
Semantic segmentation using full convolutional network

Fangzhou Ai

Department of Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA 92093
faai@eng.ucsd.edu

Yue Qiao

Department of Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA 92093
yuq021@eng.ucsd.edu

Zunming Zhang

Department of Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA 92093
zuz008@eng.ucsd.edu

Abstract

In this report, we explore the full convolutional network (fcn) and use full convolutional network to perform semantic segmentation task on Indian driving dataset. Using our baseline dataset, we archive pixel accuracy of 0.79 and average IOU of 0.33.

1 Introduction

In the report we try to solve the semantic segmentation task using full convolutional network. The goal of semantic segmentation is to label each pixel of an image with a corresponding class of what is being represented [5]. To solve the task, we utilize full convolutional network. In full convolutional network, there is no fully connected layer, only convolutional layer. In the full convolutional network structure, we use Xavier weight initialization. Deep neural network suffers from weight vanishing problem. If we use normal weight initialization methods like uniform distribution, the weight will vanish quickly during matrix calculation. Therefore, we introduce Xavier weight initialization [3], which can help us eliminate the problem. To further improve the training speed and stability of the network, we introduce batch normalization layer [4]. The batch normalization layer solves that the distribution of each layer's inputs changes during training by normalizing layer inputs.

2 Related Work

3 Methods

3.1 Baseline

The baseline model is an encoder-decoder structure. It first encode the picture using convolution layers and then decode the segmentation map using deconvolution layers. In the baseline model, we use batch normalization and use ReLU as the activation function. For the last layer, we use softmax to get the classify results.

3.2 Improved baseline

3.2.1 Dataset Augmentation

From the baseline model, we tried to do some data augmentation on the training dataset to see whether the performance could be approved. Since training is quite time-consuming, in this report, we implemented and tested two different data augmentation algorithm and tried to compare the performance with each other.

The data augmentation methods used are:

- Mirror flip: for the training dataset, we mirror flipped the dataset and fed them into the model
- Image Rotating: rotate the images slightly (for a certain degree), in this report, we tried to rotate the images by 1 degree and then fed them into the model

3.2.2 Imbalanced class problem

For the imbalanced class problem, what we usually do is to pressuring the network to categorize the infrequently seen class. We proposed to use the weighted loss method to accommodate this issue, which will weight infrequent classes more.

For the weighted loss, the idea is to set the weight of the loss of each class inversely proportion to the frequency of each class in the training dataset. For the data pre-processing, we calculated the total pixel count of each class in all the training dataset and then use this information to help us set the weight of the loss.

Suppose, the number of samples for each class is stored in $nSamples$, then we use $normedWeights = [1 - (x / \sum nSamples)]$ for x in $nSamples$ as the normalized weight of the loss for each class.

For this dataset, we calculated the number of samples for each class is:

$nSamples = [2198453584, 519609102, 28929892, 126362539, 58804972, 59032134, 94293190, 2569952, 101519794, 163659019, 105075839, 47400301, 30998126, 133835645, 135950687, 41339482, 15989829, 104795610, 8062798, 450973, 94820222, 341805135, 557016609, 71159311, 1465165566, 1823922767, 2775322]$

Then we could be able to set the weight of the loss for each class according to $nSamples$.

3.3 Custom model

Inspired by this paper[2], we replace the encoder part in FCN model with the RRCNN layer instead, the RRCNN represent better abstraction ability, thus we would expect better IoU and precision. However the RRCNN is really memory-intensive, we have to crop the image to only $256 * 256$ size and delete 3 layers from original model's encoding part, which severely influence the performance of our model. What's worse is the RRCNN encoding part contains maxpooling operation after each RRCNN block, which means the images size would shrinks even more, since we have already started from a relative small size, maxpool will almost ruin all the details in the feature map.

3.4 Transfer learning

In this part we employ the DEEPLABV3-RESNET101 model, which is a default pre-trained net work in PyTorch, from the website[1] we know that this model is constructed by a Deeplabv3 model with a ResNet-101 backbone. The pre-trained model has been trained on a subset of COCO train2017, on the 20 categories that are present in the Pascal VOC dataset. We modifier the classifier part so that it could match the number of our feature, and we freeze the parameters of the pretrained part to accelerate the whole procedure.

3.5 U-Net

Following the routine demonstrated in the U-net paper [6], we build the U-net by our self. The obvious difference is the U-net convolution block contains 2 convolution of the same size, while the FCN model only contains 1 convolution operation. We notice that the U-net model is also memory intensive while the datahub only give us a GTX 1080 ti with 11GB memory, hence there's a trade-off of between batch size, image size and model's completeness. We decide to crop the image and shrink the batch size to run the whole model, though this might not be the best solution, but this could give us a flavor of how the complete network looks like.

