The Robust Segmentation Challenge

Wei Zheng[0] [w.zheng@student.tue.nl] and Ryo Sugimura[0] [r.sugimura@student.tue.nl]

I. Introduction

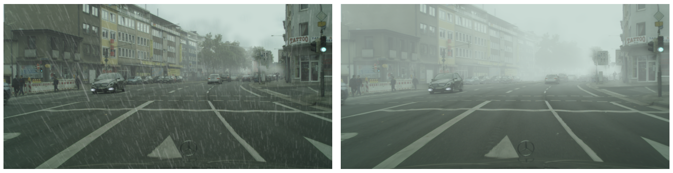
Computer vision is necessary to capture information based on images. One of the ways is semantic segmentation where you define which pixels are in which category. A semantic segmentation baseline was created using the U-Net in the previous assignment. We have tested the model using the Cityscapes dataset which is also used to train. In this paper, the model and preprocessing will be adjusted in order to account for other variances. These variances include different weathers (e.g. rain, fog, snow, etc.) and different cities. This tests the model against the degradation of image quality and generalization and the robustness of the model is considered.

II. Methodologies

A. Data

The original images are too large for us to use as the input to the model as it increases the computation time and load. So the input image is reduced to 256×512 in order to reduce the training and evaluation time. The original dataset does not contain any different weather conditions that may degrade the quality of the images. Without those data, the model cannot learn the features of those weather conditions. So the data augmentation of those weather conditions will be added to the dataset. Also, the daytime and low light adjustment will be augmented as well. The brightness of the images is adjusted in order to account for day-time differences and lighting conditions. Furthermore, Gaussian noise is added to the image for better generalization performance. Basic augmentation is also added onto the image such as flip and rotation of the images. This will hopefully prevent the model from overfitting to the training data leading to lower validation accuracy.

Additional datasets were also investigated to check the performance. The rainy and foggy images of the cityscapes were given in the Cityscapes website[1]. These two different kinds of datasets will be included in the train/validation/test set. A sample of rainy and foggy images is shown in Fig. 1.



(a) (b)

Fig. 1: Rainy image (a) and foggy image (b) from dataset

These images are the same as the original one with different weather conditions, so the mask or label is the same as the original one. Therefore, the images are added to the preprocessed data file after resizing and the same thing for the labels. Another positive point of the model is that it is computationally lighter compared to recent State-of-the-Art models. Adding these images will increase the variety of the types of images given allowing for better performance.

B. Model

The model that has been selected for semantic segmentation is the DeepLabv3 in combination with ResNet. DeepLabV3 is a model trained for semantic segmentation with good results on the Cityscapes dataset..The architecture of the model is shown in Fig. 2[2].

