ARNAB BARUA

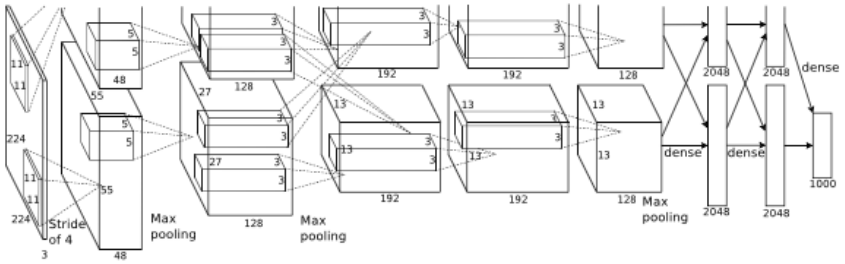
I am a learner of Machine Learning and Deep Learning. Always hungry to keep learning.

Break Down AlexNet

2019, OCT 29

Introduction:

AlexNet is one of the famous architecture of Convolutional Neural Network (CNN), and it won the ILSVRC-2012 competition. Alex Krizhevsky, along with Ilya Sutskever and Geoffrey E. Hinton, proposed AlexNet, and they used the first author name “Alex” to name it.



Prerequisite:

You have to understand how convolution, pooling and fully connected layer work. Otherwise, you will not understand. You also need to know how to calculate shape. [Click to Know How to Calculate Shape]

AlexNet has 60 million parameters and 650,000 neurons.

It has five convolutional layers, three max-pooling layers and three fully-connected layers. They use ReLu instead of the Tanh activation function to add non-linearity, and it helps to increase speed by six times. They use dropout instead of regularization to deal with overfitting.

In the training time, AlexNet divided into two parts and ran parallelly on two GTX 580 3GB GPUs and connect the output of two parts at the output layer. It takes five to six days to train. To train the model, they used a subset of the ImageNet dataset with roughly 1000 images in each of the 1000 categories. In all, there are roughly 1.2 million training images, 50,000 validation images and 150000 testing images.

The specification of the AlexNet presented below layer-wise. Input Layer: The input shape of the image is 224x224.

Layer 1: Ninety-Six (96) 55x55 feature maps in the first convolutional layer by using ninety-six (96) kernels of size 11x11 with stride of 4 pixels.

Layer 2: Ninety-Six (96) 27x27 feature maps in the first max-pooling layer by using ninety-six (96) kernels of size 3x3 with stride of 2 pixels.

Layer 3: Two hundred fifty-six (256) 27x27 feature maps in the second convolutional layer by using two hundred fifty-six (256) kernels of size 5x5 with stride of 1 pixel.

Layer 4: Two hundred fifty-six (256) 13x13 feature maps in the second max-pooling layer by using two hundred fifty-six (256) kernels of size 3x3 with stride of 2 pixels.

Layer 5: Three hundred eighty-four (384) 13x13 feature maps in the third convolutional layer by using three hundred eighty-four (384) kernels of size 3x3 with stride

of 1 pixel.

Layer 6: Three hundred eighty-four (384) 13x13 feature maps in the fourth convolutional layer by using three hundred eighty-four (384) kernels of size 3x3 with stride of 1 pixel.

Layer 7: Two hundred fifty-six (256) 13x13 feature maps in the fifth convolutional layer by using two hundred fifty-six (256) kernels of size 3x3 with stride of 1 pixel.

Layer 8: Two hundred fifty-six (256) 6x6 feature maps in the third max-pooling layer by using two hundred fifty-six (256) kernels of size 3x3 with stride of 2 pixels.

Layer 9: First fully connected layer of 4096 neurons.

Layer 10: Second fully connected layer of 4096 neurons.

Output Layer: Output layers of 1000 neurons.

