Attending to Distinctive Moments: Weakly-supervised Attention Models for Action Localization in Video

Lei Chen
Simon Fraser University
chenleic@sfu.ca

Mengyao Zhai
Simon Fraser University
mzhai@sfu.ca

Greg Mori
Simon Fraser University
mori@cs.sfu.ca

Abstract

We present a method for utilizing weakly supervised data for action localization in videos. We focus on sports video analysis, where videos contain scenes of multiple people. Weak supervision gathered from sports website is provided in the form of an action taking place in a video clip, without specification of the person performing the action. Since many frames of a clip can be ambiguous, a novel temporal attention approach is designed to select the most distinctive frames in which to apply the weak supervision. Empirical results demonstrate that leveraging weak supervision can build upon purely supervised localization methods, and utilizing temporal attention further improves localization accuracy.

1. Introduction

In this paper we present an approach for utilizing weakly supervised data to learn models for action localization in sports videos. Action localization is a core problem in video analysis – determining which person in a scene is performing an action of interest. Within the context of sports video analysis, the problem is particularly challenging. Sports scenes typically consist of multiple, interacting people. The visual appearance of people is similar because of team uniforms. Inter-person occlusion is prevalent.

However, sports videos often come with a great amount of data in corresponding media. While much of these data are only weak supervision for action localization. For example, there exists a large amount of 'play-by-play' about sports videos from corresponding websites, itemizing sequentially the events happened in a game. There are challenges in utilizing these meta-data for training action localization methods. First, the labels are weak, scene-level labels. In sports scenes there are multiple people present, and we need to disambiguate which person the label should apply to.

A second important challenge is about temporal information in labeling. Consider the example in Fig. 1. Not all moments in an action are equally distinctive. As an example, consider the basketball action labels of *layup* and *dunk*. Video clips with these labels may share a lot of similarity, with a player dribbling toward the basket surrounded by defenders. Determining which moments in time are more distinctive can help train better quality classifiers.

Figure 1: When trying to localize the player of a certain action with weak supervision, it is important to focus on the stereotypical poses that are easier to learn. Many of the player poses in the frames look similar. However, in the bottom two frames, the player taking the shot takes a distinctive pose. Our method uses an attention model to focus on the distinctive poses for learning an action model.

Our approach to address these challenges is to develop a weakly supervised deep learning model for action localization. Fully supervised training data, specifying the bounding boxes of people performing actions, is expensive to acquire. We propose a novel attention model-based loss function that is used to find the frames that are indicative of an action in weakly supervised data. As an example, we show a method for collecting and utilizing information from basketball videos and corresponding *box score* annotations that

specify the events that took place during the game. These annotations are plentiful, but are imprecise in time and do not contain spatial annotation. We show that using these weakly supervised data is effective. They can be used in conjunction with a small amount supervised data to improve the quality of action localization, showing that a little bit of supervision can go a long way toward producing accurate action localization.

2. Related Work

We develop a method for utilizing attention models for weakly supervised learning of action localization in videos. Below, we review closely related work in these areas.

Action localization: A variety of methods exist for analyzing videos according to human action labels. These methods range from video-level classification on unconstrained Internet video, to methods that spatio-temporally localize human actions. In concert with improvements in deep learning for object recognition, state of the art methods utilize deep learning approaches to learn convolutional features. In action recognition, the dense trajectories of Wang et al. [26], the best hand-crafted features, have achieved impressive performance on many tasks. However, these have yielded to deep learning approaches. In video-level action recognition, Simonyan and Zisserman [21] presented a two-stream convolutional architecture for merging image and optical flow data as input sources. Zha et al. [31] compute deep learned image-based features for each frame, and study strategies for aggregation, obtaining impressive results on TRECVID MED retrieval. Karpathy et al. [6] and Tran et al. [24] learn spatio-temporal filters in a deep network.

Temporal localization of actions has a long history in the computer vision literature. Seminal work includes Yamato et al. [29], who model actions using hidden Markov models (HMMs). A more recent example in this vein is Tang et al. [23], who extend HMMs to model the duration of each hidden state in addition to the transition parameters of hidden states. Discriminative models include those based on key poses and action grammars [13, 25, 15].

In our work we predict spatial action localizations. Classic methods include Ke et al. [7] who match templates of action to crowded video scenes. Lan et al. [8] jointly detect and recognize actions in videos based on a figure-centric visual word representation. Recent work has switched toward methods based on analyzing action tube proposals. Gkioxari and Malik [3] train SVMs for actions on top of deep learned features, and further link them in time for spatio-temporal action detection. A set of approaches have built in this direction, improving methods for producing frame-level action proposals, linking, and analyzing them to produce action labels [27, 14, 19].