4 Results

Table 1: Pixel accuracy and IOUs for different models

Model	Pixel accuracy	Mean IOU	road IOU	sidewalk IOU	car IOU	billboard IOU	sky IOU
Baseline	0.7944	0.3371	0.8632	0.2323	0.5084	0.2419	0.9403
Mirror flip	0.7546	0.2532	0.8082	0.1864	0.2262	0.2542	0.8105
Rotation	0.8041	0.3533	0.8730	0.2455	0.5111	0.3078	0.9462
weighted loss	0.7991	0.3444	0.8648	0.2406	0.5104	0.2967	0.9472
Custom	0.3696	0.0976	0.6115	0.0002	0.2023	0.0220	0.6380
Transfer learning	0.8296	0.4370	0.8687	0.1485	0.7904	0.3900	0.9427
U-Net	0.7486	0.2577	0.8266	0.0038	0.4794	0.2225	0.9125

4.1 Baseline model

Training and Validation loss for baseline model is in Figure 1. Training and Validation pixel accuracy and mean IOUs for baseline model is in Figure 2. Various IOUs for baseline model is in Figure 3. The visualization of the output is in Figure 4.

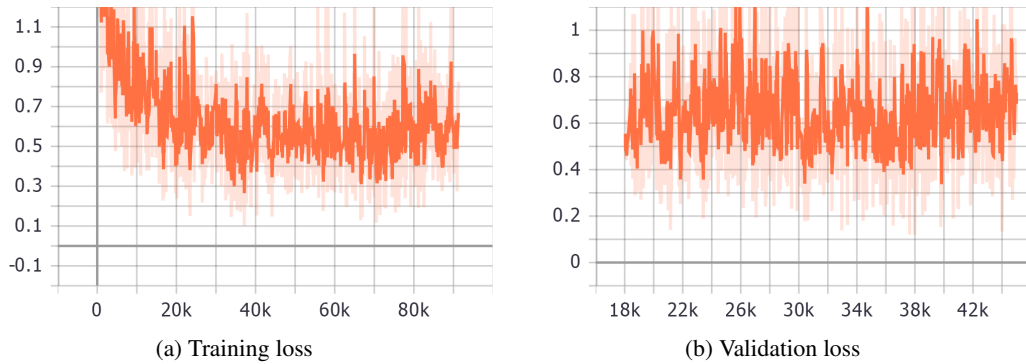


Figure 1: Training and Validation loss for baseline model

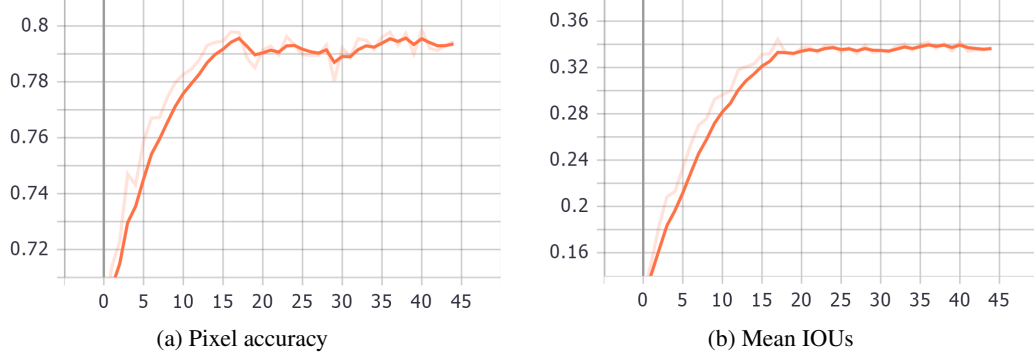


Figure 2: Training and Validation pixel accuracy and mean IOUs for baseline model

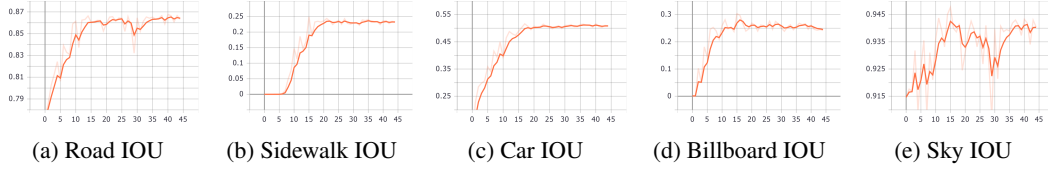


Figure 3: Various IOUs for baseline model

4.2 Improved baseline model

4.2.1 Dataset Augmentation

Table 1 shows the model performance of different augmentation method. From the table, we could see that the performance was improved by 10% simply by rotating the training dataset for a small angle.

