AlexNet

Basics：

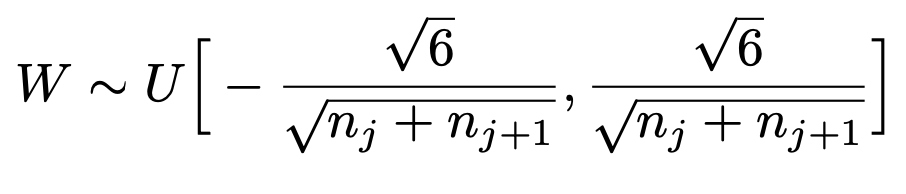
* ReLU: purpose is accelerating training (compared to traditional activation function, tanh and sigmoid
* Initialization:purpose is accelerating the early stages of learning by providing the ReLUs with positive inputs
  + weights : zero-mean Gaussian distribution with standard de- viation 0.01 (each layer)
  + biases : second, fourth, and fifth convolutional layers, as well as in the fully-connected hidden layers, with the constant 1 . Remaining layers with the constant 0

PReLU：

Comparison：We conducted comparisons on a deep but efficient model with 14 weight layers based on known paper (which was implemented successfully before). By comparing the top-1 and top-5 error rate, it can basically ensure the PReLU is better than ReLU (1.1% gain).

Justification：Based on 1000-class ImageNet 2012 dataset. He set the same total number of epochs, and the learning rates are also switched after running the same number of epochs. After analyze the table of error rate, it can justify that PReLU improves both small and large models. This improvement is obtained with almost no computational cost.

Initialization：

2010 paper：normalized initialization, uniform distribution

Delving into DL: zero-mean Gaussian distribution whose standard deviation is

Answer

Q1.

To make all the size of output image to be integers.

Q2:

The first improvement is applying PReLU as activation function rather than ReLU.

The PReLU sets the coefficient that control negative derivatives as learnable parameter.

Meanwhile, it introduces a very small number of extra parameters, which is negligible when considering total number of weights. But it can improve the performance of neural network.

Comparison:

On the one hand, he conducted comparisons on a deep but efficient model with 14 weight layers based on known paper (which was implemented successfully before). By comparing the top-1 and top-5 error rate, it can basically ensure the PReLU is better than ReLU (1.1% gain).

On the other hand, based on 1000-class ImageNet 2012 dataset. He set the same total number of epochs, and the learning rates are also switched after running the same number of epochs. After analyze the table of error rate, it can justify that PReLU improves both small and large models. This improvement is obtained with almost no computational cost.

Q3:

The improvement is setting the initialized weight by zero-mean Gaussian distribution whose standard deviation is for ReLU activation function. As for PReLU activation function, the standard deviation is . The improvement is aimed to equip the networks a robust initialization method which can remove the obstacle of training extremely deep recitifier networks.

As for Glorot and Bengio, they applied the uniform distribution with range of , which is called “Xavier” method. But its derivation is based on the assumption that the activations are linear which is invalid for ReLU and PReLU.

As for Alex paper, they applied zero-mean Gaussian distribution with standard deviation 0.01 in each layer. And the biases in second, fourth, and fifth convolutional layers, as well as in the fully-connected hidden layers, being the constant 1. Remaining layers being the constant 0. But it has difficult to converge when neural networks are extremely deep.

Comparison:

*In He et al* paper, they compared the improved initialization method with that of “Xavier” on extremely deep models with up to 30 layers. Their initialization 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. But both methods have the similar accurate of the model.

Q4:

There are four comparisons based on different bases

1. Comparisons between ReLU and PReLU

They applied the model A in paper and set same setting for two experiments. 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.

1. Comparisons of Single-model Results

Their base- line model (A+ReLU, 6.51%) is already substantially better than the best existing single-model result of 7.1% reported for VGG-19 in the latest update from paper by Simonyan and A. Zisserman

1. Comparisons of Multi-model Results

Their result is 4.94% top-5 error on the test set and was 1.7% better than the ILSVRC 2014 winner (GoogLeNet, 6.66%), which represents a ∼26% relative improvement.

1. Comparisons with Human Performance

Their result (4.94%) exceeds the reported human-level performance which is 5.1% top-5 error on the ImageNet dataset from Russakovsky *et al*.

As for Comparisons between ReLU and PReLU, they compare ReLU and PReLU on the large model A with same setting. And they found that PReLU improves both small and large models.

As for Comparisons of Single-model Results, ther compared their five models (model A with ReLU, model with PReLU, and eta.) with VGG-16 and GoogLeNet. And their baseline model outperforms VGG-19.

As for Comparisons of Multi-model Results, they combined six models and this model shows 1.7% better than ILSVRC 2014 winner.

As for Comparisons with Human Performance, their result(4.94%) exceeds the reported hunman-level performance from Russakovsky based on ImageNet dataset.