디컨볼루션 픽셀층 기반의 도로 이미지의 의미론적 분할

Deconvolution Pixel Layer Based Semantic Segmentation for Street View Images

Abdul Wahid*, Hyo Jong Lee**
*Department of Computer Science and Engineering, CAIIT, Chonbuk National University
** Department of Computer Science and Engineering, CAIIT, Chonbuk National University
abdul@jbnu.ac.kr hlee@chonbuk.ac.kr

Abstract

Semantic segmentation has remained as a challenging problem in the field of computer vision. Given the immense power of Convolution Neural Network (CNN) models, many complex problems have been solved in computer vision. Semantic segmentation is the challenge of classifying several pixels of an image into one category. With the help of convolution neural networks, we have witnessed prolific results over the time. We propose a convolutional neural network model which uses Fully CNN with deconvolutional pixel layers. The goal is to create a hierarchy of features while the fully convolutional model does the primary learning and later deconvolutional model visually segments the target image. The proposed approach creates a direct link among the several adjacent pixels in the resulting feature maps. It also preserves the spatial features such as corners and edges in images and hence adding more accuracy to the resulting outputs. We test our algorithm on Karlsruhe Institute of Technology and Toyota Technologies Institute (KITTI) street view data set. Our method achieves an mIoU accuracy of 92.04 %.

1. Introduction

Deep convolutional neural networks (DCNNs) have been very successful in terms of task involving large image datasets such as ImageNet [11], MS COCO [12] and PascalVOC [13]. Given the recent advances in computer vision, visual perception will have a leading role to play in the development of autonomous vehicles. DCNNs have an immense representation power which leads to efficient results. After there success in the task of image classification and detection DCNNs have been used extensively for semantic segmentation [8] and Action Recognition [1] in videos to classify different types of behaviors and actions among different subjects such as humans.

The objective of current segmentation networks is to serve the purpose of solving pixel-wise classification. Researchers try to design networks with as many layers as possible [10]. Layers like pooling, convolutional [2], fully connected and deconvolution, have been extensively used. Deconvolutional layers [7] are used in the networks whose features maps need to be upscaled or upsampled such as FCN [8] and SegNet [6]. However, one of the main disadvantages of deconvolutional network is the uneven overlap because of the output window size, which creates a checkerboard pattern of varying magnitudes [3].

This work proposes an efficient method which combines FCN with deconvolution pixel layer to overcome checkerboard problems in deconvolution layers. The deconvolutional pixel layer generates a pool of feature maps in a sequential manner thus addressing the need of feature maps in the latter phase of the model to depend on the initially generated feature maps [3]. Thus, connecting the adjacent pixels on the resulting feature maps.

Our experimental results show the efficiency of our method on the challenging KITTI dataset [4] and high performance in road segmentation.

2. Model

Figure 1. explains the proposed architecture in detail. The architecture consists of two parts the FCN based feature extractor and deconvolutional pixel layer.

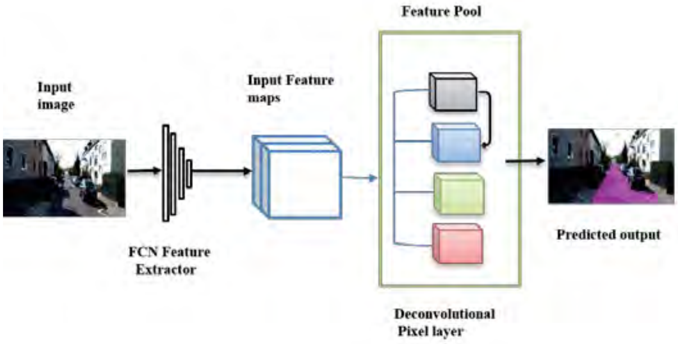


Figure 1. An end-to-end architecture with FCN and deconvolutional pixel layer for road Semantic Segmentation.

