Semantic Segmentation using Deep Learning Approaches - A Study

Salim Ahmed Ali[1], Dr. B. G. Prasad[2]

[1]Department of Computer Science & Engineering, B. M. S. College of Engineering, Bengaluru, India
salim7ali[at]gmail.com

[2]Department of Computer Science & Engineering, B. M. S. College of Engineering, Bengaluru, India
bgprasad.cse[at]bmsce.ac.in

Abstract: ***Semantic segmentation is an important technology which gets commonly used in medical imaging, autonomous driving vehicles, and also backgrounds for virtual meetings. Classes can be different real-world objects such as roads, cars, bicycles, people, trees, lanes, trucks, buildings etc. Classes can correspond to different anatomical structures and organs when considering medical images. Semantic segmentation is a broadly applicable technology. The techniques discovered to improve current semantic segmentation methods could also lend themselves to improving other dense prediction tasks. These tasks could include optical flow prediction (object motion prediction tasks), image super-resolution such as in remote gaming or in video resolution enhancement, and so on. This paper briefly presents a survey on existing work conducted to achieve semantic segmentation of image problems with the use of deep learning methods as well as image processing approaches. Deep learning provides several methods for semantic segmentation such as 2D convolution networks, 3D convolution networks, etc. This paper discusses the classification, challenges, application, and methods for semantic segmentation.***

Keywords: Deep learning, Convolutional Neural Networks, Semantic Segmentation, Deep Neural Networks, Image processing

1. Introduction

Common computer vision problems related to scene understanding [1] are image classification, object detections and segmentation and these are represented in increasing order of difficulty. Image classification is the simple task of asking whether a specific object exists in the image frame or not. Object detection is to envelop different classes of objects with bounding boxes (dog, car, road, etc). Semantic segmentation goes a step further by asking whether a given pixel belongs to a set of N classes. The result of semantic segmentation is that the objects are enclosed by exact boundaries.

Semantic segmentation plays a very important role in the development of robotics, self-driving cars, etc.

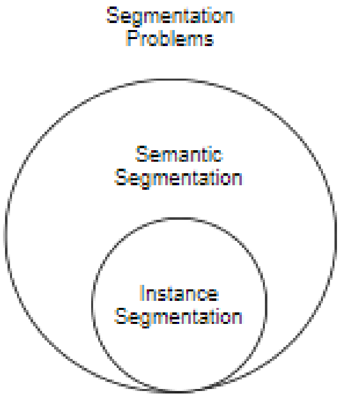

Figure 1: Types of Segmentation Problems

Semantic Segmentation is simply the labelling of every pixel which is present in an image into a set of M classes. These classes could be objects like roads, lanes, cycles, vehicles, people, buildings, etc. Another form of the semantic segmentation problem is that of Instance Segmentation. Instance segmentation associates itself with distinguishing between each object belonging to the same class (eg. vehicle_1, vehicle_2….vehicle_N based on a class of Vehicles).

Commonly used algorithms for semantic segmentation were the K-means clustering, Watershed algorithm, Graph partitioning methods, thresholding, etc. Of these, the simplest one is the thresholding approach. Even though many such algorithms exist, most of the famous ones are based on Deep Neural Nets (DNNs) such as ResNet 50. One of the main reasons being the use of skip connections in ResNet (Residual Network)[2] compared to previously used DNN's which simply used convolution layers in succession and any information passed at initial layers is very minimally remembered at the end layers. Thus, ResNet introduced the concept of skip connections which made the initial layer information directly available for the end layers and a simple element-wise addition could be applied to see this implemented. The advantage of ResNet could be seen on accuracy vs no_of_layers graph. Traditional methods yielded higher accuracies as the number of layers increased but at some point, the accuracy would start to decrease[3]. This wasn't the expected behaviour. But ResNet handled this elegantly with its skip connections. In 2015, ResNet won the ILSVRC award for Image classification and had up to 152 network layers. Residual nets have been found to achieve a 3.57% error on the ImageNet test set. For fine-grained segmentation inference, we need to provide additional information regarding the spatial location of classes. These classes have to be provided earlier and are required for localization or detection. Semantic segmentation can be thought of as a prediction task, particularly a dense prediction.

The objective of dense predictions is to create an output map that matches the size of the input image. Since we require fine-grained prediction, using semantic segmentation is an obvious choice. The goal to achieve here is for every pixel in the image to be made up of dense predictions which infer labels. This results in the pixel being labelled with the class of its bounding object or area. A decoder network preceded by an encoder network is what makes up a general semantic segmentation architecture. For the encoder component, we generally use a pre-trained classification network of the likes of ResNet/VGG. This gets followed by a decoder network. The discriminative features (low pixel density) are to be semantically projected onto the pixel space (higher resolution) by the decoder. These discriminative features are the ones learnt by the encoder and thus finally results in a dense classification.

