CSE 253: Neural Networks for Pattern Recognition Dense Semantic Image Segmentation

Akshansh Chahal
Department of Computer Science
University of California, San Diego
a3chahal@eng.ucsd.edu

Mayank Rajoria
Department of Computer Science
University of California, San Diego
mrajoria@eng.ucsd.edu

Abhishek Sen
Department of Computer Science
University of California, San Diego
asen@eng.ucsd.edu

Piyush Tayal
Department of Computer Science
University of California, San Diego
ptayal@eng.ucsd.edu

1 Main Idea

Deep Convolutional Neural Networks have broad applications in many different places, one of which is Computer Vision. In this project we aim to use Deep CNNs for the task of Semantic Image Segmentation. Image segmentation is a task of computer vision where we label specific regions of an image with a corresponding 'class' of what is being represented. And since we will be predicting for every pixel in the image, it is called 'dense'. Segmentation models are important for various tasks like autonomous driving vehicles, medical image diagnostics etc.

We plan to use the implementations of state of the art research papers in this field and build upon those to try some variations in terms of the architecture, the training & test datasets etc. We plan to use custom real life images and videos taken by us to train and test the models we create. We will also try to combine different tricks and modifications used in state of the art models together in one same model and see what effect will it have on the results.

Some of the leading state of the art models for the task of dense semantic segmentation which we will consider implementing are SegNet [1], LinkNet[2], ICNet [3].

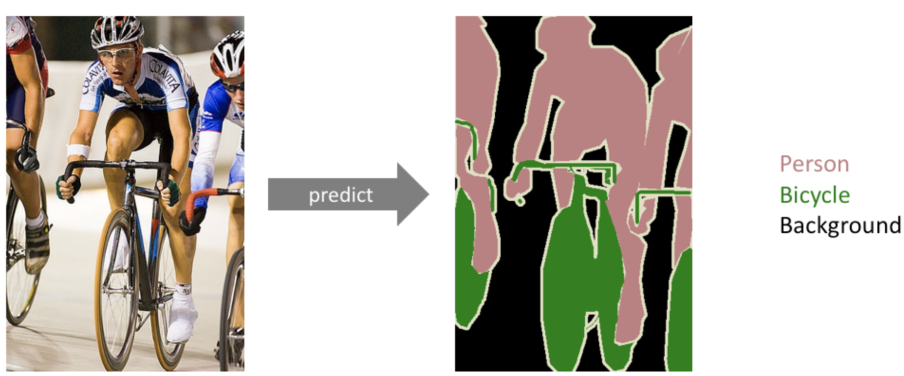

Figure 1: An example of semantic segmentation

2 Datasets

We will use the Cityscapes Dataset [4]. This dataset contains a diverse set of stereo video sequences recorded in street scenes from 50 different cities with 5,000 images with high quality annotations, 20,000 images with coarse annotations. The focus of this dataset is on the semantic understanding of urban street scenes, with semantic and instance-wise pixel-wise dense annotations. It also has a pretty rich metadata: preceding and trailing video frames, stereo, GPS, vehicle odometry. In terms of complexity there are total of 30 classes in the annotations of the images in this dataset.

We also have the Cambridge-driving Labeled Video Database (CamVid) Dataset [5], which is the first collection of videos with object class semantic labels. This provides with ground truth labels that associate each pixel with one of 32 semantic classes.

3 Architecture of the network and reference papers

An architecture for semantic image segmentation in general consists broadly of an encoder network followed by a decoder network, where encoder usually being a pre-trained classification network like ResNet or VGG. The decoder mechanism is where these architectures differ. This decoder is responsible for the task of semantically projecting the low resolution discriminative features learnt by the encoder onto a high resolution pixel space.

For example if we consider SegNet, the encoder consists of 13 convolutional layers which correspond to first 13 layers of VGG16 network. Each of the encoder layers has a corresponding decoder layer. Different decoder architecture have been tried. One of the variants upsamples the feature map without learning and then convolves with a trainable decoder filter bank. Another variant learns to deconvolve the input feature map and adds the corresponding encoder feature map to produce the decoder map. We wish to try all these variants along with some extensions.

After implementing the above architecture, we plan to experiment with different kinds of architectural changes to it. We have explored several advances of encoder-decoder fully convolutional architectures and would try adding them to the network. One of the techniques we plan to explore are skip connections. This modification was inspired by LinkNet[2] architecture. This can be done by connecting a layer from encoding network with the corresponding layer in decoding network. We hope this might provide the decoding network with finer details from the encoding part of the network. We could also try having dilated convolution layers to increase the receptive field in the encoding layer. This is inspired by Dilated Residual Networks [8]. This has been experimentally found to generate smoother feature maps instead of maxpooling and dropout

References

[1] Badrinarayanan, V., Kendall, A., & Cipolla, R. (2016). SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39, 2481-2495.

[2] Chaurasia, A., & Culurciello, E. (2017). LinkNet: Exploiting encoder representations for efficient semantic segmentation. 2017 IEEE Visual Communications and Image Processing (VCIP), 1-4.

[3] Zhao, H., Qi, X., Shen, X., Shi, J., & Jia, J. (2018). ICNet for Real-Time Semantic Segmentation on High-Resolution Images. ECCV.

[4] 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

[5] http://mi.eng.cam.ac.uk/research/projects/VideoRec/CamVid/

[6] https://paperswithcode.com/sota

[7] https://www.jeremyjordan.me/semantic-segmentation/

[8] Yu, F., Koltun, V., & Funkhouser, T.A. (2017). Dilated Residual Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 636-644.