The train and validation loss for mirror flip and rotation methods could be seen in Figure 5a and Figure 5b, respectively. The visualization of the segmented output for the first image in the test.csv overlaid on the image for mirror flip and rotation method could be seen in Figure 6a and Figure 6b, respectively.

4.2.2 Imbalanced class problem

The train and validation loss after introducing weighted loss method could be seen in Figure 7a. The performance of the model could be seen in Table 1 and the visualization of the segmented output for the first image in the testdataset overlaid on the image could be seen in Figure 7b

4.3 Custom model

The traning and validation loss are shown in Figure 8, we replace all the conv blocks with RRCNN blocks, and it shows strong overfitting signal at the first beginning, thus the result is very bad. Also, large model and limited memory force us to choose only batch size 1 and smaller picture size, which also brought significant negative effect to the result. The final segmentation picture is shown in Figure 9.

4.4 Transfer learning

With only 7 epochs, the transfer learning model got the best the result. "Standing on the shoulders of giants, discovering truth by building on previous discoveries". The result are shown in Figure 10 and Figure 11.



Figure 4: Visualization of segmented output for baseline model

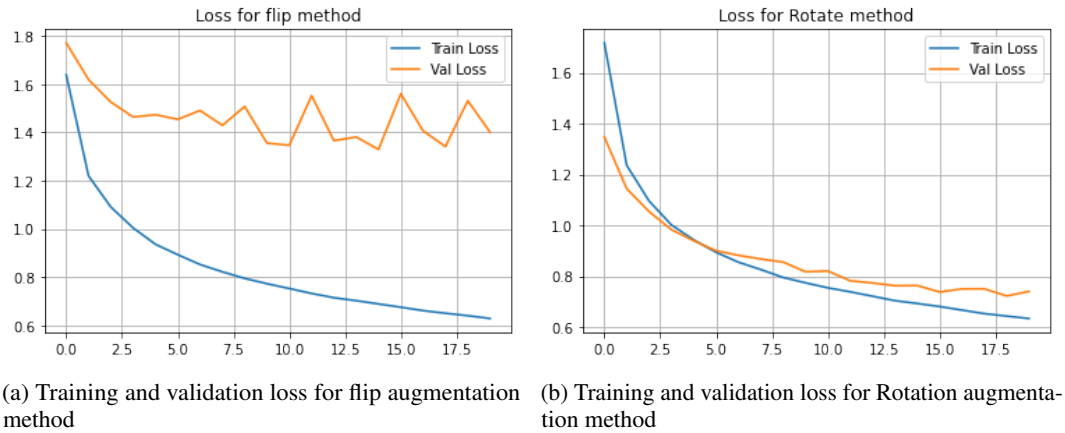


Figure 5: Training and validation loss for different augmentation methods

4.5 U-Net

The U net shows potential to beat the FCN model, however, the model is too large, we have to resize the model to $720 * 1280$ and only use SGD (mini batch size = 1) for training, and we only run 20 epochs, so the result is not that good, however we found that the precision is still increasing at the 20th epoch, so we believe that with more epochs the model would perform much better than current result. The loss is shown in Figure 12 and final result is shown in Figure 13.

5 Discussion

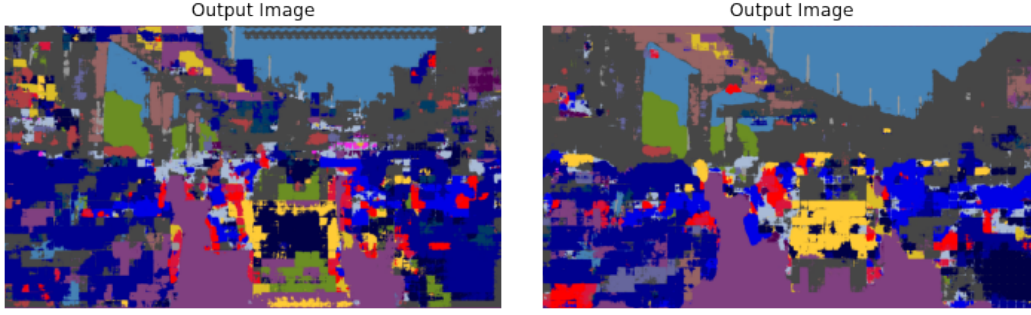
This homework gave us a flavor of how the real model works, instead of playing with simplified toy model, here we encountered the real challenging model and we made our own modifications, during this process we realize that GPU memory is really a key factor for machine learning speed-up. We came up with many ideas while due to the limit of the GPU memory, we have to made a choice between model complexity and batch size with images size. To test the full model we choose to sacrifice the latter two points, and it proved that some times make a trade off and find a balance point is more important, I would rather call it Zen of Machine learning.

