WHERE AND WHEN TO LOOK? SPATIO-TEMPORAL ATTENTION FOR ACTION RECOGNITION IN VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Inspired by the observation that humans are able to process videos efficiently by only paying attention when and where it is needed, we propose a novel spatial-temporal attention mechanism for video-based action recognition. For spatial attention, we learn a saliency mask to allow the model to focus on the most salient parts of the feature maps. For temporal attention, we employ a soft temporal attention mechanism to identify the most relevant frames from an input video. Further, we propose a set of regularizers that ensure that our attention mechanism attends to coherent regions in space and time. Our model is efficient, as it proposes a separable spatio-temporal mechanism for video attention, while being able to identify important parts of the video both spatially and temporally. We demonstrate the efficacy of our approach on three public video action recognition datasets. The proposed approach leads to state-of-the-art performance on all of them, including the new large-scale *Moments in Time* dataset. Furthermore, we quantitatively and qualitatively evaluate our model's ability to accurately localize discriminative regions spatially and critical frames temporally. This is despite our model only being trained with per video classification labels.

1 INTRODUCTION

Machine learning models been widely adopted for solving various real world tasks, ranging from visual recognition (Krizhevsky et al., 2012; Simonyan & Zisserman, 2014; He et al., 2016; Huang et al., 2017) to natural language processing (Iyyer et al., 2015; Goldberg, 2016). Although record-breaking accuracy has been reported in many of the problems, an open question remains: "why does my model make that prediction?" (Koh & Liang, 2017). By understanding the reason or logic behind the decision process of a learning algorithm, one can further improve the model, discover new science (Shrikumar et al., 2017), provide end-users with explanations and guide them on how to effectively employ the results (Ribeiro et al., 2016). In particular, for video action recognition, a proper attention model can help answer the question of *where* and *when* it needs to look at the image evidence to draw a classification decision.

The attention mechanism in the human visual system is perhaps one of the most fascinating facets of intelligence. Instead of compressing an entire image into a static representation, attention allows for salient features to dynamically come to the forefront as needed. One important property of human perception is that one does not tend to process a whole scene in its entirety at once. Instead, humans focus attention selectively on parts of the visual space to acquire information *when* and *where* it is needed, and combine information from different fixations over time to build up an internal representation of the scene (Rensink, 2000).

The attention mechanism (Xu et al., 2015; Bahdanau et al., 2015; Cho et al., 2015; Vaswani et al., 2017) provides a way to bridge the gap between the black-box decision process and human interpretability. It intuitively explains which part the model attends to when making a particular decision, which is very helpful in real applications, e.g., medical AI systems or self-driving cars.

In this paper, we propose a novel spatio-temporal attention mechanism that is designed to address these challenges. Our attention mechanism is efficient, due to its space- and time- separability, and yet flexible enough to enable encoding of effective regularizers (or priors). As such, our attention mechanism consists of spatial and temporal components shown in Fig. 1. The spatial attention component, that attenuates frame-wise CNN image features, consists of the saliency mask; regularized

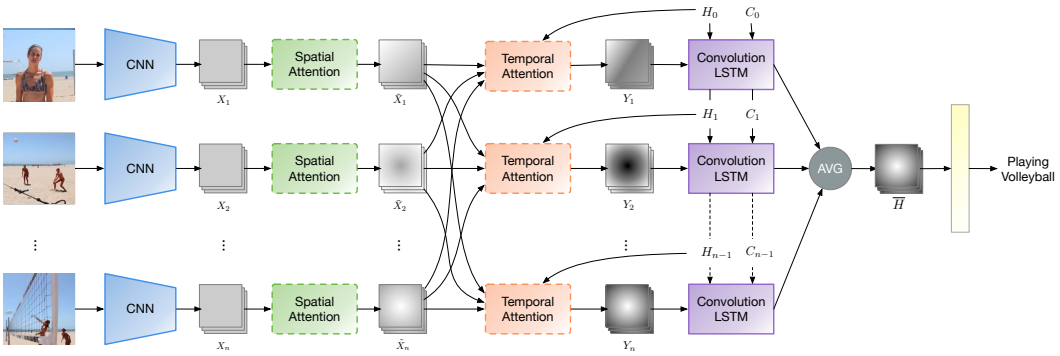


Figure 1: **Spatio-temporal attention for video action recognition.** The convolutional features are attended over both spatially, in each frame, and subsequently temporally. Both attentions are soft, meaning that the effective final representation at time t of an RNN, used to make the prediction, is a spatio-temporally weighted aggregation of convolutional features across the video along with the past hidden state from $t-1$. For details please refer to Sec. 3.

to be discriminative and spatially smooth. The temporal component consists of a uni-modal soft attention mechanism that aggregates information over the near-by attenuated frame features before passing it into Convolutional LSTM for class prediction.

Contributions: In summary, the main contributions of this work are: (1) We introduce a simple yet effective spatial-temporal attention mechanism for video action recognition; (2) We introduce three different regularizers, two for spatial and one for temporal attention components, to improve performance and interpretability of our model; (3) We illustrate state-of-the-art performance on three standard publicly available datasets. We also explore the importance of our modeling choices through ablation experiments. (4) Finally, we qualitatively and quantitatively show that our spatio-temporal attention is able to localize discriminative regions and important frames, despite being trained in a purely weakly-supervised manner with only classification labels.

