Action Recognition in the Moments-In-Time Dataset

Zihan Lin†, Hao Yin†, Jason Zhu‡

† Institute for Computational and Mathematical Engineering, Stanford University

‡ Department of Management Science and Engineering, Stanford University

Introduction

We build action recognition models for a newly released dataset of human and animal actions, the `Moments-In-Time` that permits feature localization. We explore and compare two categories of network architectures: the purely spatial model and spatio-temporal model. Measured by top-1 and top-5 accuracy, we found that the improvement of spatio-temporal models is insignificant. Looking into the accuracy in details for each classes, we found that for activities that requires temporal information to recognize (such as closing and falling), the spatio-temporal models perform significantly better. Moreover, we found that the CAM (class activation mapping) reveals accurately the time and pixels in the video that is responsible for activity recognition.

Methods

We focus on video classification models that permit feature localization. Specially, besides classifying each video by the action therein, we would like our model to highlight the time and pixels in the video that exhibits the action.

Video Classification

We investigated two categories of network architectures for video classification.

Purely spatial model This type of models classify each video based on its several single frames while ignoring the temporal information, i.e., it treats video classification as a superposition of several image classification problems.

- 2D-ResNet: we use the conventional ResNet model of depth 50, where the lower-level layers have been pre-trained on the `ImageNet`, and we apply it on 4 equal-distance frames of each video.

Spatio-temporal model This type of model is a direct generalization of the conventional 2D convolution to the 3D case, where we introduce the additional temporal dimension in video dataset. It aims to directly capture the spatiotemporal information through the 3D kernels.

We explore two models within this category:

- 3D-ResNet: We use the 3D generalization of the conventional ResNet with depth of 101, where the lower-level layers have been pretrained on the `Kinetics` video dataset, and we apply it to 16 equal-distance frames of each video.
- 3D-ResNeXt: Same as 3D-ResNet (with depth 101), with the difference that this model is a generalization of conventional ResNext.

Discriminative Feature Localization

For discriminative feature localization, we use the Class Activation Mapping (CAM) method. All fully connected layers before the output layer are removed from the model since they will mess up with location information of the features. Alternatively, a global average pooling layer is used on each channels of the last Convolutional layer (assume n channels) to produce n neurons. These n neurons are then multiplied by a matrix to produce the logits. The entries of the matrix are then used as weights for calculating a weighted average of the channels of the last Convolutional layer, which is just the heatmap to localize the features.

Data

In this course project, we use a subset of the whole `Moments-In-Time` dataset, which is provided as the dataset for the CVPR ActivityNet Challenge 2018 – Mini-Track. This dataset contains 200 action classes, each has 500 videos for training and 50 for validation. Each video is 3 seconds long with 30 FPS.

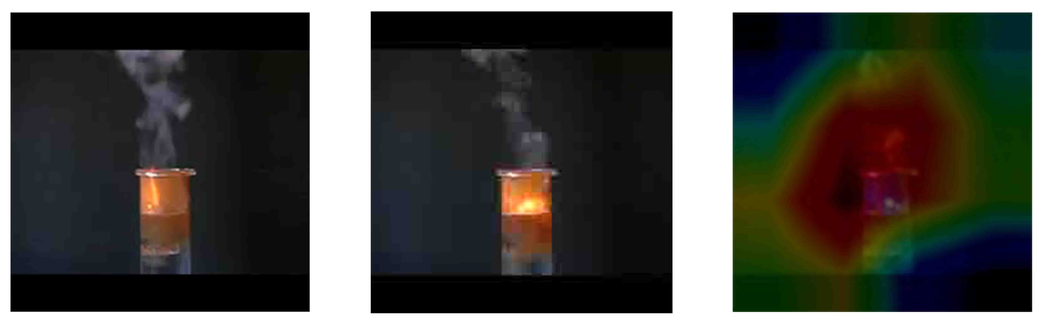
Preprocessing To apply Convolutional Neural Network to the video data, we extract all the frames from the videos, adjust the width and height to 240 by 240. apply mean subtraction for each RGB channel, and crop out only the center 112 by 112 pixels to remove the black margin for most videos.

Experiment Results

Validation Accuracy of each model

Method	Top-1 accuracy (%)	Top-5 accuracy(%)
2D-ResNet	18.0	40.7
3D-ResNet	17.5	39.8
3D-ResNeXt	**18.9**	**41.0**

Table 1: Performance Comparison between different models. 3D-ResNeXt achieves the best validation accuracy, while the relative improvement over 2D-ResNet is insignificant.



