Semantic Segmentation

Prakhar Agarwal, Cluster Innovation Centre
Dr. S.P. Mishra, DTRL, DRDO

Abstract—The abstract goes here.

Index Terms—Semantic Segmentation, Deep Learning, Pixel-wise Segmentation

1 Introduction

THIS

1.1 Semantic Segmentation

Semantic means ”meaning in language or logic”, and when we talk about semnatically segmenting an image we mean to divide the image in to meaningful parts, such as physical objects.

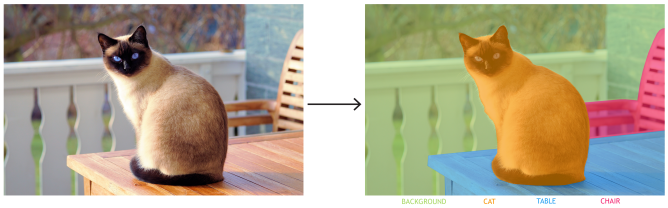

Fig. 1: Semantic Segmentation

For example, the Figure 1 has been segmented into 4 different categories (Cat, Table, Chair, and Background). As it might be evident, this task is not achievable using unsupervised methods. We need to define the desired segmentation categories, as an image with infinte detail could be segmented into an infinite number of categories.

2

2.1 Fully Convolutional Networks

Fully Convolutional Networks (FCNs) are what one gets on building a neural network using only convolutional-type layers. They keep the spatial information of an image intact.

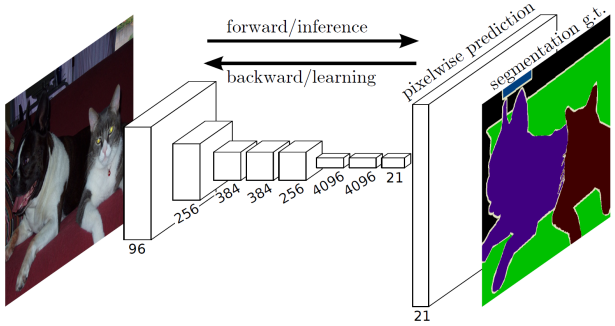

Fig. 2: Representation of a FCN (Source: J. Long et al, 2015 [1])

2.2 Deconvolution Networks

Pooling layers decrease the dimensions of an image. To scale it back up to its original size, instead of using Bilinear techniques, we have used Deconvolution layers. Deconvolution layers act as reverse convolution layers, and serve the purose of scaling up an image. Like there counterparts, they too are learnable.

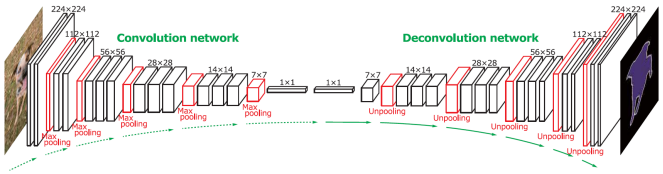

Fig. 3: Representation of a Deconvolution Network (Source: H Noh et al, 2015 [2])

3 Data Used

3.1 Cityscapes

The Cityscapes dataset [3] offers pixel-wise segmented images from stereo videos recorded in street scenes from 50 different cities. It contains 5,000 fine quality and a larger set of 20,000 weakly annotated images. The images have a resolution of 2048 x 1024 pixels. For our purpose, we use the set of finely annotated images from the left cam-era.

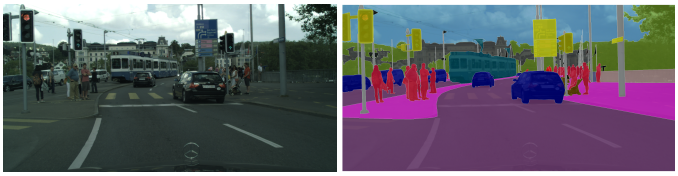

Fig. 4: Cityscapes Dataset

3.2 Stanford Scene Understanding Dataset

The Stanford Scene Understanding dataset [4], [5], [6] contains a total of 534 images (400 training and 134 test), with labels comprising of 9 classes (unknown, sky, tree/bush, road/path, grass, water, building, mountain, foreground object). Resolution of the images is 240 x 320 pixels. This served as a testing dataset to check for convergence of our models.