Weakly-supervised learning: The prevalence of partially annotated data for computer vision tasks has inspired a swath of research. This includes methods for object and action recognition. Fundamental work for the problem of action recognition was done by Laptev et al. [9], who built datasets for action recognition by considering surrogate movie script data. Rohrbach et al. [18] find corresponding regions to each object that appear as a phrase in the sentence description. Jayaraman et al. [5] learn representations based on assumptions regarding changes in neighbouring video frames. Shah et al. [20] build a generative model of video events. Ma et al. [11] extracts hierachical space time segments from videos without supervision and uses them for action recognition and localization. Mosabbeb et al. [12] proposes a matrix completion approach for weakly-supervised action recognition and localization. Siva et al. [22] presents a MIL algorithm that locates the action of interest spatially and temporally by globally optimising both inter- and intra-class distance.

Bojanowski et al. [1] explore joint localization of people and actions in movie clips. The problem setting is similar to ours where only video-level description is provided. They model the problem as assigning zero-one values to latent indicators under the constraint that paired actions and actors correspond to the same instance in the frame. In follow up work [2], temporal localization of an ordered set of descriptions corresponding to a video clip is done. A mapping is learned between text representation and image presentation and an allocation assigning frames to descriptions at the same time. Our problem and approach differ in that we take into consideration actions that are not described explicitly and conduct inference in a multi-person sports setting. Lu et al. [10] examine the problem of identifying all the players in a basketball game. A graphical model is built on top of player tracks to identify players. Although they take advantage of labeled player identities, they also add results from a supervised model into training, leading to a semi-supervised approach.

Attention models: Pioneering work on computational spatial attention models for images was done by Itti et al. [4]. Recently, such models have garnered attention for their ability to focus computational and modeling resources toward important image/video elements.

This can take many forms. One simple idea is to score a set of previously processed candidates e.g. simple dense overlapping regions or those based on objectness. An example for image captioning is the work of Xu et al. [28] where an attention model is added into a image captioner so that it will look at different parts of the image as it produces the output sentence. For video data, Yao et al. [30] develop an LSTM for video caption generation with soft temporal attention.

There is also previous work in using attention models to decide the key player in sports video. Ramanathan et

al. [16] propose a network to classify several actions in basketball videos. With the attention model, the action of the key player is paid special attention when making a prediction for a scene. Different from this work, we directly learn to perform action localization, and our attention model is utilized for training a weakly-supervised system rather than part of a frame-level predictor.

3. Method

We start by introducing a basic form of our weakly-supervised model where only clip-level supervision is provided. The training data are a set of sports video clips, each with a label specifying the key action being performed by a player. We will learn a model to localize these key actions, by finding the people performing similar, distinct actions within frames of the same action category. In such a strict weakly-supervised setting, the model will face a lot of challenges from intra-class variation and noise in the training data. These issues are tackled by our extension to the base model with semi-supervision as well as a temporal attention mechanism.

3.1. Weakly-supervised Action Localization

For each frame in a clip, we have weak supervision – essentially, we know that there exists a person in each frame of this clip who is (at some point in time) performing the specified action. We first run a player detector to obtain the top K person detections $\{\mathbf{x}_i\}_{i=1}^{K}$ in each frame. All the detections are sent to the same deep network for action classification. The categories consist of all the action classes plus a background class. In a conventional supervised case the training of a classifier will also take in the action labels a_i. Action models could be trained using standard approaches, such as negative log-likelihood loss:

$$loss(\mathbf{x}_i, a_f) = -\log(\text{softmax}(\text{CNN}_a(\mathbf{x}_i, a_f))) \qquad (1)$$

where $\text{CNN}_a(\mathbf{x}_i, a_f)$ represents the predicted score for action class a_f when feeding detection $\mathbf{x}_i$ into network, with a softmax for normalizing scores across categories.

In the weakly-supervised setting, no such instance-level action annotation is provided; action label assignments should be inferred within the training process. Since we are looking for the specific player performing the given action, only one player should score high in the given action category, while all other players should not. Moreover, the other players should score high in the background class instead.

We formulate this new weakly-supervised loss as follows. For a frame f, if we denote its corresponding (weakly-supervised) action class by a_f, the loss function takes the form:

$$F = \sum_f \min_i \{loss(\mathbf{x}_i, a_f) + \sum_{j \neq i} loss(\mathbf{x}_j, bg)\} \qquad (2)$$

where $loss(\mathbf{x}_i, a_f)$ is the loss of the i-th detection for action class a_f, and $loss(\mathbf{x}_j, bg)$ is the loss of the j-th detection for the background class. To compute error gradients for back-propagation, we first must infer which person in a scene should be assigned as performing the action. Specifically, for each frame in stochastic gradient descent we conduct inference based on the current network weights. We assign one player with the frame action label and the rest with the background label so that the sum of the above losses is minimized. This assignment is used in calculating gradients for back-propagation. This inference is computationally efficient since the assignment can be done via a simple linear search.

