Talk Track

1. The first component consists of a convolutional network. They use the VGG-16 architecture which consists of 13 layers of 3x3 convolutions along with 5 2x2 max pooling layers. They remove the last pooling layer, so an input image of shape 3WH gives a tensor of features of shape CW'H' where W' = W/16, H' = H/16 and C = 512.

2. This tensor of feature becomes the input to the next component which is the localization layer. The first part of this layer is based on Faster-RCNN. “k” anchor boxes of different aspect ratio and scale are considered in the image plane for every point in the W'H' grid of features. For each of these anchor boxes, the layer predicts offset and a confidence score. For the bounding box regression they use a parameterization similar to that of Fast-RCNN.

3. Now, for an image with dimensions W=720 and H=540 and k=12, we will have around 17K region proposals and running the recognition network and the language model on all of them will be expensive. So during training, they sample a minibatch of B=256 proposals out of which atmost B/2 are positive and the rest are negative. A proposal is postive if IoU>0.7 for some GT box and negative if IoU<0.3 with every GT box. Also, the region with maximum IoU with every GT box is a positive. During testing, they sample B=300 proposals using greedy NMS.

4.The region proposals can be of different sizes and fixed-size ft vecs must be obtained for them. If the RoI pooling layer is used, gradients can be BPed from output fvs to input fvs but not to the input proposal coordinates. So they replace it with bilinear interpolation and make the sampling grid a linear function of the proposal coordinates. Now the gradients can be Bped to the proposal coordinates.

5. These region features from the LL are then fed to a recognition network which is a FCNN and output a code for every region with dimesion D=4096 and hence compactly encoding the visual appearance. Also, one more chance is given to this RN to refine the confidence and position of each proposal region. It outputs a final confidence score and final offsets to be applied for every proposal .

6. These region codes are then used to condition a language model. For training, The word vectors appended with a START token are fed to the LSTM network. The region code is also fed at the 1st timestep. Output vectors of d=|V|+1 are obtained where V is the vocabulary and the END token. The loss function for the LM is the average cross entropy and for box regression is the smooth L1 loss. During testing, the region code is fed at the first time step, and then at every timestep the most likely next token is sampled and fed to the network in the next step, repeating till the END token is obtained.