Deep Learning Architectures for Pixel-wise Semantic Segmentation of Images

Arik Md. Isthiaque[1]*, Arifa Islam Champa[2]

[1] Department of Computer Science and Engineering, Bangladesh Army International University of Science and Technology, Cumilla, Bangladesh.
e-mail: isthiaque@baiust.edu.bd
tel.: +8801521493029

[2] Department of Computer Science and Engineering, Bangladesh Army International University of Science and Technology, Cumilla, Bangladesh.

Abstract

Detecting objects from images and making decisions form those collected data is a major part of an intelligent machine. Nowadays many real-time machines like robots, autonomous vehicle works on the basis of detecting objects from visual scenarios [1]. So, the technological advancement of computer vision is an utmost importance. The semantic segmentation of images is a part of computer vision to understand a visual scenario. In this paper, we will look deep into some of the major architecture that is developed for semantic segmentation and will look into the future of it.

Keywords- Convolutional Network, FCN8, DenseNet, SegNet, Bayesian SegNet.

1. Introduction

Image segmentation is a process where the objects of an image are classified by their pixel information. In an image, certain pixels share some common characteristics. By identifying them we can label the pixels, so they can be detected from the image. We use image segmentation to find objects in images like text (OCR), cars (automated vehicle), cancer, tumor (medical image). In recent days, deep learning algorithms got success in handwritten OCR, NLP [2], [3]. It has also a huge active interest for pixel-wise semantic segmentation. Semantic segmentation is a machine learning process where a machine can learn from a visual scenario by adding class to every pixel that contains [4]. It is not like other instance segmentations. Instance segmentation can differentiate between objects of the same classes. In semantic segmentation, it can only detect and it doesn’t care about the instances.

2. Architectures for Semantic Segmentation

2.1 Fully Convolutional Network (FCN8)

FCN8 is a deep learning algorithm for semantic segmentation in the image [5]. The architecture is built with some blocks of convolution and max pool layers. The purpose of these layers is to decompress the input image and reduce it to 1/32th of its original size. It generates some classes form this decompressed image by using some random predictions. After that, an upsampling and deconvolution layer work on that decompressed image and the image return to its original dimensions [6].

This architecture is not fully connected. It captures the contextual information via downsampling. But it doesn’t limit the input image because the original image size can be regenerated via upsampling. So, the result we got after the process is the same as the original image size. Here the architecture also uses skip connections of layers to fully regenerated the original image. In skip connection of layer architecture, the process can skip one or some of the layers before completing its process [7] [8].

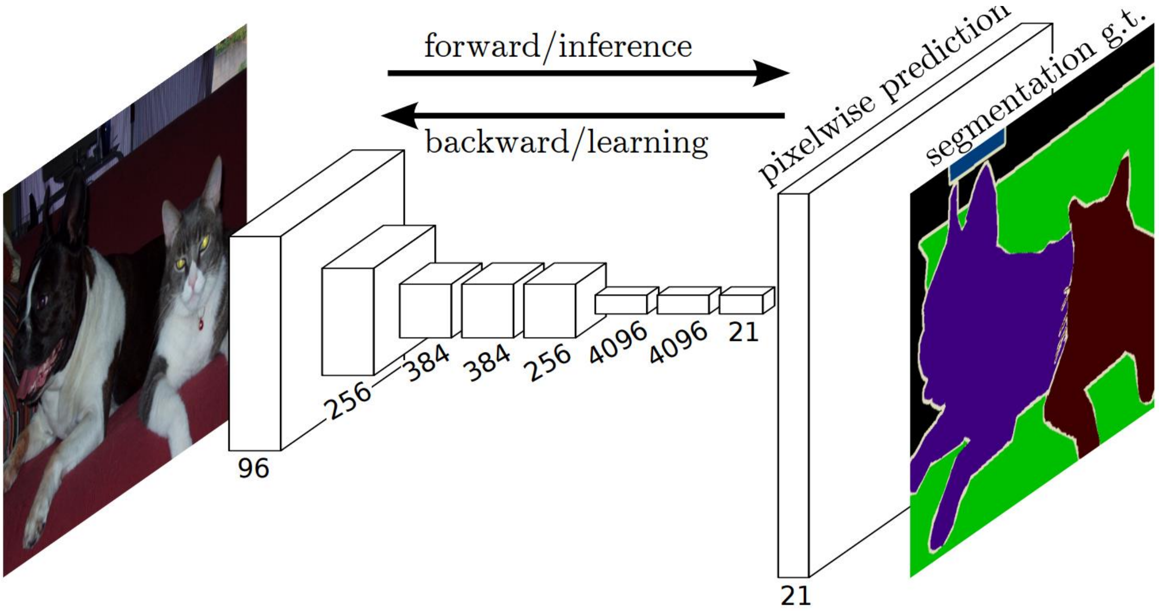

Figure 1: Fully Convolutional Network (FCN8) [5].

