Semantic and Instance segmentation

Raunak Vijan, Abhilash Kuhikar, Saurabh Mathur, Shivam Rastogi

Computer Vision Project, Indiana University, SICE

Introduction

Motivation
Robots are being used in a variety of environments from outer-space to deep seas. Vision-based navigation controls a robot's movement by analyzing image frames from the robot's camera. Thus, precise image understanding is necessary for vision-based autonomous robot navigation. Our work deals with segmenting image frames by labeling each pixel in the image as one of many classes.

Challenges
Input images are often noisy, hazy and poorly illuminated. Objects can be distorted and occluded which make it difficult to segment each instance of the object.

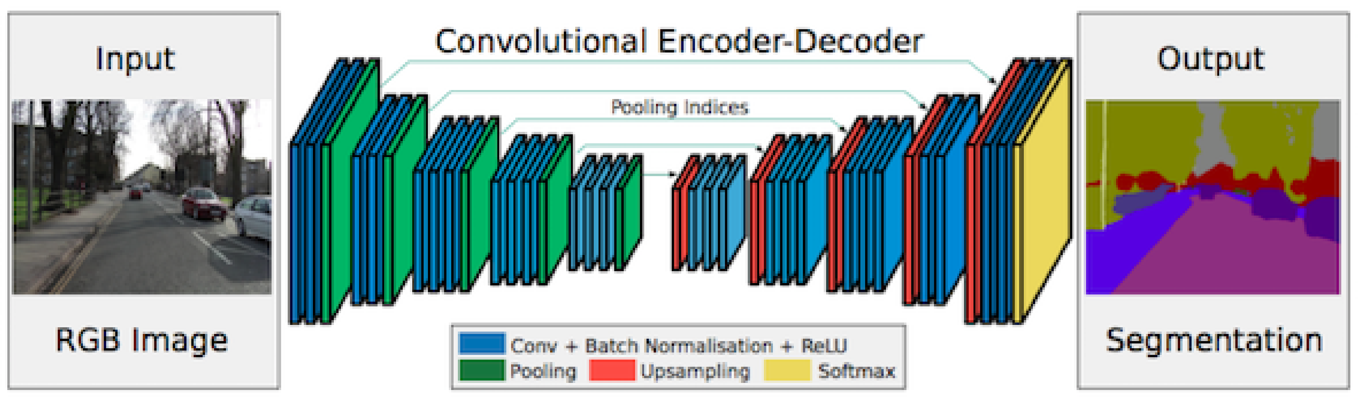
Contributions:
- We verified the accuracy of semantic segmentation on CamVid11 dataset by using different loss functions and found that the softdice loss function performed the best in terms of accuracy.
- We also re-implemented Bayesian Segnet(which gives uncertainty) and compared the accuracy to the segNet.
- We improved the accuracies by applying traditional post processing techniques such as CRF smoothing.

Data
- CamVid11 (1: building, 2: pole, 3: road, 4: sidewalk, 5: Tree, 6: SignSymbol, 7: Fence, 8: Car, 9: Pedestrian, 10:Bicyclist, 11:Void)
- Pascal VOC
- Underwater images: As part of the project, we are trying to collect and create a novel dataset for underwater object detection.

Methods and network architectures

- SegNet is an encoder-decoder architecture implemented using fully convolutional layers[1]

SegNet Architecture



Reprinted from [3]

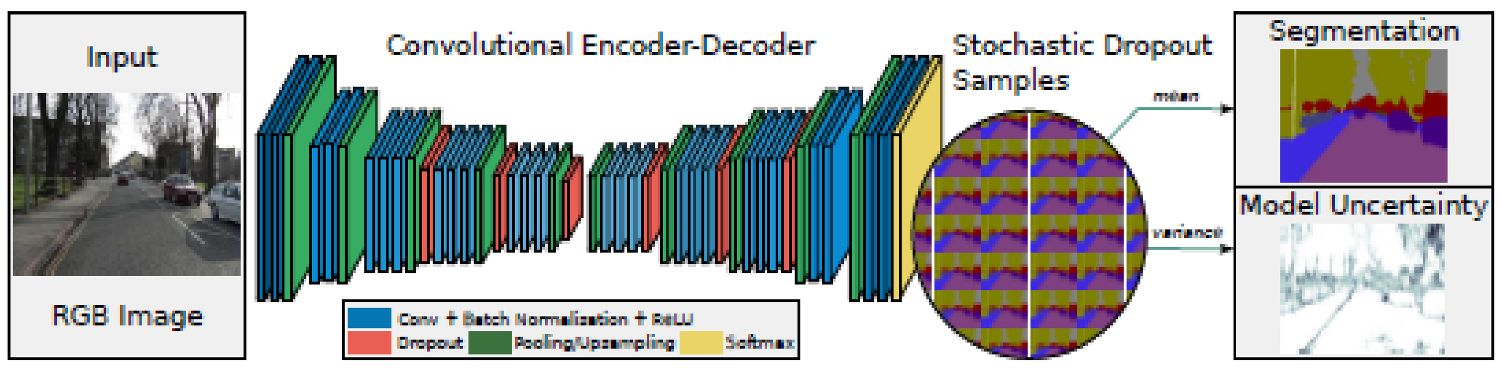
- The key difference from FCN is that this architecture uses max-unpooling instead of transposed-convolutions to decode the output.
- This architecture has an extension called Bayesian SegNet that quantifies its uncertainty in its output.
- DenseCRF is used to smoothen the segmentation masks as part of the post processing technique[2]