2 Related work

2.1 Spatial attention

Sharma et al. (2015) develop an attention-driven LSTM by highlighting important spatial locations for action recognition. However, it only focuses on the crucial spatial locations of each image, without considering the temporal importance of different video frames in a video sequence. Wang et al. (2016b) propose a Hierarchical Attention Network (HAN), which enables to incorporate attention of both spatial and motion stream into action recognition. Girdhar & Ramanan (2017) introduce an attention mechanism based on a derivation of bottom-up and top-down attention as low-rank approximations of bilinear pooling methods. However, it only uses a single frame to predict and apply spatial attention, without considering long-term temporal relations among different frames. Our method, however, uses a single frame to both predict and apply spatial attention, making it amenable to both single image and video based use cases.

2.2 Temporal attention

Visualizing where in the input the model was attending to at each output time-step produces an input-output alignment which provides valuable insight into the model's behavior. Soft attention inspects every entry of the memory at each output time-step, effectively allowing the model to condition on an arbitrary input sequence entry. Bahdanau et al. (2015) propose to a soft attention RNN model to automatically (soft-)search for parts of a source sentence that are relevant to predicting a target word. Torabi & Sigal (2017) propose an attention based LSTM model to hightlight frames in videos.

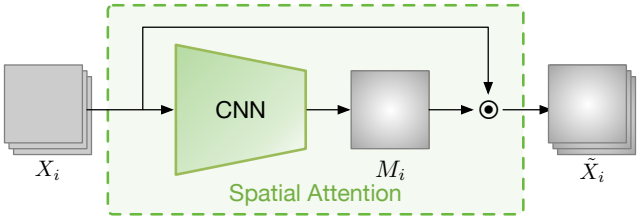


Figure 2: **Spatial attention component.** We use several layers of convolutional network to learn the importance mask $\boldsymbol{M}_i$ for the input image feature $\boldsymbol{X}_i$, the output is the element-wise multiplication $\tilde{\boldsymbol{X}}_i = \boldsymbol{X}_i \odot \boldsymbol{M}_i$. Details please refer to Sec. 3.2.

2.3 Network interpretation

Various methods have been proposed to try to explain neural networks (Zeiler & Fergus, 2014; Springenberg et al., 2014; Mahendran & Vedaldi, 2016; Zhou et al., 2016; Zhang et al., 2016; Simonyan et al., 2013; Ramprasaath et al., 2016; Ribeiro et al., 2016; 2018) in various ways. Visual attention is also one way that tries to explain which part of the image is responsible for the network's decision (Li et al., 2018; Jetley et al., 2018). Besides the explanation, Li et al. (2018) build up an end-to-end model to provide supervision directly on these explanations, specifically network's attention.

3 Spatial-temporal attention mechanism

Our overall model is an Recurrent Neural Network (RNN) that aggregates frame-based convolutional features across the video to make action predictions as shown in Fig. 1. The convolutional features are attended over both spatially, in each frame, and subsequently temporally. Both attentions are soft, meaning that the effective final representation at time t of an RNN, used to make the prediction, is a spatio-temporally weighted aggregation of convolutional features across the video along with the past hidden state from $t-1$. The core novelty is the overall form of our attention mechanism and the additional terms of the loss function that induce sensible spatial and temporal attention priors.

3.1 Convolutional frame features

We use the last convolutional layer output extracted by ResNet50 (He et al., 2016), pretrained on the ImageNet (Deng et al., 2009) dataset and fine-tuned for the target dataset, as our frame feature representation. We acknowledge that more accurate feature extractors (for instance, network with more parameters such as ResNet-154 or higher performance networks such as DenseNet (Huang et al., 2017) or SENet (Hu et al., 2018)) and/or optical flow features will likely lead to better overall performance. Our primary purpose in this paper is to prove the efficacy of our attention mechanism. Hence we kept the features relatively simple.

3.2 Spatial attention with importance mask

We apply an importance mask $\boldsymbol{M}_i$ to the i-th image features $\boldsymbol{X}_i$ to obtain attended image features by element-wise multiplication:

$$\tilde{\boldsymbol{X}}_i = \boldsymbol{X}_i \odot \boldsymbol{M}_i, \tag{1}$$

for $1 \leq i \leq n$. This operation attenuates certain regions of the feature map based on their estimated importance. Here we simply use three convolutional layers to learn the importance mask. Fig. 2 illustrates our spatial attention mechanism. However, if left uncontrolled, an arbitrarily structured mask could be learned, leading to possible overfitting. We posit that, in practice, it is often useful to attend to a few important larger regions (e.g., objects, elements of the scene). To induce this behavior, we encourage smoothness of the mask by introducing total variation loss on the spatial attention, as will be described in Sec. 3.4.

