Deconvolutional Pixel Layer Model for Road segmentation without Human Assistance

Abdul Wahid, Muhammad Intizar Ali

Data Science Institute, National University of Ireland Galway, Ireland
{a.wahid2,ali.intizar}@nuigalway.ie

Abstract. The autonomous vehicle is the evolutionary goal of designing Advanced Driver Assistance System, (ADAS) to the point where human assistance is not needed anymore. To create a broad drivable geographical area and mapping for route planning we need robust and efficient semantic segmentation algorithm. Convolutional Neural Networks (CNN) have been able to achieve state of the art performance for tasks such as Image classification, Face recognition and Detection. However semantic segmentation has remained a challenging problem in the field of computer vision. With the help of convolution neural networks, we have witnessed prolific results over time. We propose a convolutional neural network model which uses Fully Convolutional Neural Network (FCN) 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 the Karlsruhe Institute of Technology and Toyota Technologies Institute (KITTI) street view datasets. Our method achieves a mIoU accuracy of 92.04 percent.

Keywords: Autonomous Vehicle, Semantic Segmentation, Road Segmentation, Deconvolutional Pixel Layer, Convoluttional Neural Network (CNN), Fully Convolutional Network (FCN)

1 Introduction

Over the period a great research interest has been developed for the autonomous driving system. Road image segmentation is one of the essential modules in the advanced autonomous driving system. Efficient road segmentation algorithms detect the drivable roads and create a route mappings [8]. The challenging problem in an autonomous vehicle is to model or predict the movement of humans based on their behavior. To improve this we need to understand the patterns of interactions that happen between people in places like pavements, busy roads, and airports, etc [1]. Some researches have successfully addressed these problems [9] by proposing how understanding the semantics of the locations such as pavements or sidewalks can help in predicting the future movement of pedestrians more accurately.

Segmentation of road images has been a hot topic for research since the time cameras started generating high-resolution images. There are many studies in the literature that used old traditional image processing techniques such as histograms and edge detection [17] for segmentation of road images. However, the algorithms were not accurate enough when implemented in real-time and could not be extended to new and complex environments.

In recent times Convolutional Neural Networks (CNN) has attracted a lot of research interest because of their immense power and capability of solving complex problems in an efficient way. 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 [7]. Layers like pooling, convolutional [11], fully connected and deconvolution, have been extensively used. Deconvolutional layers [4] are used in the networks whose features maps need to be upscaled or upsampled such as FCN [13] and SegNet [2]. However, one of the main disadvantages of the deconvolutional network is the uneven overlap because of the output window size, which creates a checkerboard pattern of varying magnitudes [5]. This study is based on the idea of the deconvolutional network, where the resulting feature maps are nothing less than a result of recurrent shuffling of several feature maps which are actually computed from the input using the standard or independent convolutions. Therefore resulting in output feature maps with adjacent pixels which are not related directly to one another, thus leading to checkerboard problems.

In this paper, we propose an efficient model that combines FCN with deconvolutional layer to overcome checkerboard problems in deconvolutional layers. The deconvolutional pixel layer generates a pool of feature maps in the latter phase of the model to depend on the initially generated feature maps [5], thus creating a space to connect directly the adjacent pixels to each other on the resulting output feature maps. Our experimental results show the efficiency of our method on a challenging KITTI dataset [6] and high performance in road segmentation.

2 Related Work

Deep learning-based approaches have attracted a lot of attention because of their intelligent problem-solving capabilities. Tasks such as Image Classification touched new heights after the creation of AlexNet [10] and as a result nowadays deep learning is used for most of the image classification tasks. Similarly deep learning has been widely used for semantic segmentation tasks, especially to train deep networks for road segmentation [14]. One of the first models proposed for semantic segmentation was fully convolutional networks (FCN) [13] which were end-to-end trainable. Since its inception many versions of FCNs have been proposed [5, 19, 15]. Better results have been achieved between conditional random fields and FCN [3]. Decreased resolution has been a challenge in segmentation however dilated convolutions [18] addressed this problem by introducing the concept of augmenting receptive fields while preserving the resolution. As compared

to the other methods our proposed model learns more dense and rich features with the help of intermediate relationships between different feature maps. It achieves comparable accuracy with respect to the state-of-the-art methods.

3 Deconvolutional Pixel Layer Model

Figure 1 provides an overview of our proposed architecture. The architecture consists of two parts the FCN based feature extractor and the deconvolutional pixel layer. FCN generates output maps for inputs of any size, the dimensions of the output feature maps are reduced by downsampling the size of input by a factor equal to the pixel stride of the receptive fields of the output units. For associating the coarse outputs to dense pixels we use deconvolutional pixel layer. The deconvolution layer performs the periodic shuffling of various intermediate feature maps generated by different convolutional operations [16].