Note that in this learning objective, the background samples are equally important as the action samples. The background samples are abundant and provide reliable information about what the given action should *not* be like. This is important in the weakly-supervised setting and especially for categories with fewer examples. It may be hard to characterize such actions directly by looking for distinctive actions shared within the class, because any slight variation will result in a big challenge with limited examples. But the examples of the background class are shared across all categories and serve as a essential clue to find the real target action.

3.2. Semi-supervised Action Localization

The purely weakly-supervised method presents a very challenging learning problem. We have the appearance of each player in a number of frames of each action category and need to determine which player is the "correct" one in each frame. This problem is susceptible to model drift – if we believe erroneously that certain similar-posed people in many frames correspond to the "correct" action, the model will reinforce this belief and learn an incorrect model. In essence, the weakly-supervised localization above is sensitive to initialization and unfortunate co-occurrences among background poses.

This can be remedied by adding a small portion of supervision to guide the initial model to choose appropriate persons in each frame as corresponding to the action category. We utilize a similar formulation for the semi-supervised case as in the weakly-supervised one. The loss function is the same, except that for a small portion of the frames, the loss simply uses standard supervised loss:

$$\begin{aligned} F = &\sum_{f \in \mathbb{S}} \{loss(\mathbf{x}^f_{i^*}, a_f) + \sum_{j \neq i^*} loss(\mathbf{x}^f_j, bg)\} \\ &+ \sum_{f \in \mathbb{W}} \min_i \{loss(\mathbf{x}^f_i, a_f) + \sum_{j \neq i} loss(\mathbf{x}^f_j, bg)\} \end{aligned} \qquad (3)$$

where $\mathbf{x}^f_{i^*}$ is the ground truth detection for the specified action class in frame f, $\mathbb{S}$ is the collection of frames with full

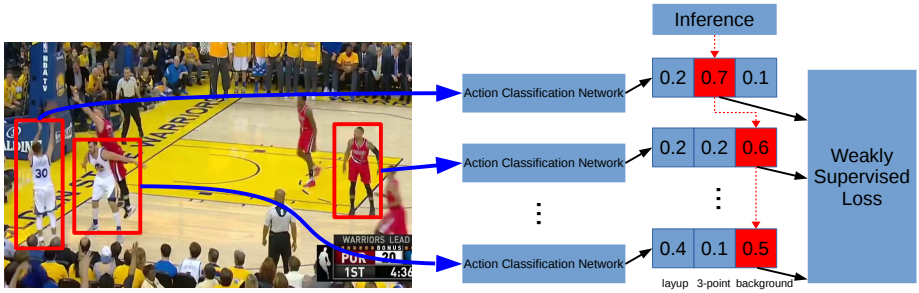


Figure 2: Shown above is the network structure for weakly-supervised action localization. The top detections from a player detector (Faster-RCNN) are fed into identical action classifier networks. This generates action scores for all action classes, plus the *background* category. For each frame, one detection should be classified into the action class of the frame. The remaining detections should be classified as *background*. In the weakly supervised setting, the player performing the action is not given at training time and has to be inferred during learning. In a semi-supervised setting, a portion of the frames come with full labeling including which player performs the action.

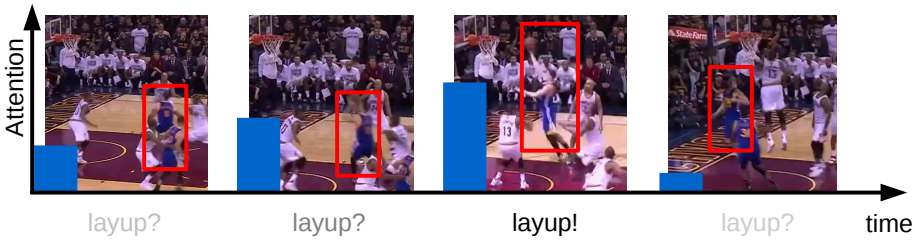


Figure 3: An illustration for the intuition behind the temporal attention model. The stereotypical action poses are very distinct from the background players and are shared by almost all clips of certain actions. Focusing on such core poses would lay a good foundation for the model to understand less distinctive cases.

ground truth supervision, and $\mathbb{W}$ contains the remainder of the frames.

3.3. Localization with Temporal Attention

The intuition behind weakly-supervised localization is the assumption that different players take similar poses when performing the same action. This is mostly true for the key moments of each action – for instance, shots, dunks, layups, etc. each contain moments of similarity within each category. However, the temporal definition of actions are vague, precise supervision is impractical, and it is unlikely that the training clips contain only those key moments. This leaves a number of frames where the target player pose varies greatly, due to the variability in actions before/after the key moments.

This might not be a problem in the supervised setting where positive examples are abundant. The ambiguous or improperly labeled data could be overcome with quantities of correctly labeled positive data. For the weakly-supervised or semi-supervised case, this problem is much harder. In a frame where none of the players has the desired pose, if the model has to choose one of the detections as a positive example of a certain action class, it will likely significantly harm performance.

