Video Segmentation within image frames for autonomous driving.

Category: Computer Vision

Team Members: Wenfei Du, Hiro Tien (Kai Ping), Mayank Malhotra

Motivation/Introduction

1. Autonomous vehicles are not able to quickly identify different kinds of moving items on the road
2. Our project helps autonomous vehicles segment movable objects
3. We're using a convolutional neural net (CNN) similar to VGG 16-layer net and then put a multilayer deconvolution network with FCN and Segnet (with Maxpool) variants

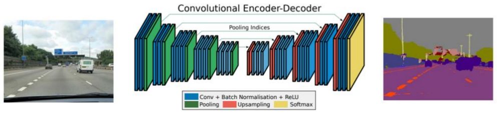
Architecture/Models

We learn the network on top of the convolutional layers adopted from VGG 16-layer net. We use two models for deconvolutional part:

1) Fully Convoluted Network (FCN) with learned upsampling layers that are connected to previous layers in the network and
2) SegNet Encoder Decoder Architecture, further specifies the un-pooling layers to achieve better efficiency and performance.

$$l = -\sum_{i} \sum_{c=1}^{8} w_c \, y_{i,c} \log(\hat{y}_{i,c})$$

$$w_c = [\underbrace{0.01}_{1\text{ class}}, \underbrace{0.99/7, \ldots, \ldots}_{7\text{ classes}}]$$


Convolutional Encoder-Decoder
Pooling Indices
Conv + Batch Normalisation + ReLU
Pooling
Upsampling
Softmax

Illustrative

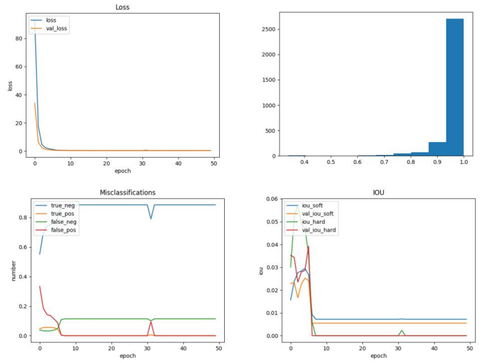
Dataset

1. The dataset is provided by Baidu through CVPR 2018 WAD Video Segmentation Challenge on Kaggle. The competition evaluated seven different instance-level annotations, which are car, motorcycle, bicycle, pedestrian, truck, bus, and tricycle.
2. The rest are combined into a background class. Images are colored, each pixel is categorized into one of the 8 categories. The final dimensions of the picture are 564px*224px.
3. Total of 44 videos provided with thousands of frames per video.
4. Most training images were more than 99% background, and we extract 244 positive examples with less than 90% background and sample 244 negative examples.
5. We train on 488 samples and validate on 380 samples.

Discussion

1. We compare FCN and Segnet based models with no regularization as the baseline.
2. We use convolutions for previous and next image in time along with current image in deconvolution.
3. We add regularization to the best performing FCN model.
4. Despite removing 90% background images and using weights, we continue to predict background after about 5 epochs of training (about 10% false negative).
5. In the first 5 epochs, the predictions are better balanced (about 5% false negative and false positive).
6. The low number of training images is a major limitation.

Results



Future Scope

1. We have tested two of the prominent architectures for image segmentation.
2. In the future:
    1. Collect more positive example images from videos
    2. Tune number and size of layers
    3. Tune regularization
    4. Better model correlations in time and in video information (camera side and road)
    5. Add object detection model prior to segmentation

References

[1]. M. Ranzato, F. J. Huang, Y. Boureau, and Y. LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in CVPR, 2007.

[2] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in CVPR, pp. 2528–2535, IEEE, 2010.

[3] K. Kavukcuoglu, P. Sermanet, Y. Boureau, K. Gregor, M. Mathieu, and Y. LeCun, "Learning convolutional feature hierarchies for visual recognition," in NIPS, pp. 1090–1098, 2010.

[4] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in ECCV, pp. 184–199, Springer, 2014.