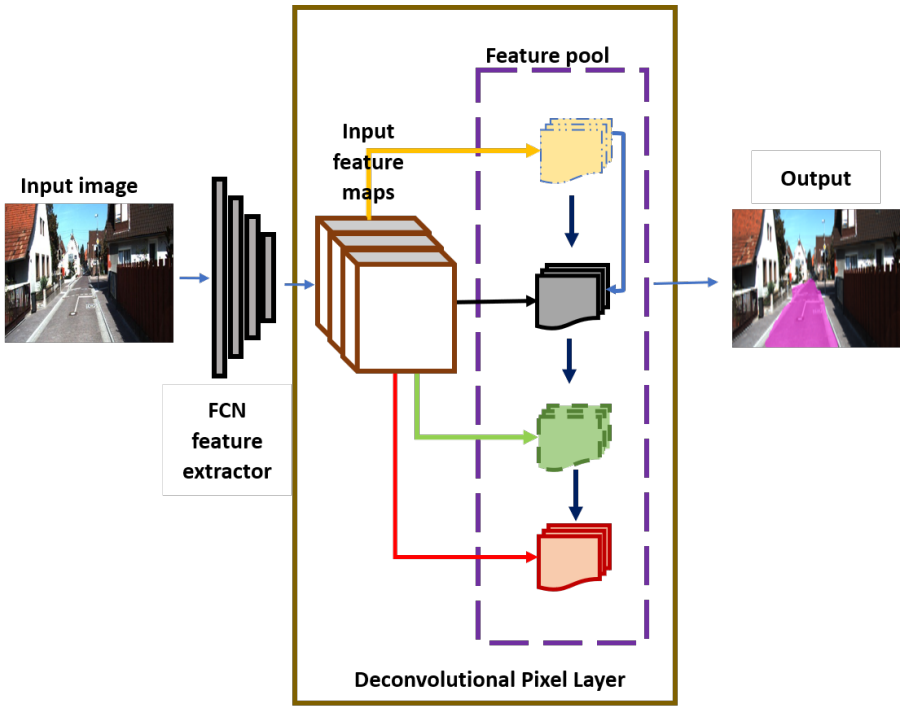

Fig. 1. An end-to-end architecture with FCN and deconvolutional Pixel layer for road semantic segmentation

The deconvolutional pixel layer up samples the 4 x 4 feature map to 8 x 8 feature map as shown in figure 2. The feature map in yellow color in the deconvolutional pixel layer is obtained by a 3 x 3 convolution. Another 3 x 3

convolutional is performed on the yellow feature map in order to generate the feature map of grey color. The yellow and grey feature maps are added together because they are dilated to form a single large feature map. Additionally, we use masked 3x3 convolution on the last two feature inputs, thus they are combined together into one big feature map. The last two feature maps are combined because of their missing relationships. In pixel upsampling layer the feature maps are divided into four groups, while as in other upsampling techniques the feature maps are increased only by the factor of two. One of the advantages of the proposed method is that the information from the input feature map is passed on to other intermediate feature maps, therefore, making it easy for the model to learn more dense features and the relationship between the pixels. Thus reducing the dependencies for training the model efficiently as a result of a collective intermediate relationship between the feature maps [5].

4 Experimental Implementation and Evaluation

The architecture designed uses a VGG-16 Convent trained on Imagenet as the encoder and a decoder based on FCN-8 [13] with a deconvolutional pixel layer [5]. We performed a series of experiments on KITTI [6] 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. We used mean intersection over union (mIoU) as the metrics to calculate the accuracy of road segmentation.
Displayed equations are centered and set on a separate line.

$$IoU = 100 * \frac{true-positive}{true-positive+false-negatives+false-positives} \tag{1}$$

In the above equation true-positives represent the correctly predicted pixels of a specific class While the false-positives are those pixels which are predicted incorrectly. In other words pixels which belong to a different class other than the correct class. Since the training set consists of only 298 training samples, 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 of three categories of street views. The model was trained using TensorFlow framework on Pascal NVIDIA 1080 Ti GPU. Figure 2 shows the training loss of the proposed model.

1. UU – Urban unmarked (98/100)
2. UM – Urband unmarked (95/96)
3. UMM – Urban multiple marked lines (96/94)

Figure 3 and 4 show the output generated by the proposed model for road segmentation on the test set of KITTI dataset. It demostrates how well the proposed model performs in classifying the road pixels versus non road pixels. We