To alleviate such problems, we introduce the following temporal attention model. The attention model encourages the localization network to put more emphasis on the easily recognizable or distinguishable examples by assigning a weight to the loss of every frame. During training, the attention model will learn to assign low weights to frames incurring high error. Since stereotypical action poses are easier to distinguish, this should focus training on more appropriate examples.

The attention value is computed for each frame, and then normalized over each clip with a softmax. The attention value is computed from holistic frame-level features and the responses of all action classifiers from all players. A multi-layer perceptron takes both inputs to generate the final attention value, using the loss function:

$$F = \sum_{f\in\mathbb{S}} w_f \cdot [loss(\mathbf{x}^f_{i^*}, a_f) + \sum_{j\neq i^*} loss(\mathbf{x}^f_j, bg)] + \sum_{f\in\mathbb{W}} w_f \cdot [\min_i \{loss(\mathbf{x}^f_i, a_f) + \sum_{j\neq i} loss(\mathbf{x}^f_j, bg)\}] \tag{4}$$

where w_f is the attention of f-th frame, given by

$$w_f = \text{softmax}(\phi(\mathbf{f}_f, \mathbf{r}_f)) \tag{5}$$

$$\mathbf{f}_f = \text{CNN}_f(I_f) \tag{6}$$

$$\mathbf{r}_f = [\text{CNN}_a(\mathbf{x}_1, a_1), \text{CNN}_a(\mathbf{x}_1, a_2), \ldots, \text{CNN}_a(\mathbf{x}_2, a_1), \ldots, \text{CNN}_a(\mathbf{x}_K, a_N)] \tag{7}$$

where $\mathbf{f}_f$ is the holistic frame-level feature generated from

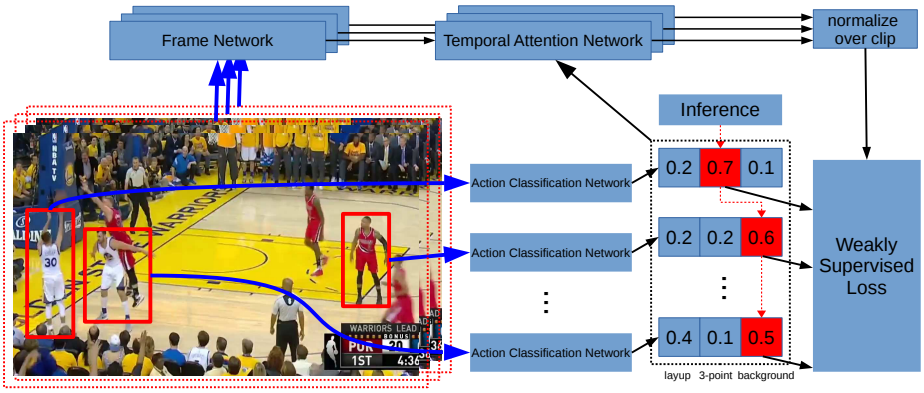


Figure 4: Shown above is the temporal model structure. On top of the base model, a frame feature extraction network is introduced. Both frame feature and action scores are sent to the temporal networks and generate a weight for the frame indicating its importance in the training.

CNN_f, a frame-level network jointly trained with the localization model, $\mathbf{r}_f$ is the responses of all action classifiers for all players in the frame and ϕ is a multi-layer perceptron.

In summary, our approach learns an action model by utilizing fully and weakly-labeled data. For the weakly-labeled data, we define a loss function that selects the highest scoring person in each frame according to the specified action category, balanced against background labels for all other people. Further, an attention model is applied to each video clip, allowing the model to focus on the most distinctive poses for each category.

4. Experiments

We conduct experiments on action grounding in sports video. We collect a novel dataset, mining structured text descriptions of basketball games along with associated video footage. Experiments evaluate the effectiveness of our method in a semi-supervised setting, and verify the effectiveness of the attention model.

Dataset: We collected a new dataset to test our action grounding system. The training set contains 746 clips from 13 NBA basketball games; a separate test set has 398 clips from 6 games. Clips are extracted according to corresponding play-by-play descriptions from *espn.com*. Each clip is one second long and covers the action described in the play-by-play. All clips are actions performed by a player of the Golden State Warriors. We sample 9 frames from each clip, for a total of 6714 frames in the training set and 3582 in the testing set. The clips fall into 5 categories: free-throw, layup, dunk, two-point and three-point. The label assignment is purely according to the play-by-play description with no manual adjustment. We will release the data annotations to enable comparisons.

Pre-processing: We train a player detector using the Faster-RCNN network[17] on the NCAA dataset [16]. Note that the camera angle and resolution of NCAA games differ from that of NBA games, leading to some erroneous detections. Frames of 40 NCAA games are used in training the detector. We run the detector over on our dataset and take the top 10 detections, resized to 256×256, as the input to our localization network.

