Urban Scene Segmentation for Autonomous Vehicles

Team: Hsiao-Chen Huang, Eddie Tseng, Ping-Chun Chiang, Chih-Yen Lin

Department of Electrical and Computer Engineering, ECE 228, University of California, San Diego, La Jolla, CA 92093, USA

UC San Diego
Jacobs School of Engineering

Background & Motivation

With the rapid developing and the continuing evolution of autonomous vehicles, people begin to increase its reliability. To guarantee the safety of autonomous vehicles, we need to make AVs classify objects as quick as possible (e.g. 24 frames per second). Hence, we decided to solve semantic segmentation problem to localize objects in the images using **Fully Convolutional Networks (FCNs)** [1] and evaluate the advantages of different networks.

Dataset

We use **Cityscapes Dataset**[2] to train our network, and the dataset contains 19998 images which is consist of street scenes of 50 cities and corresponding annotation images. The annotations represent 30 classes in different RGB color for each pixel. However, we just choose 19 classes in our task, and an example is shown in *Figure 1*.

Figure 1. Example of dataset

Features

Input:

$$\{x_1, x_2, ..., x_N\} \in X \qquad (1)$$

Annotation:

$$\{y_1, y_2, ..., y_N\} \in Y$$
$$y_n = \{(c_0, c_1, ..., c_{N_c}, c_{N_c+1}) | c_i \in \{0,1\}\}, \quad \forall n \qquad (2)$$

Sigmoid activation function:

Use sigmoid function (eq 3) as our activation function so that all the outputs will lie between 0 and 1.

$$S(x) = \frac{1}{1+e^{-x}} \qquad (3)$$

Cross-entropy loss function:

Evaluate cross entropy as loss function (eq 4).

$$L[y, \hat{y}] = -\sum_i y_i \log_2(\hat{y}_i) \qquad (4)$$

Methods

Fully Convolutional Network can efficiently learn to make dense prediction in pixel-wise classification task.

- Using 1x1 convolutions to transfer feature map to pixel-to-pixel prediction.
- Deconvolutional layers are used to capture different level of shape details and expand the prediction size to input size.
- Skip connection allows lower level information to reach top level by adding deconvolutional layers to the previous layer.

The main goal is to design a network as a bunch of convolutional layers with downsampling and upsampling. The architecture of **AlexNet** and **VGG net** with skip connection we used are shown in *Figure 2* and *Figure 3*.

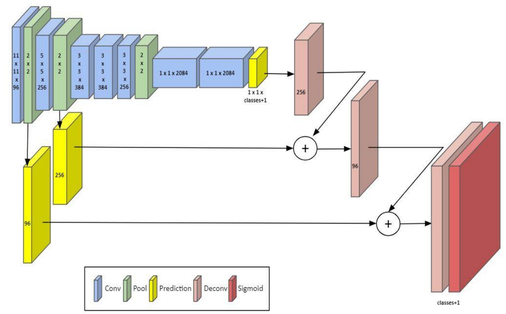

Figure 2. Alex Net [3]

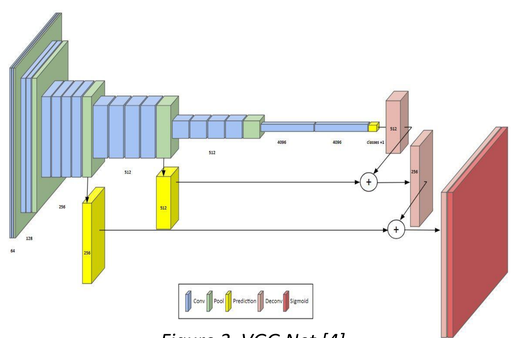

Figure 3. VGG Net [4]

Results

The dataset contains 18,000 training and 1,998 testing images. AlexNets were trained in 13k steps and VGG was trained in 93k steps(skip) vs 41k steps(non-skip).