2.2. SegNet

SegNet is an encoder-decoder architecture built with a deep convolutional neural network for pixel-wise semantic segmentation for images and videos. It is generally built for roads, buildings, cars, pedestrians and for differentiated the context between roads and side-walks. SegNet extracts low-resolution pixel information from the input image and classifies them as per pixel. This extraction must generate some patterns which can be used in boundary localization [4].

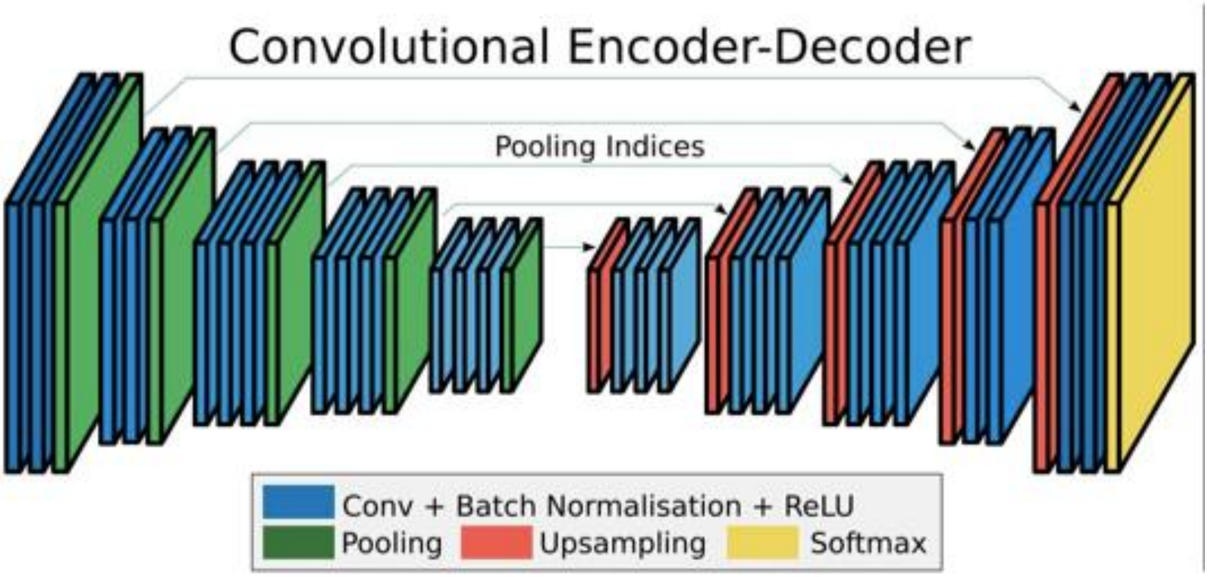
Convolutional Encoder-Decoder
Pooling Indices
Conv + Batch Normalisation + ReLU
Pooling
Upsampling
Softmax

Figure2: SegNet Architecture [4].

SegNet built with an encoder layer and a decoder for a pixel-wise segmentation. The model is in Figure 1. The encoder layer (is) built with 13 convolutional layers that follow the VGG16 network [3]. The training process can be initialized from weights trained for classification with large datasets [9].

SegNet is built with a stack of encoders with a similar number of decoders. The decoders made the output the same size as the input. This is an important drawback for the architecture as the increasing number of layers increase the parameters of the model [4].

2.3. Bayesian SegNet

This is an extension of the main SegNet architecture. Its main job is to find the uncertainty in the SegNet architecture [4] [10]. It consists of the same sequence of non-linear processing layers called encoders and a corresponding set of decoders followed by a pixel-wise classifier. In this Bayesian architecture, the performance is improved for some baseline datasets. It can probabilistic segmentation by finding the posterior distribution over convolutional weights. Here Monto Carlo dropout is used to approximate the inferences [11].

2.4. Fully Convolutional DenseNet

FCN8 is built with downsampling, an upsampling and some skip layers. In this FCN8 there is a major drawback that is the features explanations in the upsampling model. The performance of FCN8 could be extended by avoiding features explorations in the upsampling model. This problem is solved in this new densely connected network which is called the Fully Convolutional DenseNet. In the Fully Convolutional DenseNet, the convolutional operations are substituted by a new operation which is referred to as the transition up. It consists of several convolutional modules that upsample the compressed image. Input images are concatenated by the skip layer with a dense layer [12].