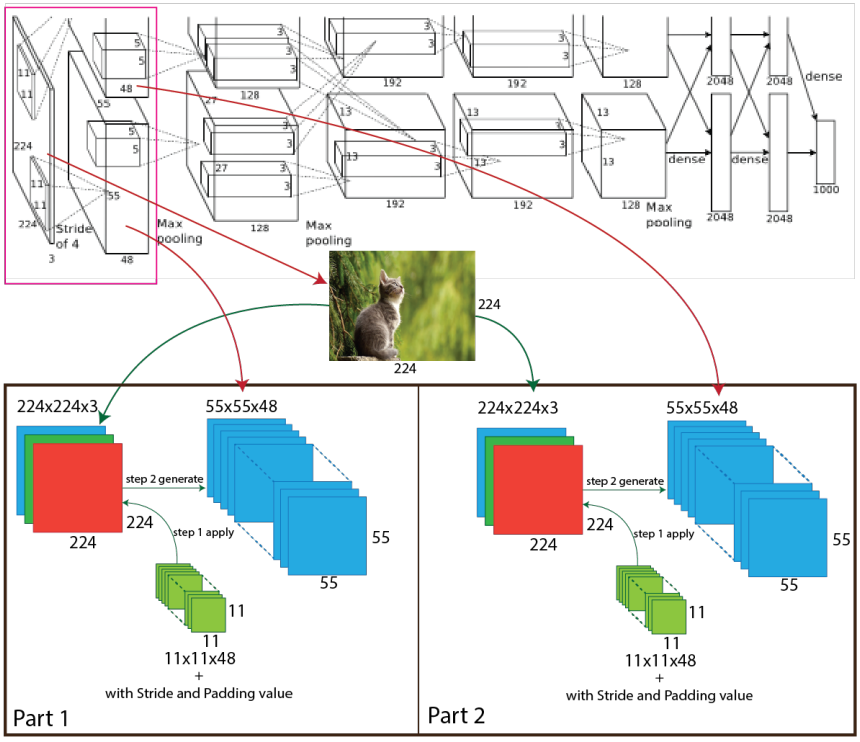
The units of the output layer depend on the number of classes. Suppose one thousand classes in the dataset, then number units in the output layer must be 1000.

Now, I am going to explain how to generate outputs in each layer.

Layer 1:

At the beginning of the AlexNet architecture, the RGB input image goes through 96 filters having size 11x11 with stride value four and padding value is 2, and the output shape change from 224x224x3 to 55x55x96. Originally at the training time, the model divided into two parts. For that reason, 96 filters divided into two parts and the output shape will be 55x55x48 for each part (55x55x96/2 or 55x55x48*2). For better understanding,

a picture of the process presented below.



Now we use an equation to know how the process work, which helps to calculate output shape. The parameters for calculation given below:

1. Input size 224x224x3 (W=224).
2. The total number of kernels or filters 96. For each part number of kernels or filters 48.
3. A single kernel of filter size 11X11 (K=11)
4. Number of padding 2 (P=2)
5. Stride 4 (S=4)

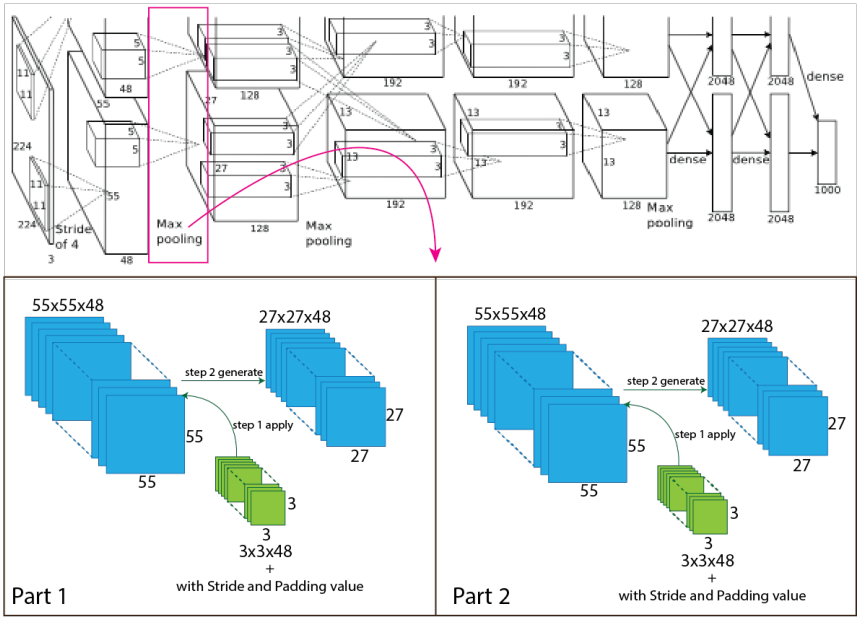
Now insert all the values in the equation.

$$Output = \frac{W - K + 2(P)}{S} + 1 = \frac{224 - 11 + 2(2)}{4} + 1 = 55$$

So, after the calculation, the result is as same as the image, and the output shape of the first convolution layer turns into 55x55x48 for each part and a totally 55x55x96.

