Semantic Segmentation

By-
Divyang Teotia, Daniel Uvaydov, Satish Kumar Anbalagan, Varun Sahasrabudhe

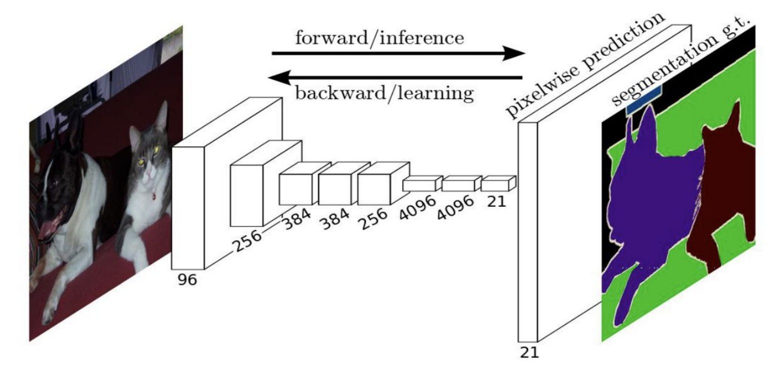
Agenda

- Overview & Challenges
- Implementation Details
- Training
- Results
- Conclusion & Possible Improvements
- Team members contribution
- References

Overview & Challenges

Semantic Segmentation: Overview

- What ? Classify and clustering each and every pixel in the image which belong to same object class
- How ? Includes segmenting by a feature extraction network trained for image classification like VGGNet, ResNets, DenseNets, MobileNets, NASNets etc
- Why ? autonomous driving, medical, HCI, photo editing tools, robotics vision and understanding
- Cityscapes fine labels data sets used



Challenges

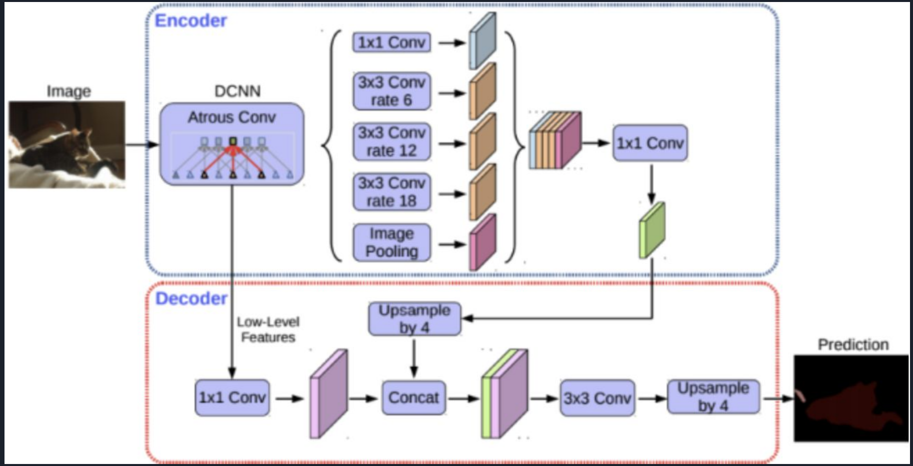
- Tradeoffs between accuracy vs speed per memory, while maintaining the efficiency of the network during classification
- Network model encounters objects of many different sizes that require features processing at different scales
- Improving localization of object boundaries
- Segmenting and existence of objects at multiple scales
- Reduced feature resolution caused by a repeated combination of max-pooling and downsampling
- Better refining of feature map using Channel attention

Implementation Details

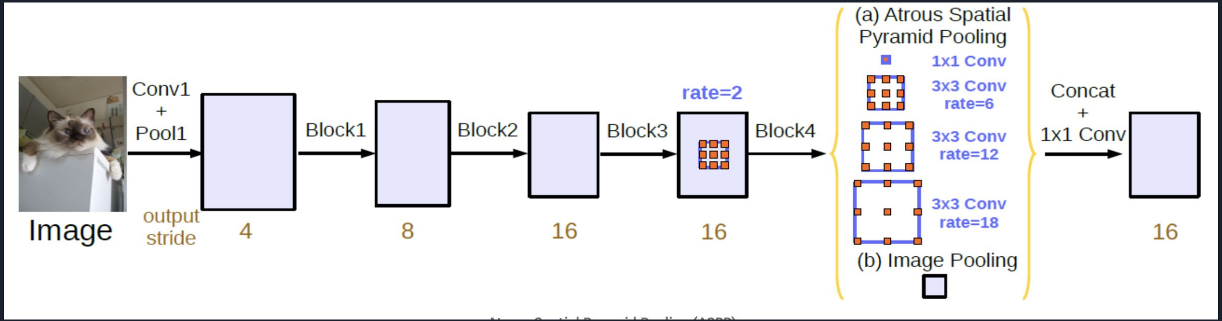
DeepLabs v3

Encoder-Decoder Architecture:

- Deeplabs prevents signal decimation and learns multi scale contextual features
- Uses an ImageNet pretrained Resnet as its main feature extractor with atrous conv in the last block
- Uses Atrous Spatial Pyramid Pooling (ASPP) on top of Resnet to classify regions of an arbitrary scale and decoder upsamples the output in stages



ASPP



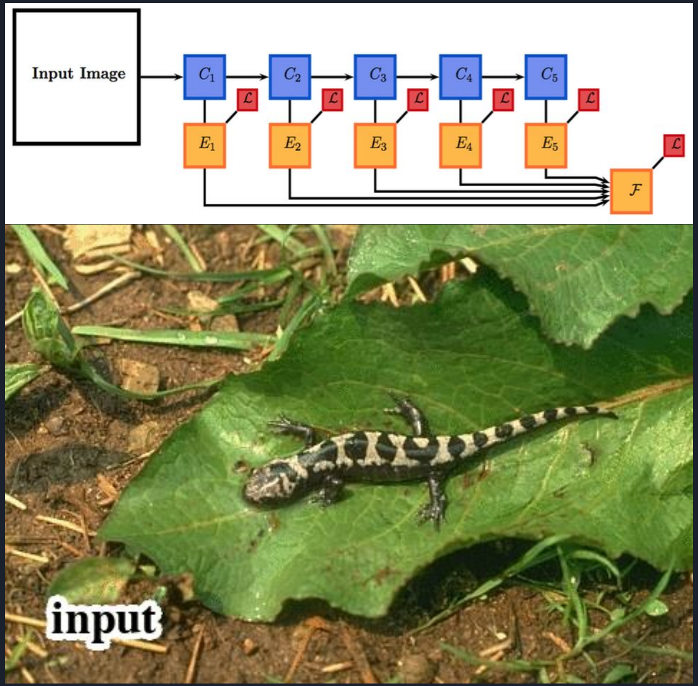
- Provides the model with multi scale information using a series of atrous convolutions with different dilation rates to capture long range context.
- To add global context information, ASPP incorporates image level features via Global Average Pooling
- Finally, all the multiple scales are concatenated along with global features and followed by a 1x1 convolution to feed to the decoder.

Holistic Edge Detection Preprocessing

- Uses Deep Supervised Network training to fine tune VGG for the task of boundary detection

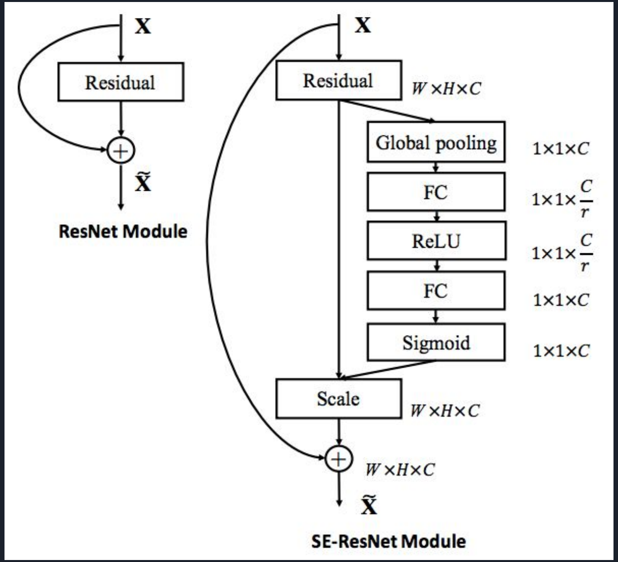
- We will supplement our input with an extra channel using the output of a pre-trained HED

- Add a long skip connection from input to decoder and reiterate boundary information right before decoder output.