6 Individual contributions to the project

Fangzhou Ai Custom model, Transfer learning and U-Net model.

Yue Qiao Evaluation metrics, Dataset loading and Baseline Model.

Zunming Zhang Improved baseline model with mirror flip, rotate and weighted loss for imbalanced classes.



(a) Visualization of segmented output for Mirror flip augmentation method (b) Visualization of segmented output for Rotation augmentation method

Figure 6: Visualization of segmented output for different augmentation methods

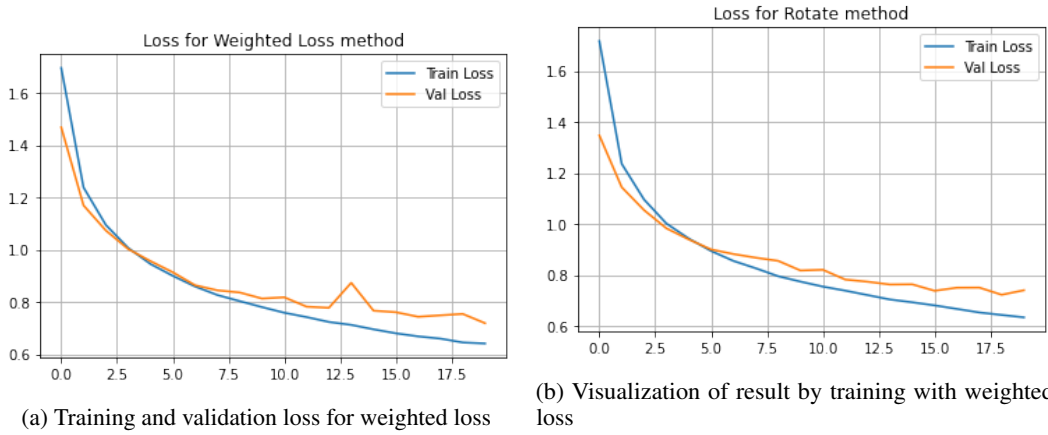


Figure 7: Weighted loss

References

- [1] Deeplabv3-resnet101 | pytorch. https://pytorch.org/hub/pytorch_vision_deeplabv3_resnet101/. (Accessed on 02/15/2021).
- [2] Md Zahangir Alom, Mahmudul Hasan, Chris Yakopcic, Tarek M. Taha, and Vijayan K. Asari. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation, 2018.
- [3] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterton, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. JMLR Workshop and Conference Proceedings. URL <http://proceedings.mlr.press/v9/glorot10a.html>.
- [4] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015.
- [5] Jeremy Jordan. An overview of semantic image segmentation. <https://www.jeremyjordan.me/semantic-segmentation/>, 5 2018. (Accessed on 02/15/2021).
- [6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

