Deep Learning on Images and Signals

U-Nets in a Self-Driving Segmentation Task

Nils Fahrni

BSc Data Science Student

December 3, 2024

Contents

Chapter 1

Task

1.1 Research Question

The research question which I aim to answer with this mini-challenge is:

“How does the performance of a U-Net semantic segmentation model differ between scenes of city streets and non-city streets in the BDD100K dataset?”

This research question is interesting because it addresses the practical challenges of deploying semantic segmentation models in real-world applications, such as autonomous driving, where environmental variability is a key concern. City and non-city environments differ significantly in terms of visual characteristics, object density, and lighting conditions, which can impact the performance of computer vision models.

1.2 Dataset

I decided for the BDD100K dataset. The BDD100K dataset is the largest driving video dataset, featuring 100,000 videos and supporting 10 tasks for evaluating and advancing multitask learning in autonomous driving. It offers diverse geographic, environmental, and weather conditions, making it a benchmark for studying heterogeneous multitask learning and training robust computer vision models [1].

For this Mini Challenge I use the "10k" subset, which is made up of 10,000 RGB images and are sampled from the 100,000 videos' frames. This subset is intended for semantic segmentation tasks. These 10,000 images have already been pre-partitioned into a train, validation and test partition. The train partition consists of 8000 images, the validation 1000 and the test 1000.

This smaller subset does unfortunately not have scene attributes but the larger video dataset does. Since the semantic segmenatiton subset is derived from the video dataset, I retrieve the scene attributes through the larger datasets metadata JSON. The issue here is that not all images in the semantic segmentation subset seem to be in the video dataset. I will therefore only use the small overlap of images that has scene attributes and exists both in the video and semantic segmentation dataset. This overlap consists of 3426 images.

1.3 Methodology and Procedure

To answer the research question, I will train a **U-Net** model on the BDD100K dataset. The U-Net architecture is a common choice for segmentation tasks as it contains an Encoder-Decoder structure:

1. **Encoder**: Extracts features from the input image using a series of convolutional and downsampling operations, capturing contextual information. The downsampling path in the encoder captures features at multiple scales, enabling the model to understand both local and global context. This is essential in self-driving tasks to distinguish between small objects (like traffic cones) and large areas (like the road).
2. **Decoder**: Gradually upsamples the feature maps and uses convolutions to predict dense segmentation maps. This structure is ideal for segmenting objects in self-driving scenarios, such as lanes, vehicles, pedestrians, and road signs.

As a second model I will modify the U-Net model to include an attention mechanism, which I will implement myself according to [2]. I expect the addition of attention to be another improving factor because:

1. **Driving Reality**: Different objects and their spatial relationships often define the context. For instance:
   (a) A cyclist is more likely to be found near a bike lane or the edge of a road.
   (b) A pedestrian might be near a crosswalk but not in the middle of a highway. Cars and trucks are expected on roads but not sidewalks.
2. **Attention Benefit**: The attention mechanism allows the model to focus not just on isolated objects but also on the relationships between them. It helps the network infer that the presence of a bike lane increases the likelihood of a cyclist or that highway lanes imply the absence of pedestrians and cyclists.

I will then evaluate the models on the test set and compare the performance between city street and non-city street scenes. To measure and evaluate the performance between different model complexities numerically, I will use the mean Intersection over Union (mIoU) as the evaluation metric.

Both models will be explored within the schema of "A Recipe for Training Neural Networks" according to Andrej Karpathy [3].

Chapter 2

Discussion

Chapter 3

Reflection

Bibliography

[1] F. Yu, H. Chen, X. Wang, *et al.*, *BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning*, Apr. 2020. arXiv: 1805.04687. (visited on 11/25/2024).

[2] O. Oktay, J. Schlemper, L. L. Folgoc, *et al.*, *Attention U-Net: Learning Where to Look for the Pancreas*, May 2018. DOI: 10.48550/arXiv.1804.03999. arXiv: 1804.03999. (visited on 12/03/2024).

[3] A. Karpathy, *A Recipe for Training Neural Networks*, https://karpathy.github.io/2019/04/25/recipe/, Apr. 2019. (visited on 12/03/2024).