PRE-REPORT: Fully Convolutional Networks for Semantic Segmentation

SongSeungWu

1 ABSTRACT

FCN can take input of any size and produce correspondingly-sized output. By adapting existing classification networks (AlexNet, VGG, GoogLeNet) for segmentation through fine-tuning, the model combines deep and shallow layer information to achieve high accuracy on datasets like PASCAL VOC, NYUDv2, and SIFT Flow, with efficient inference times.

2 INTRODUCTION

ConvNets have progressed from image classification to finer tasks like object detection, part and keypoint prediction, and local correspondence, now moving towards pixel-level predictions in semantic segmentation. This paper proposes the first end-to-end, pixel-wise prediction method using supervised pre-training. By applying upsampling within the network, it produces output matching the input size without additional pre- or post-processing. A skip architecture combines global and detailed information, allowing for accurate and refined predictions.

3 FULLY CONVOLUTIONAL NETWORKS

Fully Convolutional Networks are designed to perform pixel-wise predictions by accepting inputs of various sizes and preserving spatial information through a 3D array structure in each layer. Each higher layer position is path-connected to specific receptive fields in the input image. FCNs are adapted from traditional classification networks to generate dense outputs across all layers, restoring resolution to match the input size via in-network upsampling layers. This setup allows efficient feedforward computation and backpropagation, surpassing the efficiency of patch-based learning. FCNs also leverage OverFeat's shift-and-stitch trick or use deconvolution layers for pixel-wise prediction.

- **Adapting classifiers for dense prediction:** Typical classification networks (LeNet, AlexNet) process fixed-size inputs for nonspatial outputs, but transforming their fully connected layers into convolutions allows these networks to handle inputs of any size and output classification maps.

- **Shift-and-stitch is filter rarefaction:** Shift-and-stitch, introduced by OverFeat, creates dense predictions by shifting inputs and interlacing outputs, avoiding interpolation. For an output downsampled by factor f, the input is shifted and padded, producing predictions aligned with the receptive field centers. Adjusting filter size and stride can replicate this output, but at the cost of smaller receptive fields and longer computation.

- **Upsampling is backwards strided convolution:** Another way to generate dense predictions is through upsampling, which can be viewed as convolution with a fractional stride of 1/f. This process, known as deconvolution or backward convolution, can be implemented by reversing the standard convolution passes. Deconvolution layers can use learned filters instead of fixed ones, and a stack of deconvolution layers with activation functions can even learn nonlinear upsampling.

- **Patchwise training is loss sampling:** Both patchwise and fully convolutional training can use loss sampling to generate various distributions, with whole-image fully convolutional training performing the same loss calculation as patchwise training but more efficiently.

4 SEGMENTATION ARCHITECTURE

ILSVRC classifiers are converted into FCNs with in-network upsampling and pixelwise loss for dense prediction. A novel skip architecture combines coarse semantic and local appearance information to improve prediction. Training and validation were performed on the PASCAL VOC 2011 segmentation challenge dataset, using per-pixel multinomial logistic loss for training and mean intersection over union as the validation metric.

- **From Classifier to Dense FCN:** Proven classification architectures (AlexNet, VGG, GoogLeNet) were adapted to FCNs for experimentation. The VGG 16-layer network was chosen, with the final classifier layer removed and fully connected layers converted to convolutions. A 1×1 convolution layer and a deconvolution layer were added for dense predictions. FCN-VGG16 achieved a mean IU of 56.0, surpassing previous performance, and further improved to 59.4 with additional data. GoogLeNet did not match the performance of VGG in segmentation.

- **Combining what and where:** A new FCN architecture was designed to enhance spatial precision by combining feature layers. To counteract the coarseness of 32-pixel predictions, skip connections were added, linking higher layers with finer, lower layers to create FCN-16s and FCN-8s. This approach improved mean IU to 62.7 and enhanced fine structure detail. Further fusion efforts reached diminishing returns, so no additional layers were fused.

- **Experimental Framework:** Models were trained with SGD and momentum, using fixed learning rates for FCN-AlexNet, FCN-VGG16, and FCN-GoogLeNet. All layers were fine-tuned, requiring 3 days for FCN-32s and 1 day each for FCN-16s and FCN-8s on a single GPU. Full image training was employed without patch sampling, and class imbalance was managed without adjustments. FCN-VGG16 showed improved performance with additional training data from PASCAL VOC 2011.

5 RESULTS

Tests were conducted on the PASCAL VOC, NYUDv2, and SIFT Flow datasets to evaluate the FCN skip architecture. FCN-8s surpassed previous performance by over 20% on the PASCAL VOC 2011 and 2012 test sets while significantly reducing inference time. On NYUDv2, experiments with RGB-D input and various fusion methods were compared for performance. For SIFT Flow, a two-headed FCN-16s was employed to perform semantic and geometric classifications simultaneously, achieving state-of-the-art results in both tasks.