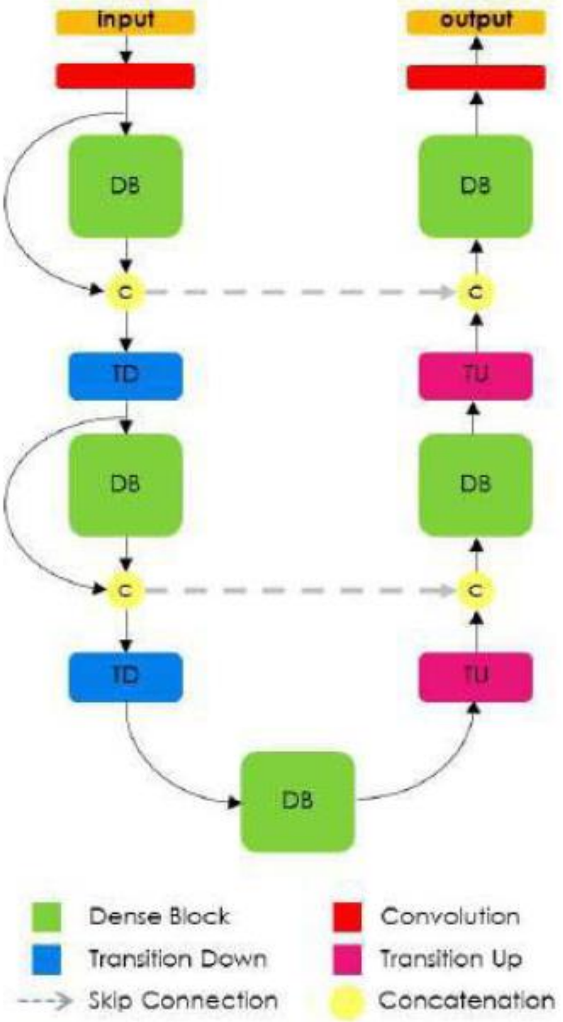

Figure 3: Fully Convolutional DenseNet Architecture [12].

Fully Convolutional DenseNet extended the FCN8 architecture to tackle the problems of image semantic segmentation by making it fully convolutional and by building dense blocks and add them with the extracted feature maps.

3. Comparisons Between the Architectures

Architecture	Global Accuracy	
	CamVid	**Pascal VOC**
FCN8	57.0	62.2
SegNet	62.5	51.9
Bayesian SegNet	86.9	-
FC-DenseNet56	86.8	-
FC-DenseNet67	88.9	-
FC-DenseNet103	90.8	-

Figure 4: Global accuracy comparison of all the architectures tested with the “CamVid” dataset and the “Pascal VOC” dataset [10] [12].

4. Discussion

The “Pascal VOC” and the “CamVid” dataset contain standardized images that can be used for object detection. “Pascal VOC” mainly contains outdoor scenes. On the other hand, “CamVid” contains images related to the city roads. So, from the above comparison of the datasets, we can see that the FCN8 which is the base of most of the deep learning algorithms of semantic segmentation doesn’t have that global accuracy but from time to time the model is updated with so many extensions. Now the recent trending architecture that is the Fully Conventional DenseNet is developing day by day. It was first trained with the 56 layers which is known as the FC-DenseNet56 [12]. As the layers are being increased the accuracy has a significant change. Now the most recent architecture has 103 layers which is known as FC-DenseNet103.

5. Conclusions

Semantic segmentation of the image is one of the key processes of machine automation like robotics and autonomous vehicle. So, it is important to develop these basic processes more efficient to make our machines smarter. From FCN8 to Fully Convolutional DenseNet many things are changed for achieving state-of-the-art accuracies and make the process more efficient. The main motivation behind this review paper is to make an overview of the deep learning architectures for semantic segmentation of images for researchers who are interested in working with this process.

References

[1] P. Luc, N. Neverova, C. Couprie, J. Verbeek, Y. LeCun; The IEEE International Conference on Computer Vision (ICCV), 2017, pp. 648-657C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “*Going deeper with convolutions*,” in CVPR, pp. 1–9, 2015.

[2] K. Simonyan and A. Zisserman, “*Very deep convolutional networks for large-scale image recognition*,” CoRR, vol. abs/1409.1556, 2014.

[3] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, “*Learning hierarchical features for scene labeling*,” IEEE PAMI, vol. 35, no. 8, pp. 1915–1929, 2013.

[4] V. Badrinarayanan A. Handa R. Cipolla "*SegNet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labeling*" arXiv preprint arXiv:1505.07293 2015.

[5] J. Long, E. Shelhamer, T. Darrell, “*Fully Convolutional Networks for Semantic Segmentation*” The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3431-3440

[6] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. *DeCAF: A deep convolutional activation feature for generic visual recognition*. In ICML, 2014.

[7] K. He, X. Zhang, S. Ren, and J. Sun. *Spatial pyramid pooling in deep convolutional networks for visual recognition*. In ECCV, 2014.

[8] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. *Caffe: Convolutional architecture for fast feature embedding*. arXiv preprint arXiv:1408.5093, 2014.

[9] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “*ImageNet Large Scale Visual Recognition Challenge,*” International Journal of Computer Vision (IJCV), pp. 1–42, April 2015.

[10] V. Badrinarayanan A. Handa R. Cipolla *Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene* Understanding (Submitted on 9 Nov 2015 (v1), last revised 10 Oct 2016 (this version, v2))

[11] V. Badrinarayanan, F. Galasso, R. Cipolla. *Label propagation in video sequences*. In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pages 3265–3272. IEEE, 2010.

[12] S. Jegou, M. Drozdzal, D. Vazquez, A. Romero, Y. Bengio; *The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation* The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2017, pp. 11-19