Fully Convolutional Networks for Semantic Segmentation

Authors: Jonathan Long, Evan Shelhamer and Trevor Darrell
Publisher: IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
Year: 2015

Keywords: Semantics Segmentation, Fully Convolutional Networks, Image Segmentation, Deconvolution

1. **Abstract:**

"Convolutional networks are powerful visual models that yield hierarchies of features. We show that convolutional networks by themselves, trained end-to-end, pixels-to-pixels, exceed the state-of-the-art in semantic segmentation. Our key insight is to build "fully convolutional" networks that take input of arbitrary size and produce correspondingly-sized output with efficient inference and learning. We define and detail the space of fully convolutional networks, explain their application to spatially dense prediction tasks, and draw connections to prior models. We adapt contemporary classification networks (AlexNet, the VGG net, and GoogLeNet) into fully convolutional networks and transfer their learned representations by fine-tuning to the segmentation task. We then define a novel architecture that combines semantic information from a deep, coarse layer with appearance information from a shallow, fine layer to produce accurate and detailed segmentations. Our fully convolutional network achieves state-of-the-art segmentation of PASCAL VOC (20% relative improvement to 62.2% mean IU on 2012), NYUDv2, and SIFT Flow, while inference takes less than one fifth of a second for a typical image."

2. **Introduction:**

- Many of the applications of Convolutional Neural Networks had been implemented for object detection and semantic segmentation without pretraining which resulted in the limited results.
- The motivation of the paper is to define an unique architecture which combine the external shallow and fine layers of appearance information to the coarse and deep layers of semantic information to increase the accuracy of the image segmentation.
- This approach deals with the supervised pre training of Fully Convolutional networks for pixel-to-pixel prediction and application of transfer learning into the predefined neural networks to increase the precision of results. The state of the art results were experimented on various datasets such as SIFT FLOW, NYUDv2 etc.,

3. **Related works:**
    - The advantages like altering only the parameters and layers without interacting with total Neural Networks, had been increased the application of Transfer learning.
    - Formerly, the FCN were used for the recognition of numbers by Matan et al., as their input consists only single dimensional strings and further the operations of FCN had been extended to detect maps based on the postal addresses.
    - Both the applications of FCN defined above used the inference. Even some of the
    - following approaches like semantic segmentation, sliding window and image restoration would have to implement convolutional inference.
    - As the prior methods implemented superpixel projection, filtering, input shifting and multi-scale pyramid processing for their process of semantic segmentation, the authors eliminated all those procedure, instead they adapted supervised pretraining with the help of image classification.
    - In addition, None of the previous approaches pre trained the Fully Convolutional Networks end to end.
4. **Content and Approach:**
    - The fundamental factors of CNNs depend on relative spatial coordinates and they function on the limited input regions.
    - The layers with the deep networks with certain functions calculates the deep filter(also called as Fully Convolutional Networks), which are non linear in structure.
    - As the authors state that if the loss function is equal to the sum of the spatial dimensions of the end layer, such that $l(x; \theta) = \Sigma_{ij,}\ l'\ (x_{ij}\ ; \theta)$, then the SGD of the total images of on loss function 'l'will be equal to the SGD of l' considering the end layer receptive images as a mini batch.
    - General networks of recognition such as AlexNet and the extended successors will receive only the fixed size input sand result in the nonspatial outputs.
    - By considering the fully connected layers as the convolutions with kernels which would shield it's entire input layers, would result in taking up the variable inputs.
    - The variable inputs would result in the spatial classification maps which would be the general choice for semantic segmentation. The dimensions of the output maps can also be curtailed by subsampling.
    - The x pixels were shifted to right and y pixels to down (known as padding), if we decrease the sample by certain factor.
    - Reducing the sample size, a kind of tradeoff, will result in computational expense however increases the accuracy of the information. This in fact creates smaller receptive fields.
    - On the other hand the shift and Stitch trick is also a type of trade off, in which we does not reduce the size of receptive field, but instead we stop accessing

information from finer scale compared to the normal design. this result in the denser output.

- The other way of relating the denser pixels to the course output is by interpolation.
- Patchwise and FCN training can be resulted in any of the distribution, despite that the efficiency relies on thee minibatch size and the overlapping.
- Although the process of FCN training can be done for every batch with the corresponding receptive fields, with respect to the image loss, is more effective than the uniform sampling, it decreases the number of achievable batches.
- AlexNEt, GoogLeNet and VGG architectures, which had been awarded to architectures at ILSVRC14 were considered for the author's experimentation.
- For each of the Convolutional Layer considered, the authors had removed the final layer and will modify all the CNN to FCNs.
- The training is done by SGD optimizer and for the class scoring, they had initialized with zero so that it would not result in faster convergence and better performance.
- All the layers of the network had also been fine tuned as the training of total network from scratch is impossible by considering the limited amount of time.
- Patch Sampling and Augmentation were also implemented on the training data which had resulted in no improvement.
- Caffe on single NVIDIA Tesla K40c had been used for the purpose of both training and testing all the adapted models.

5. Results:

- The FCN results were tested on the datasets like PASCAL, SIFT flow, NYUDv2 based on scene parsing and the semantic segmentation.
- Various parameters such as mean accuracy, pixel accuracy, mean IU were also tested and compared for various datasets and the architectures.
- As expected, Fully Convolutional Networks had resulted best outputs of mean IU with 62.7% compared to R-CNNs with 47.9%.
- In all the parameters compared from pixel accuracy to mean IU, FCN-16s had outperformed all the other architectures.
- For the SIFT flow, dataset the model had resulted in 94.3% of geometric accuracy exceeding all the previous approaches.

6. Conclusion and Future works:

- It can be concluded that the Fully Convolutional Networks can result in the outstanding outputs with pretraining and transfer learning. The results projects the efficiency of the model.
- As the total network is not being modified, the approach can be implemented without any difficulty. It is also defined to be simple and yet feasible.

- The work can be extended to implement into multi-resolution layer which would substantially improves the learning speed and inference with the state of the art.