## **Research paper-5: SegNet Deep Convolutional Encoder-Decoder Architecture for Image Segmentation**

Source: <https://arxiv.org/pdf/1511.00561.pdf>

## **Summary**

Introduction:

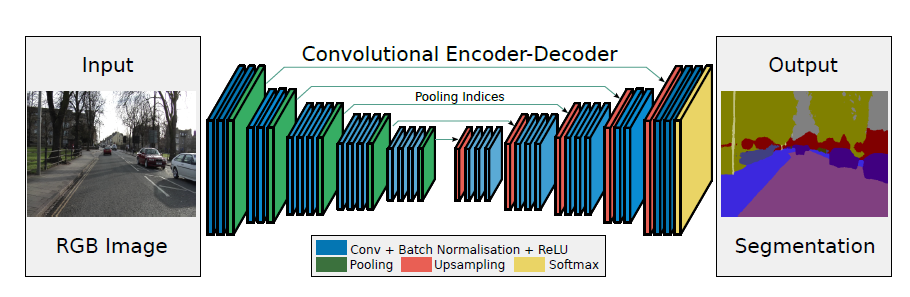
* SegNet is primarily built for scene understanding by Semantic segmentation where it has a wide array of applications.
* Deep learning has seen huge success lately in handwritten digit recognition, speech, categorizing whole images, and detecting objects in images, Now there is an active interest in semantic pixel-wise labeling.
* Early methods that were used for Semantic segmentation had relied on low-level vision cues by using popular machine learning algorithms.
* SegNet focuses on the need to map low-resolution features to input resolution for pixel-wise classification where mapping produces features that are useful for accurate boundary localization.
* This idea of SegNet was inspired by an architecture designed for unsupervised feature learning where it involves Reusing max-pooling indices in the decoding process.

Related Work:

* Semantic pixel-wise segmentation is an active topic of research, fuelled by challenging many datasets. Before the arrival of deep networks, the best-performing methods mostly relied on hand-engineered features fed into a classifier like Random Forest or Boosting to predict the class probabilities.
* [Deep Convolutional Networks for Scene Parsing](http://david.grangier.info/scene_parsing/abstract/grangier-deep-cnn.pdf) by David Grangier, Leon Bottou, Ronan Collobert for Deep convolutional network for modeling the complex scene label structures.
* [Semantic image segmentation with deep convolutional nets and fully connected crfs (DeepLab)](https://arxiv.org/pdf/1412.7062.pdf) from researches at google by C. Liang-Chieh, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille where Conditional Random Field (CRF) is used for segmentation.
* [Learning deconvolution network for semantic segmentation (DeconvNet)](https://arxiv.org/pdf/1505.04366.pdf) by H. Noh, S. Hong, and B. Han which uses deconvolution and unpooling layers.
* [Very deep convolutional networks for large-scale image recognition](https://arxiv.org/pdf/1409.1556.pdf) by K. Simonyan and A. Zisserman which deals with the network for large-scale image recognition.

**SegNet Architecture**

SegNet has an encoder-decoder architecture that has a decoder that upsamples its input using the transferred pool indices from its encoder to produce a sparse feature map and then performs convolution with a trainable filter bank to densify the feature map. Segnet does not contain fully connected layers and hence it only has convolutional layers. The final decoder output feature maps are fed to a soft-max classifier for pixel-wise classification.



**Implementation Detail:**

Segnet captures and stores boundary information in the encoder feature maps before sub-sampling is performed. The more efficient way to store this information is to store only the max-pooling indices, i.e, the locations of the maximum feature value in each pooling window are memorized for each encoder feature map.

* SegNet has an encoder network and a corresponding decoder network, followed by a final pixel-wise classification layer.
* The encoder network consists of 13 convolutional layers that correspond to the first 13 convolutional layers in the VGG16 network designed for object classification and hence initialize pre-trained weights.
* Each encoder in the encoder network performs convolution with a filter bank to produce a set of feature maps that are batch normalized with ReLU activation applied.
* Followed by Max-pooling with a 2x2 window and stride 2 to get sub-sampled output by a factor of 2 where meanwhile max pool indices are memorized.
* The Decoder network upsamples its input feature map using the memorized max-pooling indices from the corresponding encoder feature map which results in sparse feature map.
* Followed by feature maps are then convolved with a trainable decoder filter bank to produce dense feature maps with batch normalization.
* Segnet consists of less number of parameters when compared to other architectures.
* The final decoder output is fed to a multi-class softmax classifier to produce class probabilities for each pixel independently.