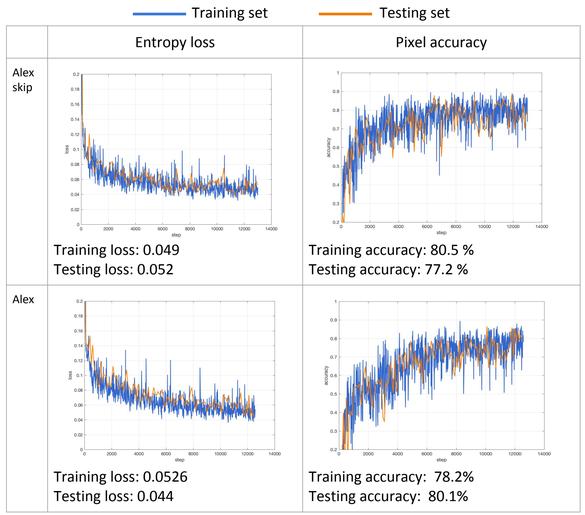

Table 1.

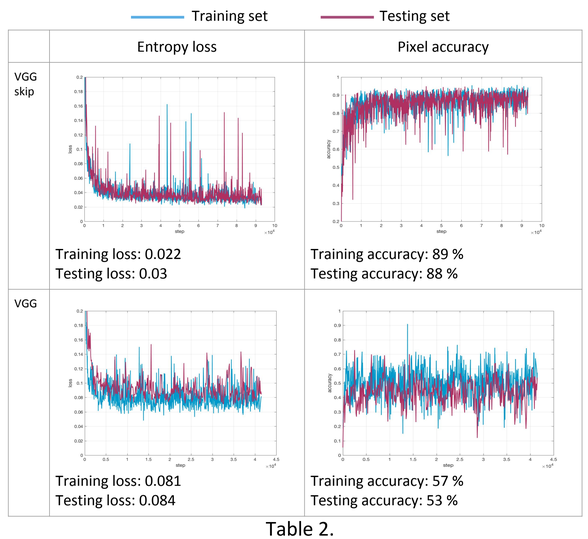

Table 2.

Discussion

Skip Connection:

The accuracy of the model with skip connection should be higher and improve the segmentation detail because of fusing information we loss during pooling operation.

AlexNet vs. VGG net:

VGG is similar to AlexNet, but more filters. Thus, VGG can extract higher features. That is why it is currently the most popular model, deep and simple.

Batch Size:

Due to the limited GPU and RAM size, we trained mini-batch gradient descent with batch size equal to 5 for AlexNet and 2 for VGG. Therefore, this is the reason for large fluctuations.

Learning Rate:

In this project we use learning rate = 0.0001. For the future work, we will tune learning rate to optimize the training speed.

As the prediction result shown in *Figure 4,* green indicates road, blue indicates cars, etc..

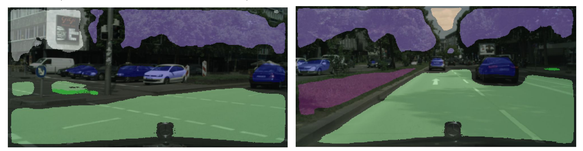

Figure 4. Results of VGG with skip connection

Future Work

- Real-time segmentation
- Implement different model (RNN, LSTM) [5]
- Try more classes or use different dataset.
- Calculate class mean accuracy which consider the performance of each object.
- Tuning parameters.

Reference

[1] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651, April 2017.

[2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. *The cityscapes dataset for semantic urban scene understanding*.CoRR, abs/1604.01685, 2016.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[4] Sugata, T. L. I., and C. K. Yang. "Leaf App: Leaf recognition with deep convolutional neural networks." *IOP Conference Series: Materials Science and Engineering*. Vol. 273. No. 1. IOP Publishing, 2017.

[5] Valipour, Sepehr, et al. "Recurrent fully convolutional networks for video segmentation." *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*. IEEE, 2017.