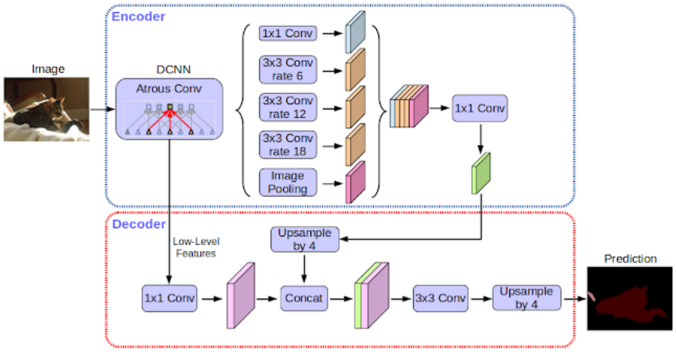


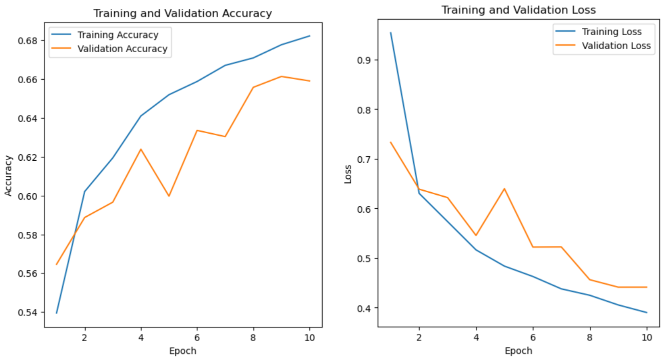
Fig. 2: Architecture of DeepLabV3+

The model incorporated in this assignment incorporates the ResNet to this architecture. It has the encoder-decoder structure with ResNet101 convolution as the backbone of the model. There also is atrous convolution which allows the convolution of a larger receptive field without increasing the number of parameters. Similar to the UNet model which was used as the baseline, there are skip connections that connect the encoder and decoder allowing the model to preserve the spatial information obtained in the earlier layers while the deeper layers extract the content information. The skip connections are for the stacked feature maps with 1×1 convolution.

C. Training

The accuracy measure for this task will be the Intersection Over Union (IOU) and the loss measure is cross-entropy loss. These measures will be used in order to evaluate the model quantitatively and qualitatively. For the quantitative evaluation, the training/validation loss and accuracy of the model are plotted over epochs.

[0] *dept. of Electrical Engineering*, *University of Technology Eindhoven*, Eindhoven, The Netherlands



(a) (b)

Fig. 3: Accuracy (a) and loss (b) of baseline (UNet)

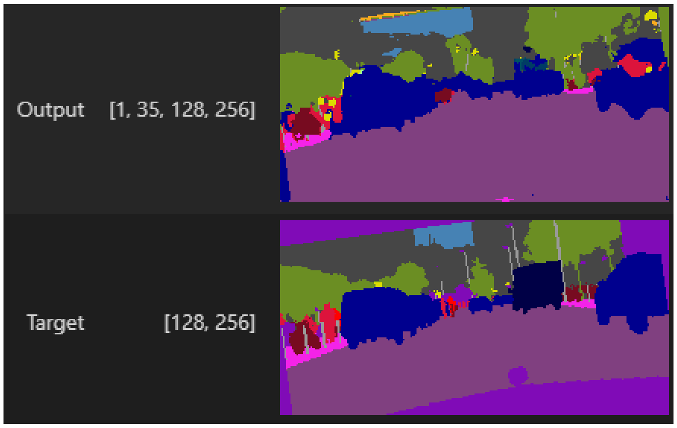


Fig. 4: Segmentation results (UNet)

III. Results

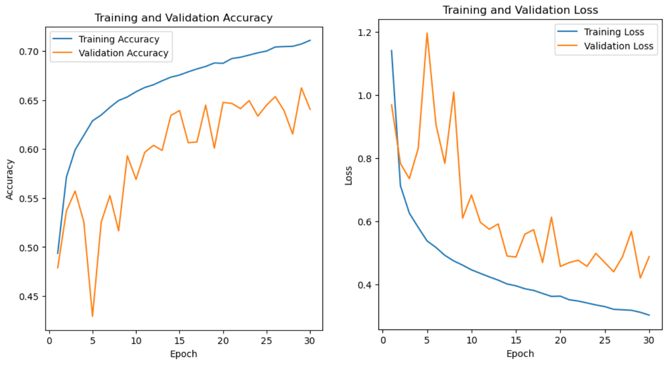
A. Baseline (UNet)

For the baseline, the basic UNet model was selected. The hyperparameters of the model were optimized manually for the baseline. The results of UNet are shown in Fig. 5. Fig. 3a shows the accuracy of the model, while Fig. 3b shows the loss.

The accuracy of training and validation goes up as the training goes on while the loss goes down and they are close. The final accuracy of the training was about 70% and validation was 65%. However, the accuracy and loss have not yet converged due to the limitation of the hardware. In addition, the segmented images and the actual mask (truth) are shown in Fig. 4. Although the smallest details are not correctly segmented in the predictions, the broad segmentation seems to be in the correct places.

B. Modified Model (DeepLabV3+ResNet)

Given the model, accuracy, criterion and optimizer explain in Sec. II-B. The hyperparameters of the model are fine-tuned by manually. The results of training and evaluation are given in Fig. 6. Fig. 5a shows the accuracy of the model, while Fig. 5b shows the loss.



