SUPPLEMENTARY MATERIALS

SUPPLEMENTARY MATERIAL 1

U-Net

The U-Net is one of the simplest segmentation models that are widely used for biomedical image segmentation. It is widely used when the dataset size is relatively small.

The network architecture consists of the contracting path as the encoder and the expansion path as the decoder. The contracting path consists of repeatedly applying two $3 \times 3$ unpadded convolutions, each followed by a ReLU activation function and a $2 \times 2$ max pooling operation with stride 2 for downsampling. At every downsampling step, the feature channels are doubled. Every step in the expansion path consists of an upsampling of the feature map followed by a $2 \times 2$ convolution that halves the number of feature channels and then concatenating with the corresponding cropped feature map from the encoder path, and followed by two $3 \times 3$ convolutions, each followed by a ReLU. The cropping is performed due to the lost pixels at every convolution. Finally, at the last layer, a $1 \times 1$ convolution is used to map each 64-component feature vector to the required number of classes.

It is suggested that data augmentation should be performed only when the dataset size is relatively small. The data augmentation includes different image transformation steps such as shift, rotation, and other elastic deformations.

MASK R-CNN

The Mask R-CNN is an extension to the popular object detection algorithm Faster R-CNN by adding an extra mask head. Mask R-CNN adds a key element, the pixel-to-pixel alignment that is not present in the Fast R-CNN. Similar to the Fast R-CNN, Mask R-CNN framework consists of two stages. The first stage generates the object proposals, while the second stage performs the classification of proposals to generate the bounding boxes and masks.

Mask R-CNN uses feature pyramid network (FPN) as a backbone network. The Faster R-CNN with the FPN backbone extracts region of interest features from different levels of the feature pyramid. It is shown that using a ResNet-FPN backbone for feature extraction in Mask R-CNN produces gains in both speed and accuracy.

The feature maps generated are passed through a $3 \times 3$ convolution layer. The outputs are then further passed to two branches, one that calculates the box scores and another to compute the bounding box regressors. The region proposals in an image are done by selective search and then use a CNN to extract a 2048-feature vector. This vector is then passed on to the linear SVM classifier. The generation of proposals is itself a complex algorithm and is not within the scope of this article.

PSPNet

PSPNet exploits the ability to capture global context information by different region-based context aggregation through the pyramid pooling module together with the PSPNet. The global representation is effective to produce good quality results on the scene parsing task, while PSPNet provides a framework for pixel-level prediction.

The pyramid pooling module consists of four subregions’ average pooling. The first level performs global average pooling over each feature map to generate a single bin output. The second level module divides the feature map into $2 \times 2$ subregions, which then performs average pooling for each subregion. The third level divides the feature map into $3 \times 3$ subregions and then performs average pooling for each subregion while the fourth module splits the feature map into $6 \times 6$ subregions and then performs pooling for each subregion. In this pyramid pooling module, bilinear interpolation is used for upsampling while concatenation is used for context aggregation. Finally, a convolution layer is used to generate the final segmentation map.
YOU ONLY LOOK AT COEFFICIENTS

You Only Look at Coefficients (YOLACT) is an instance segmentation model that achieved state-of-the-art results in real-time instance segmentation. The model is divided into two subtasks that can be run parallelly. The two tasks are: generating prototype masks and predicting per-instance mask coefficients. Finally, the instance masks are produced by a linear combination of the prototypes with mask coefficients. YOLACT uses Residual Networks as its backbone. The three stages are:

a. Prototype generation: This stage predicts a set of k prototype masks for an image. Protonet is implemented as an FCN by having k channels in the last layer and is attached to a backbone feature layer

b. Mask coefficients: YOLACT simply adds a third extra head to the object detection branch to predict a vector of “mask coefficients” for each anchor that encodes an instance’s representation in the prototype

c. Mask assembly: For each instance that remains after the NMS, a mask for that instance is built by linear combination of the work done by the above two stages. A sigmoid is applied for non-linearity.

YOLACT uses three loss functions for training. The classification loss and the box regression loss are calculated in the same manner as other object detection algorithms. In order to compute the mask loss, the pixel-wise binary cross entropy between assembled masks and the ground truth masks is calculated.

DEEPLABV3

The core idea behind this segmentation network is the atrous convolution that is able to adjust the filter’s field of view and also control the feature responses in CNNs. The atrous convolutions run in cascade or in parallel to capture multi-scale context by implementing multiple atrous rates. Multi-grid methods are employed in a hierarchy of grids of different sizes and various atrous rates are used in the proposed model.

The Atrous Spatial Pyramid Pooling (ASPP) includes a batch normalization layer that is taken from the Inception-v2 network. Here, four parallel atrous convolutions with different rates are applied on top of the feature maps. This effectively captures multi-scale information. Therefore, the ASPP is able to robustly segment objects with filters at multiple sampling rates and field of views, augmented with image level features to capture context information.