Layer 2:

Then in the AlexNet architecture, the output of the first convolution goes through the max-pooling, and the number of the filter is 96 having size 3x3 with stride value 2. After applying the max-pooling output shape change from 55x55x96 to 27x27x96 and for each part 55x55x48 to 27x27x48. For better understanding, a picture of the process presented below.



Now we use an equation to know how the process work, which helps to calculate output shape. The parameters for calculation given below:

1. Input size 55x55x96 (W=55).
2. The total number of kernels or filters 96. For each part number of kernels or filters 48.
3. A single kernel of filter size 3x3 (K=3)
4. Number of padding 0 (P=0)
5. Stride 2 (S=2)

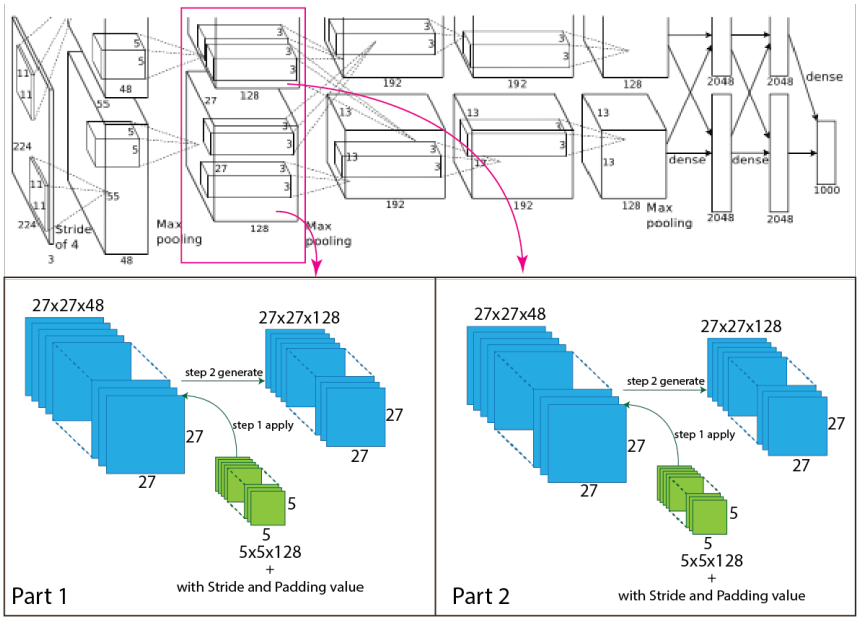
Now insert all the values in the equation.

$$Output = \frac{W - K + 2(P)}{S} + 1 = \frac{55 - 3 + 2(0)}{2} + 1 = 27$$

So, after the calculation, the result is as same as the image, and the output shape of the first max-pooling layer turns into 27x27x48 for each part and totally of 27x27x96.

Layer 3:

After that, in the AlexNet architecture, the output of the first max-pooling layer goes through again convolution, and the number of the filter is 256 having size 5x5 with stride value one and padding value is 2. After applying, the second convolution output will be 27x27x96 to 27x27x256 and for each part output will be 27x27x48 to 27x27x128. For better understanding, a picture of the process presented below.



Now we use an equation to know how the process work, which helps to calculate output shape. The parameters for calculation given below:

1. Input size 27x27x96 (W=27).
2. The total number of kernels or filters 256. For each part number of kernels or filters 128.
3. A single kernel of filter size 5x5 (K=5)
4. Number of padding 2 (P=2)
5. Stride 1 (S=1)