Squeeze and Excitation Blocks

- Squeeze and excitation blocks attempt to map the channel interdependencies
- Squeezes all feature maps to single values (per channel), extracts channel features through FC layers, and weights each channel value
- Original feature map is then scaled with weighted channel values that are continuous



Training

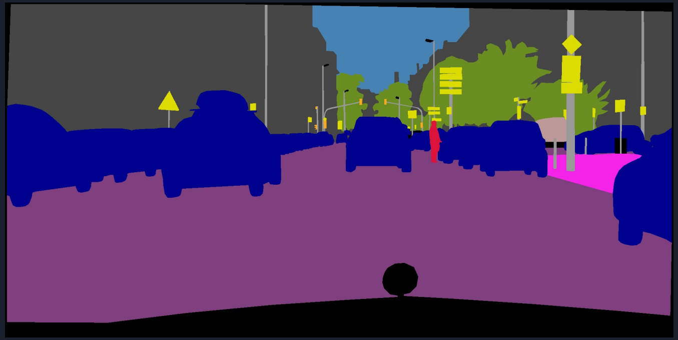
Training

- Models
  - DeepLabsv3 (Vanilla)
  - DeepLabs v3 with Squeeze and Excitation (DLSE)
  - DeepLabs v3 with Squeeze and Excitation using softmax (DLSE-SF)
  - DeepLabs v3 with Squeeze and Excitation and HED Preprocessing (DLSE-HED)
  - Note - Trained with progressive input sizes :
    - Size 1 - 128 x 224
    - Size 2 - 256 x 448
    - Size 3 - 512 x 912
- Params
  - Loss : Categorical Cross Entropy
  - Optimizer: Adam (lr = 1e-5)
  - Batch_size:
    - Size 1 - 16
    - Size 2 - 8
    - Size 3 - 4

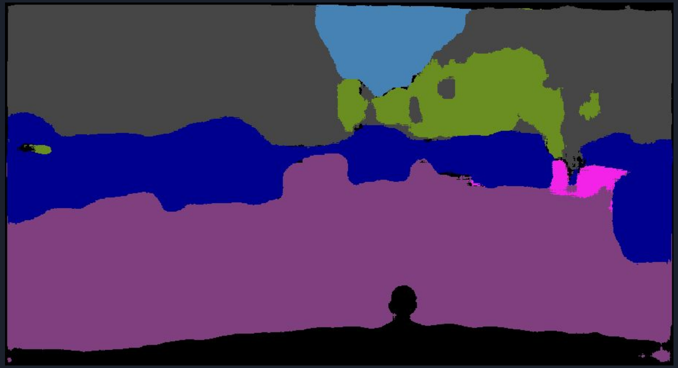
DeepLabs v3 (Vanilla)

- Traditional DeepLabsV3 with ResNet-50 as backbone

- Loaded with pre-trained weights from ImageNet training



Target

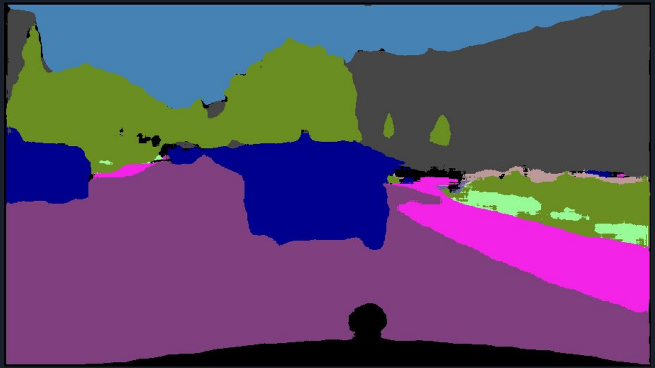


Predicted

DeepLabs v3 with Squeeze and Excitation (DLSE)

- Squeeze and excitation blocks added to ResNet-50 backbone
- Added after each branch of the residual block except atrous block.
- Channel weights squeezed by a factor of 16 before being excited back to normal

Target

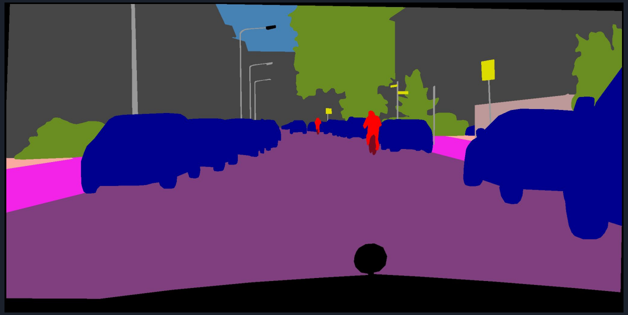


Predicted

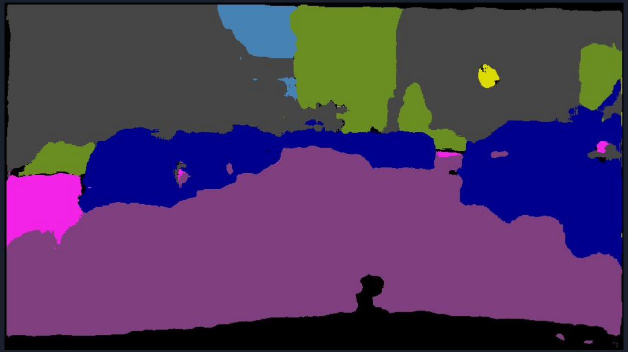
DeepLabs v3 with Squeeze and Excitation using softmax (DLSE-SF)