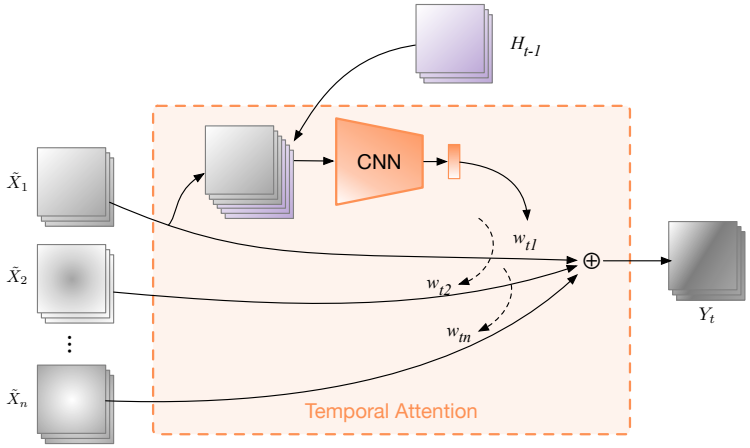


Figure 3: **Temporal attention component.** The temporal attention learns a temporal attention weight w_{ti} at each time step t. The final feature map $\boldsymbol{Y}_t$ at time t to the ConvLSTM is a weighted sum of the feature from all the previous masked frames. Details please refer to Sec. 3.3.

3.3 Temporal attention

Inspired by attention for neural machine translation (Bahdanau et al., 2015), we introduce the temporal attention mechanism which generates *energy* for each attended frame $\tilde{\boldsymbol{X}}_i$ at each time step t,

$$e_{ti} = \Phi(\boldsymbol{H}_{t-1}, \tilde{\boldsymbol{X}}_i), \tag{2}$$

where $\boldsymbol{H}_{t-1}$ represents the ConvLSTM hidden state at time $t-1$ that implicitly contains all previous information up to time step $t-1$, $\tilde{\boldsymbol{X}}_i$ represents the i-th frame masked features. $\Phi = \Phi_H(\boldsymbol{H}_{t-1}) + \Phi_X(\tilde{\boldsymbol{X}}_i)$, where Φ_H and Φ_X are feedforward neural networks which are jointly trained with all other components of the proposed system.

This temporal attention model directly computes soft attention weight for each frame at each time t as shown in Fig. 3. It allows the gradient of the cost function to be backpropagated through. This gradient can be used to train the entire spatial-temporal attention model jointly.

The importance weight w_{ti} for each frame is:

$$w_{ti} = \frac{\exp(e_{ti})}{\sum_{i=1}^{n}(\exp(e_{ti}))}, \tag{3}$$

for $1 \leq i \leq n, 1 \leq t \leq T$. This importance weighting mechanism decides which frame of the video to pay attention to. The final feature map $\boldsymbol{Y}_t$ to the ConvLSTM is a weighted sum of the feature from all of the frames as the ConvLSTM cell inputs:

$$\boldsymbol{Y}_t = \frac{1}{n}\sum_{i=1}^{n} w_{ti}\tilde{\boldsymbol{X}}_i, \tag{4}$$

where $\tilde{\boldsymbol{X}}_i$ denotes the i-th masked frame of each video, n represents the total number of frames for each video.

For RNN, instead of using conventional LSTM (Graves, 2013), we use Convolutional LSTM (ConvLSTM) (Shi et al., 2015) instead. The drawback of conventional LSTM is its use of full connections in the input-to-state and state-to-state transitions in which no spatial information is encoded. In contrast, in ConvLSTM, each input $\boldsymbol{Y}_t$, cell output $\boldsymbol{O}_t$, hidden state $\boldsymbol{H}_t$ and gate $\boldsymbol{I}_t, \boldsymbol{F}_t, \boldsymbol{O}_t$ are 3D tensors whose last two dimensions are spatial dimensions (rows and columns).

$$\begin{aligned}
\boldsymbol{I}_t &= \sigma(\boldsymbol{W}_{xi} * \boldsymbol{Y}_t + \boldsymbol{W}_{hi} * \boldsymbol{H}_{t-1} + \boldsymbol{b}_i), \\
\boldsymbol{F}_t &= \sigma(\boldsymbol{W}_{xf} * \boldsymbol{Y}_t + \boldsymbol{W}_{hf} * \boldsymbol{H}_{t-1} + \boldsymbol{b}_f), \\
\boldsymbol{C}_t &= \boldsymbol{F}_t \circ \boldsymbol{C}_{t-1} + \boldsymbol{I}_t \circ \tanh(\boldsymbol{W}_{xc} * \boldsymbol{Y}_t + \boldsymbol{W}_{hc} * \boldsymbol{H}_{t-1} + \boldsymbol{b}_c), \\
\boldsymbol{O}_t &= \sigma(\boldsymbol{W}_{xo} * \boldsymbol{Y}_t + \boldsymbol{W}_{ho} * \boldsymbol{H}_{t-1} + \boldsymbol{b}_o), \\
\boldsymbol{H}_t &= \boldsymbol{O}_t \circ \tanh(\boldsymbol{C}_t).
\end{aligned}$$

We use the following initialization strategy for the ConvLSTM cell state and hidden state for faster convergence:

$$\boldsymbol{C}_0 = g_c(\frac{1}{n}\sum_{i=1}^{n}\tilde{\boldsymbol{X}}_i), \qquad \boldsymbol{H}_0 = g_h(\frac{1}{n}\sum_{i=1}^{n}\tilde{\boldsymbol{X}}_i) \tag{5}$$

where g_c and g_h are two layer convolutional networks with batch normalization (Ioffe & Szegedy, 2015).

We calculate the average hidden states of ConvLSTM over time length T,

$$\overline{\boldsymbol{H}} = \frac{1}{T}\sum_{i=1}^{T}\boldsymbol{H}_i, \tag{6}$$

and send it to a fully connected classification layer for the final video action classification.