FCN generates output maps for inputs of any size, the dimensions of the output feature maps are reduced by down sampling. As a result, reducing it from the size of the input by a factor equal to the pixel stride of the receptive fields of the output units.

The proposed method is used to associate coarse outputs to dense pixels is by deconvolution pixel layer. The deconvolution layer performs the periodic shuffling of various intermediate feature maps generated by different convolutional operations [5].

The deconvolutional pixel layer up samples the 4 x 4 feature map to 8 x 8 feature map. The feature map in grey color in deconvolutional pixel layer is obtained by a 3 x 3 convolution. Another 3×3 convolutional is performed on the grey feature map in order to generate the feature map of blue color. The grey and the blue feature maps are added together because they are dilated to form a single big feature map. Also, we use masked 3 x 3 convolution on the last two feature inputs, thus they are combined into one big feature map. The last two feature maps are combined because of their missing relationship. In pixel upsampling layer the feature maps are divided into four groups while as in other up sampling techniques the feature maps are increased only by the factor of two. Also because of this reason the training efficiency while implementing pixel deconvolution increases because of the collective relationship between feature maps [3].

3. Experimental Analysis

The architecture designed uses a VGG-16 Convnet pre-trained on ImageNet as the encoder and a decoder based on FCN-8 [8] with deconvolutional pixel layer [3].

We performed a series of experiments on KITTI [4] road segmentation dataset which consists of 289 training images and 290 test images. It consists of three categories of road scenes. It is a very well-known fact that, proper selection of hyperparameters is a very crucial process for training DNNs. In our experiment we used Adam Optimizer and a series of experiments proved that a learning rate of 0.00001 works best for our model. Since the training set consists only 298 training sample, hence we run our experiment for only 25 epochs. Also, we selected a batch size of 8 based on our training dataset. We used the fixed input size of 160 x 576 x 3. The dataset consists three categories of street views.

- UU – Urban unmarked (98/100)
- UM – Urban marked (95/96)
- UMM – Urban multiple marked lines (96/94)

The model was trained using Tensorflow framework on Pascal NVidia 1080 Ti GPU. Figure 2 shows the training loss while we train our model.

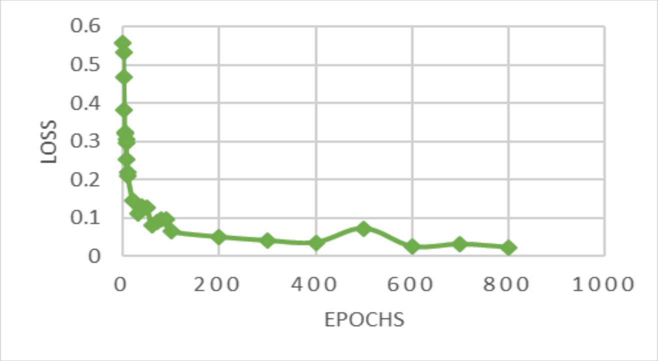
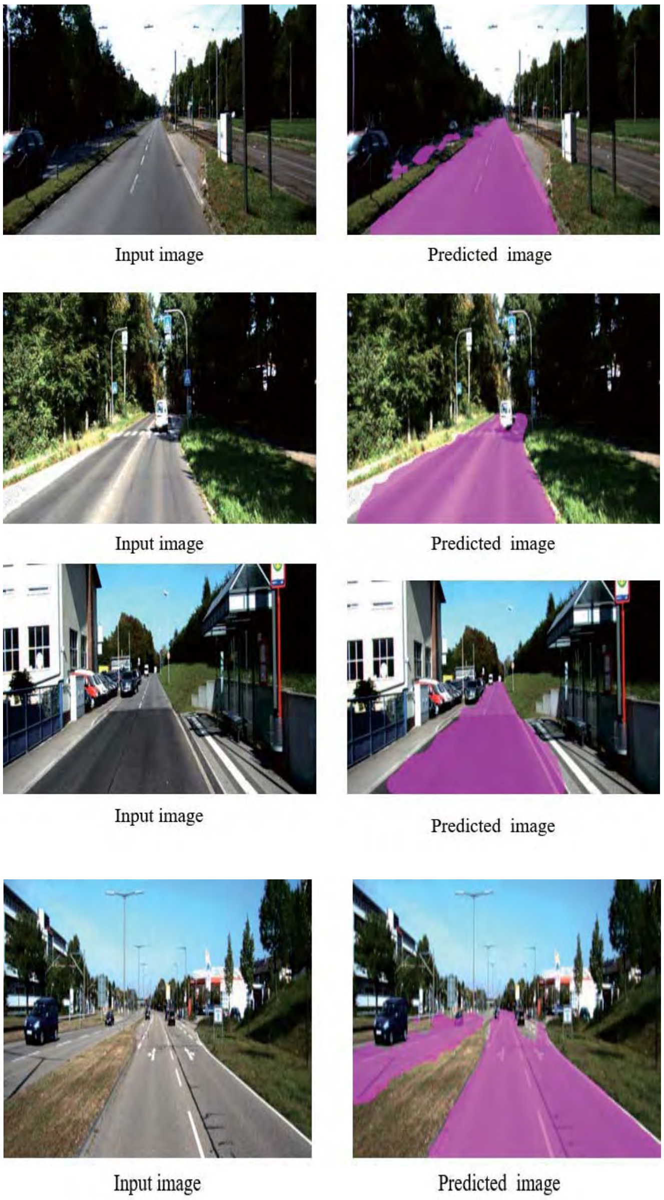