Semi-supervision: Our experiments are done in a semi-supervised setting where the grounding label (key player location) of a small subset is provided in training. For frames without a detection whose IoU with ground-truth is greater than 0.5, the loss function will take all candidates as background examples. In the experiments below, number of supervision means the number of clips per action category whose grounding labels are provided in training. Up to 5 supervised clips per category (225 frames in total) are used in the experiments. Since the choice of fully supervised clips influences the performance, we run all experiments for 5 repeats with different fully supervised clips, reporting mean and standard deviation.

Network structure details: The action classification and frame feature networks both use the Alexnet structure. The input to the temporal attention model are the fc7 layer of the frame network and the fc8 layer of the localization network. In the attention model, both inputs are first each sent to a fully connected layer resulting in vectors of the same dimension. The two vectors are added, fed into two fully connected layers to produce a scalar before being normalized across frames from the same clip. The frame network is initialized from the Caffe ImageNet pre-trained model. For the action localization network, we pre-train on our fully supervised data for 1000 iterations and then fine-tune on the whole dataset in the semi-supervised setting. All models are trained with a learning rate of 0.001, with 27 frame mini-batches (3 clips).

Evaluation: We annotate the ground truth bounding boxes for frames where the player performing the action is visible. These bounding box annotations are independent

Model	Supervised Only					Semi-Supervised					Semi-Supervised with Attention				
#supervision	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
free-throw	0.541	0.773	0.837	0.895	0.912	0.635	0.940	0.942	0.949	0.948	0.756	0.942	0.943	0.944	0.944
dunk	0.231	0.356	0.397	0.448	0.497	0.167	0.422	0.533	0.547	0.654	0.207	0.516	0.571	0.612	0.646
layup	0.213	0.273	0.313	0.375	0.407	0.200	0.402	0.472	0.521	0.579	0.275	0.510	0.535	0.563	0.582
two-point	0.214	0.260	0.331	0.385	0.406	0.218	0.405	0.477	0.523	0.585	0.305	0.529	0.574	0.598	0.609
three-point	0.216	0.333	0.364	0.416	0.437	0.256	0.444	0.525	0.616	0.659	0.261	0.565	0.618	0.656	0.673
overall	0.268	0.378	0.426	0.481	0.504	0.297	0.508	0.573	0.630	0.675	0.353	0.607	0.646	0.673	0.687
overall std	0.059	0.030	0.010	0.025	0.020	0.132	0.122	0.140	0.093	0.003	0.125	0.048	0.031	0.018	0.015

Table 1: Action grounding accuracy for all models on ***transductive set*** with different number of supervision. Both semi-supervised model and semi-supervised with attention model outperforms their initialization model trained on only on supervised data, successfully extracting information from weakly-supervised data. Semi-supervised models with attention demonstrate best performance among all models in both grounding accuracy and stability.

Model	Supervised Only					Semi-Supervised					Semi-Supervised with Attention				
#supervision	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
free-throw	0.644	0.857	0.939	0.966	0.958	0.616	0.986	0.985	0.988	0.988	0.799	0.986	0.987	0.987	0.986
dunk	0.356	0.422	0.554	0.585	0.605	0.320	0.573	0.713	0.728	0.802	0.405	0.764	0.771	0.791	0.795
layup	0.229	0.306	0.366	0.409	0.464	0.277	0.509	0.593	0.630	0.667	0.330	0.630	0.659	0.677	0.674
two-point	0.295	0.367	0.445	0.522	0.543	0.321	0.539	0.627	0.683	0.748	0.417	0.710	0.729	0.742	0.742
three-point	0.284	0.431	0.453	0.524	0.545	0.380	0.570	0.630	0.723	0.797	0.399	0.756	0.774	0.788	0.786
overall	0.384	0.517	0.579	0.636	0.652	0.415	0.670	0.727	0.775	0.822	0.508	0.793	0.807	0.818	0.817
overall std	0.065	0.058	0.028	0.024	0.026	0.192	0.124	0.135	0.082	0.005	0.168	0.022	0.011	0.003	0.002

Table 2: Action grounding accuracy for all models on ***inductive set*** with different number of supervision. Conclusions drawn from transductive set still applies with no obvious over-fitting.

of the candidate bounding boxes from the automated player detector. We follow the previously constructed loss function to find the target player, specifically, we take the candidate detection $\mathbf{x}_i^*$ as the prediction from the model, where i^* is defined by

$$i^* = \arg\min_i \{loss(\mathbf{x}_i^f, a_f) + \sum_{j \neq i} loss(\mathbf{x}_j^f, bg)\} \qquad (8)$$

During testing, we ignore frames where the desired target is not visible. For the rest of the frames, if the highest score candidate $\mathbf{x}_i^*$ in the frame has an IoU greater than 0.5 with the ground truth box we take it as correct. Since we provide localization labels for only a small portion of the training data (up to 225 frames), we first test on the training data where the instance labels are not provided (transductive setting). In addition, the trained models are also tested on the testing set whose frames are not used in the training phase (inductive setting). All of the 225 frames that are potentially used for supervision are excluded from test for fair comparison.

