Semantic Segmentation in the Traffic Environment with the DeepLabv3+ Model

Sheng Li, Jin Woo Park, Soyeon Jung

SP18 CS231N Final Project

{lisheng, jinwoop, soyeonj}@stanford.edu

Objectives

- Investigate the state-of-the-art semantic segmentation model DeepLabv3+ [1]
- Apply the DeepLabv3+ model to the newly-released Mapillary Vistas Dataset [2]
- Study the atrous convolution, the atrous spatial pyramid pooling (ASPP), and the encoder-decoder module in DeepLabv3+
- Train and analyze performance of neural network application in an iterative process
- Understand and perform semantic segmentation in the traffic environment that relates with autonomous driving

Mapillary Vistas Dataset

- Authenticity: road-side and street scene from multiple cities around the world with a variety of weather, season, time and light condition
- High Quality: 66 object categories with pixel-wise semantic annotations
- Diversity: diverse set of resolutions, aspect ratios and viewpoints from different imaging equipment

Pole	Street-light	Car	Traffic-sign
Road	Side-walk	Curb	Person
Billboard	Traffic-light	Building	Vegetation

Table 1: The most annotated objects

The 20,000 labeled images in the academically available dataset is pre-processed and split into:

Training	Validation	Test
18,000	1,000	1,000

Table 2: The splits of the dataset

Methodology

The DeepLabv3+ Model

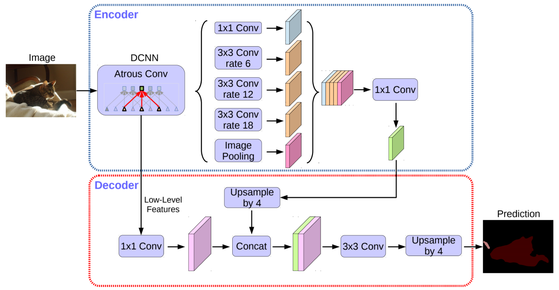


Fig. 1: DeepLabv3+ Network Architecture [1]

1. **Atrous Convolution**: convolution with atrous rate r, the stride determining the filter's field-of-view

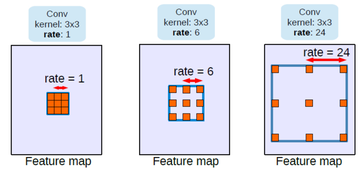
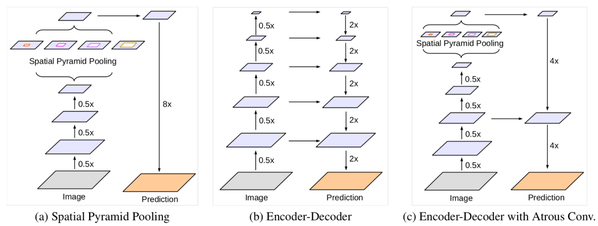


Fig. 2: Atrous convolution with various rates on a 3×3 filter (i.e. kernel)

2. **Atrous Spatial Pyramid Pooling (ASPP)**: multiple atrous convolutions in parallel with different r's; augmented with global average pooling (GAP), 1×1 convolution with 256 filters, and batch normalization

3. **Encoder-Decoder Modules**: downsample (encoder) and then upsample (decoder), to preserve the input dimension



(a) Spatial Pyramid Pooling (b) Encoder-Decoder (c) Encoder-Decoder with Atrous Conv.

Fig. 3: Encoder-decoder structure with ASPP from DeepLabv3 [1]

Implementation Details

Transfer Learning

- Using *Xception 65* [1] as the backbone of the CNN model
- Using DeepLabv3+ model pre-trained on the Cityscapes Dataset as the initial checkpoint
- Retraining the classifier, fine-tuning the weights and the batch norm parameters
- Resizing the input images for data augmentation
- Using 4 x NVIDIA Tesla K80 GPUs

Hyperparameter Tuning

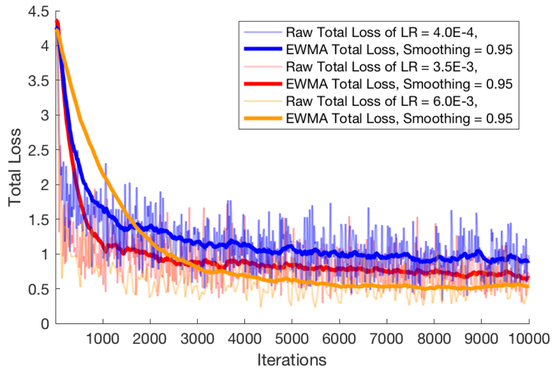


Fig. 4: Loss curves with various learning rates

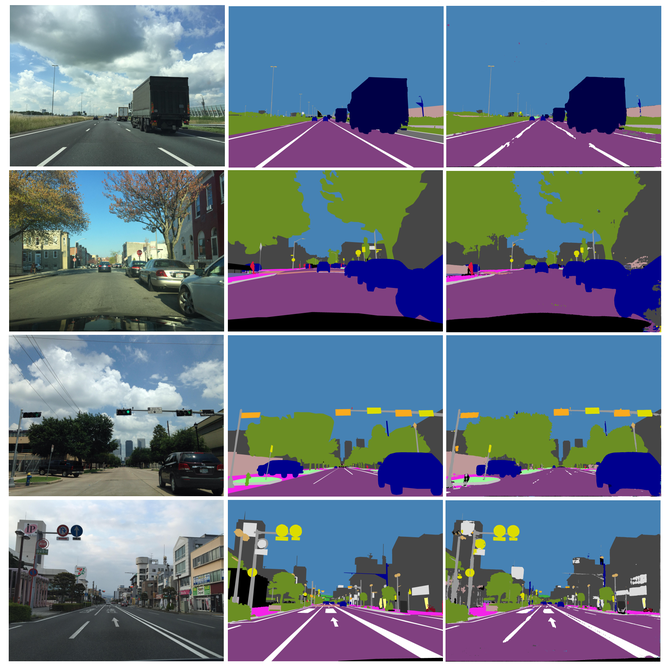
Hyperparameter	Value
Learning Rate	6E-3
Learning Rate Decay	Poly. 0.9 / 200 iter.
L2 Reg. Strength	1E-6
Image Crop Size	500x500
Batch Size	16

Table 3: Best performing hyperparameters (partial)

Results

Exp. No.	mIoU on **Val**. Set (%)	mIoU on **Test** Set (%)
Baseline	17.1	13.2
Second Best	29.1	27.2
Best	32.2	**31.1**

Table 4: mIoU of baseline and the best performing models



(a) Image (b) Ground Truth (c) Prediction

Discussion

- From practical point of view, many of the 66 classes is not so important. We can downsample the categories.
- The result rank ~No. 6 on the Mapillary Leaderboard. Still can improve by training longer with more exploration on hyperparameters

References:

[1] Chen, Liang-Chieh, et al. "Encoder-decoder with atrous separable convolution for semantic image segmentation." *arXiv preprint arXiv:1802.02611* (2018).

[2] Neuhold, Gerhard, et al. "The mapillary vistas dataset for semantic understanding of street scenes." *Proceedings of the International Conference on Computer Vision (ICCV), Venice, Italy*. 2017.