3.4 Loss function

Considering the spatial and temporal nature of our video action recognition; we would like to learn (1) a sensible attention mask for spatial attention, (2) reasonable importance weighting scores for different frames, and (3) improve the action recognition accuracy at the same time. Therefore, our loss function L:

$$L = L_{\mathrm{CE}} + \lambda_{\mathrm{TV}} L_{\mathrm{TV}} + \lambda_{\mathrm{contrast}} L_{\mathrm{contrast}} + \lambda_{\mathrm{unimodal}} L_{\mathrm{unimodal}}, \tag{7}$$

where L_{CE} is the cross entropy loss for classification, L_{TV} represents the total variation regularization (Rudin et al., 1992); L_{contrast} represents the mask and background contrast regularizer; and L_{unimodal} represents unimodality regularizer. λ_{TV}, $\lambda_{\mathrm{contrast}}$ and $\lambda_{\mathrm{unimodal}}$ are the weights for corresponding regularizers.

The total variation regularization L_{TV} of the learnable attention mask encourages spatial smoothness of the mask and is defined as:

$$L_{\mathrm{TV}} = \sum_{i=1}^{n}\left(\sum_{j,k}|\boldsymbol{M}_i^{j+1,k} - \boldsymbol{M}_i^{j,k}| + \sum_{j,k}|\boldsymbol{M}_i^{j,k+1} - \boldsymbol{M}_i^{j,k}|\right), \tag{8}$$

where $\boldsymbol{M}_i$ is the mask for the i-th frame, and $\boldsymbol{M}_i^{j,k}$ is entry at the (j,k)-th spatial location of the mask. Different from the total variation of the mask of using L_2 loss in Dabkowski & Gal (2017), we use L_1 loss instead. The contrast regularization L_{contrast} of learnable attention mask is to suppress the irrelevant information and highlight important information:

$$L_{\mathrm{contrast}} = \sum_{i=1}^{n}\left(-\frac{1}{2}\boldsymbol{M}_i \odot \boldsymbol{B}_i + \frac{1}{2}\boldsymbol{M}_i \odot (1 - \boldsymbol{B}_i)\right), \tag{9}$$

where $\boldsymbol{B}_i = \mathbb{I}\{\boldsymbol{M}_i > 0.5\}$ represents the binarized mask, $\mathbb{I}$ is the indicator function applied element-wise.

The unimodality regularizer L_{unimodal} encourages the temporal attention weights to be unimodal, biasing against spurious temporal weights. This stems from our observation that in most cases only one activity would be present in the considered frame window, with possible irrelevant information on either or both sides. Here we use the log concave distribution to encourage the unimodal pattern of temporal attention weights:

$$L_{\mathrm{unimodal}} = \sum_{t=1}^{T}\sum_{i=2}^{n-1}\max\{0, w_{t,i-1}w_{t,i+1} - w_{t,i}^2\}, \tag{10}$$

where T represents the ConvLSTM time sequence length and n is the number of frames for each video. More details on this log concave sequence please refer to Appendix A.

4 Experiments

In this section, we first conduct experiments to evaluate our proposed method on video action recognition task on three public available datasets. Then we evaluate our spatial attention mechanism on the spatial localization task and our temporal attention mechanism on the temporal localization task respectively.

Model	HMDB51	UCF101
Visual attention (Sharma et al., 2015)	41.31	84.96
ResNet50-ImageNet	47.78	82.30
Ours	**49.93**	**86.02**
Ablation Experiments		
Ours w/o spatial attention	49.21	84.83
Ours w/o temporal attention	49.22	85.06
Ours w/o L_{TV}	49.02	85.05
Ours w/o $L_{contrast}$	49.15	85.15
Ours w/o $L_{unimodal}$	49.35	85.57

Table 1: **Top-1 accuracy (%) on HMDB51 and UCF101 dataset.**

4.1 VIDEO ACTION RECOGNITION

We first conduct extensive studies on the widely used HMDB51 and UCF101 datasets. The purpose of these experiments is mainly for ablation study to examine the effects of different sub-components. Then we show that our method can be applied to the challenging large-scale Moments in Time dataset.

Datasets. HMDB51 dataset (Kuehne et al., 2011) contains 51 distinct action categories, each containing at least 101 clips for a total of 6,766 video clips extracted from a wide range of sources. These videos include general facial actions, general body movements, body movements with object interaction, body movements for human interaction.

UCF101 dataset (Soomro et al., 2012) is an action recognition dataset of realistic action videos, collected from YouTube, with 101 action categories.

Moments in Time Dataset (Monfort et al., 2018) is a collection of one million short videos with one action label per video and 339 different action classes. As there could be more than one action taking place in a video, action recognition models may predict an action correctly yet be penalized because the ground truth does not include that action. Therefore, it is believed that top 5 accuracy measure will be more meaningful for this dataset.

Experimental setup. We use the same parameters for HMDB51 and UCF101: single Convolutional LSTM layer with hidden-state dimension 512, sequence length $T = 50$, $\lambda_{TV} = 10^{-5}$, $\lambda_{contrast} = 10^{-4}$, $\lambda_{unimodal} = 1$. For the Moments in Time dataset, we use time sequence length $T = 15$. For more details on the experimental setup please refer to Appendix B.1.

