Contextual Reasoning in Scene Understanding

Raj Shah, Dr Alex Kendall

June 18, 2018

UNIVERSITY OF CAMBRIDGE
DEPARTMENT OF ENGINEERING

Introduction

Scene understanding is often tackled by Semantic Segmentation, a fundamental Computer Vision task. In such tasks, scene geometry and scene semantics play an important role in developing context-aware models.

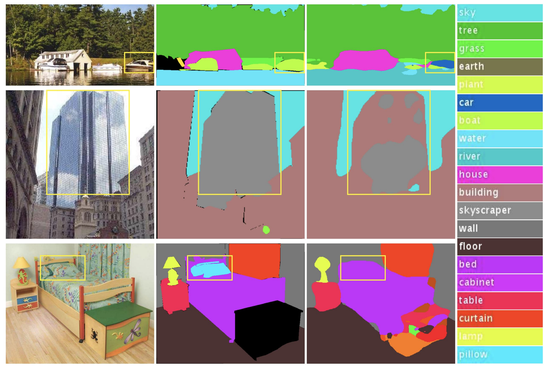

Figure 1: FCN [1] - not context aware!

Semantic Segmentation

Semantic segmentation involves classifying different parts of a visual input into semantically interpretable classes. This is done by assigning each pixel in the input space an object/semantic class, or simply, labeling each pixel in an input image.

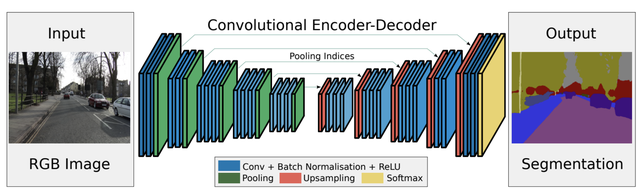

Figure 2: SegNet [2]

- Encoder-decoder networks for dense pixel-wise classification.
  - Encoder performs classification in the input space, gradually reducing the input's spatial dimension.
  - Decoder maps features learned by the encoder to input space while recovering spatial dimension and object details.
  - Skip-connections between the encoder and decoder stages to recover finer object details.
- Dilated or atrous convolutions to allow for exponential increase in receptive field/FOV without decrease in spatial dimensions.
- Pyramid pooling and ASPP modules to aggregate multi-scale context.

Saliency

A model's receptive field allows us to assess its ability to incorporate context in its predictions. One way to measure its impact on the output predictions is saliency. Saliency represents the importance of regions in an image for a vision system to understand the image.

This idea has been explored in classification tasks, where the output is a single prediction.

- *Occlusion sensitivity* [3] - occlude part of an image and monitor posterior, activations, and classifier output.

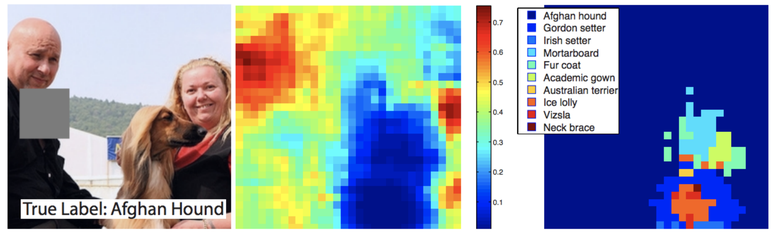

Figure 3: Occlusion ablation in classification.

- *Saliency map* - image that signifies parts of the input image the classifier is most confident about in its classification.
  - Gradient methods [4] - compute gradient of class score with respect to input image so as to rank pixels of input image based on their influence on class scores.
  - Masking model [5] - create a mask that represents the tightest rectangular crop of an input image that contains the entire salient region.

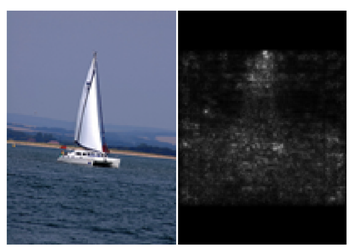

Figure 4: Saliency map.

Goal

- Extend ideas of saliency to semantic segmentation, where output is an input-sized image with per-pixel classification.
- Design a method to quantify a segmentation model's theoretical and effected receptive field so as to link context in scene understanding tasks.

Occlusion and Semantic Segmentation

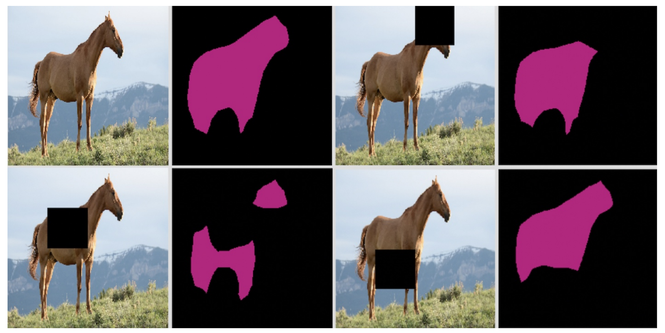

Figure 5: Binary segmentation on non-occluded and occluded images.

Figure 6: Occlusion sensitivity in semantic segmentation. Heatmap is a function of a sliding window across the input image.

References

[1] Evan Shelhamer, Jonathan Long, Trevor Darrell *Fully Convolutional Networks for Semantic Segmentation*. arXiv preprint arXiv:1411.4038 (2014)

[2] Vijay Badrinarayanan, Alex Kendall, Roberto Cipolla *SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation*. arXiv preprint arXiv:1511.00561 (2015)

[3] Matthew D Zeiler, Rob Fergus *Visualizing and Understanding Convolutional Networks*. arXiv preprint arXiv:1311.2901 (2013)

[4] Karen Simonyan, Andrea Vedaldi, Andrew Zisserman *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*. arXiv preprint arXiv:1312.6034 (2013)

[5] Piotr Dabkowski, Yarin Gal *Real Time Image Saliency for Black Box Classifiers*. arXiv preprint arXiv:1705.07857 (2017)