Semantic segmentation on CityScapes dataset

Bierenbroodspot S.A.K.
dept. Electrical Engineering, Technische Universiteit Eindhoven

Eindhoven, the Netherlands
s.a.k.bierenbroodspot@student.tue.nl

***Abstract*—**

I. Introduction

The CityScapes dataset is created to capture the complexity of real-world urban scenes. This dataset is a benchmark suite and large-scale dataset to train and test approaches for pixel-level and instance-level semantic labeling [1]. It contains dash cam frames from a vehicle driving in 50 different German cities.The frames are manually selected to obtain a large number of dynamic objects, a varying scene layout, and a varying background. The subset we are using contains a total of 2975 images.

There are multiple challenges to face for a semantic segmentation model on this dataset. The first of which is the high resolution of the images. Since the images have a resolution of $(1024, 2048)$ they are very large compared to other datasets, for example, it is more than 10 times larger than ImageNet which has an average resolution of $(469, 387)$ [2]. This can be computationally expensive during the training of a model. Another challenge are the fine-grained object boundaries, urban scenes often contain a lot of small details and a lot of details. A good model should be able to capture all details of varying sizes and positions. Since a model application such as autonomous driving would require (near) real-time inference the efficiency of the model is also an important challenge for this problem. The final challenge this paper will focus on is the generalization to unseen environments. Since the application of the model requires understanding of the input, it is important that the model generalizes well so it will perform as expected when it is exposed to new environments.

The first published model implementation on this dataset is DeepLab [3]. This model combines the strenght of deep convolutional networks with a "Conditional Random Field (CRF)" in order to solve the poor localization of these deep convolutional networks. In 2017, the ResNet-38 model was published [4]. The authors addressed gradient vanishing in deep residual networks and optimized the structure in a way it can be trained end-to-end. The current state of the art model is VLTSeg [5]. VLTSeg combines image based features as well as text based features for the segmentation.

II. Baseline implementation and results

A. The Unet model

For the baseline implementation a Unet architecture is used. This architecture features downsampling using convolutional layers, after which the image is upsampled again to the original image size. Each of the down- and upsampling layers also includes a skip connection from the downsampling side towards the upsampling side to help the model identify both high and low level features of the image. More details can be found in the original paper [6].

B. Training the model

To get a working model, inspiration was drawn from the blog of Andrej Karpathy (one of the co-founders of OpenAI). One of his blog posts describes a method for training a neural network with useful tips and hyperparameter guidelines [7]. All the images are transformed to a tensor and resized to (128,128). The reason for the smaller size is that the original large image size requires too much memory to train the model on, especially with a decent batch size for faster training time. The data is split into a training set and a validation set. The sizes of these sets are 80% and 20% of the original dataset size respectively. The split is generated with a manual seed equal to: "2147483647" for reproducibility. There is an external test set available with 500 images. There are limited possibilities to test on this test set, therefore some values are missing in the results.

The loss function of choice is cross-entropy loss due to its natural suitability for classification problems. The SGD optimizer was used with a learning rate of "0.001". The batch size is equal to 64 which is determined by the amount of GPU memory available. The model is trained for a total of 60 epochs.

C. Evaluation

The average DICE score is considered as the main evaluation metric for this research. This DICE score, also known as the Sørensen-Dice coefficient [8], consists of a value between 0 and 1 for each of the classes, where 1 means a perfect segmentation for this class and 0 means there is no overlap between the prediction and the ground truth label. The equation for this metric is given by

$$DICE(A, B) = \frac{2|A \cap B|}{|A| + |B|} \tag{1}$$

where A and B represent a mask of whether the pixel belongs to the current class. Repeat this for all of the classes and take the average and this is the average DICE score this paper uses.

A secondary form of evaluation is manually visualizing the prediction and assessing its properties. For a comparison, all predictions are displayed together with the original image and the ground truth label.

D. Results

The model is trained in 3 sessions and the intermediate result is evaluated for its DICE score. Table I shows the average DICE scores for all different datasets at different time points. There is no overfitting behaviour since the validation score is not lower than the training score. It is possible that when training for a longer period of time the validation DICE score will still go up. Note that the DICE score on the external test set is considerably lower. This is due to the difference in evaluation, at the external test location the predictions are resized to the original image size whereas on the training and validation set the smaller size is evaluated.