- Use softmax at last layer instead of sigmoid
  - Might improve convergence

- Take channel with highest value from softmax and scale it by constant
  - Promotes stronger hierarchy



Target



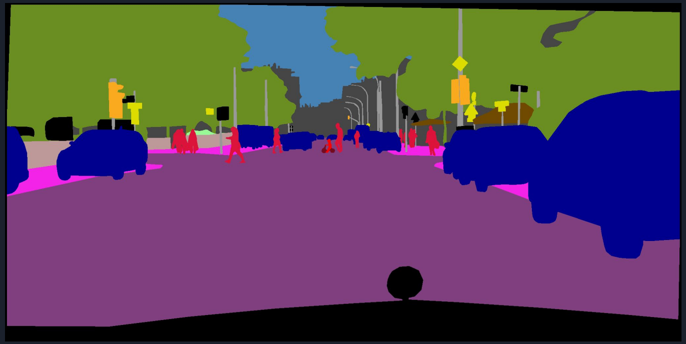
Predicted

DeepLabs v3 with Squeeze and Excitation and HED Preprocessing (DLSE-HED)

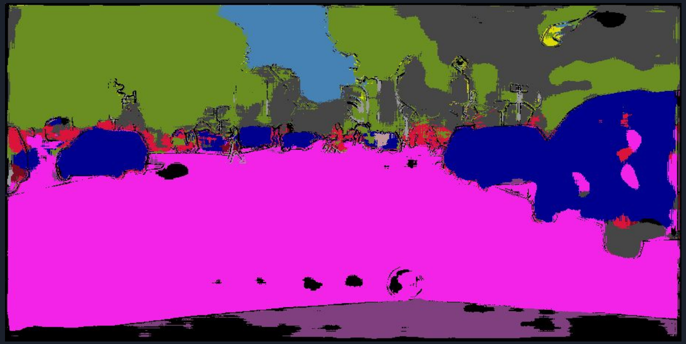
- Dual input model with ResNet-50 as backbone

- Task of improving boundary detection with pre-trained HED

- Takes a long skip connection from input plus HED output to decoder
  - reiterated boundary information
  - supplemented the input with an extra channel



Target



Predicted

Results

Model	***mIOU (Mean Intersection over Union)***
Vanilla DL	60.24
DLSE	57.5
DLSE-SF	52.79
DLSE-HED	38.83

Conclusion and Possible Improvements

Conclusion

- The Squeeze and Excitation block was able to assist in establishing a channel hierarchy, gave comparable results to the original deeplabs model .

- Switching sigmoid to softmax in addition to scaling highest output of softmax proved to be slightly detrimental with little impact on convergence time

- The incorporation of HED output to our model deteriorated the model's performance significantly and points to the need of fusing this information via a custom loss function instead.

Possible Improvements

- Train on coarse labels in addition to fine labels
    - As is done on benchmark approaches
- Implement a custom loss function
- Using a dual headed network for improving localization and boundary detection through a unified architecture

Team Member Contributions

- Divyang Teotia: Developing and training the vanilla and DLSE models
- Daniel Uvaydov: Developing and training DLSE-SF and pre-processing with HED
- Satish Kumar Anbalagan: Developing and training DLSE-HED model and pre-processing with HED
- Varun Sahasrabudhe: Pre-processing HED data, compiling the resources and making the presentation

References

- https://heartbeat.fritz.ai/a-2019-guide-to-semantic-segmentation-ca8242f5a7fc
- https://towardsdatascience.com/semantic-segmentation-with-deep-learning-a-guide-and-code-e52fc8958823
- https://towardsdatascience.com/squeeze-and-excitation-networks-9ef5e71eacd7
- Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
- Chen, Long, et al. "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
- https://www.analyticsvidhya.com/blog/2019/02/tutorial-semantic-segmentation-google-deeplab/
- Xie, Saining, and Zhuowen Tu. "Holistically-nested edge detection." Proceedings of the IEEE international conference on computer vision. 2015.