Loss functions: Weighted CrossEntropy and SoftDice

$$D = \frac{2\sum_i^N p_i g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2}$$

$$WCE = -\sum_{i=1}^{N} w_i y_i \log(p_i)$$

Bayesian Segnet

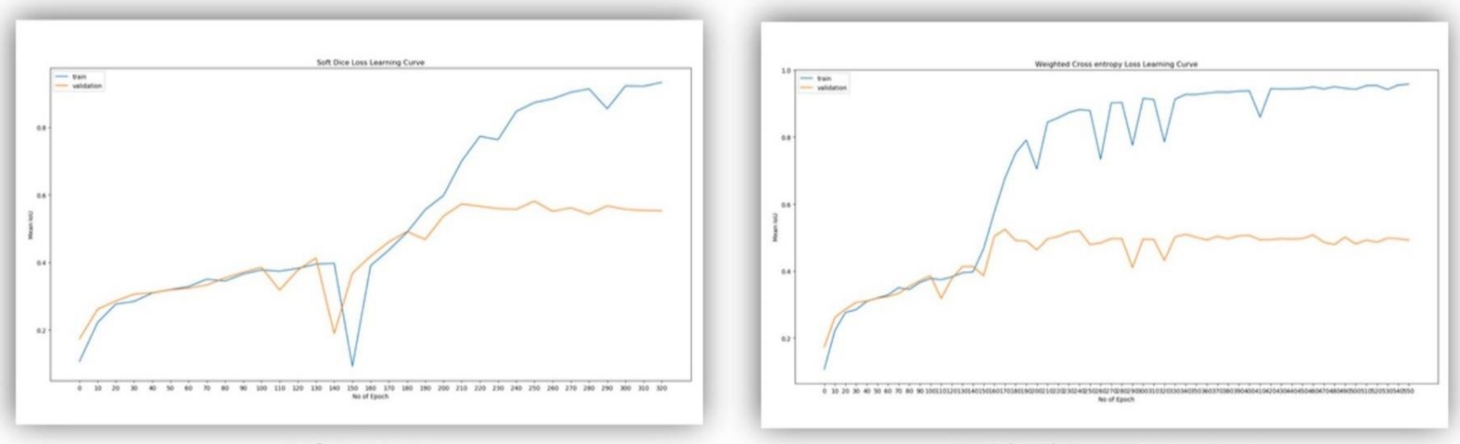


Reprinted from [4]

- Bayesian Segnet applies dropout at test time to create an ensemble. We sample 40 images and take mean and variance of all the images.

Data and Results

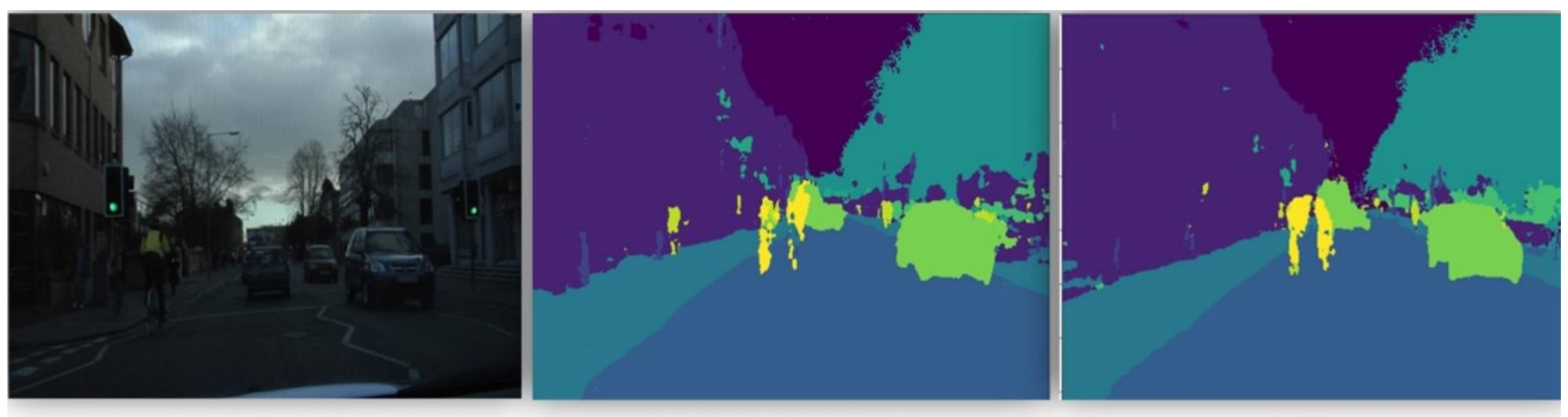
Learning curves for Segnet model for different loss functions



