Question1: Explain why the size of the input image should be changed from 224x224 (as it is given in Figure 2 of the AlexNet paper) to 227x227.?

If the input image size is 224x224, the size of output image is 54.25x54.25 in the first convolution layer. In this case, the output image cannot be represented by a tensor, so the size of image must be integer. However, if we choose 227x227 as the input image size, the output image size is integer and, in the rest, layers the size of output images is also integer.

Question 2: Explain in less than 15 lines what the first proposed improvement was, which was about the activation function used. How did it compare to the AlexNet approach and what was its justification? (20 points)

Answer2:

Using PReLU instead if ReLU. Conducting comparisons on a deep but efficient model with 14 weight layers and using the ImageNet dataset.They trained this model with ReLU applied in the convolutional layers and the first two fully-connected layers. Then they trained the same architecture from scratch, with all ReLUs replaced by PReLUs. The top-1 error of PReLUs is lower than that of ReLUs(1.2% gain).

For fair comparisons, both ReLU/PReLU models are trained using the same total number of epochs, and the learning rates are also switched after running the same number of epochs. In addition, they used the same model A to train with three different single scale and multi-scale. Combing with the comparisons above, they concluded that PReLU is better than ReLU and PReLU improves both small and large models. This improvement is obtained with almost no computational cost.

For PReLU, it just replaces the parameter-free ReLU activation by a learned parametric activation. Actually, ReLU is just a special case for PReLU.

Question 3: Explain in less than 25 lines what this second improvement is, how different it is from the approaches of AlexNet and of Glorot and Bengio, and what is its justification. (20 points)

Answer3:

Deriving a theoretically sound initialization method by taking ReLU/PReLU into account, which helps with convergence of very deep models (*e.g*., with 30 weight layers) trained directly from scratch. While Glorot and Bengio proposed “Xavier” initialization method, which is based on the assumption that the activations are linear, is invalid for ReLU/PReLU, and it is difficult to converge for extremely deep models.

For the initialization in ReLU case,the improvement is to set a zero-mean Gaussian distribution whose standard deviation is . For PReLU case, its standard deviation is . As for “Xavier” case, its weights in each layer follow the uniform distribution in the interval .

Question 4: Explain in less than 10 lines how He and al quantify the performance of their two above improvements on ImageNet as compared to the Alexnet approach. (20 points)

Answer 4:

1. Comparisons between ReLU and PReLU : For the multi-scale combination, PReLU reduces the top- 1 error by 1.05% and the top-5 error by 0.23% compared with ReLU. The results in Table 2 and Table 4 consistently show that PReLU improves both small and large models. This improvement is obtained with almost no computational cost.

2. Comparisons of Single-model Results: For comparing single-model results. They first show 10-view testing results which show that model (A+PReLU) is better than model(A+ReLU). Then for comparisons single-model results, which are all obtained using multi-scale and multi-view test, we can also discover that PReLU is better than ReLU.

3. Comparisons of Multi-model Results: their result (top-5 error 4.94%) is 1.7% better than the ILSVRC 2014 winner(6.66%), and also a ~17% relative improvement over Baidu(5.98%)

4. Comparisons with Human Performance from Russakovsky’s test: their result (4.94%) exceeds the reported human-level performance(5.1% reported in Russakovsky’s paper )