Convolutional Neural Networks for Classification of Noisy Sports Videos

Joey Asperger and Austin Poore
{joey2017, hapoore}@stanford.edu

Background

The goal of our project is the automatic classification of noisy, user-generated sports videos. Activity classification in general is an important problem with wide-ranging applications, from allowing YouTube videos to be indexed and searched to more efficient processing of surveillance footage. Substantial work has been published in this area, including the following:

- 3-dimensional convolutions, which incorporate temporal information [1]
- Treating video frames as stand-alone images, and applying an image classifier
- Breaking the video into chunks, classifying each chunk individually, and then combining the chunk-level predictions into a single output prediction for the video as a whole [2]

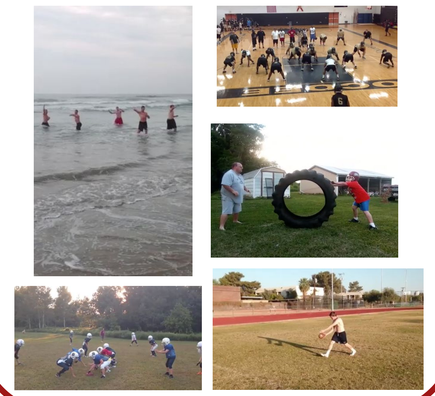
Dataset and Problem Statement

For our project, we used the *Sports Videos in the Wild* (SVW) dataset from Michigan State [3]. The dataset consists of roughly 4200 videos (average length 11.6 seconds) encompassing 44 different activities across 30 different sports. The videos are all filmed on mobile devices, and many feature strange locations, odd angles, and other real-world imperfections.

The problem we sought to solve was that of classification at the sport level. Given a video, our models would predict which of the thirty sports the clip showed. We evaluated our models based on overall accuracy, dividing the number of videos predicted correctly by the total number of videos.

Data Samples

To give a sense of how challenging this dataset is, the following frames all come from football videos:



Approaches and Methods

Initially, our strategy was to try basic versions of several different models. Then, once we figured out which seemed to do well, we could concentrate on optimizing our most promising candidates.

Here are some of the techniques we tried:

- Basic CNN with two convolutional layers (tried with and without batch normalization) and one or two affine layers.
- Basic CNN with two layers of 3-dimensional convolutions, where we stack contiguous frames into a cube, then pass a 3D filter over it (with dimensions of height, width, and time).
- Pretrained Inception-ResNet-V2 model, which combines parts of GoogleNet and ResNet, with an affine layer to make predictions and an LSTM to make predictions.

Results

Model	Validation Accuracy
1	43.3%
2	41.7%
3	47.7%
4	72.3%
5	71.0%
6	85.6%
7	74.7%

Model 1: Two convolutional layers (with ReLU activation), batch normalization, and dropout (25%), followed by an affine layer. 30 frames sampled from each video.

Model 2: Two 3D convolutional layers with ReLU and max pooling, with affine layer.

Model 3: Broke videos into 10 chunks, classified each chunk using basic model (Model 1 without dropout), then combined.

Model 4: Pretrained Inception-Resnet-V2 model fine-tuned on our data, using single frame only.

Model 5: Model 4, only backpropagating through top half of pretrained model

Model 6: Model 4 averaged across 10 frames.

Model 7: Model 4 with LSTM prediction layer across 16 frames.

Discussion

Our experiments showed that fine-tuning a pre-trained imagenet model performs far better that training a model from scratch. Considering the small size of our dataset, this is not surprising. In addition, Inception-Resnet-v2 is very sophisticated and much more complex than anything we tried from scratch.

In general, we had more success by treating frames as still images to be classified than by attempting to capture complex temporal features. Intuitively, this makes some sense, because with a few exceptions, a person can likely distinguish among all of these sports pretty easily given a handful of frames. That said, simple techniques like breaking videos into chunks did appear to yield some performance gains, suggesting that temporal information is still useful.

The most successful previous result on this dataset that we could find hovered around 80% accuracy [4], but it relied on some handcrafted features. While we have not run our model on the test set yet, we have already achieved 85.6% validation accuracy, and will continue fine-tune our model in the next week.

Future Work

Our immediate goals for future work are to continue fine-tuning our most successful Inception-Resnet-V2 based models to achieve high performance, such as trying the video-chunking technique or different prediction layers on the top.

There's also room for many more experiments to be done on different model architectures and feature selection techniques. Given time and large computational resources, we would be interested in continuing to explore how best to capture temporal features.

References

[1] Ji, Shuiwang, et al. "3D convolutional neural networks for human action recognition." *IEEE transactions on pattern analysis and machine intelligence* 35.1 (2013): 221-231.
[2] Wang, Limin, et al. "Temporal segment networks: towards good practices for deep action recognition." *European Conference on Computer Vision*. Springer International Publishing, 2016.
[3] Safdarnejad, Seyed Morteza, et al. "Sports videos in the wild (SVW): A video dataset for sports analysis." *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*. Vol. 1. IEEE, 2015.
[4] Rachmadi, Reza Fuad, Keiichi Uchimura, and Gou Koutaki. "Combined Convolutional Neural Network for Event Recognition." *Korea-Japan Joint Workshop on Frontiers of Computer Vision*. 2016.