Semantic Image Segmentation for Autonomous Driving Scenarios Combining FCNs, DeepLab, and Attention

Peggy (Yuchun) Wang
wangyuc@stanford.edu

Khalid Ahmad
kahmad@stanford.edu

Introduction

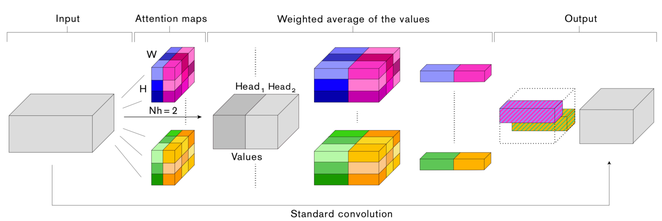
- **Motivation**: Semantic segmentation is important for autonomous driving scenarios, since it provides accurate road information for use in navigation and planning
- **Traditional approach:** Deep CNN architectures, such as DeepLab
- **Our new approach:** Multi-headed self-attention (popularized in NLP) augmented Deep CNNs to DeepLab and FCNs

Dataset

- **Mapillary Vistas** = 25,000 high-resolution street-level images
- 18,000/2000/5000 = training/validation/test split
- **65** object categories with wide range of geography, camera viewpoints, time of day, and weather
- **Input:** RGB images resampled to 512 × 1024 resolution
- **Output:** segmentation map of image (512 × 1024)

Approach

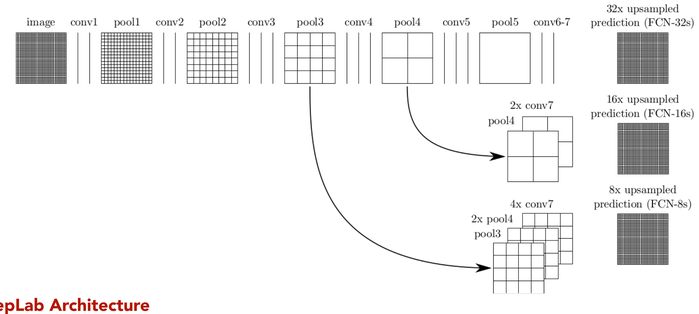
- **Attention - Augmented Convolution Block**
- A convolution layer and a multi-head attention layer
- Carried out over a given input activation map and combined to produce an output state of shape equal to that of the convolutional layer's output
- ReLU activation, Cross-entropy Loss, SGD w/ Momentum



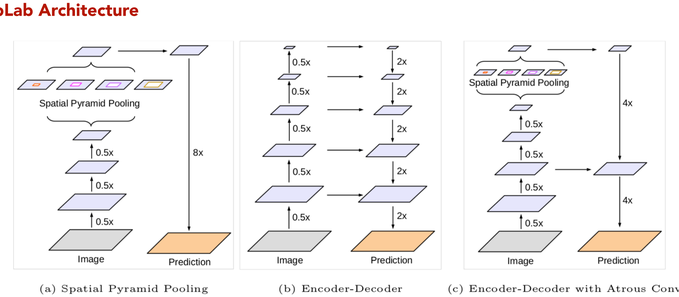
Experiments

- Four models: **FCN, FCN with attention, DeepLab3+, and DeepLab3+ with attention**
- Trained for **40,000** iterations each

FCN Architecture



DeepLab Architecture



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

Hyperparameters

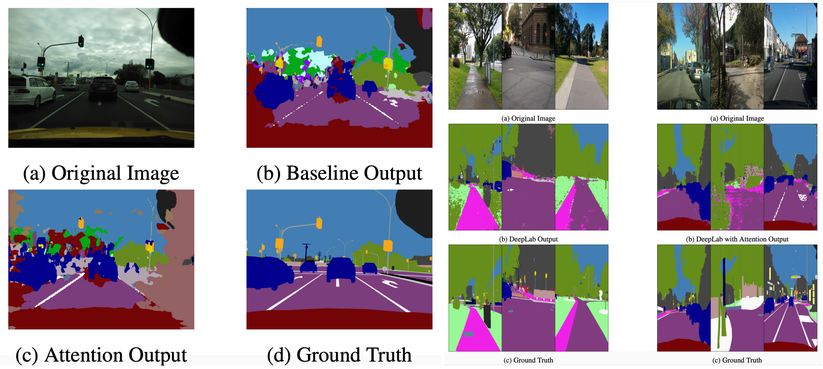
Name	FCN Values	DeepLab Values
Iterations	40,000	40,050
Batch Size	4	4
Validation Interval	1000	4500
Optimizer	SGD w/ Momentum	SGD w/ Momentum
Learning Rate	1E-04	0.01
Weight Decay	0.0005	0.0005
Momentum	0.99	0.9
Loss Type	Cross-entropy	Cross-entropy
Backbone	-	Xception

Results

Name	FCN	AA FCN	DeepLab	AA DeepLab
Overall Acc	0.829990	0.868064	0.842424	0.832322
Mean Acc	0.286447	0.304605	0.202086	0.169805
FreqW Acc	0.728482	0.780485	0.740195	0.724240
Mean IoU	0.217500	0.253031	0.160526	0.135338
Final Loss	0.4496	0.6804	0.697	0.626

Output

FCN Output **DeepLab Output**



(a) Original Image (b) Baseline Output (a) Original Image (a) Original Image

(b) DeepLab Output (b) DeepLab with Attention Output

(c) Attention Output (d) Ground Truth (c) Ground Truth (c) Ground Truth

Conclusion + Future Work

- **Promising results** but would need to validate with **more training**
- Mean IoU metric is 53.37% on Vistas benchmark
- Top Model (AA FCN) has mean IoU of **25.50%**
- AA FCN > FCN, DeepLab > AA DeepLab
- FCN performed better than DeepLab, may be resolved with more training + hyperparameter tuning
- Future work includes:
  - **Training for 200,000 iterations**
  - Hyperparameter tuning
  - Augmenting images
  - Add Cityscapes dataset