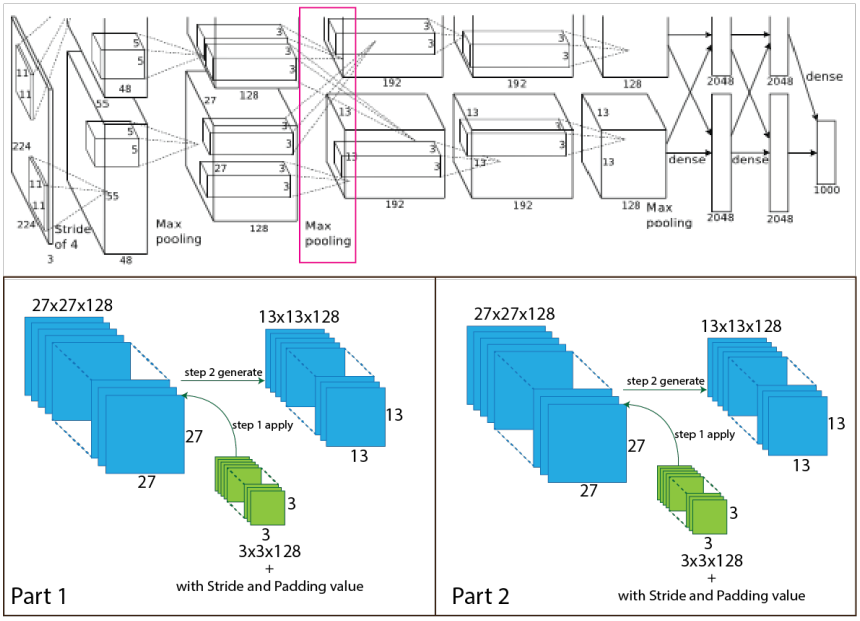
Now insert all the values in the equation.

$$Output = \frac{W - K + 2(P)}{S} + 1 = \frac{27 - 5 + 2(2)}{1} + 1 = 27$$

So, after the calculation, the result is as same as the image, and the output shape of the second convolution layer turns into 27x27x128 for each part and totally of 27x27x256.

Layer 4:

Then in the AlexNet architecture, the output of the second convolution goes through the max-pooling, and the number of the filter is 256 having size 3x3 with stride value 2. After applying the max-pooling output shape change from 27x27x256 to 13x13x256 and for each part 27x27x128 to 13x13x128. For better understanding, a picture of the process presented below.



Now we use an equation to know how the process work, which helps to calculate output shape. The parameters for calculation given below:

1. Input size 27x27x256 (W=27).
2. The total number of kernels or filters 256. For each part number of kernels or filters 128.
3. A single kernel of filter size 3x3 (K=3)
4. Number of padding 0 (P=0)
5. Stride 3 (S=3)

Now insert all the values in the equation.

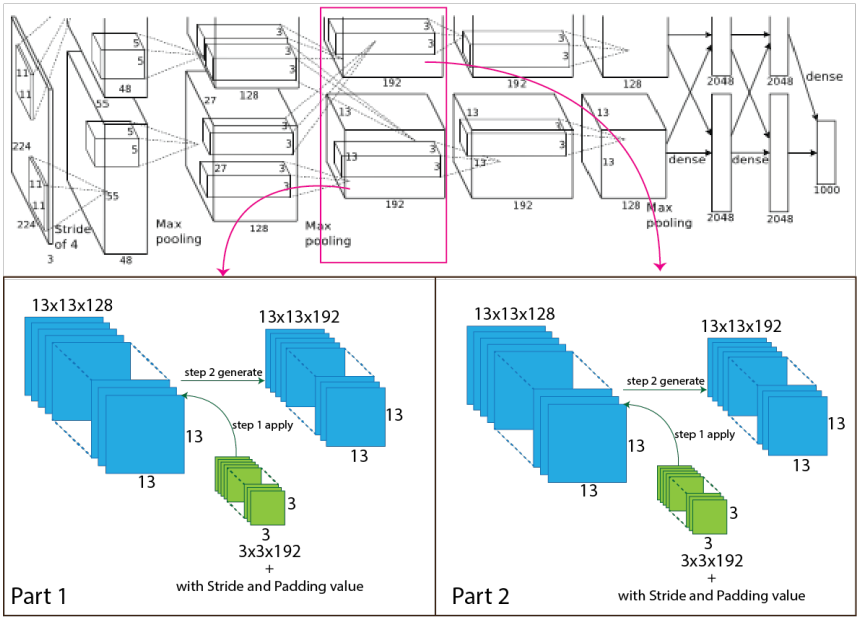
$$Output = \frac{W - K + 2(P)}{S} + 1 = \frac{27 - 3 + 2(0)}{2} + 1 = 23$$

So, after the calculation, the result is as same as the image, and the output shape of the second max-pooling layer turns into 13x13x128 for each part and totally 13x13x256.

Layer 5:

Then in the AlexNet architecture, the output of the

second max-pooling layer goes through the convolution, and the number of the filter is 384 having size 3x3 with stride value one and padding value is 1. After applying the third convolution layer, the output shape change from 13x13x256 to 13x13x384 and for each part 13x13x128 to 13x13x192. For better understanding, a picture of the process presented below.