Quantitative results. We show the top-1 video action classification accuracy for HMDB51 and UCF101 dataset in Table 1. It is clear from the table that the proposed method outperforms the conventional ResNet50-ImageNet and visual attention (Sharma et al., 2015). From the ablation experiments, it demonstrates that all the sub-components of the proposed method contribute to improving the final performance.

The results on the Moments in Time dataset are reported in Table 2. Our method achieves the best accuracy comparing to other single-modality-based methods, and obtains better or comparative results comparing to the methods which uses more than one modality. TRN-Multiscale (Zhou et al., 2018), which uses both RGB and optical flow images, has better performance than ours, however, extracting optical flow images for such large datasets is very time-consuming and needs the same order of magnitude of storage as RGB images.

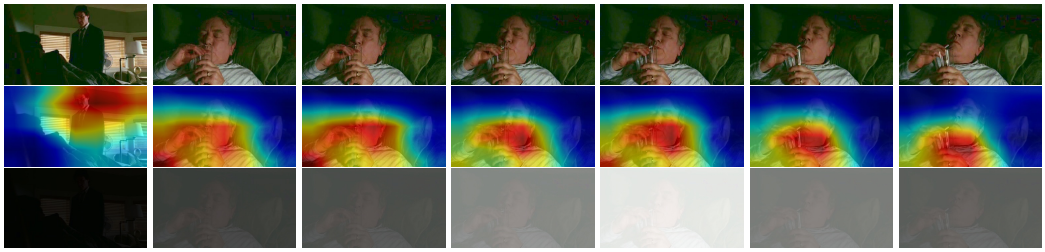
Qualitative results. We visualize the spatial attention and temporal attention results in Fig. 4. We can see that the spatial attention can correctly focus on important spatial area of the image, and the temporal attention shows a unimodal distribution for the entire action from starting the action to completing the action. More results are shown in Appendix C.1.

4.2 WEAKLY SUPERVISED LOCALIZATION

Due to the existence of spatial and temporal attention mechanisms, our model can not only classify the action of the video, but also give a better interpretability of the results, *i.e.* telling which region and frames contribute more to the prediction. In other words, our proposed model can also localize

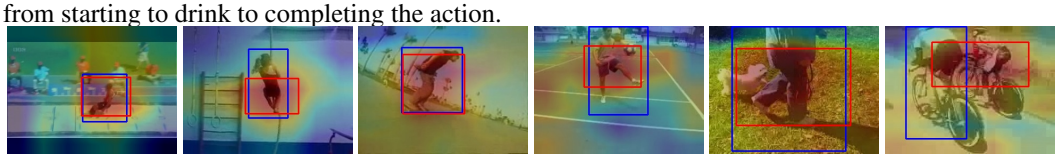
Model	**Modality**	**Top-1** (%)	**Top-5** (%)
ResNet50-ImageNet (Monfort et al., 2018)	RGB	26.98	51.74
TSN-Spatial (Wang et al., 2016a)	RGB	24.11	49.10
TRN-Multiscale (Zhou et al., 2018)	RGB	27.20	53.05
BNInception-Flow (Monfort et al., 2018)	Optical flow	11.60	27.40
ResNet50-DyImg (Monfort et al., 2018)	Optical flow	15.76	35.69
TSN-Flow (Wang et al., 2016a)	Optical flow	15.71	34.65
TSN-2stream (Wang et al., 2016a)	RGB+Optical flow	25.32	50.10
TRN-Multiscale (Zhou et al., 2018)	RGB+optical flow	**28.27**	**53.87**
Ours	RGB	27.55	53.52

Table 2: **Results on Moments in Time dataset.** ResNet50-ImageNet and TRN-Multiscale spatial results reported here are based on authors' (Monfort et al., 2018) released trained model.



(a) frame01 (b) frame11 (c) frame29 (d) frame55 (e) frame59 (f) frame67 (f) frame89

Figure 4: **Examples of spatial temporal attention.** (Best viewed in color.) A frame sequence from a video of *Drink* action in HMDB51. The original images are shown at the top row, spatial attention is shown as heatmap (red means important) in the middle row, and temporal attention score is shown as the gray image (the brighter the frame is, the more crucial the frame is) at the bottom row. It shows that spatial attention can focus on important areas while temporal attention can attend to crucial frames. The temporal attention also shows a unimodal distribution for the entire action from starting to drink to completing the action.



(a) (b) (c) (d) (e) (f)

Figure 5: **Examples of spatial attention for action localization.** (Best viewed in color.) Blue bounding boxes represent ground truth while the red ones are predictions from our learned spatial attention. (a) long jump, (b) rope climbing, (c) skate boarding, (d) soccer juggling (e) walking with dog, (f) biking.

the most discriminant region and frames at the same time. To verify this, we conduct the spatial localization and temporal localization experiments.

4.2.1 Spatial action localization

Dataset. UCF101-24 is a subset of 24 classes out of 101 classes of UCF101 that comes with spatio-temporal localization annotation, released as bounding box annotations of humans with THUMOS2013 and THUMOS2014 challenge (Jiang et al., 2014).

