Dangerous Action Recognition of Pedestrians with Convolutional Networks

Binghao Wang, Qianyao Yang

February 6, 2017

1 Introduction

Recognizing dangerous human actions is important to the safety of autonomous vehicles on the street. Currently, object detection techniques are widely used in the industry, however it doesn't predict potential dangerous motion from pedestrians. We will implement the state-of-the-art two-stream convolutional network model[1], of which the spatial stream is mainly trained on the standard video action benchmarks(UCF-101, HMDB-51) while the temporal stream is trained on bi-directional optical flow. In this proposal, we will talk about our motivation, briefly summarize our goal and list some details of the project.

2 Motivation

The action recognition problem can be explained as taking some sequential data as inputs and giving a single classification as output. Comparing to object detection on 2D images, more information can be extracted from the additional temporal domain in 3D. For instance, only detecting the location of the pedestrians in front is not sufficient for a self-driving car. The sequential actions may contain crucial information for it to make a decision, which can be lifesaving for both passengers and pedestrians.

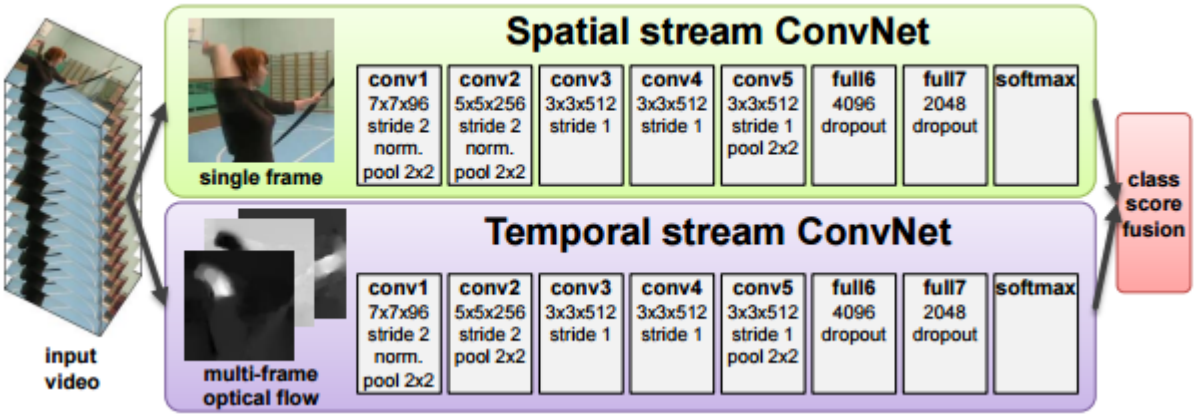

Figure 1: Two-stream architecture for video classification(from [1])

[1] proposed a two-stream architecture(Figure 1), which achieved 88.0% mean accurary on UCF-101 benchmark. The problems of our implementation are twofold: 1)the temporal ConvNet requires gigantic amount of training data(multiple TBs) and 2)the recognition speed is not satisfying since we want to run it at real time. Possible improvement approaches would be: 1)using more hand-crafting features on optical flow and 2)reconstructing the model architecture or using other features.

3 Project Summary

We will build up an architecture based on [1] and train and finetune the model for recognizing dangerous actions of pedestrians. Different features such as optical flow, trajectory, motion vectors will be tested to achieve better result. Hopefully, an improvement or compromise will be figured out to deploy this model on real time recognition. LSTM network will also be tested.

4 Project Details

Archtecutre and Environment

Most of training and test procedures will be using Torch and MXNet on a Nvidia Titan X GPU, and the final model will be deployed in Torch.

Implementation Issue and Chanllenges

Most tricky issues are the implementation of multimodal network and the search for optimal temporal stream feature extraction method. The insufficience of training data will also restrict our model performance.

Timeline

- Feb 6 - Feb 20 Implement the basic architecture of the network, also test LSTM approaches(if time allows);
- Feb 20 - Mar 6 Implement and test feature extraction methods, submit the progress report;
- Mar 6 - Mar20 Tinker the network, make demo and prepare for presentation;
- Apr 3 - Apr 24 Submit final project report.

5 Conclusion

We want to implement a two-stream convolutional model especially for recognizing dangerous actions of pedestrians, which will benefit significantly for autonomous driving system. Considering of the shortage of computational power of embedded system, we will also try to accelerate the architecture by different model structures and feature extraction techniques.

Statement

We read the leaflet "Beware of Plagiarism" and understand that plagiarism can lead to a failing mark in the project.

References

[1] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pp. 568-576. 2014.

[2] Zhang, Bowen, Limin Wang, Zhe Wang, Yu Qiao, and Hanli Wang. *Real-time action recognition with enhanced motion vector CNNs.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2718-2726. 2016.

[3] Wang, Heng, and Cordelia Schmid. *Action recognition with improved trajectories.* In Proceedings of the IEEE International Conference on Computer Vision, pp. 3551-3558. 2013.