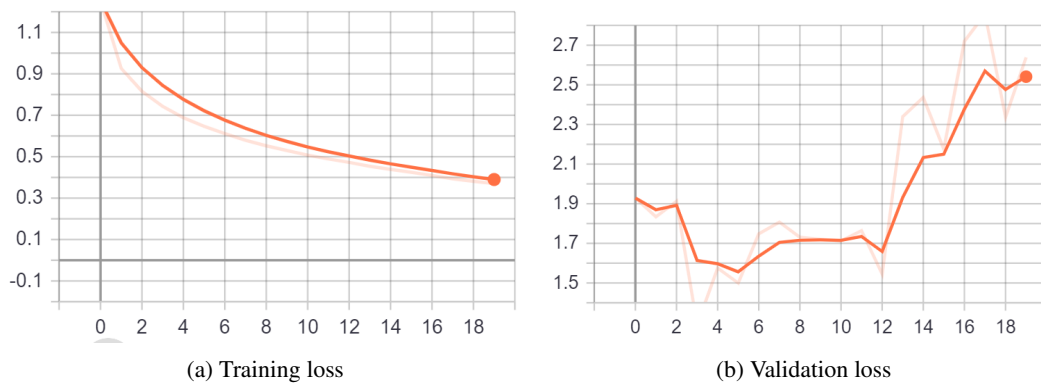


Figure 8: Training and Validation loss for custom model

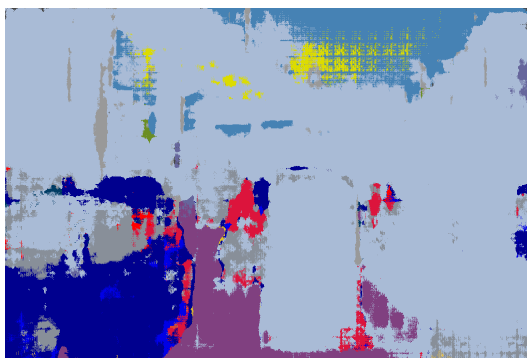


Figure 9: Visualization of segmented output for custom model

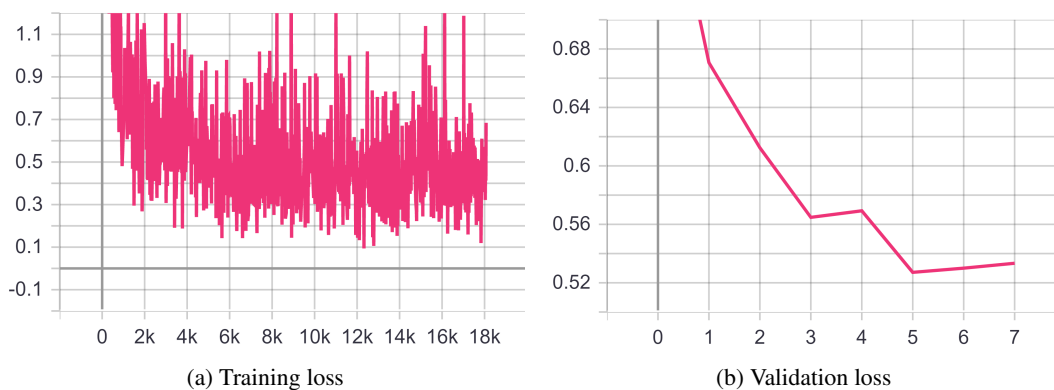


Figure 10: Training and Validation loss for Transfer learning model

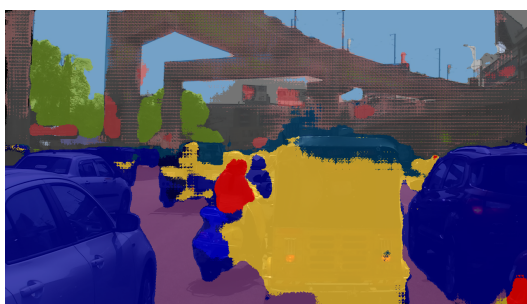


Figure 11: Visualization of segmented output for Transfer learning model

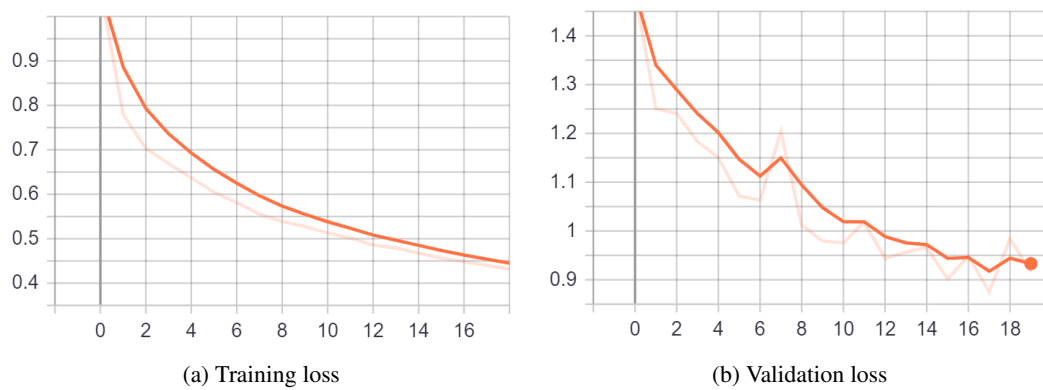


Figure 12: Training and Validation loss for Unet model

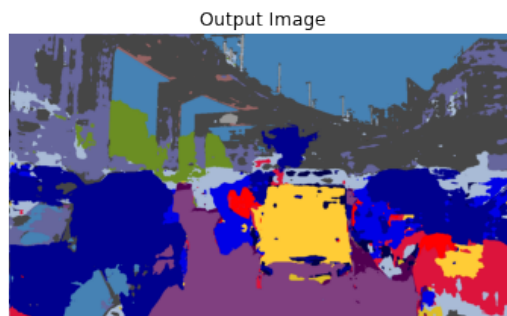


Figure 13: Visualization of segmented output for Unet model