Experimental setup. We use the same hyper-parameters for UCF101-24 as HMDB51 and UCF101 in the previous section. The model we used for feature extraction is our pretrained ResNet50 model on UCF101 in the previous experiment. For training, we only use the classification labels without spatial bounding box labels. For evaluation, we threshold the produced saliency mask at 0.5 and the tightest bounding box that contains the thresholded saliency map is set as the predicted localization box for each frame. Then these predicted localization boxes are compared with the ground truth bounding boxes at different Intersection Over Union (IOU) levels.

Methods	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$
Fast action proposal * (Yu & Yuan, 2015)	42.8%	–	–	–
Learning to track * (Weinzaepfel et al., 2015)	54.3%	51.7%	**47.7**%	**37.8**%
Ours	**67.0**%	**56.1**%	34.1%	17.7%

Table 3: **Spatial action localization results on UCF101-24 dataset** measured by mAP at different IoU thresholds α. * The baseline methods are strongly supervised spatial localization methods.

Method	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$
(Yeung et al., 2016)	48.9%	44.0%	36.0%	26.4%	17.1%
(Wang et al., 2017)	44.4%	37.7%	28.2%	21.1%	13.7%
Ours	**70.0**%	**61.4**%	**48.6**%	**32.6**%	**17.9**%

Table 4: **Temporal action localization results on THUMOS'14 dataset** measured by mAP at different IoU thresholds α.

Qualitative results. We show some qualitative results in Fig. 5. Our spatial attention can attend to important action areas. The ground truth bounding boxes include all the entire human actions, while our attention could attend to crucial parts of an action such as in Fig.5 (d) and (e). Furthermore, our attention mechanism is able to attend to areas with multiple human actions. For instance, in Fig.5 (f) the ground truth only includes one person bicycling, but our attention can include both people bicycling. More qualitative results including failure cases are included in Appendix C.2.

Quantitative results. With our spatial-temporal attention mechanism, the action recognition accuracy for UCF101-24 is 93.04%, which is a 6.12% performance boost compared with 86.92% by averaging the predictions of ResNet-50 over 50 frames. Table 3 shows the quantitative results for UCF101-24 spatial localization results. Our attention mechanism works better compared with the baseline methods when the IoU threshold is lower mainly because our model only focuses on important spatial areas rather than the entire human action annotated by bounding boxes. Compared with the baseline methods training with ground truth bounding boxes, we only use the action classification label, no ground truth bounding boxes are used.

4.2.2 TEMPORAL ACTION LOCALIZATION

Dataset. The action detection task of THUMOS14 (Jiang et al., 2014) consists of 20 classes of sports activities, and contains 2765 trimmed videos for training, while 200 and 213 untrimmed videos for validation and test respectively. More details on this dataset and pre-processing are included in Appendix B.1.

Experimental setup. We use the same hyperparameters for THUMOS14 as HMDB51, UCF101 and UCF101-24. For training, we only use the classification labels without temporal annotation labels. For evaluation, we threshold the normalized temporal attention importance weight at 0.5. Then these predicted temporal localization frames are compared with the ground truth annotation at different IoU thresholds.

Qualitative results. We first visualize some examples of learned attention weights on the test data of THUMOS14 in Fig. 6. We see that our temporal attention module is able to automatically highlight important frames and to avoid irrelevant frames corresponding to background or non-action human poses. More qualitative results are included in Appendix C.3.

Quantitative results. With our spatial temporal attention mechanism, the video action classification accuracy for the THUMOS'14 20 classes improved from 74.45% to 78.33%: a 3.88% increase. Besides improving the classification accuracy, we show our temporal attention mechanism is able to highlight discriminative frames quantitatively in Table 4. Compared with strongly supervised method (Yeung et al., 2016) and weakly supervised method (Wang et al., 2017), our method achieves the best accuracy in terms of different levels of IoU thresholds.

Figure 6: **Examples temporal localization with temporal attention from THUMOS14.** The upper two rows show *Volleyball* action original images and imposed with temporal attention weights respectively. The lower two rows show *Throw Discus* action. Our temporal attention module can automatically highlight important frames and avoid irrelevant frames corresponding to non-action poses or background.

5 Conclusion

In this work, we develop a novel spatial-temporal attention mechanism for the task of video action recognition, demonstrating the efficacy and showing state-of-the-art performance across three publicly available datasets. Also, we introduce a set of regularizers that ensure our attention mechanism attends to coherent regions in space and time, further improving the performance and increasing the model interpretability. Moreover, we qualitatively and quantitatively show that our spatio-temporal attention is able to localize discriminative regions and important frames, despite being trained in a purely weakly-supervised manner with only classification labels.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *ICLR*, 2015.

Kyunghyun Cho, Aaron Courville, and Yoshua Bengio. Describing multimedia content using attention-based encoder-decoder networks. *IEEE Transactions on Multimedia*, 2015.

Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. In *NIPS*, 2017.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Rohit Girdhar and Deva Ramanan. Attentional pooling for action recognition. In *NIPS*, 2017.

Yoav Goldberg. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420, 2016.

Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. Deep unordered composition rivals syntactic methods for text classification. In *ACL*, volume 1, pp. 1681–1691, 2015.

Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip HS Torr. Learn to pay attention. *ICLR*, 2018.

Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. `http://crcv.ucf.edu/THUMOS14/`, 2014.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. *ICML*, 2017.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pp. 1097–1105, 2012.

Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, 2011.

Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. *CVPR*, 2018.

