the ImageNet, and fine-tune it on VAAD with the STSM. Note that we do not add the IRN into the pipeline at the very beginning in that IRN requires a relative stable spatial features. Therefore, after fine-tuning the 3D convolution network and STSM, we add the IRN into the whole pipeline.

## Training Schemes

**Pure BA and Action Recognition** In this training scheme, all the models carry out the pure BA and action recognition task, in other words, one model takes charge of only one task and we do not merge information between action and adverb. This setting is treated as the basis to show how well can the models deal with the new semantics without any other extra information and evaluate our BAUN.

**Transfer Learning between BA and Action** In order to further demonstrate the value of BA, we conduct experiments to prove that BA and action can boost the performance of each other by sharing information between different semantics. Under this setting, we pre-train models on BA or action and then fine-tune them on the other task.

**Multi-Task** Besides pre-training scheme, we also use multi-task scheme to share information between the two semantics. We add a classifier for each task on top of models, calculate a total loss by weighting losses of the two tasks.

Table 2: Results on VAAD. **Left three columns**: results in order of BA recognition without action information, pre-trained on action (PTA), and through multi-task; **Right three columns**: results in order of action recognition without adverb information, pre-trained on BA (PTB), and through multi-task. Here we use mAP, accuracy (hit@1), and Hit@5 as metrics. Note that only for Multi-Task scheme, we add conditional distributions between action and BA as constraints for the final results.

| | BA Recg | | BA Recg (PTA) | | Multi-Task BA | | Act Recg | | Act Recg (PTB) | | Multi-Task Act | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | mAP | Hit@5 | mAP | Hit@5 | mAP | Hit@5 | mAP | Acc | mAP | Acc | mAP | Acc |
| ConvN-LSTM | 19.4 | 68.7 | 21.7 | 71.1 | 22.3 | 71.7 | 27.4 | 32.1 | 30.2 | 36.3 | 34.1 | 39.2 |
| Two-Stream-RGB | 15.8 | 65.6 | 19.3 | 73.0 | 20.0 | 75.4 | 29.7 | 33.3 | 31.3 | 38.3 | 33.4 | 38.8 |
| Two-Stream-Flow | 16.4 | 71.9 | 22.1 | 77.3 | 23.1 | 78.6 | 31.4 | 34.5 | 35.3 | 38.0 | 36.3 | 39.3 |
| Two-Stream-Fusion | 18.3 | 75.2 | 25.4 | 80.0 | 26.3 | 81.8 | 35.4 | 37.3 | 37.2 | 40.6 | 37.9 | 41.2 |
| RGB-I3D | 24.7 | 73.7 | 26.9 | 77.4 | 27.5 | 78.3 | 51.1 | 51.8 | 53.7 | 54.5 | 53.5 | 55.8 |
| Flow-I3D | 26.4 | 77.3 | 29.3 | 82.1 | 30.4 | 82.3 | 55.6 | 56.0 | 58.3 | 58.9 | 59.2 | 59.8 |
| Two-Stream I3D | 28.6 | 78.2 | 31.8 | 83.0 | 32.8 | 83.8 | 58.3 | 59.2 | 61.4 | 60.9 | 62.2 | 61.6 |
| RGB SlowFast R50 | 27.7 | 77.5 | 30.2 | 82.4 | - | - | 56.8 | 57.9 | - | - | - | - |
| RGB-BAUN | 28.1 | 77.9 | 30.8 | 82.6 | 31.9 | 83.1 | 52.8 | 53.2 | 54.3 | 55.2 | 55.3 | 57.2 |
| Flow-BAUN | 30.9 | 82.7 | 34.0 | 85.5 | 35.5 | 86.7 | 56.9 | 57.9 | 59.1 | 59.8 | 60.8 | 62.2 |
| Two-Stream-BAUN | **33.8** | **84.3** | **35.5** | **87.8** | **37.9** | **89.6** | **59.8** | **61.4** | **62.2** | **61.9** | **63.7** | **63.4** |

Table 3: List of 5 easiest and 5 hardest BA classes sorted by AP of the three models (RGB input). Note that physical related adverbs outperform psychological ones.

| | ConvNet-LSTM | | Two-Stream | | Two-Stream I3D | |
|---|---|---|---|---|---|---|
| | adverb | AP | adverb | AP | adverb | AP |
| 1 | professionally | 63.2 | professionally | 56.0 | easily | 73.6 |
| 2 | sweetly | 60.2 | solemnly | 50.0 | professionally | 70.6 |
| 3 | easily | 53.3 | promptly | 38.1 | sweetly | 63.4 |
| 4 | slowly | 45.8 | sweetly | 37.8 | promptly | 50.4 |
| 5 | promptly | 41.5 | slowly | 35.2 | slowly | 49.0 |
| -5 | hesitantly | 4.5 | gracefully | 2.9 | ironically | 6.4 |
| -4 | painfully | 4.4 | intently | 2.8 | hesitantly | 6.0 |
| -3 | ironically | 3.6 | surprisedly | 1.7 | surprisedly | 4.2 |
| -2 | surprisedly | 1.5 | ironically | 1.7 | solemnly | 2.7 |
| -1 | weakly | 0.7 | weakly | 0.7 | weakly | 1.3 |