4 Segmentation Architecture

We use a hybrid architecture of a Fully Convolutional Network, a Deconvolution Network, and skip connections.

4.1 Skip Connections

With increasing network depths, vanishing gradients cause the accuracy to decrease rapidly. Deeper networks also have slower training, as gradient changes become minute.

Skip connections solve these problems by joining the output of an earlier layer to a later one, skipping all those in between. These shortcut connections help in keeping the effect of gradients intact during backpropagation.

When using pooling layers, information gets lost at each step. This limits the network's ability to resize the image back to its original size. Skip connections help pass the high resolution information to later layers.

4.2 Architecture

5 RESULTS

APPENDIX A
MODEL ARCHITECTURE - CITYSCAPES DATASET

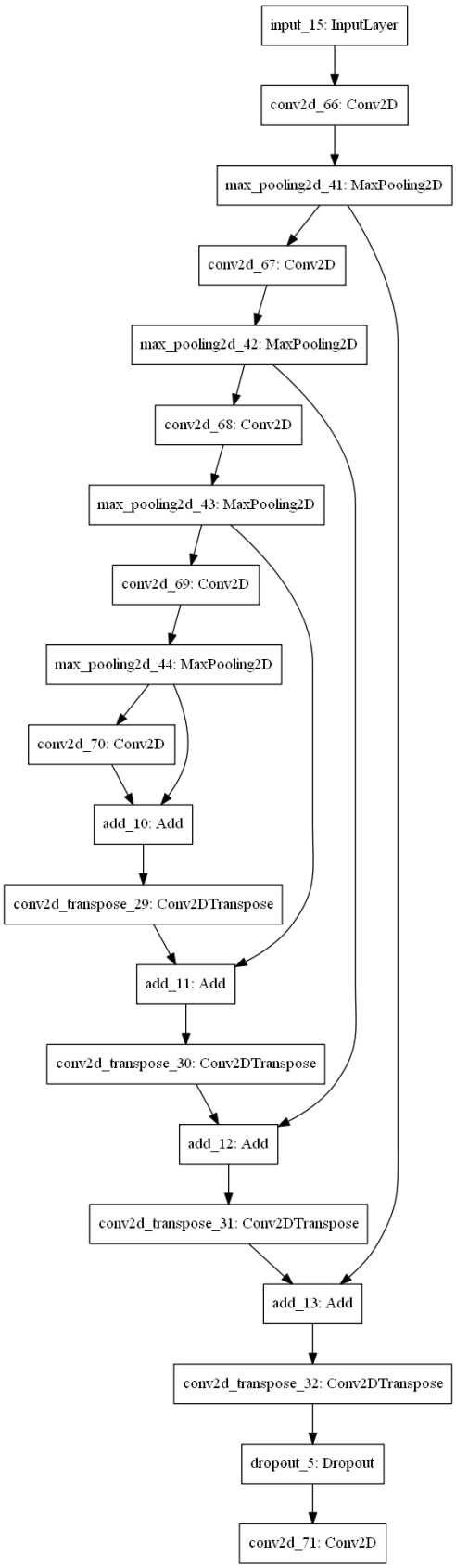

Fig. 5: Model Architecture - Cityscapes Dataset

ACKNOWLEDGMENTS

The authors would like to thank...

REFERENCES

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
[2] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1520–1528.
[3] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
[4] B. Liu, S. Gould, and D. Koller, "Single image depth estimation from predicted semantic labels," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 1253–1260.
[5] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning depth from single monocular images," in *Advances in neural information processing systems*, 2006, pp. 1161–1168.
[6] A. Saxena, M. Sun, and A. Y. Ng, "Make3d: Learning 3d scene structure from a single still image," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 5, pp. 824–840, 2009.

PLACE PHOTO HERE

Michael Shell Biography text here.

John Doe Biography text here.

Jane Doe Biography text here.