4.1. Semi-supervised Localization

In the semi-supervised setting, we explore the influence of the number of supervised examples. Supervision from 1 clip per action to a maximum of 5 clips per action is provided. To combat model drift, we "burn in" the network by pre-training with only the fully-labeled data. The model is first trained only on the supervised part for 1000 epochs and later fine-tuned with all the data.

Results are presented in Tab. 1. The semi-supervised models consistently outperform their fully-supervised equivalents, demonstrating the utility of our weakly-supervised approach.

4.2. Temporal Attention

The same settings are used for the temporal attention model (1 to 5 fully labeled examples). Models are fine-tuned for 30000 iterations from the fully-supervised model trained on the supervised part only. Results in Tab. 1 indicate that such temporal guidance is helpful in all cases for both performance and stability, especially with little supervision provided. The larger performance gap at lower supervision demonstrates the ability of the temporal attention model to extract the right information from complicated and challenging situations. As the amount of the amount supervision grows, advantages moderate but remain. The likely reason is that when both model can easily distinguish the easy targets with the help of more supervision, the gap between performance relies more on their ability to recognize less distinguishable examples the attention model does not explicitly emphasize. It is worth noting that the temporal attention model does not over-fit on the easy examples, with more supervision its performance also grows like the semi-supervised model.

4.3. Inductive Test

We also present inductive results on new data not used as weak supervision. The results in Tab. 2 demonstrates the same conclusions as the transductive ones. Note that the

Model	Supervised Only					Semi-Supervised from Supervised					Semi with Attn from Supervised				
#supervision	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
action accuracy	0.326	0.384	0.421	0.482	0.493	0.603	0.659	0.664	0.692	0.703	0.623	0.707	0.714	0.711	0.720
bbox accuracy	0.393	0.495	0.562	0.624	0.633	0.402	0.649	0.709	0.763	0.812	0.496	0.784	0.798	0.808	0.810
overall accuracy	0.179	0.270	0.323	0.388	0.397	0.305	0.514	0.541	0.590	0.620	0.383	0.608	0.624	0.632	0.637

Table 3: Action grounding on inductive test with no frame label provided. Action labels(from five action category) and bounding box are infered simultaneously. Bounding box accuracy is similar to results with frame label provided, indicating model's ability to distinguish key player. Due to the training procedure and confusion in the label set, the action accuracy is not as good.

clips in training	transductive	inductive
200	0.566 $\pm$0.023	0.723 $\pm$0.026
400	0.633 $\pm$0.020	0.790 $\pm$0.013
full dataset	0.687 $\pm$0.015	0.817 $\pm$0.002

Table 4: Grounding accuracy for semi-supervised different number of clips used in traing, all with 5 clips per action category as supervision

#supervision	1	2	3	4	5
overall	0.238	0.278	0.316	0.337	0.354
overall std	0.103	0.071	0.041	0.021	0.026

Table 5: Overall grounding accuracy and std for models without background term in loss trained on the fully supervised data only with different number of supervision.

higher performance (for all methods) is due to the higher detection recall and higher percentage of (easier) free-throw frames in the inductive test set.

In the experiments above, frame level labels are provided to perform the grounding. We evaluate the ability of our inference criterion to predict simultaneously the action class as well as the bounding box as follows:

$$i^*, a_f^* = \arg\min_{i,a_f} \{loss(\mathbf{x}_i^f, a_f) + \sum_{j \neq i} loss(\mathbf{x}_j^f, bg)\} \quad (9)$$

where a_f is chosen from the five action categories. Presented in Tab. 3 are: action accuracy measuring the percentage of frames where the action prediction is correct, bounding box accuracy measuring the percentage of frames where the predicted bounding box has an IoU over 0.5 with groundtruth, and overall accuracy measuring the percentage of frames where both the action and bounding box are correctly predicted. Results show that bonding box accuracy is close to the grounding performance with frame labels provided, indicating the model's ability to distinguish the key player. The action accuracy is not as good since the model is explicitly trained for such a task. Note that the two-point classes are not carefully distinguished from the dunk and layup classes in the play-by-play texts, leading to certain confusion across categories.

4.4. Amount of Weakly-supervised Data

The advantage we expect from semi-supervised learning is its ability to learn from the weakly-supervised data. It would be beneficial if the model's performance can improve by just taking in more weakly-supervised data which is less demanding to collect. Here we test the model's ability to utilize weakly-supervised data by training with 200 clips, 400 clips and all 746 clips, all with fully supervised grounding labels for 5 clips per action category. The results present in Tab. 4 indicate sustained improvement with more weakly-labeled data, validating the potential from semi-supervised learning.

