1. First, 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.

Not integer

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

The first improvement is to use a Parametric Rectiﬁed Linear Unit (PReLU) that generalizes the traditional rectiﬁed unit instead of ReLu. PReLU improves model fitting with nearly zero extra computational cost and little overfitting risk. For PReLU, the coefficient of the negative part is not constant and is adaptively learned.

They conducted comparisons on a small model E with 14 weight layers on ImageNet 2012, using 10-view testing with same size image and epoch. Firstly, as a baseline, they train this model with ReLU that is AlexNet approach applied in the convolutional (conv) layers and the first two fully connected (fc) layers. Then, they train the same architecture from scratch, with all ReLUs replaced by PReLUs.

Moreover, they conducted comparisons on a large model A in ImageNet 2012 using dense testing with same epoch and way of switching learn rates. They tested with ReLu and PReLu seperately.

Finally, compared results of top-1 and top-5 error rates, 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.

(14)

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

The second improvement is to use a robust initialization method that particularly considers the rectiﬁer nonlinearities. This method enables train extremely deep rectiﬁed models directly from scratch and to investigate deeper or wider network architectures. This gives more flexibility to explore more powerful network architectures.

The AlexNet approach initialize random weights drawn from Gaussian distributions with fixed standard deviations (0.01). The Glorot and Bengio approach is to adopt a properly scaled uniform distribution for weight initialization in the interval. This is called “Xavier” initialization. Its derivation is based on the assumption that the activations are linear.

In He et al paper, no linear activations are used, so they used new initialization methods for ReLu and PReLu separately. For ReLU, standard deviation is to lead to a zero-mean Gaussian distribution. Similarly, for PReLu, standard deviation is .

He et al take model B in the VGG team’s paper to explain why the constant standard deviation of 0.01 makes some deeper networks stall. Then, compared with “Xavier” initialization, they used a 22-layer large model and a 33-layer small model using ReLu as the activation for both initialization methods. “Xavier” initialization sets standard deviation and He et al initializations sets s standard deviation. Figure 2 comparing the convergence of a 22-layer large model shows that both methods are able to make them converge and He’s method starts reducing error earlier. Besides, the “Xavier” initialization method leads to 33.90/13.44 top1/top-5 error, and He’s method leads to 33.82/13.34 for the model in Table 2. Figure 3 comparing the convergence of a 30-layer small model shows that He’s method is able to make the extremely deep model converge. On the contrary, the “Xavier” method completely stalls the learning, and the gradients are diminishing as monitored in the experiments.

(22)

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)

1. Comparisons between ReLU and PReLU: 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. The results in Table 2 and Table 4 consistently show that PReLU improves both small and large models.
2. Comparisons of Single-model Results: Comparing with models of A, ReLU, A, PReLU, B, PReLU and C, PReLU, we know that PReLU improves results and increasing the width can still improve accuracy.
3. Comparisons of Multi-model Results: Their result is 4.94% top-5 error on the ImageNet 2012 test set, and it is compared with other five models, including MSRA, VGG, GoogLeNet, VGG (arXiv v5), and Baidu. Also, we can know that their MSRA PReLU-nets has the best results.
4. Comparisons with Human Performance from Russakovsky et al: Their result (4.94%) exceeds the reported human-level performance in Russakovsky et al (5.1%) top-5 error on the ImageNet dataset.