ACSE-8 Coursework 1

Question 1.

Because in our simplified version of Alexnet, we don't split the data into two GPUs, means that we need the input size n to satisfy:

$$\frac{n+2p-f}{s}+1=55, \text{ where } p=0, f=11, s=4,$$

55 is the first layer of Alex net (on one GPU), this gives us $n=227$.

For the blanks in the spreadsheet, please see the Excel file in the ~~dk~~ repository.

Question2.

As the author argued in the Introduction, he claims that CNNs have much fewer connections and parameters, so they are easier to train comparing to standard feed forward neural networks. As we know, the parameters in a standard feed forward neural network (number of) strongly depend on the size of the input data, while CNNs not.

The main remaining obstacle to using even larger CNNs is that they are prohibitively expensive (in time) and ~~also~~ the limitation of the amount of memory available on current GPUs.

Question 3

According to Section 3.1, as demonstrated in the plot, the deep convolutional neural networks work with ReLUs train several times faster than equivalents with tanh units (using less epochs to achieve same error rate.) To conclude, training

ReLU (non-saturating non linearity) does much better job than tanh(x) or sigmoid (saturating nonlinearities) in the sense of fast learning, which is very important for training large datasets.

Question 4.

The last layer is a fully connected convolutional neuron network with output size is 1000X1, 1 channel. It just convert the input to a length 1000

vector, with an appropriate activation function, represents the probability of the input for each category (because the Original ILSVRC dataset has 1000 categories). ~~Dfc~~ Thus we could make a prediction for a certain input as the category corresponding which is the largest in this length 1000 vector.

Question 5.

The loss function should be a cross-entropy, V-net

$$J(w) = \sum_{X \in Image} w(X) \log p_k(X).$$

where X is a pixel in the image and $w(X)$ is the weighting function. also $p_k(X)$ is a probability provided by Softmax function, and l is the true category where X is (our y).

Question 6.

Because the ILSVRC data set has 1000 categories, of which is just the labels for our supervised learning (similar to constraints in solving PDEs or functional extrema), and also $10\ bit = 2^{10}\ bytes = 1024 \approx 1000$, so after compressing, the constraints has size of 10 bits.

Question 7

We initialize the weights in each layer to form a Gaussian distribution with mean of 0 and standard derivation of 0.01. Also we initialize the neuron bias in the 2nd, 4th and 5th convolutional layers with constant 1, initialize the neuron bias in the remaining layers with constant 0.

Question 8.

Not quite, by our formula,

Output size $= \frac{n+2p-f}{s} + 1$, after extracting 224×224 from 256×256, we have output size $= \frac{256+0-224}{1} + 1 = 33$, so the pre-filter has a size of 33×33, 2 channels (because adding horizontal reflections), then we have the total factor is $33^2 \times 2 = 2178 \neq 2048$.

Question 9:

From Section 5, the author stated "We trained the network for roughly 90 cycles through the training set of 1.2 million images." So the epochs used is just 90 and the batch size ~~$\frac{1.2 \times 10^6}{90} \approx 13333$~~ 128 as he stated in the beginning of Section 5.

Question 10.

The top-1 and top-5 errors on the dataset are 67.4% and 40.9% respectively attained by the net in Fall 2009 version of Imagenet with 10,184 categories and 8.9 million images, but with an additional, sixth convolutional layer over the last pooling layer.