Soft Dice Loss

Weighted Cross Entropy Loss

Segmentation results on CamVid11

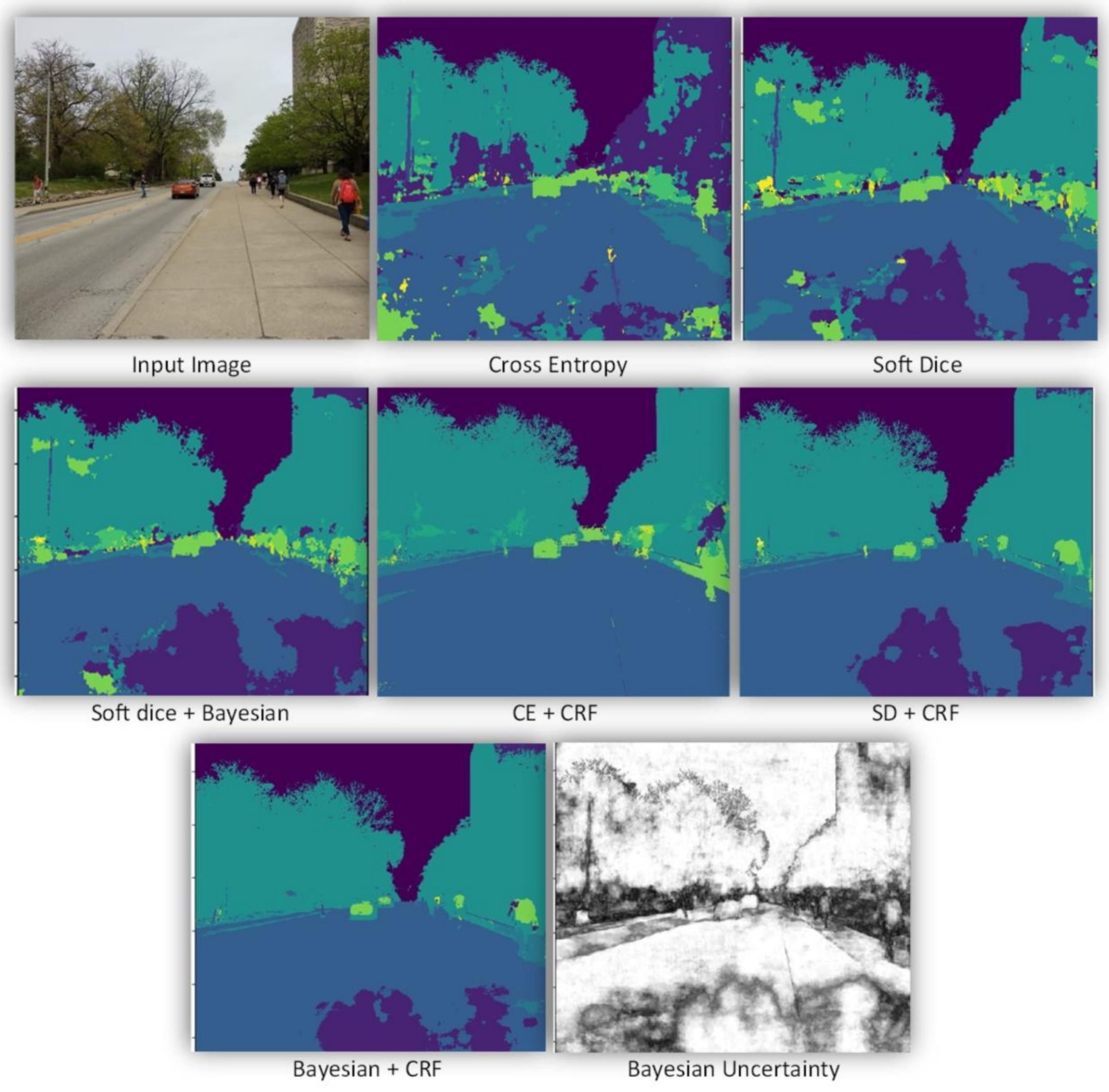


Input CamVid

Weighted Cross Entropy

Soft Dice

Segmentation results on IU campus-street images



Model Evaluation

	Mean IoU	Overall Acc	Mean Acc	FreqW Acc
CrossEntropy	0.5244	0.8312	0.7076	0.7512
SoftDice	0.5818	0.9135	0.6767	0.8508
Bayesian	0.6023	0.9234	0.6809	0.8472

Conclusions and Future Work

- Complete collecting a new dataset of underwater images as our original motivation and apply instance and semantic segmentation on them.
- Temporal smoothening on existing video data.
- Train end to end model for video semantic segmentation with spatio-temporal smoothing.
- Compress the segnet for real time inference using student-teacher network.

Bibliography

1. Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence, 39(12), 2481-2495.
2. Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding - Alex Kendall, Vijay Badrinarayanan, Roberto Cipolla
3. http://mi.eng.cam.ac.uk/projects/segnet/
4. Kaiming He, Georgia Gkioxari, et al. Mask R-CNN