4.5. Background Class in the Loss Function

One observation we take advantage of when proposing the approach is that only one of the players should be performing the labeled action while others should be performing none of the possible actions. In comparison, conventional multiple instance learning methodology would use a weaker assumption that at least one of the players is performing the target action.

We evaluate the effect of the background model. Tab. 5 shows results of a fully-supervised model without the background term in the loss function (c.f. Tab. 2 left); semi-supervised models fail to learn in this setting. The reduced performance is likely due to the lack of use of background examples in correcting weak supervision, combating data noise, and shaping the decision boundary with positive examples.

5. Conclusion

We demonstrated that attention models can be used to select distinctive frames for learning action localization from weakly supervised data. We created a dataset of weakly supervised action data by combining basketball video data with action labels extracted from text descriptions of the games. The weak supervision lacks spatial and precise temporal localization of action, but our model is capable of overcoming these challenges.

Our experiments explored the use of weakly supervised action data in isolation. Further, the weakly supervised data was used to augment a fully supervised dataset to improve action localization results. Ablation studies showed that the attention model and learned weak supervision are effective means of increasing action localization performance. Further study demonstrates the effectiveness of the semi-

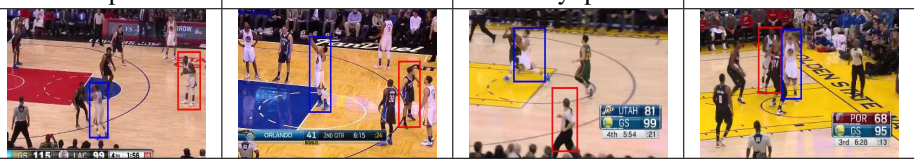
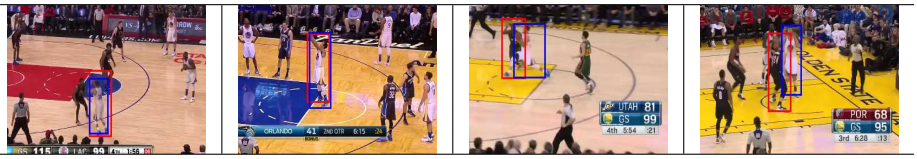
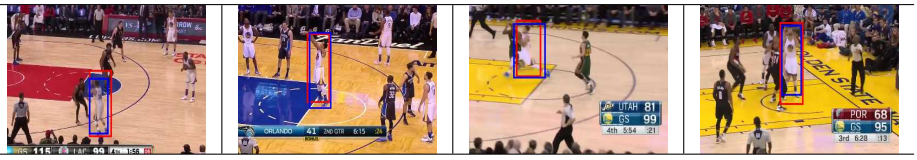
	three point	free-throw	layup	dunk
supervised only				
semi-supervised				
semi-supervised attention				

Table 6: Visualization of player detection from different models. Blue box is the ground truth and red box is the highest score candidate. Adding semi-supervision and temporal attention allows the model to extend knowledge from supervised examples to the whole dataset and recognize distinctive moments that vary from the limited cases from supervising clips.

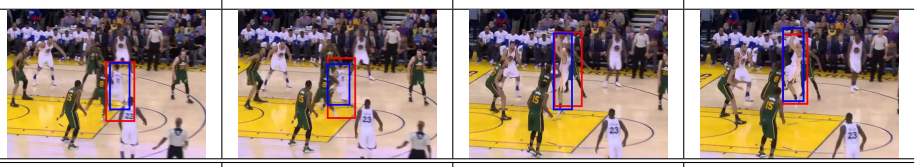
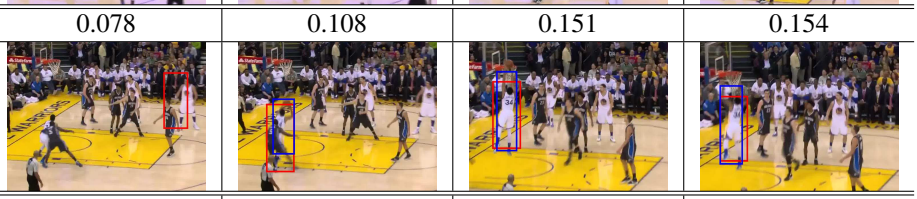
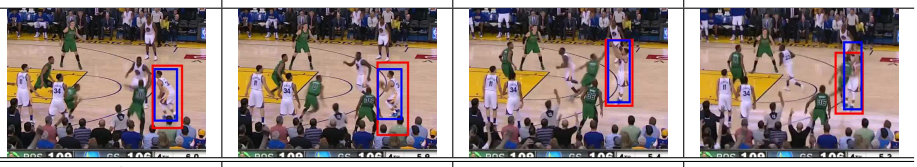
Attention	0.108	0.078	0.145	0.141
Frame				
Attention	0.078	0.108	0.151	0.154
Frame				
Attention	0.059	0.133	0.172	0.078
Frame				
Attention	0.167	0.170	0.091	0.088
Frame				

Table 7: Visulization of attention values for frames. Blue box is the ground truth and red box is the highest score candidate. 4 frames from each clip are shown with their frame attention value on top. The attention scores demonstrate the model's ability to attend to distinctive moments. The last row also indicates that low attention values are assigned to cases where the model is confused, which helps the model to be robust to uncertainty in the training process.

supervised approach to extract information from weakly-supervised data.