Ground Truth: combusting

2D ResNet Top 5 Predictions:
1: pouring (0.264) 2: combusting (0.226) 3: burning (0.112) 4: launching (0.088) 5: smoking (0.048)

3D ResNet Top 5 Predictions:
1: burning (0.226) 2: combusting (0.185) 3: juggling (0.102) 4: spinning (0.045) 5: pouring (0.023)

Ground Truth: cooking

2D ResNet Top 5 Predictions:
1: grilling (0.620) 2: barbecuing (0.320) 3: baking (0.028) 4: frying (0.018) 5: combusting (0.004)

3D ResNet Top 5 Predictions:
1: grilling (0.664) 2: barbecuing (0.151) 3: baking (0.117) 4: frying (0.033) 5: cooking (0.013)

Figure 1: Examples of missed detections. Some ground-truth labels are confusing, which exhibits the difficulty of this action recognition challenge.

Confusing Categories

2D-ResNet

Freq	Actual	Predicted
0.460	barbecuing	grilling
0.320	waking	sleeping
0.300	planting	gardening
0.280	emptying	filling
0.260	handwriting	drawing
0.260	boiling	frying
0.200	studying	reading
0.200	slicing	chopping
0.200	exercising	stretching

3D-ResNet

Freq	Actual	Predicted
0.687	barbecuing	grilling
0.448	gardening	planting
0.367	frying	stirring
0.306	sailing	boating
0.285	closing	opening
0.285	barking	howling
0.244	boiling	stirring
0.244	digging	planting
0.229	cooking	stirring

Table 2: Most common confusions between categories for 2D-ResNet and 3D-ResNext. It gives an intuition about the difficulty of the task, and show that the most common failures come from fine-grained recognition, such as confusing frying versus stirring

Improvement with using 3D-ResNeXt

3D-ResNeXt is better

Freq	Actual
0.420	bulldozing
0.400	gardening
0.276	clinging
0.260	chopping
0.260	frying
0.240	filling
0.240	closing

2D-ResNet is better

Freq	Actual
0.400	tattooing
0.331	folding
0.320	juggling
0.300	planting
0.288	swimming
0.280	stirring
0.280	peeling

Table 3: Comparison between the performance of 2D-ResNet and 3D-ResNeXt on extreme action categories. Numbers in the table are differences in top-1 validation accuracy. On the left are the categories where 3D-ResNeXt performs better than 2D-ResNet, while on the right is contrary. We see that actions that require temporal information to recognize, such as filling and closing, spatial-temporal methods obtains more accuracy. However, for actions that is easily recognized by a single picture such as tatooing and juggling, 2D-ResNet performs better.

Feature Localization using CAM

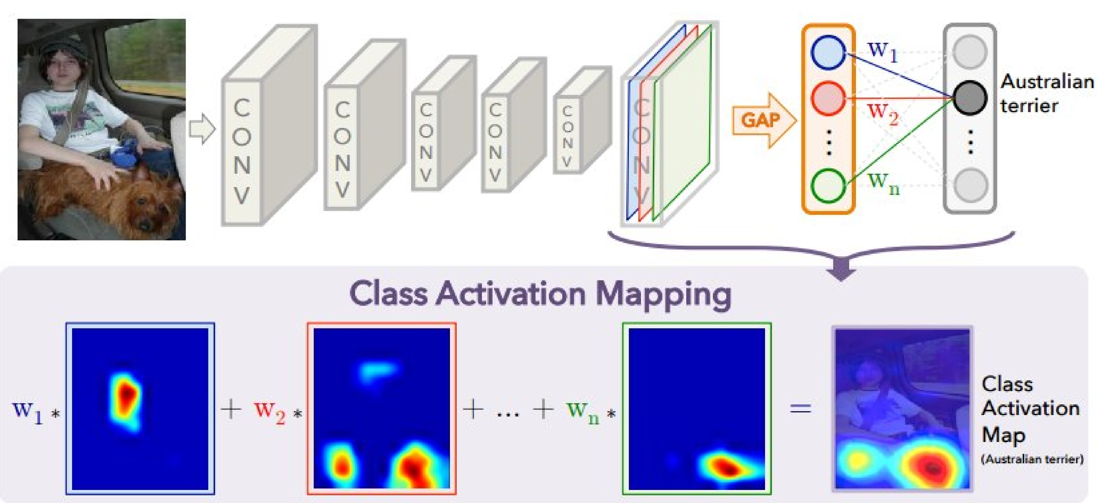
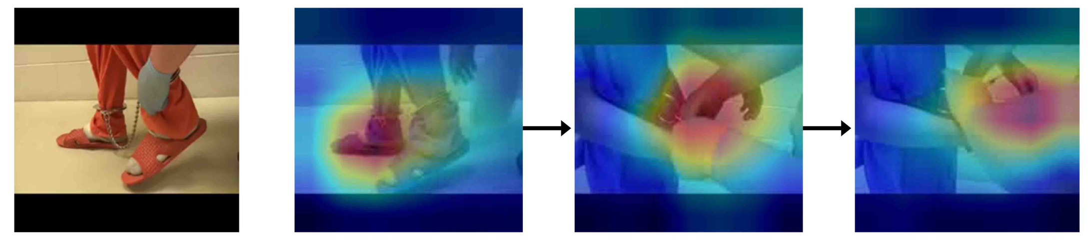


