Fully Convolutional Neural Networks for Semantic Segmentation

CVPR 2015

Summary

The authors of this paper propose a network architecture that extends conventional CNNs used for image classification to take arbitrarily-sized input images and output pixel-level classification while preserving the dimensionality of the input. The architecture requires no fully-connected layers since the classification is performed inherently in the convolutional filters of the network. The input dimensionality is retained by upsampling the low-dimensional feature representation. Addressing the loss of spatial feature information induced by repeated subsampling, skip connections are employed tp fuse *"deep, coarse, semantic information and shallow, fine, appearance information"*.

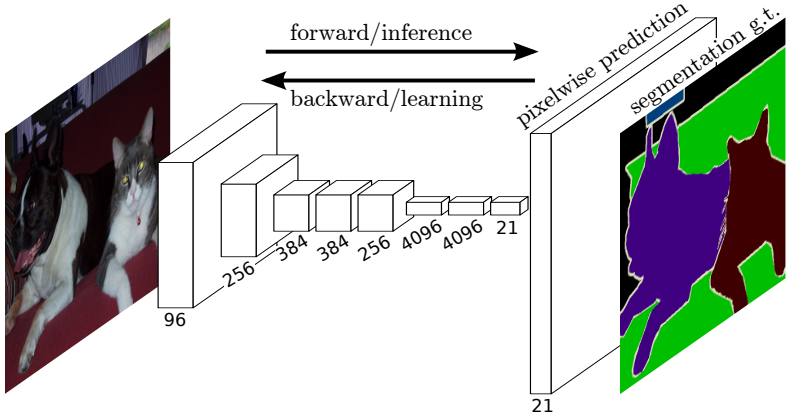

Figure 1: Overview of the FCN architecture. The 21 class scores are computed in a fully-convolutional manner before upsampling back to the input dimensions. The basic framework doesn't include any skip connections.

Main Contributions

- **Extension of Classification Networks** pre-trained AlexNet, GoogLeNet and VGG net are adapted to the FCN architecture and fine-tuned to the task of semantic segmentation. According to the authors, the paper constitutes the first attempt to perform end-to-end training of FCNs for semantic segmentation using pre-trained feature extraction networks.
- **Simplified Segmentation Pipeline** The proposed networks process whole image inputs and ground truths and compute the segmentation output without further pre- or post-processing techniques.
- **Skip Connections** Due to the loss of spatial information during the feature extraction, standard FCN models frequently produce coarse segmentation outputs. The authors observe that incorporating higher-dimensional information from earlier layers into the upsampling process significantly increases the quality of the segmentation.

Implementation Details

- **Transform Classifier to FCN** All three backbone networks are designed to output single class scores. Consequently, to transform them into an FCN, the final classification layer is removed and fully-connected layers are replaced by convolutional layers. In order to obtain pixel-level classification, a convolutional layer with 1x1 kernels and a channel dimension equal to the number of classes is appended. To retain the input dimensionality in the output, an upsampling layer is added as a final layer.
- **Upsampling** The upsampling of the feature maps is performed inside the network using what is referred to as deconvolution. Unlike commonly used schemes such as bilinear upsampling, deconvolution allows for learning the filter parameters of the upsampling operation as part of the network's end-to-end training.
- **Information Fusion** The authors test two different schemes for fusing information from different layers using skip connections. Both variants are implemented by computing class scores from the respective layers' activations and performing element-wise summation. The basic version combines the output of the two final convolutional blocks, while the advanced version uses information taken from three different layers. The outputs are fused at the dimensionality of the most shallow layer, requiring upsampling of the deeper feature maps before fusing the computed class scores.

Evaluation

- **Baseline** To find a suitable baseline model the authors test the standard FCN version of each of the three classifiers without skip connections. Among these, the FCN-VGG16 performs best with performance almost on a par with previous state-of-the-art methods.
- **PASCAL VOC** The main evaluation of the proposed method is carried out on the PASCAL VOC 2011 and 2012 test sets. The FCN-VGG16 is tested as a baseline model (FCN-32s) as well as with both skip connection schemes, referred to as FCN-16s and FCN-8s based on the upsampling factor of the final layer. Exemplary segmentation outputs of all three models are shown in Figure 2. In comparison to previous state-of-the-art approaches, the FCN-8s achieves a 30% relative performance increase while also being considerably faster (Figure 3).

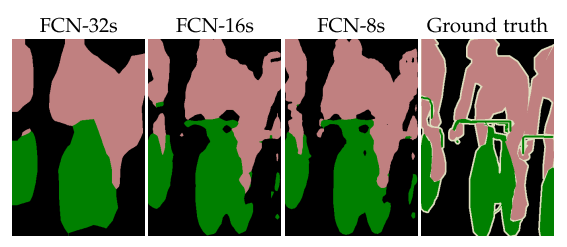

Figure 2: Exemplary network outputs. We can observe that the incorporation of spatial information via skip connections allows the network to output fine-grained segmentation masks.

	mean IU VOC2011 test	mean IU VOC2012 test	inference time
R-CNN [5]	47.9	-	-
SDS [14]	52.6	51.6	$\sim$ 50 s
FCN-8s	**67.5**	**67.2**	$\sim$ **100 ms**

Figure 3: Segmentation Performance of three different architectures. The FCN-8s performs considerably better in terms of quality of the segmentation as well as speed than previous state-of-the-art networks.

References

This summary is solely based on my understanding of the original paper. All images used here are taken from the original paper as well. The paper can be found under the following link:
`https://people.eecs.berkeley.edu/~jonlong/long_shelhamer_fcn.pdf`