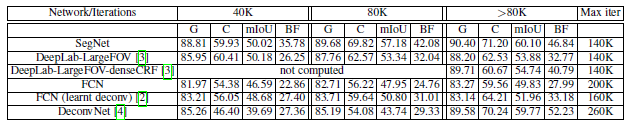
Training:

* CamVid road scenes dataset is used to benchmark the performance, the dataset is small, consisting of 367 training and 233 testing RGB images
* The Challenge is to segment 11 classes such as road, building, cars, pedestrians, signs, poles, side-walk, etc.
* The input is normalized and all weights are initialized using the HE-Normal initialization.
* Used Stochastic gradient descent (SGD) with a fixed learning rate of 0.1 and a momentum of 0.9 using the Caffe implementation of SegNet.
* The training set is shuffled and each mini-batch is then picked in order thus ensuring that each image is used only once in an epoch.
* Used the cross-entropy loss as the objective function for training the network. The loss is summed up over all the pixels in a mini-batch.
* Also experimented with training the different variants without class balancing or equivalently using natural frequency balancing.

Benchmarks:

**CamVid road scene Dataset:**

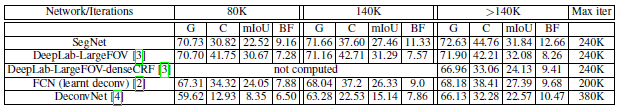
SegNet benchmarked against several other wells adopted deep architectures for segmentation such as FCN, DeepLab, and DeconvNet and have tried to understand the performance of these architectures by training end-to-end on the CamVid road scene segmentation datasets.



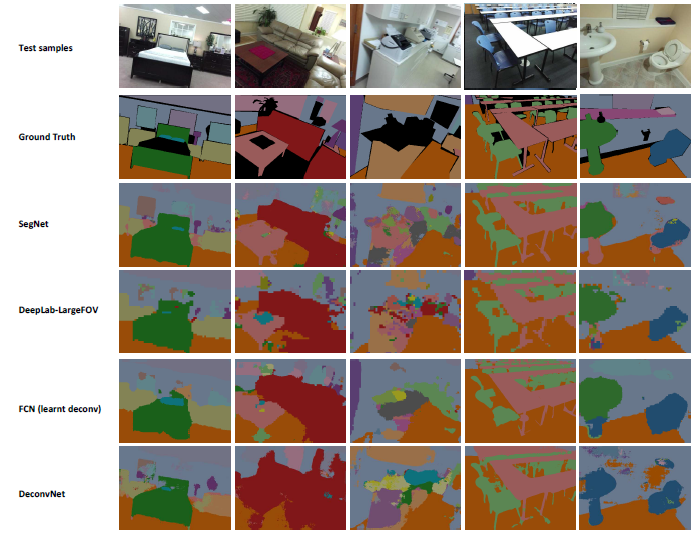
* SegNet predictions are more accurate in 8 out of the 11 classes. It also shows a good 10% improvement in class average accuracy when trained on a large dataset of 3.5K images when compared to CRF based methods.
* SegNet outperforms all the other methods with more information in the research paper.

**SUN RGB-D Dataset:**

SUN RGB-D is an Indoor Scenes Dataset that is a very challenging and large dataset of indoor scenes with 5285 training and 5050 testing images. The images are captured by different sensors and hence come in various resolutions.



* The task is to segment 37 indoor scene classes including the wall, floor, ceiling, table, chair, sofa, etc. This task is made hard by the fact that object classes come in various shapes, sizes, and in different poses.
* More information on research papers about the benchmark scores and all comparisons.



* segmentation quality is better when object classes are reasonably sized but is very noisy when the scene is more cluttered.
* Some parts of an image of a scene do not have ground truth labels and shown in black color. These parts are not masked in the corresponding deep model predictions.
* Results are from the model corresponding to the highest mIoU accuracy.

**Conclusion:**

* SegNet is a deep convolutional network architecture for semantic segmentation and the main motivation behind SegNet is a need for an efficient architecture for road and indoor scene understanding which is efficient both in terms of memory and computational time
* SegNet stores the max-pooling indices of the feature maps and uses them in its decoder network to achieve good performance.
* Segnet offers flexibility to reuse pre-trained weights of the VGG16 backbone network.
* SegNet is compared with other important variants that revealed the practical trade-offs involved in designing architectures for segmentation, particularly training time, memory versus accuracy.