Now we use an equation to know how the process work, which helps to calculate output shape. The parameters for calculation given below:

1. Input size 13x13x256 (W=13).
2. The total number of kernels or filters 384. For each part number of kernels or filters 192.
3. A single kernel of filter size 3x3 (K=3)
4. Number of padding 1 (P=1)
5. Stride 1 (S=1)

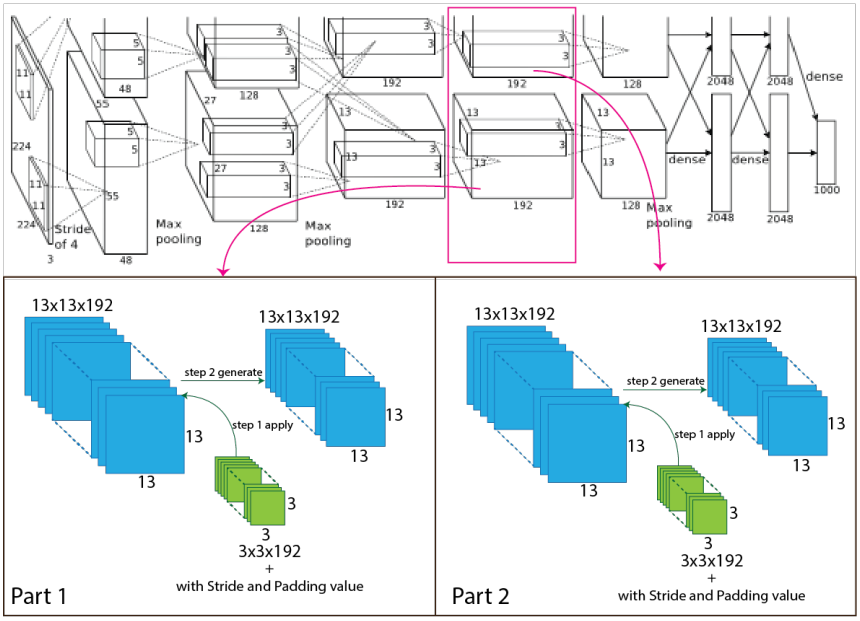
Now insert all the values in the equation.

$$Output = \frac{W - K + 2(P)}{S} + 1 = \frac{13 - 3 + 2(1)}{1} + 1 = 13$$

So, after the calculation, the result is as same as the image, and the output shape of the third convolution layer turns into 13x13x192 for each part and totally 13x13x384.

Layer 6:

Then in the AlexNet architecture, the output of the third convolutional layer goes through the convolution again, and the number of the filter is 384 having size 3x3 with stride value one and padding value is 1. After applying the third convolution layer, the output shape change from 13x13x384 to 13x13x384 and for each part 13x13x192 to 13x13x192. For better understanding, a picture of the process presented below.



Now we use an equation to know how the process work, which helps to calculate output shape. The parameters for calculation given below:

1. Input size 13x13x384 (W=13).
2. The total number of kernels or filters 384. For each part number of kernels or filters 192.
3. A single kernel of filter size 3x3 (K=3)

4. Number of padding 1 (P=1)
5. Stride 1 (S=1)

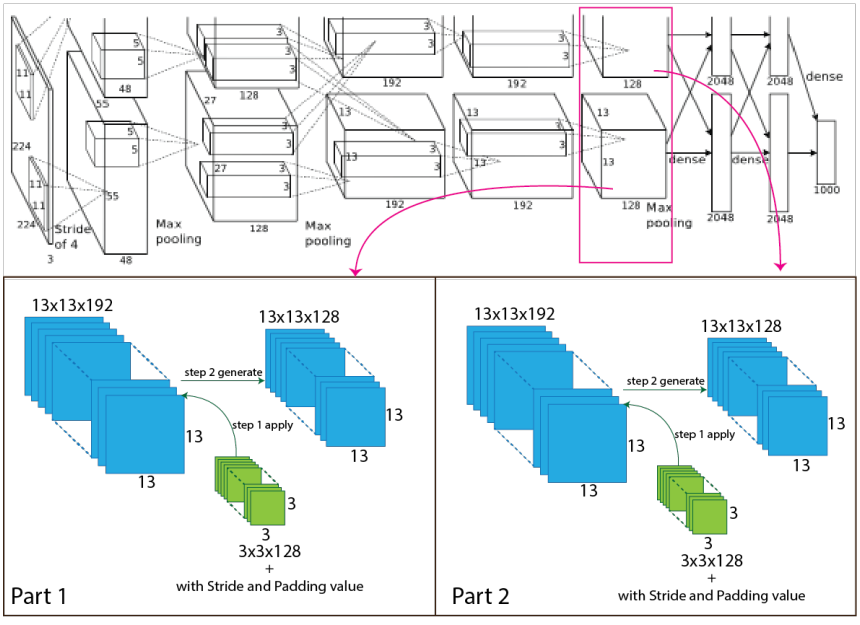
Now insert all the values in the equation.