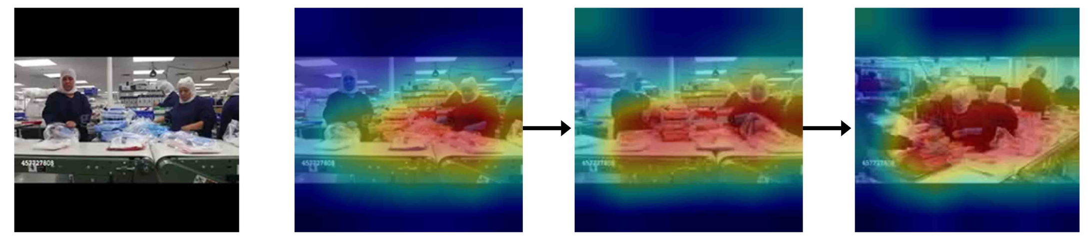
Figure 2: Methodology of Class Activation Mapping (CAM) on images.



Ground Truth: arresting

2D ResNet Top 5 Predictions:
1: tying (0.506) 2: taping (0.084) 3: wrapping (0.064) 4: folding (0.062) 5: sanding (0.031)

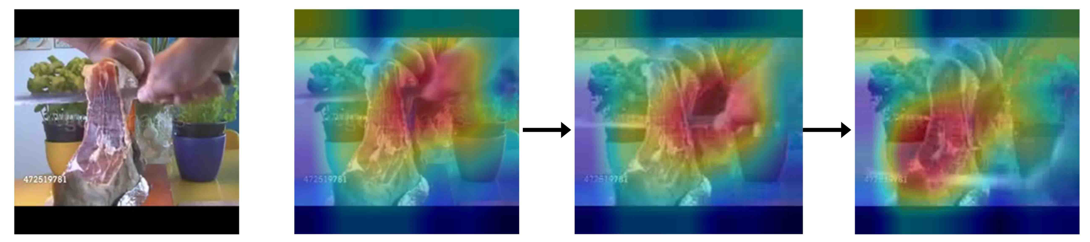
3D ResNet Top 5 Predictions:
1: slicing (0.258) 2: chopping (0.180) 3: folding (0.105) 4: cutting (0.046) 5: wrapping (0.038)



Ground Truth: assembling

2D ResNet Top 5 Predictions:
1: sewing (0.122) 2: studying (0.118) 3: dining (0.086) 4: crafting (0.066) 5: placing (0.064)

3D ResNet Top 5 Predictions:
1: tying (0.506) 2: taping (0.084) 3: wrapping (0.064) 4: folding (0.062) 5: sanding (0.031)



Ground Truth: cutting

2D ResNet Top 5 Predictions:
1: emptying (0.205) 2: filling (0.150) 3: stirring (0.048) 4: pouring (0.048) 5: gardening (0.045)

3D ResNet Top 5 Predictions:
1: tying (0.506) 2: taping (0.084) 3: wrapping (0.064) 4: folding (0.062) 5: sanding (0.031)

Figure 3: Examples of CAM applied to videos. We see, even though we have wrong classification, the feature is localized accurately.

Conclusion

All three methods explored in our report give reasonable performance on the `Moments-In-Time` dataset mini-track. We found that the spatio-temporal methods performs much better in recognizing activities that requires extensive temporal information, such as falling and closing. The great performance of CAM reveals that our model can accurately focus on the important time and pixel in the video, showing the great potential of the model if given larger data volume and computing power.

Selected Reference

- K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. CVPR, 2016.
- B.Zhou, A.Khosla, L.A., A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. CVPR, 2016.
- K. Hara, H. Kataoka, and Y. Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? CVPR, 2018.
- M. Monfort et al. Moments in time dataset: one million videos for event understanding.