Aravindh Mahendran and Andrea Vedaldi. Salient deconvolutional networks. In *ECCV*, 2016.

Mathew Monfort, Bolei Zhou, Sarah Adel Bargal, Alex Andonian, Tom Yan, Kandan Ramakrishnan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *arXiv preprint arXiv:1801.03150*, 2018.

RS Ramprasaath, D Abhishek, V Ramakrishna, C Michael, P Devi, and B Dhruv. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CVPR*, 2016.

Ronald A Rensink. The dynamic representation of scenes. *Visual cognition*, 7(1-3):17–42, 2000.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. AAAI, 2018.

Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 1992.

Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. Action recognition using visual attention. *arXiv preprint arXiv:1511.04119*, 2015.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NIPS*, 2015.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. *ICML*, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

Atousa Torabi and Leonid Sigal. Action classification and highlighting in videos. *arXiv preprint arXiv:1708.09522*, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016a.

Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *CVPR*, 2017.

Yilin Wang, Suhang Wang, Jiliang Tang, Neil O'Hare, Yi Chang, and Baoxin Li. Hierarchical attention network for action recognition in videos. *arXiv preprint arXiv:1607.06416*, 2016b.

Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. Learning to track for spatio-temporal action localization. In *CVPR*, 2015.

Huijuan Xu, Abir Das, and Kate Saenko. R-C3D: region convolutional 3d network for temporal activity detection. In *ICCV*, 2017.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pp. 2048–2057, 2015.

Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. In *CVPR*, 2016.

Gang Yu and Junsong Yuan. Fast action proposals for human action detection and search. In *CVPR*, 2015.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014.

Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. In *ECCV*, 2016.

Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *ICCV*, 2017.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.

Bolei Zhou, Alex Andonian, and Antonio Torralba. Temporal relational reasoning in videos. In *ECCV*, 2018.

A LOG-CONCAVE SEQUENCE

In probability and statistics, a unimodal distribution is a probability distribution that has a single peak or mode. If a distribution has more modes it is called multimodal. The temporal attention weights are a univariate discrete distribution over the frames, indicating the importance of the frames for the task of classification. In the context of activity recognition, it is reasonable to assume that the frames that contain salient information should be consecutive, instead of scattered around. Therefore, we would like to design a regularizer that encourages unimodality. To this end, we introduce a mathematical concept called the log-concave sequence and define the regularizer based on it.

We first give a formal definition of the unimodal sequence.

Definition 1. *A sequence $\{a_i\}_{i=1}^n$ is unimodal if for some integer m,*

$$\begin{cases} a_{i-1} \leq a_i & \text{if } i \leq m, \\ a_i \geq a_{i+1} & \text{if } i \geq m. \end{cases}$$

A univariate discrete distribution is unimodal, if its probability mass function forms a unimodal sequence. The log-concave sequence is defined as follows.

Definition 2. *A non-negative sequence* $\{a_i\}_{i=1}^n$ *is log-concave if* $a_i^2 \geq a_{i-1}a_{i+1}$.

This property gets its name from the fact that if $\{a_i\}_{i=1}^n$ is log-concave, then the sequence $\{\log a_k\}_{i=1}^n$ is concave. The connection between unimodality and log-concavity is given by the following proposition.

Proposition 1. *A log-concave sequence is unimodal.*

Proof. Rearranging the defining inequality for log-concavity, we see that

$$\frac{a_i}{a_{i-1}} \geq \frac{a_{i+1}}{a_i},$$

so the ratio of consecutive terms is decreasing. Until the ratios decrease below 1, the sequence is increasing, and after this point, the sequence is decreasing, so it is unimodal. □

Given the definition of log-concavity, it is straightforward to design a regularization term that encourages log-concavity:

$$R = \sum_{i=2}^{n-1} \max\{0, a_{i-1}a_{i+1} - a_i^2\}. \tag{11}$$

By Proposition 1, this regularizer also encourages unimodality.

B More datasets and implementation details

B.1 More details on the dataset and experimental setup

HMDB51 and UCF101 The dataset pre-processing and data augumentation are the same as the ResNet ImageNet experiment (He et al., 2016). All the videos are resized to 224×224 resolution and fed into a ResNet-50 pretrained on ImageNet. The last convolutional layer feature map size is $2048 \times 7 \times 7$. The experimental setup for the Moments in Time dataset is the same as HMDB51 and UCF101 except time sequence and image resolution.

Moments in Time For the Moments in Time dataset (Monfort et al., 2018), the videos only have 3 seconds, much shorter than HMDB51 and UCF101. We extract RGB frames from the raw videos at 5 fps. Therefore, the sequence length is $T = 15$. Following the practice in (Monfort et al., 2018) to make all the videos uniform resolution, we resize the RGB frames to 340×256 pixels. When extracting features, we use the ResNet-50 pretrained on ImageNet model using resized images with resolution of 256×256 pixels. The data augmentation is the same as the ResNet ImageNet experiment (He et al., 2016). The feature map size of the last convolutional layer is $2048 \times 8 \times 8$.