$$Output = \frac{W - K + 2(P)}{S} + 1 = \frac{13 - 3 + 2(1)}{1} + 1 = 13$$

So, after the calculation, the result is as same as the image, and the output shape of the fourth convolution layer turns into 13x13x192 for each part and totally 13x13x384.

Layer 7:

Then in the AlexNet architecture, the output of the fourth convolutional layer goes through the convolution again, and the number of the filter is 256 having size 3x3 with stride value one and padding value is 1. After applying the third convolution layer, the output shape change from 13x13x384 to 13x13x256 and for each part 13x13x192 to 13x13x128. For better understanding, a picture of the process presented below.



Now we use an equation to know how the process work, which helps to calculate output shape. The parameters for calculation given below:

1. Input size 13x13x384 (W=13).
2. The total number of kernels or filters 256. For each part number of kernels or filters 128.
3. A single kernel of filter size 3x3 (K=3)
4. Number of padding 1 (P=1)
5. Stride 1 (S=1)

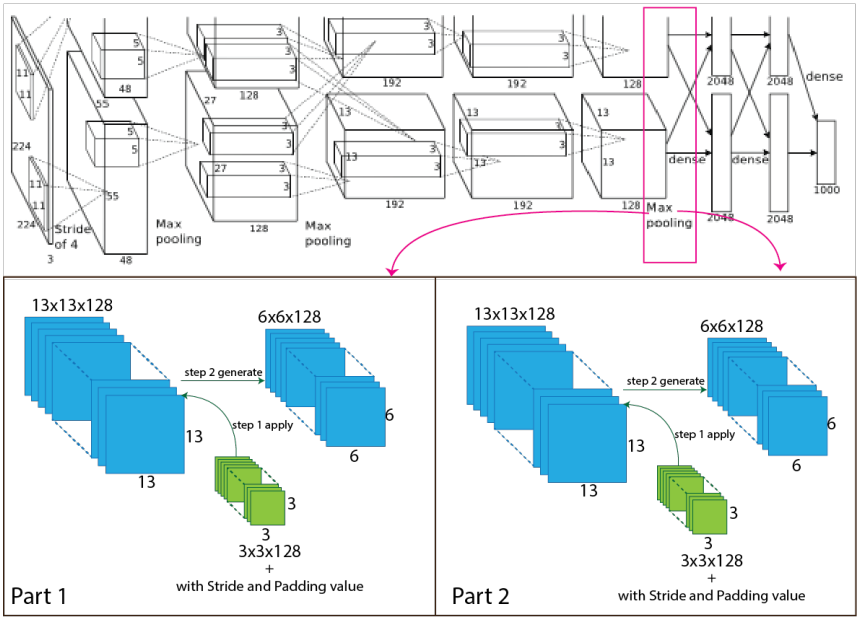
Now insert all the values in the equation.

$$Output = \frac{W - K + 2(P)}{S} + 1 = \frac{13 - 3 + 2(1)}{1} + 1 = 13$$

So, after the calculation, the result is as same as the image, and the output shape of the fifth convolution layer turns into 13x13x128 for each part and totally 13x13x256.

Layer 8:

Then in the AlexNet architecture, the output of the fifth convolutional layer goes through the max-pooling layer again, and the number of the filter is 256 having size 3x3 with stride value two and padding value is 0. After applying the third max-pooling layer, the output shape change from 13x13x256 to 6x6x256 and for each part 13x13x128 to 6x6x128. For better understanding, a picture of the process presented below.



Now we use an equation to know how the process work, which helps to calculate output shape. The parameters for calculation given below:

1. Input size 13x13x256 (W=13).
2. The total number of kernels or filters 256. For each part number of kernels or filters 128.
3. A single kernel of filter size 3x3 (K=3)
4. Number of padding 0 (P=0)
5. Stride 2 (S=2)

Now insert all the values in the equation.