TABLE I
AVERAGE DICE SCORE OF THE TRAINING, VALIDATION, AND EXTERNAL TEST DATASET FOR THE BASELINE MODEL IMPLEMENTATION.

Epochs trained	**Average DICE score**		
	Training	***Validation***	***Test***
25	0.21858	0.22154	-
35	0.25815	0.26083	-
60	0.31615	0.31719	0.24736

Figure 1 shows an image from the validation set, the ground truth label (Target), and the model prediction. The prediction shows the models focus on the classes that appear a lot. Most of the image is assigned to road, building, or car. Overall, the model is able to distinguish low-level features. Crucially, the model does not segment the pedestrians in the image. For practical application it can be assumed that this would be a fatal error in the model and improvements need to be made. In the next section the model will be improved to reach peak performance.

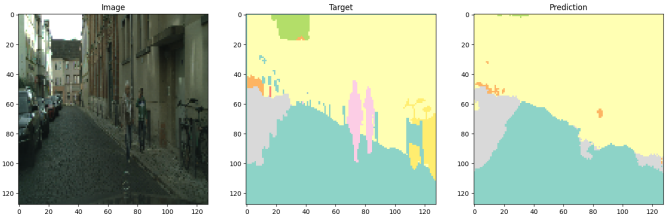

Fig. 1. A sample image from the validation set displayed as the original image, the ground truth target, and the prediction of the baseline model. All images have been resized to (128,128)

III. PEAK PERFORMANCE

To reach this peak performance there are two versions to be discussed in this paper. The first version will be called: ”Modern Unet”, the second version will be called: ”Modern Unet PT”.

A. The Modern Unet model

In order to increase the performance of the implementation on this problem a new model is implemented. This model is referred to as a ”Modern Unet” and is introduced in the paper ”Denoising Diffusion Probabilistic Models” [9]. The implementation used in this paper is largely based on GitHub pdearena which is largely based on GitHub labml.ai

TABLE II
AVERAGE DICE SCORE OF THE TRAINING, VALIDATION, AND EXTERNAL TEST DATASET FOR THE MODERN UNET IMPLEMENTATION.

Epochs trained	**Average DICE score**		
	Training	***Validation***	***Test***
100	0.57843	0.55278	0.43405

TABLE III
AVERAGE DICE SCORE OF THE TRAINING, VALIDATION, AND EXTERNAL TEST DATASET FOR THE BASELINE MODEL IMPLEMENTATION.

Epochs trained	**Average DICE score**		
	Training	***Validation***	***Test***
50	0.62800	0.60973	-
100	0.65855	0.62446	-
180	0.68704	0.63450	0.48619
200	0.68698	0.63213	-

IV. OUT OF DISTRIBUTION DETECTION

REFERENCES

[1] Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, et al.. The Cityscapes Dataset for Semantic Urban Scene Understanding;. Available from: www.cityscapes-dataset.net.

[2] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A Large-Scale Hierarchical Image Database;. Available from: http://www.image-net.org.

[3] Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. 2014 12. Available from: http://arxiv.org/abs/1412.7062.

[4] Wu Z, Shen C, van den Hengel A. Wider or Deeper: Revisiting the ResNet Model for Visual Recognition. 2016 11. Available from: http://arxiv.org/abs/1611.10080.

[5] Hümmer C, Schwonberg M, Zhou L, Cao H, Knoll A, Gottschalk H. VLTSeg: Simple Transfer of CLIP-Based Vision-Language Representations for Domain Generalized Semantic Segmentation. 2023 12. Available from: http://arxiv.org/abs/2312.02021.

[6] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. 2015 5. Available from: http://arxiv.org/abs/1505.04597.

[7] Karpathy A. A Recipe for Training Neural Networks; 2019. Available from: https://karpathy.github.io/2019/04/25/recipe/.