Figure 2. Training loss

Figure 3 shows the output generated by our proposed method for road segmentation.



Input image

Predicted image

Input image

Predicted image

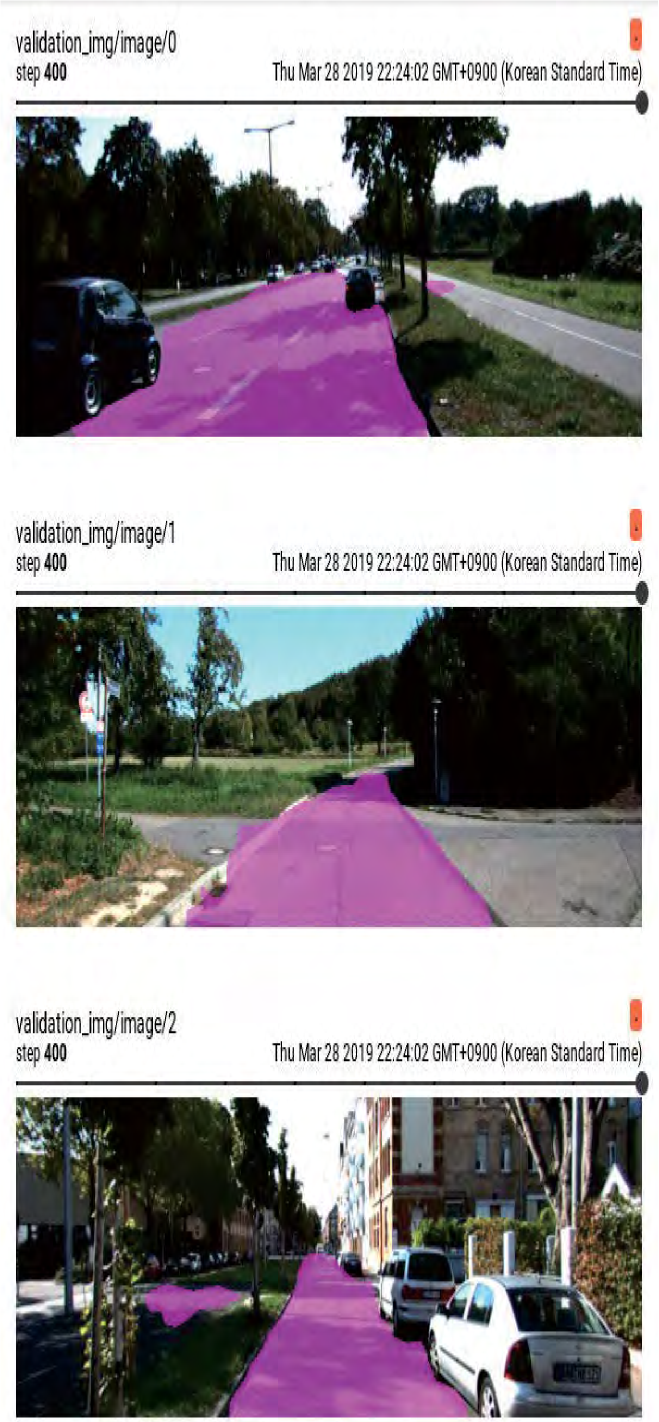
Input image

Predicted image

Input image

Predicted image

(a)


validation_img/image/0
step 400
Thu Mar 28 2019 22:24:02 GMT+0900 (Korean Standard Time)
validation_img/image/1
step 400
Thu Mar 28 2019 22:24:02 GMT+0900 (Korean Standard Time)
validation_img/image/2
step 400
Thu Mar 28 2019 22:24:02 GMT+0900 (Korean Standard Time)

(b)

Figure 3. Visualization of results (a) results on the test set of KITTI dataset (b) visualization of segmentation mask generation during training

Figure 3 explains it quite well that the proposed model performs well in classifying the road pixels versus non road pixels. We use standard evaluation metric mean intersection over union (mIoU) to evaluate our model. Our proposed model yields a test accuracy of 92 .04 % on KITTI road segmentation dataset.

4. Conclusion

In this study we designed a deep architecture which can investigate how fully convolution networks can be used with deconvolutional pixel layer for semantic segmentation. The proposed model is tested on KITTI road segmentation dataset. It is proved to be capable of differentiating road pixels from the rest. It can be seen from our experiments that our method produces good results with less data as well. However, the model does not perform well in some cases where the lighting condition is poor because it has an impact on accuracy. The accuracy of model can be increased by performing data augmentation as it can be used to increase the size of training examples with different lighting conditions.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (GR 2016R1D1A3B03931911). This research was supported by “Research Base Construction Fund Support Program” funded by Chonbuk National University in 2019.