Table 4: Benchmark results on action datasets. Here we use accuracy (hit@1) as the metric. "orig" means training the model without BA pre-training and "trans" means the models are pre-trained on BA and transferred to action.

| | UCF-101 | | HMDB-51 | | Kinetics | |
|---|---|---|---|---|---|---|
| | orig | trans | orig | trans | orig | trans |
| ConvN-LSTM | 54.2 | 56.3 | 18.3 | 22.5 | 53.9 | 57.3 |
| Two-Stream-RGB | 72.8 | 74.1 | 40.5 | 45.3 | 57.9 | 60.5 |
| Two-Stream-Flow | 81.0 | 83.7 | 46.6 | 48.3 | 49.6 | 52.1 |
| Two-Stream-Fusion | 83.6 | 85.5 | 47.1 | 50.2 | 62.8 | 64.3 |
| RGB-I3D | 95.1 | 96.7 | 74.3 | 76.4 | 68.4 | 69.2 |
| Flow-I3D | 96.5 | 97.0 | 77.3 | 78.4 | 61.5 | 62.5 |
| Two-Stream I3D | 97.8 | 98.0 | 80.9 | 81.3 | 71.6 | 72.2 |
| RGB-BAUN | 95.8 | 97.1 | 75.3 | 78.3 | 69.2 | 71.6 |
| Flow-BAUN | 97.0 | 97.6 | 78.0 | 79.1 | 62.9 | 64.0 |
| Two-Stream-BAUN | 98.1 | 98.4 | 81.7 | 83.7 | 72.4 | 73.2 |

This setting can better synthesize the two semantics than pre-training scheme because transferring may lose some key information from the first task.

## Results

**Pure BA and Action Recognition** We first analyze the pure BA and action recognition training scheme. Results in Tab. 2 show that the performance of BA recognition is significantly lower than action, which verifies that compared with action, BA recognition is really more challenging, mainly due to the intrinsic deeper knowledge of adverb. From another perspective, BA recognition offers promising potentials to build powerful video understanding models by understanding more exquisite and abundant semantic.

Additionally, our BAUN model exhibits superior performance on BA recognition tasks due to its specific-designed structures. On VAAD, it outperforms I3D model by 11% and achieves competitive results compared with the latest video baseline: SlowFast (Feichtenhofer et al. 2018) when adopting RGB images as input. Besides, it also performs well on action recognition tasks, achieving on average 2.5% improvements on action datasets (see Tab. 2 and Tab. 4). Improvements on action are relatively low as BAUN is an adverb model, not tailored for action. Note that, using optical flow as input delivers better performance on VAAD since the expression features are extracted from RGB input, which leads to accessing to both the RGB and flow information in optical flow model.

In order to demonstrate the relative difficulty of each adverb, we list the easiest and hardest 5 recognized adverbs in Tab. 3. It's evident that physical related adverbs like slowly, easily, promptly achieve much better performance than adverbs describing inner activities like painfully, hesitantly. This reveals that our BAUN model still does not solve the recognition of all kinds of BA, especially those inner related ones which serve as a challenging part in BA recognition and we need more ingenious models to deal with it.

**Two Information Sharing Schemes** In this section, we compare the above two information sharing schemes. As expected, for all models, these schemes can boost the model's performance for both action and BA recognition. The multi-task scheme adopts the conditional distributions between action and BA, hence its results are better than the transfer learning scheme. These reveal that BA is a valuable semantic and its features have universality. Moreover, we transfer the models pre-trained on VAAD to other action datasets and the results are listed in Tab. 4. All the models benefit from the transferred semantics and gain 3.4% increase on average.