(a) (b)

Fig. 5: Accuracy (a) and loss (b) of modified model (DeepLabV3+ResNet)

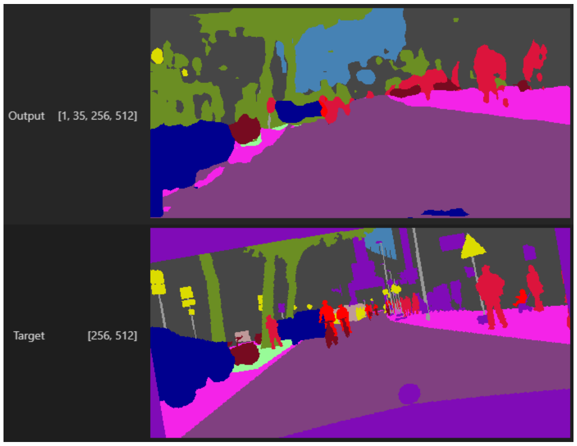


Fig. 6: Segmentation results (DeepLabV3+ResNet)

The modified model's training and validation have similar characteristics to the baseline as epochs increase the accuracy increase and loss decreases. However, the accuracy and loss of the validation have more variance or ups and downs. The modified model has also a larger difference between the training and validation loss. Fig. 6 shows the semantic segmentation of the images. Details such as signs and small objects did not have correct predictions while the larger objects had better segmentation results.

Although there is extra data with rain and fog images, the model performance on image quality and generalization wasn't better compared to the baseline in CodaLab.

IV. Discussions

A. Comparison of the results

In this work, the Deeplabv3+ model with ResNet101 as the backbone was employed to address the semantic segmentation problem on the Cityscapes dataset. While the model demonstrated comparable generalization performance to the baseline U-Net, it exhibited weaker performance in terms of image quality on Codalab.

B. Limitations of the model

The input image size of (512x256) might not be sufficient to fully exploit the capabilities of the DeepLabV3+ model. If computational resources permit, using higher-resolution images during training could help the model capture more fine-grained details and improve the image quality of the segmentation results.

With only 5,000 images in the training dataset, the model might not have enough data to learn complex features and generalize well. The small dataset size could also contribute to overfitting. Increasing the size of the training dataset or applying appropriate data augmentation techniques can help the model generalize better and reduce the risk of overfitting. However, it is important to ensure that the chosen data augmentation methods do not introduce excessive distortions or artificiality. Some inappropriate data augmentation can lead to bad training performance.

ResNet101, as the backbone of the DeepLabV3+ model, might not be the best choice for this task. Although it is a powerful feature extractor, the trade-off between its complexity and the benefits it provides should be considered, as it may lead to overfitting or reduced generalization performance. It might be too deep for the input image size of the given task. So exploring different backbone architectures, or even a custom-designed backbone, may yield better performance in terms of both generalization and image quality.

Atrous convolutions used in the DeepLabV3+ model may lead to the artificiality or reduced image quality in the segmentation results. This may be caused by the large dilation rates used for atrous convolutions, which can result in checkerboard-like patterns or discontinuities in the segmented regions (As shown in Fig. 6. Adjusting the dilation rates of atrous convolutions or employing other convolution techniques, such as depth-wise separable convolutions or dilated spatial pyramid pooling, might help alleviate artificiality and improve image quality.

V. Conclusion

The methods that were chosen were to obtain a better model performance with different image quality and generalization. However, better performance of the model was not achieved due to multiple reasons. The reasons include hardware restrictions like small datasets and the simplicity of models. Furthermore, having a pre-trained model for the backbone of the model and realistic augmentation of the data to avoid unnecessary artificiality may improve the model's performance. For the model, there are hyperparameters that had to be fine-tuned in order for the model to perform better. These small reasons added up to result in worse performance compared to the baseline.

References

[1] *The cityscapes dataset*. [Online]. Available: https://www.cityscapes-dataset.com/.

[2] M. 1. Monday and C. V. LearningTensorFlow, *Semantic image segmentation with deeplab in tensorflow*. [Online]. Available: https://ai.googleblog.com/2018/03/semantic-image-segmentation-with.html.