This paper presents various segmentation architectures with a comparison of standard datasets. An early architecture for this task is the U-Net[4] architecture which has formed the basis for other segmentation architectures.

2. Literature Survey

a) U-Net

U-Net[4] is a network architecture made up of encoder-decoder segments which propagates essential features of an image to the final layers for classification tasks. U-Net adds functionality on top of the FCN architecture. It is a deep learning model made up of convolution and max-pooling layers. The encoder is made up of convolution and max-pooling layers which shrink up to a point. This is followed by a decoder which is implemented similarly but in reverse order where we use up-convolution. Incorporates concatenation layers between the encoder and decoder network path. It uses a kernel size of 3x3 whereas the up-convolution and max-pooling are specified by a 2x2 kernel. The end layer is a 1x1 convolution layer that has the task of generating the class probability map.

Achievements

- Improved performance compared to other segmentation techniques at the time.

Limitations

- Fails in the extraction of tiny features.
- A large number of training parameters are required.
- Skip connections force aggregation only at the same-scale feature maps of encoder and decoder subnetworks.[5]

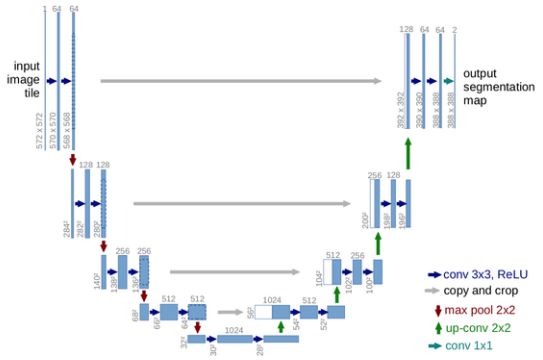

Figure 2: U-Net architecture

b) Multi-Scale Attention [6]

Better features can be obtained from using filters differing in scale and orientation. When for example we take a 2x scaled image (zoomed in), the finer features of the image are better mapped to the segmentation model. This could include cases of far off street lights.

It also allows us to visualize important features at different positions and scales using an attention-based approach which learns to weigh the multi-scale features at each pixel location. Using multi-scale inputs produces greater performance compared to a single scale input (very commonly used).

When the final output of the network is added with extra supervision for every scale, performance improves for the model.

Achievements:

- Merging multi-scale (zoomed in and zoomed out)[7][8] features with an attention model increases the performance compared to max-pooling or average baselines.

Limitations

- Adds redundant information usage as similar low-level features are obtained multiple times at varying scales. i.e. encoder-decoder architectures[9]

c) DeepLab[10]

For FCNs, when consecutive downsizing is done by pooling operations, a lot of information is lost. Also, this is computationally expensive during deconvolution to upsample by 32x. Normal convolutions thus lose a lot of fine details due to these repeated convolutions. Thus, DeepLab proposed the use of Atrous convolutions which are filters with dilations. A dilation of 1 would result in a normal convolution. Whereas a dilation of 2, one hole is introduced between every parameter resulting in a filter to look like a 5x5 one while having 3x3 convolution parameters. Thus, we can test with increased dilations to increase context for a filter.

DeepLab also incorporated Spatial Pyramidal Pooling (SPP) which was introduced by SPPNet[11] which is used to capture multi-scale information from a feature map. While SPP allows for capturing multi-scale information with a single image, ASPP takes this same concept and applies it to Atrous convolutions. The DeepLab system is able to achieve

79.7% mIOU for the PASCAL VOC-2012 semantic image segmentation problem.

Achievements:

- Atrous spatial pyramid pooling (ASPP) is utilized for segmenting objects at differing scales.
- Improves localization of object boundaries using DCNNs.

d) Fully Convolutional Network

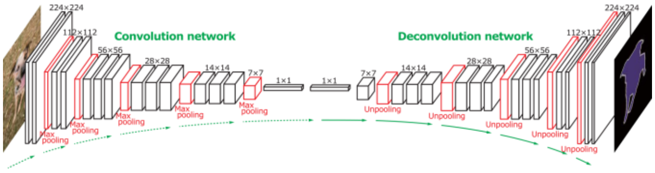

Figure 3: Typical FCN architecture

Fully Convolutional Network (FCN)[12] were an early Deep Convolutional Neural Networks (DCNNs) that introduced the idea that the fully connected layer at the end of a CNN and can be thought of as doing a 1x1 convolution over the entire region. Thus, this final fully connected layer (dense layer) is replaced by a convolution layer in FCNs. This has the benefit of removing the constraint that input size needs to be fixed. Also, this results in a feature map being produced at the output, rather than a class output for a normal input sized image. FCN also proposes using learned upsampling during deconvolution which allows for learning non-linear upsampling as well.

Achievements:

- Input image size need not be fixed.

