AI6126 Homework2

Name: *BAIWEN* Matric number: *G1903363K*

Question 1:

Explain why is scaling and shifting often applied after batch normalization?

Answer1:

Normalizing the mean and standard deviation of a unit can reduce the expressive power of the neural network containing that unit. Replacing the batch of hidden unit activations H with $\gamma H + \beta$ can maintain the expressive power of the network. The newly introduced parametrization can represent the same family of functions of the input as the old parametrization. Besides, in the new parameters, the mean of $\gamma H' + \beta$ determined solely by β and will not impacted by the complicated interaction between the parameters in the layers below H

Question 2:

Describe the main changes, e.g., architecture change or new losses, made in Faster R-CNN in comparison to Fast R-CNN

Answer2：

Architecture: Faster R-CNN introduce Region Proposal Networks (RPN), which is a deep convolutional neural network to computing the proposals. RPN will share the feature map generated by the common set of convolutional layers with the object detection network. To generate region proposals, it sides network over the convolutional feature map output by the last shared convolutional layer. This small network takes as input an $n \times n$ spatial window of the input convolutional feature map. Each sliding window is mapped to a lower-dimensional feature. This feature is fed into two sibling fully connected layers—a box-regression layer (reg) and a box-classification layer (cls). At each sliding-window location, it simultaneously predict multiple region proposals, where the number of maximum possible proposals for each location is denoted as k. So the reg layer has $4k$ outputs encoding the coordinates of k boxes, and the cls layer outputs $2k$ scores that estimate probability of object or not object for each proposal. **Loss:** Each anchor will be assigned with a binary class label (of being an object or not). The anchor/anchors with the highest Intersection-overUnion (IoU) overlap with a ground-truth box, or an anchor that has an IoU overlap higher than 0.7 with any ground-truth box will be assigned the positive labels. if anchor's IoU ratio is lower than 0.3 for all ground-truth boxes, it will be assigned with negative label. Anchors that are neither positive nor negative do not contribute to the training objective. Thus the training is to minimize the objective function following the multi-task loss as follows: $L(p_i, t_i) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(p_i, p_i^*)$

Question 3:

Describe the differences between the operation of RoIPool and RoIAlign. Explain why RoIAlign is preferred over RoIPool

Answer 3:

RoIPool It first quantizes a floating-number RoI to the discrete granularity of the feature map, this quantized RoI is then subdivided into spatial bins which are themselves quantized, and finally feature values covered by each bin are aggregated (usually by max pooling). Quantization is performed on RoI boundaries or bins division. **RoIAlign** It use bilinear interpolation to compute the exact values of the input features at four regularly sampled locations in each RoI bin, and aggregate the result. **Prefer Reason** The quantizations of RoIPool introduce misalignments between the RoI and the extracted features. While this may not impact classification, which is robust to small translations, it has a large negative effect on predicting pixel-accurate masks. RoIAlign are not sensitive to the exact sampling locations, or how many points are sampled, as long as no quantization is performed.

Question 4:

1. Explain why the encoder-decoder architecture is widely used in semantic segmentation tasks.
2. Does the plain encoder-decoder architecture have potential drawbacks? If so, how can we fix them?

Answer 4:

3. Auto-encoder can learn a compressed, distributed representation (encoding) for a set of input data.
4.
    - Train an autoencoder needs a lot of data, processing time, hyperparameter tuning, and model validation before building the real model.
    - An autoencoder learns to capture as much information as possible rather than as much relevant information as possible. If the information most relevant to the problem makes up only a small (in magnitude) part of the input, the autoencoder may miss it.

Question 5:

When we apply consecutive 1-dilated, 2-dilated, 4-dilated and 8-dilated 3x3 convolution, what is the final receptive field?

Answer 5:

- 1-dilated: 3x3 receptive field
- 2-dilated: 7x7 receptive field
- 4-dilated: 15x15 receptive field
- 8-dilated: 31x31 receptive field

Question 6:

1. Even though dilated convolution improves upon standard convolution, what are the potential hard cases for dilated convolution?
2. How will you further improve upon dilated convolution?

Answer 6:

3. Gridding Problem
    - Local information is completely missing when use large dilate
    - Long-ranged information might be not relevant
    - The receive information in a nearby rxr regions may impair the consistency of local information
4.
    - Make sure dilation rate within a group should not have a common factor relationship (like 2,4,8, etc.), otherwise the gridding problem will still hold
    - Assign dilation rate follows a sawtooth wave-like heuristic for each layer

Reference

1. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks
2. 一文读懂Faster RCNN
3. 一文读懂目标检测：R-CNN、Fast R-CNN、Faster R-CNN、YOLO、SSD
4. Mask R-CNN
5. 如何理解空洞卷积
6. Understanding Convolution for Semantic Segmentation
7. Rethinking Atrous Convolution for Semantic Image Segmentation