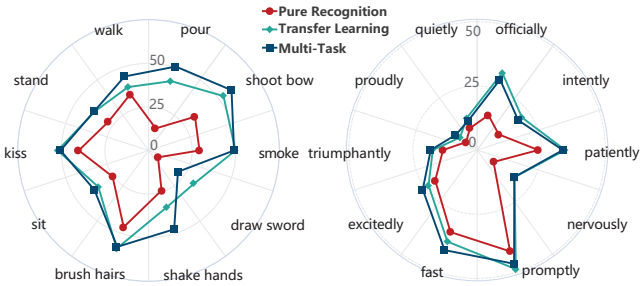


Figure 4: Ten categories' results of ConvNet-LSTM model with three training schemes. We pick out the categories with most performance improvements. **Left:** Top 10 action categories. **Right:** Top 10 BA categories. (Metric: AP)

Table 5: Ablation studies results on VAAD with RGB as input. We adopt mAP, hit@5 and hit@1(accuracy) as metrics. "w/o X" denotes without X module in the implementation.

| Method | Act Reg | | BA Reg | |
|---|---|---|---|---|
| | mAP | Acc | mAP | hit@5 |
| BAUN | 52.8 | 53.2 | 28.1 | 77.9 |
| BAUN w/o STSM | 52.1 | 52.7 | 26.0 | 74.6 |
| BAUN w/o IRN | 51.6 | 52.4 | 25.7 | 74.2 |
| BAUN w/o expression features | 52.2 | 53.0 | 27.1 | 76.8 |

In Fig. 4, we pick out each 10 categories of action and BA that obtain most improvements from the two information sharing schemes on ConvNet-LSTM model and visualize them. We can see that the top 10 action categories contain many tiny actions such as pour, smoke, kiss, and shake hands, proving that with the aid of adverbs, action models can pay more attention to the details to a certain extent. For adverbs, 7 of the top 10 categories with most improvements are related to inner activities such as intently, proudly, and nervously which are difficult to recognize just as the discussion mentioned before. These results reveal actions and adverbs will help each other to overcome the difficulties since the original results on these categories are not satisfactory and the features obtained by the information sharing schemes are more abundant and wide-ranging.

## Ablation Studies

**STSM module** We first visualize the long-term temporal features in Fig. 5. For each time stamp, the mean values of the activation after sigmoid function are presented. We can see that it successfully amplifies the prominent video fragments after a global analysis. If we ablate STSM from the whole pipeline (results shown in Tab. 5), the performance decays by 1.0% for action recognition and 5.9% for BA recognition, which further validates that our STSM module aiming at separating the spatial and temporal features does possess a stronger capability to process the temporal and spatial information.

**IRN module** IRN module explores the context relations in the videos and endows the model with a larger spatial receptive field in the low layers to make the model understand the global spatial features in an early stage. After removing the IRN module, the performance drops by 1.9% on action recognition and 6.7% on BA recognition.

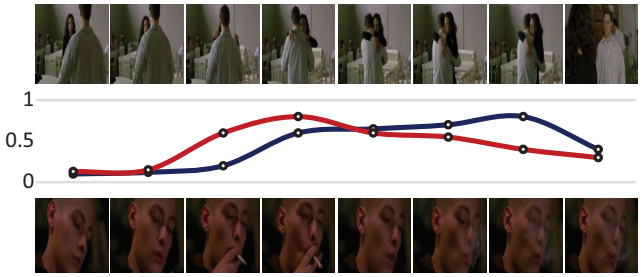


Figure 5: Visualization on the activations of ConvLSTM in STSM. First row is a hug video and the third row is a smoking one. The second row shows the mean of the activation values of every frame, where the blue line corresponds to the hug video and the red is for the smoking video.

**Expression Features** The expression features are incorporated to better represent the mood and attitude information of humans in order to better model the mood related BAs. When removing this feature inputs, the performance decays by 0.6% on action recognition and 2.5% on BA recognition, which further showcases that expression information is more effective for BA than action.

## Conclusions

This paper proposes a new video semantics: Behavior Adverb (BA), a corresponding model: BA Understanding Network (BAUN) and constructs a new benchmark: VAAD. We evaluate our BAUN and other three representative video understanding models on four video understanding datasets: VAAD, UCF-101, HMDB-51 and Kinetics. Results highlight the challenge of the BA recognition problem, demonstrate promising potentials of BA for video understanding community and verify that our BAUN receives best performance on BA recognition. Meanwhile, utilizing both BA and action information together facilitates the models to achieve better performance on the two recognition tasks.

For future work, although our IRN and expression model achieve noticeable progress on the psychologically relevant and relatively subjective BAs, the still unsatisfactory results on them underscore the need for more powerful video understanding approaches.

## Acknowledgements

This work is supported in part by the National Key R&D Program of China, No. 2017YFA0700800, National Natural Science Foundation of China under Grants 61772332.

## References

Abu-El-Haija, S., et al. 2016. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint*.

Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; and Carlos Niebles, J. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 961–970.

Carreira, J., and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 4724–4733.

Davies, M. 2008. *The corpus of contemporary American English*. BYE, Brigham Young University.