Limitations:

- Generated results have rough and fuzzy boundaries.

e) Mask R-CNN

Faster RCNN is an excellent algorithm that mostly gets utilized for tasks such as object detection. Faster R-CNN comprises of two phases. Candidate object bounding boxes are proposed by the principal stage which is called Region Proposal Network (RPN). The subsequent stage, which is generally Fast R-CNN, extricates highlights utilizing RoIPool from every candidate box. It also performs regression of bounding boxes and classification as well. The highlights utilized by the two phases can be shared for quicker inference.

Faster R-CNN generates two results for each of the probable objects. These are class label and bounding-box offset. Another third branch is combined which generates this object mask. This object mask is a binary mask indicating the bounding box with the use of pixels. Finer spatial layout is extracted using FCN. Mask RCNN merges Faster R-CNN with a FCN to achieve results. Mask RCNN has a few extra upgrades that make it considerably more precise than FCN.

Achievements:

- Allows to eliminate artefacts that exist on overlapping instances
- Is easily extensible to other problems such as pose estimation

Limitations:

- Can produce false positives.[13]
- Labels may not be generated.

3. Evaluation Metrics and Comparison

$$IOU = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

mIOU is the IOU measured over the entire set of classes.

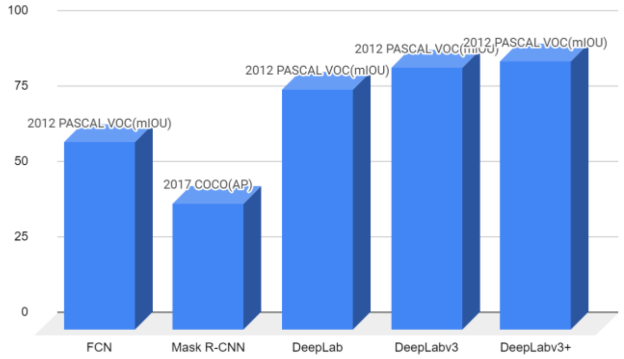

Figure 4: Scores of models over the 2012 PASCAL VOC dataset (mIOU) and 2017 COCO dataset (AP)

4. Conclusion

Semantic Segmentation is an important computer vision problem that has its uses in medical image diagnosis, self-driving cars, scene cropping, etc. The myriad of solutions that semantic segmentation has come to solve have made great contributions to the field of Machine Learning and Computer vision. In this paper, we have discussed the U-Net architecture for its instrumental nature and contributions to other architectures and techniques. We have discussed the merits and demerits of its applications and also talked of the changes and improvements it has undergone.

Other techniques such as multi-scale data have been very practical with most new age architectures implementing a form of scale aware imaging in segmentation problems. Granular details and large scale details have been the focus of the scale approach.

In this paper, we have discussed the different architectures and techniques hinged on Deep Learning models which have provided major contributions to the field of Semantic segmentation.

References

[1] Everingham, M., Eslami, S.M.A., Van Gool, L. et al. The PASCAL Visual Object Classes Challenge: A Retrospective. Int J Comput Vis 111, 98–136 (2015). https://doi.org/10.1007/s11263-014-0733-5

[2] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern

Recognition (CVPR), Las Vegas, NV, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[3] Andrew Ng. Residual Networks - Case Studies. https://www.youtube.com/watch?v=ZILIbUvp5lk&t=311s . Nov, 2017

[4] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. MICCAI.

[5] Zhou, Z., Siddiquee, M., Tajbakhsh, N., & Liang, J. (2020). UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. IEEE transactions on medical imaging, 39 (6), 1856–1867. https://doi.org/10.1109/TMI.2019.2959609

[6] L. Chen, Y. Yang, J. Wang, W. Xu and A. L. Yuille, "Attention to Scale: Scale-Aware Semantic Image Segmentation," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 3640-3649, doi: 10.1109/CVPR.2016.396.

[7] Florack, L., Ter Haar Romeny, B., Viergever, M. et al. The Gaussian scale-space paradigm and the multiscale local jet. Int J Comput Vision 18, 61–75 (1996). https://doi.org/10.1007/BF00126140

[8] P. Arbeláez, M. Maire, C. Fowlkes and J. Malik, "Contour Detection and Hierarchical Image Segmentation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 5, pp. 898-916, May 2011, doi: 10.1109/TPAMI.2010.161.

[9] A. Sinha and J. Dolz, "Multi-scale self-guided attention for medical image segmentation," in IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2020.2986926.

[10] "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs" Liang-Chieh Chen*, George Papandreou*, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille (*equal contribution) arXiv preprint, 2016

[11] K. He, X. Zhang, S. Ren and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904-1916, 1 Sept. 2015, doi: 10.1109/TPAMI.2015.2389824.

[12] J. Long, E. Shelhamer and T. Darrell, "Fully convolutional networks for semantic segmentation," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 3431-3440, doi: 10.1109/CVPR.2015.7298965.

[13] Jiageng Zhang, Jingyao Zhan, Yunhan Ma. Mask R-CNN. https://cseweb.ucsd.edu/classes/sp18/cse252C-a/CSE252C_20180509.pdf .