Spring 2021: Deep Learning

Assignment 3 : Video Semantic Segmentation

Background

An autonomous driving system is a real-world application that would greatly require reli- ability. Unexpected incidents may occur at any time, and they require immediate, and yet appropriate, responses. In order to make a good decision, the system should fully understand the situation, therefore semantic segmentation might be a better solution to introduce AI into the field of autonomous vehicle design and development.

Semantic segmentation is fundamentally a classification task. What differentiates semantic segmentation from image classification is that semantic segmentation requires class prediction for entire pixels in the given image. Therefore, the semantic segmentation task also requires a subtle understanding of local relationships whereas image classification focuses on abstracting the given image globally.

The Highway driving dataset

Let's consider the highway driving dataset[1] here. The dataset consists of 20 sequences of 60 frames with a 30Hz frame rate. Therefore, a total of 1200 frames with annotations are provided. Originally, longer clips were recorded, and 2 seconds from each clip were trimmed. It was done with a hypothesis that a correlation between adjacent frames is key information in semantic video segmentation, every sequence is carefully an- notated while maintaining consistency. The frames in a single sequence were annotated in chronological order. Each annotator was asked to annotate adjacent frames, and formerly annotated results for prior frames were provided as their reference. On average, 2.2 annotators had annotated a single sequence.

There are 10 classes commonly appear in driving scenarios:

1. road,
2. lane,
3. sky,
4. fence,
5. construction,
6. traffic sign,
7. car,
8. truck,
9. vegetation, and
10. unknown.

[1] Kim, B., Yim, J., & Kim, J. (2020). Highway driving dataset for semantic video segmentation. arXiv preprint arXiv:2011.00674. [Link to the paper: https://arxiv.org/abs/2011.00674] [Link to the dataset: https://sites.google.com/site/highwaydrivingdataset/Download]

The ***unknown*** class includes undefined objects, the bonnet of the data-collecting vehicle, and ambiguous edges. The most relevant classes to autonomous driving were selected from the high-speed driving standpoint. As the majority of selected classes have an intuitively interpretable definition, we only define some classes here.
The ***lane*** class is literally the lane on the road. Other marks printed on the road in order to inform the drivers are excluded.
We define the ***fence*** as the structures on both side of the road. The fence class can be considered as a sub-class of the ***construction*** class. However, we have separated this class from the construction class as the fence class is one of the most notice-worthy structures observed during driving.
In Figure 1, the ***fence*** class is in red. The construction class contains every man-made structure except for the road and fence. It is indicated in purple in Figure 1. More detailed information regarding the dataset can be found in the supplementary material.

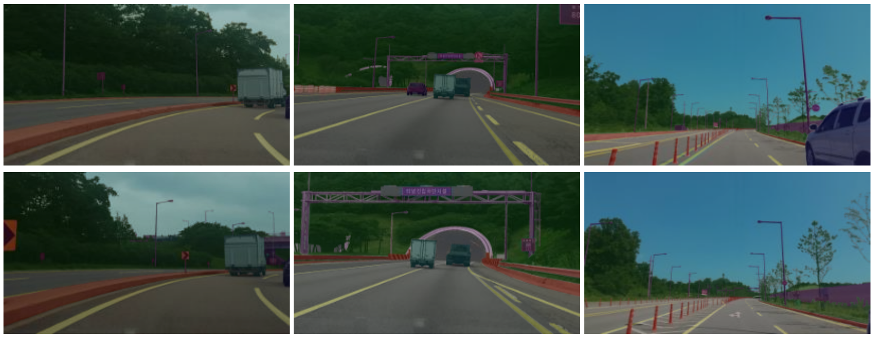

Figure 1: Samples from the collected dataset. Each image is overlaid with its annotation. The first row presents the first frames of sequences while the second row presents the last frames

Evaluation metric to be used

In order to evaluate the labeling performance for each class, we typically use the intersection over union (IoU). A pixel that is annotated as an unknown class is not considered as a performance measure.
However, as a performance measure for the whole dataset, the IoU metric is considerably biased to certain classes that cover a large area. That is undesirable as the classes covering a relatively small area are not less important. Therefore, as a metric for the whole dataset, we use the mean IoU (mIoU), which is the IoU averaged over all the classes, so that every class contributes equally to the performance measure.

We split the dataset into training and test sets. The training set consists of 15 sequences while the test set consists of the remaining five sequences. Rather than randomly splitting the sequences, we split the training and test sets to have a similar class distribution. More detailed statistics are presented in the supplementary material.

Tasks for this assignment

1. ***[For everyone enrolled***] Build a semantic segmentation model (using the 15 training video sequences) to be able to segment the 10 classes in the given 5 test video sequences.
    a. Please report both training and test mIoU.
    [Hint: You may try implementing various baseline algorithms outlined in the paper. Also, why not try U-Net? Also, several variants of U-nets might come in handy]
2. ***[For only graduate students in the class]*** Please think about deploying your program in a way that it will be able to perform semantic segmentation on real time video streams, not necessarily image frames only. Consider speed of segmentation and infrastructure requirement. Outline a reasonable budget. Also, show a demonstration in terms of Zoom recording shared with me or in a one-to-one meeting as part of the assignment.