References

[1] Chéron, G., Laptev, I. and Schmid, C., 2015. P-cnn: Pose-based cnn features for action recognition. In Proceedings of the IEEE international conference on computer vision (pp. 3218-3226).

[2] LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), pp.2278-2324.

[3] Gao, H., Yuan, H., Wang, Z. and Ji, S., 2017. Pixel deconvolutional networks. arXiv preprint arXiv:1705.06820.

[4] Geiger, A., Lenz, P. and Urtasun, R., 2012, June. Are we ready for autonomous driving? the kitti vision benchmark suite. In 2012 IEEE Conference on Computer Vision and Pattern Recognition (pp. 3354-3361). IEEE.

[5] Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1874–1883, 2016.

[6] Badrinarayanan, V., Handa, A. and Cipolla, R., 2015. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. arXiv preprint arXiv:1505.07293.

[7] Deconvolutional networks. Matthew D Zeiler, Dilip Krishnan, Graham W Taylor, and Rob Fergus. In Computer Vision and Pattern Recognition (CVPR), 2010

IEEE Conference on, pp. 2528–2535. IEEE, 2010.

[8] Long, J., Shelhamer, E. and Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).

[10] He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[11] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., 2009, June. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.

[12] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C.L., 2014, September. Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Springer, Cham.

[13] Everingham, M., Van Gool, L., Williams, C.K., Winn, J. and Zisserman, A., 2010. The pascal visual object classes (voc) challenge. International journal of computer vision, 88(2), pp.303-338.