Dhall, A.; Goecke, R.; Joshi, J.; Hoey, J.; and Gedeon, T. 2016. Emotiw 2016: Video and group-level emotion recognition challenges. In *ICMI*, 427–432.

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

Douglas-Cowie, E.; Cowie, R.; Sneddon, I.; Cox, C.; Lowry, O.; Mcrorie, M.; Martin, J.-C.; Devillers, L.; Abrilian, S.; Batliner, A.; et al. 2007. The humaine database: addressing the collection and annotation of naturalistic and induced emotional data. *Affective compfuting and intelligent interaction* 488–500.

Fan, Y.; Lu, X.; Li, D.; and Liu, Y. 2016. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *ICMI*.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2018. Slowfast networks for video recognition. *arXiv preprint*.

Girshick, R. 2015. Fast r-cnn. In *ICCV*, 1440–1448.

Gu, C.; Sun, C.; et al. 2017. Ava: A video dataset of spatio-temporally localized atomic visual actions. *CoRR* 4.

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

Hu, H.; Gu, J.; Zhang, Z.; Dai, J.; and Wei, Y. 2018. Relation networks for object detection. In *CVPR*, 3588–3597.

Ioffe, S., et al. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint*.

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

Johnson, J.; Hariharan, B.; van der Maaten, L.; Fei-Fei, L.; Lawrence Zitnick, C.; and Girshick, R. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2901–2910.

Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; and Fei-Fei, L. 2014. Large-scale video classification with convolutional neural networks. In *CVPR*, 1725–1732.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. 2017. The kinetics human action video dataset. *arXiv preprint*.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint*.

Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Carlos Niebles, J. 2017. Dense-captioning events in videos. In *ICCV*.

Kristan, M.; Matas, J.; Leonardis, A.; Felsberg, M.; Cehovin, L.; Fernandez, G.; Vojir, T.; Hager, G.; Nebehay, G.; and Pflugfelder, R. 2015. The visual object tracking vot2015 challenge results. In *ICCV Workshops*.

Kuehne, H.; Jhuang, H.; et al. Hmdb: a large video database for human motion recognition. In *ICCV*.

Laptev, I. 2005. On space-time interest points. *IJCV* 64(2-3).

Maji, S.; Bourdev, L.; and Malik, J. 2011. Action recognition from a distributed representation of pose and appearance. In *CVPR*.

Nam, H., and Han, B. 2015. Learning multi-domain convolutional neural networks for visual tracking. *CoRR* abs/1510.07945.

Ng, H.-W.; Nguyen, V. D.; Vonikakis, V.; and Winkler, S. 2015. Deep learning for emotion recognition on small datasets using transfer learning. In *ICMI*, 443–449.

Pang, B.; Zha, K.; Cao, H.; Shi, C.; and Lu, C. 2019. Deep rnn framework for visual sequential applications. In *CVPR*.

Petrosino, A., and Gold, K. 2010. Toward fast mapping for robot adjective learning. In *2010 AAAI Fall Symposium Series*.

Redmon, J., and Farhadi, A. 2016. Yolo9000: better, faster, stronger. *arXiv preprint*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 91–99.

Santoro, A.; Raposo, D.; Barrett, D. G.; Malinowski, M.; Pascanu, R.; Battaglia, P.; and Lillicrap, T. 2017. A simple neural network module for relational reasoning. In *NeurIPS*, 4967–4976.

Sigurdsson, G. A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I.; and Gupta, A. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*.

Simonyan, K., and Zisserman, A. 2014. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, 568–576.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint*.

Srivastava, N.; Mansimov, E.; and Salakhudinov, R. 2015. Unsupervised learning of video representations using lstms. In *ICML*.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *CVPR*, 1–9.

Valstar, M., and Pantic, M. 2010. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Workshop on EMOTION (satellite of LREC)*, 65.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*, 5998–6008.

Wang, H.; Kläser, A.; Schmid, C.; and Liu, C.-L. 2011. Action recognition by dense trajectories. In *CVPR*, 3169–3176.

White, L. 1991. Adverb placement in second language acquisition: Some effects of positive and negative evidence in the classroom. *Interlanguage studies bulletin (Utrecht)* 7(2):133–161.

Wu, Z.; Wang, X.; Jiang, Y.-G.; Ye, H.; and Xue, X. 2015. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In *ACM MM*, 461–470.

Xingjian, S.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; and Woo, W.-c. 2015. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NeurIPS*, 802–810.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*.

Yue-Hei Ng, J.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; and Toderici, G. 2015. Beyond short snippets: Deep networks for video classification. In *CVPR*, 4694–4702.

Zhou, B.; Andonian, A.; Oliva, A.; and Torralba, A. 2018. Temporal relational reasoning in videos. In *ECCV*, 803–818.