After reading various papers on image segmentation, I find that they try to solve following problems:

1. CNNs contain semantic information in later layers but finer details are not available, so how to localize? i.e. how to capture both global and local context
2. Different scales of objects in the image
3. Spatial transformation – rotation of objects
4. How to account for label compatibility among neighboring pixels and objects?

Deeplab models

It uses dilated convolution to address the first problem as well as second problem. Dilated convolution is done in parallel with different dilation rates and combined. Then, bilinear interpolation is performed to reach original image size.

Instead of providing scaled images to model, sampling of the image is done using dilation rates to make sure even if object is scaled it would be captured by some rate.

The first paper[6] mentions that they solve 3rd problem by applying CRF to finetune. Also, it mentions that in other work, researchers used hypercolumn[7] but I am not sure how either solves spatial transformation of an object in images; have to explore more.

In an updated version of the model, authors use an encoder-decoder architecture, where encoder is same as before(uses ASPP) and decoder is made up of deconvolution layers instead of bilinear interpolation. This makes sense, as more parameters could learn more refined segmentation.

Apart from the first model, Deeplab the authors drop CRF. It seems that better training achieves the same results. I haven’t found a concrete reason for that in the paper[8]. There is one more update [9], where authors tweak the decoder and use depth wise convolution to save parameters.

Fully convolution models

[10]FC layers from dense classifiers are removed, lower layers are upsampled using learnt deconv layers and combined successively with predictions from higher layer upsampling and finally prediction is at image size.

Parsenet[11] argues that though theoretically higher encoder layers in FCNN do have large receptive fields, effective RFs are smaller and do not capture global context. This model captures a global context feature at every layer, unpools and concatenates it to every feature layer and forwards.

Effective RF may be smaller in case of classification, as shown in [12] but why is it smaller in case of segmentation, for a fully trained network?

Pix2pix and conditional GAN

The basic advantage of GAN I understand is that they don’t need labels and data distribution is learnt better with use of a discriminator. Does it hold for conditional GAN? Why should unet learnt through a discriminator with a L1 loss perform better?

[1] offers an answer. It says discriminator can detect mismatches in higher order stats between model prediction and ground truth which pixel level loss cannot, and we don’t have to define those losses. Most of the CRF models use pairwise potentials. Some of them do explore higher order potentials but this is set explicitly based on expert understanding of the problem. Using adversarial learning long range and higher order functions without explicitly defining them.

What is the connection of patch wise discriminator training with Markov models?

Markov property- Given the neighborhood labels and Image, pixel label is independent of the rest of the labels.

In the original GAN paper, the proof of convergence is not clear; trying to understand.

Bayesian Segnet

[20] provides measure of uncertainty in predictions and reports higher uncertainty at object boundaries. It says it uses fewer parameters. It uses dropout at test time. [21] explains how dropout in NN can be interpreted as probabilistic gaussian process. This explains why dropout works and allows to obtain uncertainty estimates. The math here is not clear; working on it.

Pyramid based models

[13] attempts to solve 2nd problem. FPN model was used in region proposal for object detection. It was modified to do segmentation. From the paper it looks like it predicts square regions for objects; will visit the implementation. The authors mention that it works for arbitrary image size.

PSPNet[14] uses a pyramid pooling structure and suggests a global context prior for better classification of objects. As effective RF in CNN is smaller and doesn’t capture the scene, authors propose a fix. I am not sure if this is applicable to our problem, as context is same for our images.

RNN models

There are some graph LSTM models which have been proposed. To be studied.

Attention based models

To be studied

Panoptic segmentation

Facebook approach of combining semantic and instance segmentation. The authors propose a new metric which is a highlight of [15]. This is not a new architecture; not applicable to our problem.

Active contour models

To be studied

CNN+CRF models

CRF tries to find joint distribution of image and labels over the pixels. In [3], authors use a pairwise potential CRF, meaning the graph is factored using functions over connected pixels only. The pairwise function used is a constant,1. I have not yet understood why. In [4] author uses appearance and smoothness functions, which assign larger value for colour similarity and closeness.[5] uses a function, not clear. Also, instead of maximizing the MAP or minimizing the energy, this paper formulates a SVM error.

I went through a lot of papers on CRF for image segmentation. I understood the linear chain CRFs which use forward-backward algorithm, which is used for NER tagging in NLP. Also, I understood the variational inference and mean field inference which is used as approximation for inference in CRF models. I haven’t yet understood how this is implemented in image segmentation; working on it.

How is CRF as a ‘finetuner’ implemented?

‘Super pixel’ is a term used in many papers. What is that?

Approximations made for ‘log linear models’?

“Densely connected models” term appears in Markov model papers. What are these?

Dense random field models are those MRF in which every node is connected to every other. Advantage over sparse models is that sparse models cause smoothing of object boundaries and long range inconsistencies. – from [2]

This makes sense as we should model image in graph model as dense and correlation should be figured out by parameters. Why did the previous models use parse nodes for image tasks?

Unary and pairwise potentials?

When you restrict the potentials to edges instead of maximal cliques, it’s called pairwise.

I referred [16],[17],[18],[19] to understand basics of MRF and CRF. I am going through implementation to understand this better.

References

1. <https://arxiv.org/pdf/1611.08408.pdf>
2. <https://proceedings.mlr.press/v28/kraehenbuehl13.pdf>
3. <https://arxiv.org/pdf/1801.08268.pdf>
4. <https://github.com/roym899/crf_segmentation_tutorial>
5. <https://arxiv.org/pdf/1503.08263.pdf>
6. <https://arxiv.org/pdf/1412.7062.pdf>
7. <https://arxiv.org/pdf/1411.5752.pdf>
8. <https://arxiv.org/pdf/1706.05587.pdf>
9. <https://arxiv.org/pdf/1802.02611.pdf>
10. <https://arxiv.org/pdf/1411.4038.pdf>
11. <https://arxiv.org/pdf/1506.04579.pdf>
12. <https://arxiv.org/pdf/1412.6856.pdf>
13. <https://arxiv.org/pdf/1612.03144.pdf>
14. <https://arxiv.org/pdf/1612.01105.pdf>
15. <https://openaccess.thecvf.com/content_CVPR_2019/papers/Kirillov_Panoptic_Segmentation_CVPR_2019_paper.pdf>
16. <https://cseweb.ucsd.edu/~elkan/250Bwinter2012/loglinearCRFs.pdf>
17. <https://towardsdatascience.com/implementing-a-linear-chain-conditional-random-field-crf-in-pytorch-16b0b9c4b4ea>
18. <https://homepages.inf.ed.ac.uk/csutton/publications/crftut-fnt.pdf>
19. <https://bjlkeng.github.io/posts/variational-bayes-and-the-mean-field-approximation/>
20. <https://arxiv.org/pdf/1511.02680.pdf>
21. https://arxiv.org/pdf/1506.02142.pdf