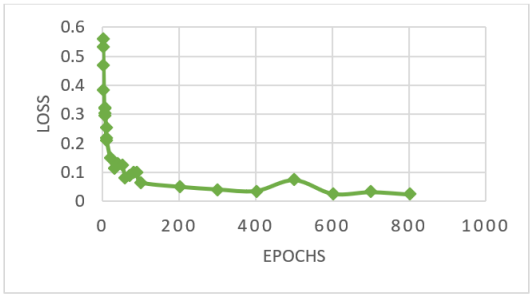

Fig. 2. Training Loss

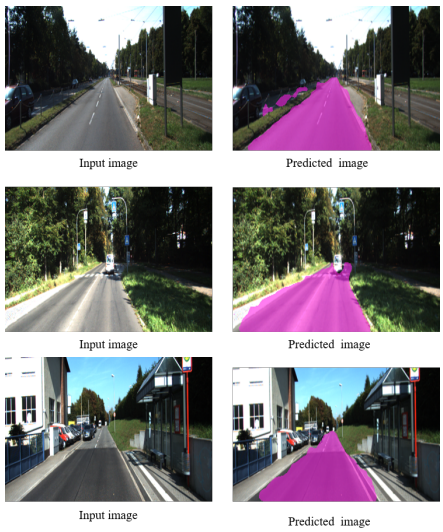

Fig. 3. Represents the raw input images and their corresponding predicted output

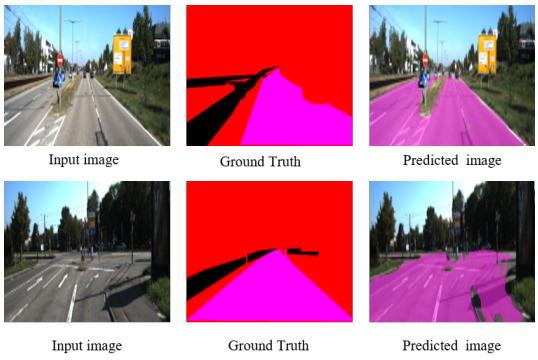

Fig. 4. Shows the predicted images with their respective ground truths and the input image.

used standard evaluation metrics mean intersection over union (mIoU) to evaluate our model. Our proposed model yields a test accuracy of 92.02 % accuracy on KITTI road segmentation dataset when compared to [12].

5 Conclusion

In this study, we confer the application of our proposed deep end-to-end architecture which investigates how fully convolution networks can be used jointly with deconvolutional pixel layer for semantic segmentation. The proposed model is tested on a benchmark KITTI road segmentation dataset. We also discussed the improved capability of this joint network to efficiently differentiate road pixels from the rest. It can be seen from our experiments that our method produces considerable results with fewer data. However, the model struggles with the scenarios with poor lighting conditions. In future, we would like to explore visual perception with detection and fusion of several cameras for more diverse data with different illumination effects.

Acknowledgement. This publication has emanated from research supported in part by a research grant from Science Foundation Ireland under Grant Numbers SFI/16/RC/3918 (Confirm SFI Research Centre for Smart Manufacturing), SFI/12/RC/2289_P2 (Insight SFI Research Centre for Data Analytics) and Enable Spoke, co-funded by the European Regional Development Fund.

References

1. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S.: Social lstm: Human trajectory prediction in crowded spaces. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 961–971 (2016)
2. Badrinarayanan, V., Handa, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. arXiv preprint arXiv:1505.07293 (2015)
3. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062 (2014)
4. Escalante, H.J., Escalera, S., Guyon, I., Baró, X., Güçlütürk, Y., Güçlü, U., van Gerven, M.: Explainable and interpretable models in computer vision and machine learning. Springer (2018)
5. Gao, H., Yuan, H., Wang, Z., Ji, S.: Pixel deconvolutional networks. arXiv preprint arXiv:1705.06820 (2017)
6. Geiger, A., Lenz, P., Urtasun, R.: 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 (2012)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
8. Hillel, A.B., Lerner, R., Levi, D., Raz, G.: Recent progress in road and lane detection: a survey. Machine vision and applications 25(3), 727–745 (2014)

9. Kitani, K.M., Ziebart, B.D., Andrew, J.: Bagnell, and martial hebert. activity forecasting. In: European Conference on Computer Vision. Springer. vol. 59, p. 88 (2012)
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
11. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al.: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11), 2278–2324 (1998)
12. Levi, D., Garnett, N., Fetaya, E., Herzlyia, I.: Stixelnet: A deep convolutional network for obstacle detection and road segmentation. In: BMVC. pp. 109–1 (2015)
13. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
14. Ma, W.C., Wang, S., Brubaker, M.A., Fidler, S., Urtasun, R.: Find your way by observing the sun and other semantic cues. In: 2017 IEEE International Conference on Robotics and Automation (ICRA). pp. 6292–6299. IEEE (2017)
15. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
16. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1874–1883 (2016)
17. Yoo, H., Yang, U., Sohn, K.: Gradient-enhancing conversion for illumination-robust lane detection. IEEE Transactions on Intelligent Transportation Systems 14(3), 1083–1094 (2013)
18. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
19. Zeiler, M.D., Krishnan, D., Taylor, G.W., Fergus, R.: Deconvolutional networks. In: 2010 IEEE Computer Society Conference on computer vision and pattern recognition. pp. 2528–2535. IEEE (2010)