6 CONCLUSION

Fully convolutional networks (FCNs) extend modern classification networks to segmentation tasks, achieving significant performance improvements through multi-resolution layer combinations. This approach enhances learning and inference speed, setting a new state-of-the-art.

PRE-REPORT: Learning Deconvolution Network for Semantic Segmentation

SongSeungWu

1 ABSTRACT

This paper proposes a novel semantic segmentation algorithm by learning a deconvolution network built on VGG 16-layer convolutional layers. The deconvolution network, consisting of deconvolution and unpooling layers, predicts pixel-wise class labels and segmentation masks. The trained network is applied to each proposal in an input image, and results are combined to produce the final segmentation map. The proposed algorithm overcomes limitations of fully convolutional network-based methods by integrating deep deconvolution networks and proposal-wise prediction, effectively capturing detailed structures and handling multi-scale objects. The network achieves a high accuracy of 72.5% on the PASCAL VOC 2012 dataset.

2 INTRODUCTION

CNNs have achieved excellent performance in various visual tasks such as image classification, object detection, and semantic segmentation. However, traditional FCN-based segmentation methods face limitations with multi-scale issues and loss of object details. This paper addresses these issues by proposing a multi-layer deconvolution network, comprising deconvolution, unpooling, and ReLU layers, applied to individual object proposals for detailed segmentation. The model achieved outstanding results on the PASCAL VOC 2012 dataset, and further improved through ensemble with FCN, setting a new state-of-the-art.

3 ARCHITECTURE

The network consists of two parts: a convolution network and a deconvolution network. The convolution network, based on a modified VGG 16-layer net, extracts multidimensional features from the input image. The deconvolution network, a mirrored version of the convolution network, generates object segmentation using unpooling, deconvolution, and ReLU layers. The final output is a probability map of the same size as the input, indicating each pixel's class probability.

4 Deconvolution Network for Segmentation

- **Unpooling:** Pooling in convolution networks filters noisy activations by summarizing them, but it can lose spatial details. To address this, unpooling layers in the deconvolution network restore the original activation size. This is achieved by using switch variables that record maximum activation locations during pooling, placing each activation back to its original position, aiding in accurate object structure reconstruction.

- **Deconvolution:** The output from the unpooling layer is enlarged but sparse, and deconvolution layers densify this using learned filters. Unlike convolution, which maps multiple inputs to a single output, deconvolution maps a single input to multiple outputs, creating a dense activation map. These learned filters help reconstruct object shapes, capturing different levels of detail across layers, thereby directly incorporating class-specific shape information for semantic segmentation.

- **Analysis of Deconvolution Network:** The deconvolution network in this algorithm is essential for precise object segmentation. Unlike simple deconvolution, it uses unpooling, deconvolution, and ReLU to produce a dense pixel-wise probability map. Unpooling restores detailed structures, while deconvolution filters capture class-specific shapes, suppressing noise for accurate segmentation. Compared to FCN-8s, this network generates denser and more precise activation maps.

5 TRAINING

The network is very deep, containing numerous parameters, and the number of segmentation training examples is relatively small, making training challenging. Several strategies were applied to effectively utilize the limited number of examples for successful network training.

- **Batch Normalization:** Deep neural networks are challenging to optimize due to the internal covariate shift problem. To address this, batch normalization was applied, standardizing the input distributions of each layer. Batch normalization layers were added to all convolutional and deconvolutional layers, proving essential for optimal network performance.

- **Two-stage Training:** While batch normalization aids optimization, the semantic segmentation space remains large, so a two-stage training method was applied. In the first stage, easy examples were used for training, and in the second stage, more challenging examples were introduced using object proposals. This approach enhances the network's robustness to variations in object position and scale.

6 INFERENCE

The network is trained for instance-wise segmentation. Candidate proposals are generated from an input image, with the network applied to obtain segmentation maps for each proposal. The results are then aggregated for full-image segmentation by using pixel-wise maximum or average values to suppress noise, followed by CRF for final pixel-wise labeling. An ensemble with FCN leverages the network's ability to capture fine details along with FCN's overall shape extraction, enhancing performance by combining complementary features.

7 EXPERIMENTS

The proposed network was evaluated on PASCAL VOC 2012, and performance was notably improved through ensemble with FCN. Training was done in two stages with appropriate data for each stage, using SGD with momentum and batch normalization instead of dropout. Results showed that DeconvNet captures fine object details better than FCN and handles multi-scale objects effectively. The ensemble method further enhanced performance, and adding CRF provided minor additional improvements.

8 CONCLUSION

A novel semantic segmentation algorithm is proposed using a deconvolution network that generates dense, precise segmentation masks. The instance-wise prediction approach effectively manages object scale variations, and the ensemble with FCN further enhances performance. This network achieved state-of-the-art results on the PASCAL VOC 2012 benchmark without using external data.