[8] Zou KH, Warfield SK, Bharatha A, Tempany CMC, Kaus MR, Haker SJ, et al.. Statistical Validation of Image Segmentation Quality Based on a Spatial Overlap Index 1 : Scientific Reports; 2004. Available from: http://www.slicer.org.

[9] Ho J, Jain A, Abbeel P. Denoising Diffusion Probabilistic Models;. Available from: https://github.com/hojonathanho/diffusion.

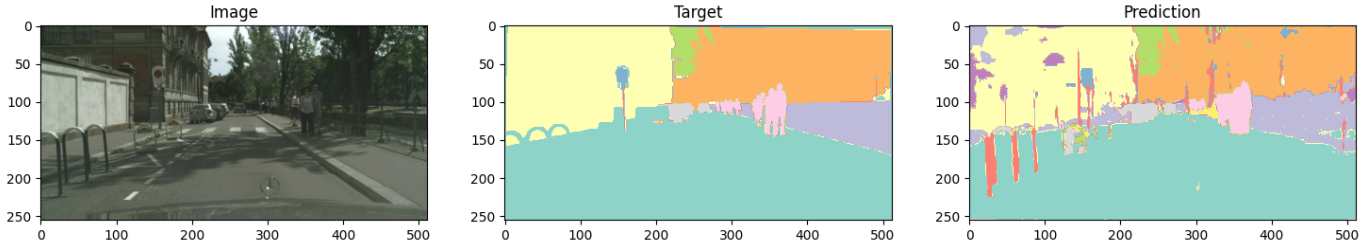

Fig. 2. A sample image from the validation set displayed as the original image, the ground truth target, and the prediction of the Modern Unet model. All images have been resized to (256,512)

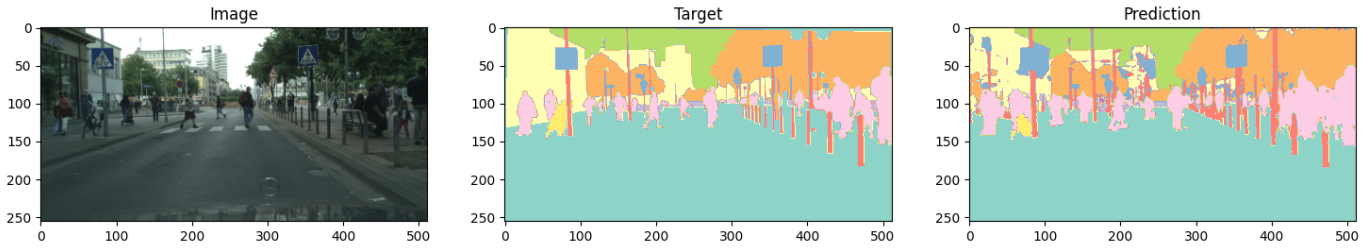

Fig. 3. A sample image from the validation set displayed as the original image, the ground truth target, and the prediction of the Modern Unet model. All images have been resized to (256,512)

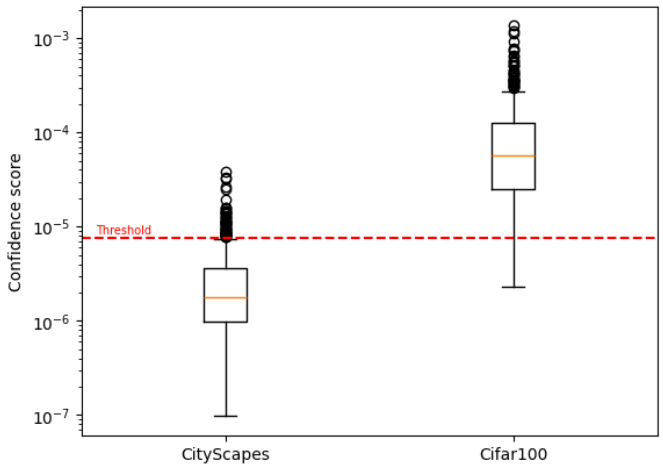

Fig. 4. A sample image from the validation set displayed as the original image, the ground truth target, and the prediction of the baseline model. All images have been resized to (128,128)