THUMOS14 The action detection task of THUMOS'14 (Jiang et al., 2014) consists of 20 classes of sports activities, and contains 2765 trimmed videos for training, while 200 and 213 untrimmed videos for validation and test respectively. Following the standard practice (Yeung et al., 2016; Zhao et al., 2017), we use the validation set as training and evaluate on the testing set. 220 videos in validation set and 212 videos in test set have temporal annotations in 20 classes. This dataset is particularly challenging as it consists of very long videos (up to a few hundreds of seconds) with multiple activity instances of very small duration (up to few tens of seconds). Most videos contain multiple activity instances of the same activity class. In addition, some videos contain activity segments from different classes. Following the standard practice (Xu et al., 2017) to avoid the training ambiguity, we remove the videos with multiple labels. We extract RGB frames from the raw videos at 10 fps. We use the ResNet-50 pretrained on ImageNet and fine-tuned on Thumos14, and the preprocessing and data augmentation of Thumos14 is the same as ImageNet. The last convolutional layer feature map size is $2048 \times 7 \times 7$.

B.2 More implementation details

All the experiments are evaluated on machines with a single Nvidia GeForce GTX 1080Ti GPU. The networks are implemented using the Pytorch library and our code will be publicly available with this paper.

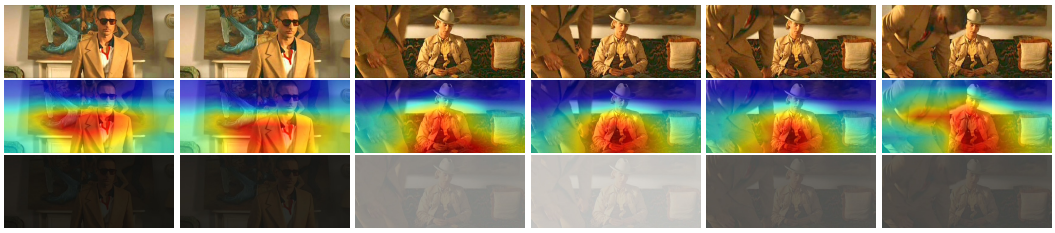


Figure 7: **Multiple actions in one image for video action recognition** The *Sit* action from HMDB51. In the first two frames, there is no sitting action while the spatial attention capture the important area, but the temporal attention can effectively ignore them as the background information. It is interesting that in the last few frames, there is another person trying to sit down, but the visual attention can only capture one sitting person.

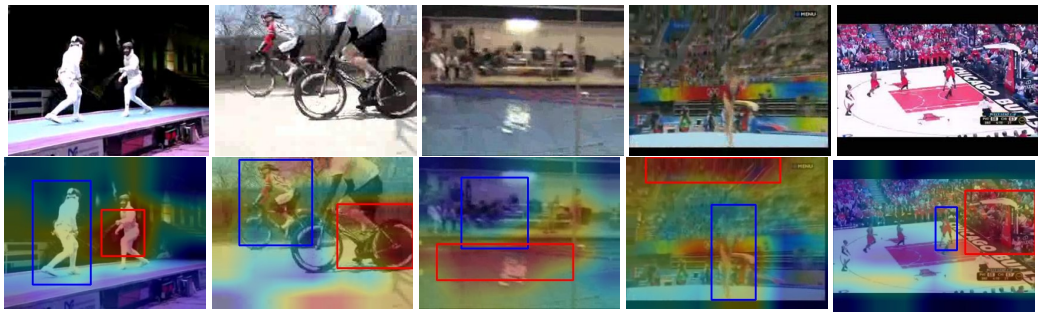
C MORE RESULTS

C.1 MORE SPATIAL TEMPORAL ATTENTION RESULTS

Fig. 7 shows more results on spatial temporal attention.

Cliff Diving Floor Gymnastics Ice Dancing Horse Riding Pole Vault

Figure 8: **Examples of spatial attention for action localization.** (Best viewed in color.) Blue bounding boxes represent ground truth while the red ones are predictions from our learned spatial attention. Our spatial attention mechanism is able to focus on important part of the action, while the ground truth bounding boxes labels focus on the entire human pose. As in the training stage, the ground truth bounding boxes are not used, and the model can only depend on crucial spatial area rather than the entire action to make prediction. For actions with object interactions, such as *Horse Riding* and *Pole Vault*, the ground truth box focuses on human pose while the model focuses on objects (such as Horse, Pole) as well.



Fencing Biking Diving Floor Gymnastics Basketball Dunk

Figure 9: **Failure cases for spatial localization.** (Best viewed in color.) In the ground truth bounding boxes, there is only one bounding box in human action for each frame but there may be more than one person performing the same action. Typical IOU=0 case is that our attention focuses on the unlabeled human action, such as *Fencing* and *Biking* shown here. Strong motion blur also leads to failure cases, such as the *Diving* and *Floor Gymnastics* frames shown here. The diving and gymnastics poses are highly motion blurred so the spatial attention focuses on the swimming pool and audiences respectively. Some of the important background information also leads to failure cases, such as the basketball frame in the *Basketball Dunk* shown here.

(a) (b) (c) (d) (e) (f)

Figure 10: **Failure example for temporal attention localization.** A sequence of *Tennis Swing* action from one video of THUMOS14. All temporal localization is correctly localized except the frame (b). The labeled action starts from frame (b) but our temporal attention module still assigns a low importance score.

C.2 More spatial localization results

Fig. 8 shows more spatial localization results. Fig. 9 shows some failure cases.

C.3 More temporal localization results

Fig. 10 shows more results on temporal localization with our temporal attention.