References

[1] P. Bojanowski, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic. Finding actors and actions in movies. In *Interna-*

tional Conference on Computer Vision (ICCV), 2013.

[2] P. Bojanowski, R. Lajugie, E. Grave, F. Bach, I. Laptev, J. Ponce, and C. Schmid. Weakly-supervised alignment of video with text. In *International Conference on Computer Vision (ICCV)*, 2015.

[3] G. Gkioxari and J. Malik. Finding action tubes. In *Computer Vision and Pattern Recognition (CVPR)*, 2014.

[4] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *PAMI*, 20(11):1254–1259, 1998.

[5] D. Jayaraman and K. Grauman. Slow and steady feature analysis: Higher order temporal coherence in video. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[6] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2014.

[7] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. In *International Conference on Computer Vision (ICCV)*, 2007.

[8] T. Lan, Y. Wang, and G. Mori. Discriminative figure-centric models for joint action localization and recognition. In *International Conference on Computer Vision (ICCV)*, 2011.

[9] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition (CVPR)*, 2008.

[10] W.-L. Lu, J.-A. Ting, J. J. Little, and K. P. Murphy. Learning to track and identify players from broadcast sports videos. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1704–1716, 2013.

[11] S. Ma, J. Zhang, N. Ikizler-Cinbis, and S. Sclaroff. Action recognition and localization by hierarchical space-time segments. In *International Conference on Computer Vision (ICCV)*, 2013.

[12] E. A. Mosabbeb, R. Cabral, F. De la Torre, and M. Fathy. Multi-label discriminative weakly-supervised human activity recognition and localization. In *Asian Conference on Computer Vision (ACCV)*, 2014.

[13] J. C. Niebles, C.-W. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *European Conference on Computer Vision (ECCV)*, 2010.

[14] X. Peng and C. Schmid. Multi-region two-stream R-CNN for action detection. In *European Conference on Computer Vision (ECCV)*, 2016.

[15] H. Pirsiavash and D. Ramanan. Parsing videos of actions with segmental grammars. In *Computer Vision and Pattern Recognition (CVPR)*, 2014.

[16] V. Ramanathan, J. Huang, S. Abu-El-Haija, A. Gorban, K. Murphy, and L. Fei-Fei. Detecting events and key actors in multi-person videos. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[17] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems (NIPS)*, 2015.

[18] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision (ECCV)*, 2016.

[19] S. Saha, G. Singh, M. Sapienza, P. H. S. Torr, and F. Cuzzlion. Deep learning for detecting multiple space-time action tubes in videos. In *British Machine Vision Conference (BMVC)*, 2016.

[20] S. Shah, K. Kulkarni, A. Biswas, A. Gandhi, O. Deshmukh, and L. S. Davis. Weakly supervised learning of heterogeneous concepts in videos. In *European Conference on Computer Vision (ECCV)*, 2016.

[21] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576, 2014.

[22] P. Siva and T. Xiang. Weakly supervised action detection. In *British Machine Vision Conference (BMVC)*, 2011.

[23] K. Tang, L. Fei-Fei, and D. Koller. Learning latent temporal structure for complex event detection. In *Computer Vision and Pattern Recognition (CVPR)*, 2012.

[24] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *International Conference on Computer Vision (ICCV)*, 2015.

[25] A. Vahdat, B. Gao, M. Ranjbar, and G. Mori. A discriminative key pose sequence model for recognizing human interactions. In *Eleventh IEEE International Workshop on Visual Surveillance*, 2011.

[26] H. Wang, A. Kläser, C.Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR)*, 2011.

[27] P. Weinzaepfel, Z. Harchaoui, and C. Schmid. Learning to track for spatio-temporal action localization. In *International Conference on Computer Vision (ICCV)*, 2015.

[28] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, 2015.

[29] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. In *Computer Vision and Pattern Recognition (CVPR)*, 1992.

[30] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Video description generation incorporating spatio-temporal features and a soft-attention mechanism. *CoRR arXiv:1502.08029*, 2015.

[31] S. Zha, F. Luisier, W. Andrews, N. Srivastava, and R. Salakhutdinov. Exploiting image-trained cnn architectures for unconstrained video classification. In *British Machine Vision Conference (BMVC)*, 2015.