Semantic segmentation

Semantic segmentation is one of the key problems in the field of computer vision, as inferring knowledge form images plays a very important role in medical analysis, object detection in satellite images, iris recognition, autonomous vehicles, and many more tasks. It is probably amongst the best approaches towards complete scene understanding.

Image segmentation means assigning a semantic annotation label to each pixel in the image. Therefore, image segmentation is also categorized as a dense prediction task. In itself it does not distinguish between object instances, only object categories.

After some research in the field and considering other architectures (like FCNs and U-net) as well, we chose SegNet [paper ref] as our main inspiration for our project.

Segnet:

The SegNet architecture also follows the encoder-decoder pattern, as most semantic segmentation architectures. The encoder of the network contains a fully convolutional VGG16, while the decoder has repeated upsampling layers followed by convolutional layers. The decoder network is the inverse of the encoder: it has the same type of convolutions as the encoder network, regarding filter sizes, and the channels of corresponding layers.

In SegNet upsampling does not take part in the learning process. It is a sort of backward max-pooling operation.

Residual connections:

During the forward pass in the encoder, while downsampling, the max-pooling indices are stored – meaning the location of the highest value pixel In the pax-pooling window at each sliding position of the layer. These are then used when the corresponding upsampling layer is working in the decoder. The remaining pixels in the upsampled output are set to zero.

“Hence, the decoder network of SegNet consists of a hierarchy of decoders, one corresponding to each encoder and the appropriate decoder uses the max-pooling indices from the corresponding encoder to perform non-linear upsampling of their input feature maps.

One thing to note is that the decoder corresponding to the first encoder (closest to the input image) produces a multi-channel feature map, although its encoder input has 3 channels (RGB). This is unlike the other decoders in the network which produce feature maps with the same number channels and size as their encoder inputs. This is done for generating the segmentation masks for each class plus background.”

https://mohitjain-me.cdn.ampproject.org/c/s/mohitjain.me/2018/09/30/a-look-at-image-segmentation/amp/