SHAPE-BIASED SEMANTIC SEGMENTATION

BŁAŻEJ OSIŃSKI, FLORIN GOGIANU

Abstract

This document should work as a lab notebook of sorts containing expositions of the ideas being tested, experimental results and their interpretation.

1 DEEPLABV3 BASELINE

Saturday, Jun 22, 2019

1.1 Short description of the model

`DeepLabV3` L. Chen et al. 2017 uses `ResNet101` for feature extraction resulting in a spatial resolution up to 16 times smaller than the input. These features are then further processed at different scales using Atrous Spatial Pyramid Pooling in which dilated convolutions are employed in order to keep the feature resolution unchanged (and therefore stop losing details by further downsampling). The prediction is then simply upsampled to the original dimensionality using a bilinear transformation.

For context, this model achieves 0.75 - .80 mIoU on *Pascal VOC* in its various incarnations and up to .86 when pre-trained on other semantic segmentation datasets.

1.2 Preliminary results

Virtual Kitti Gaidon et al. 2016 is not suggesting a train/test split and until establishing the train/test protocol I worked mostly towards showing that I can overfit the train set.

For the time being I was able to achieve only ~ 0.75 mean IoU on the train set, with the model currently not being able to demonstrate segmentation of fine details such as `TrafficLight` or `TrafficSign` poles and `Poles`, especially if in the background, as illustrated in Figure 1.

I initially thought this is because of i) the lower maximum resolution of input images (375px vs the recommended 513px) and ii) the fewer number of optimization steps in the first batch of experiments. While addressing i) improved the things a bit, IoU didn't increase significantly. Currently waiting for results on 60k optimization steps vs 30k as trained until now.

My current hypothesis is that the `output_stride`, the ratio between input and output spatial resolutions is too high right now, impeding the model from learning fine details. I'll give some details below on how I plan to address this.

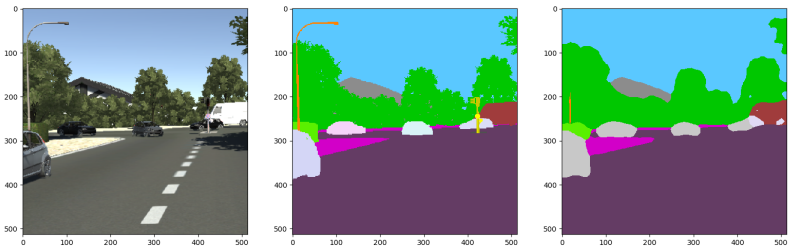

Figure 1: Initial segmentation result. Sample `0001/clone/00340`.

1.3 Training setup

The latest training setup which we will build on is described here:

1. Model definition https://pytorch.org/hub/pytorch_vision_deeplabv3_resnet101/
2. Pre-trained `ResNet-101` backbone.
3. Polynomial learning rate schedule $(1 - \frac{\text{step}}{\text{total\_steps}})^{\alpha}$
4. The `output_stride` (ratio of input to output) is 16. In L. Chen et al. 2017 this is 16 for the first 30k optimization steps then changed to 8 for the final 30k steps in order to increase the resolution of the feature maps. See the paper for details.
5. Augmentation:
   - `RandomScale` to min=1.4, max=2.0
   - `RandomCrop` of size 513px
   - `RandomHorizontalFlip`
6. Images are normalized with the ImageNet statistics.
7. BatchNorm layers are trained during the entire run. The Pascal-VOC protocol trained these layers for the first 30k steps only.
8. This configuration allows for batches of size 10.

1.4 Further steps

1. Start using the Real Kitti validation split and report the IoU.
2. Experiment with training `output_stride` of 8. This has the disadvantage of decreasing the batch size enough to mess with the `BatchNorm` layers. I am considering doing this:
   (a) Train 30k iterations with `output_stride=16, batch_size=10` and tuning `BatchNorm`
   (b) Set `output_stride=8` and freeze `BatchNorm`, thus mirroring the protocol in the original paper.
   (c) Adjust the protocol to make the comparison possible with Y. Chen et al. 2018
3. Start evaluating with `output_stride=8`, even if trained at a smaller resolution.

REFERENCES

Chen, Liang-Chieh, George Papandreou, Florian Schroff, and Hartwig Adam

2017 "Rethinking Atrous Convolution for Semantic Image Segmentation", *CoRR*, abs/1706.05587, arXiv: 1706.05587, http://arxiv.org/abs/1706.05587.

Chen, Yuhua, Wen Li, Xiaoran Chen, and Luc Van Gool

2018 "Learning Semantic Segmentation from Synthetic Data: A Geometrically Guided Input-Output Adaptation Approach", *CoRR*, abs/1812.05040, arXiv: 1812.05040, http://arxiv.org/abs/1812.05040.

Gaidon, Adrien, Qiao Wang, Yohann Cabon, and Eleonora Vig

2016 "Virtual Worlds as Proxy for Multi-Object Tracking Analysis", *CoRR*, abs/1605.06457, arXiv: 1605.06457, http://arxiv.org/abs/1605.06457.