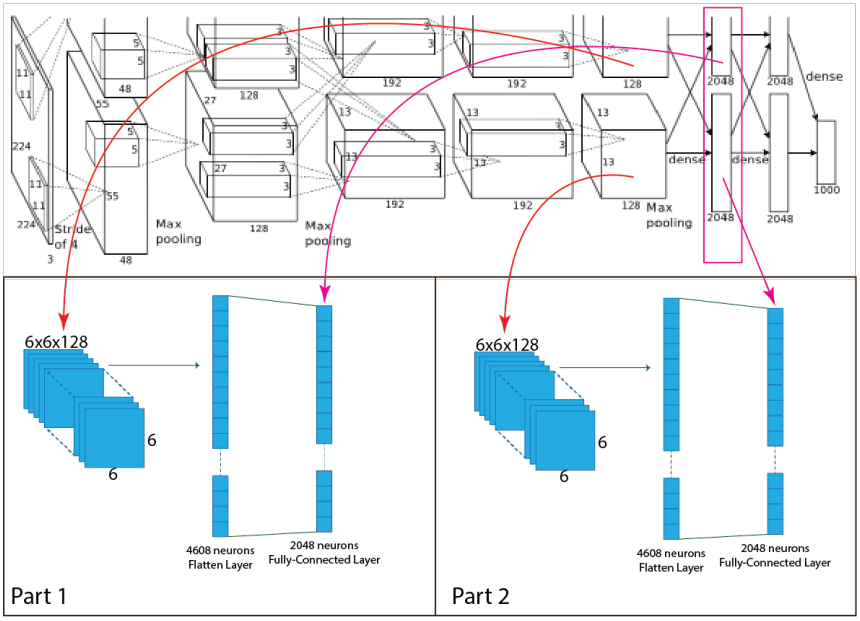
$$Output = \frac{W - K + 2(P)}{S} + 1 = \frac{13 - 3 + 2(1)}{1} + 1 = 13$$

So, after the calculation, the result is as same as the image, and the output shape of the third max-pooling layer turns into 6x6x128 for each part and totally 6x6x256.

Layer 9:

This layer of the AlexNet architecture is a fully connected

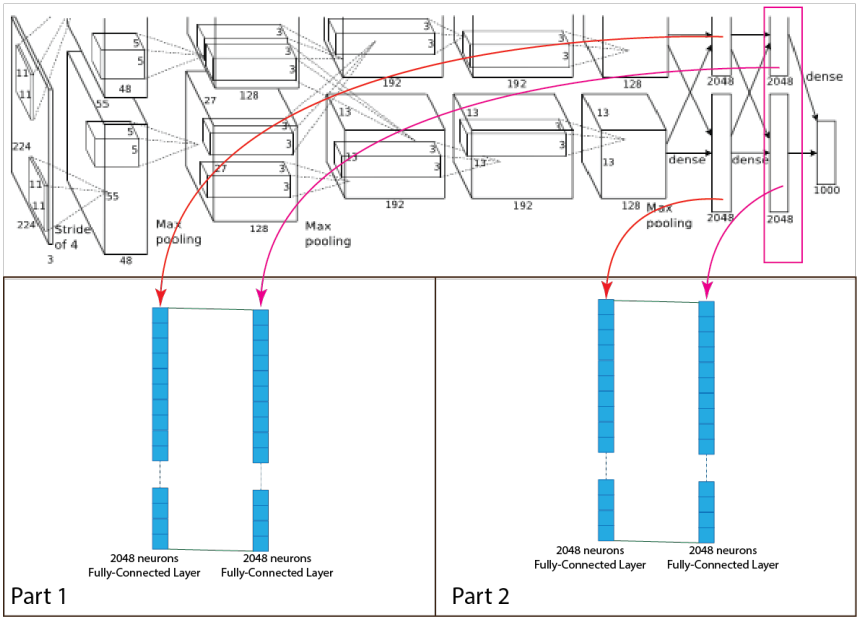
layer with 4096 feature maps and each of size 1x1. For each part, a fully connected layer has 2048 feature maps in a total of 4096 feature maps. After the max-pooling layer, the output of the max-pooling layer turns into a flatten layer, and that flatten layer connects with the fully connected layer. So, the output shape of the third max-pooling layer of each part 6x6x128 (total 6x6x256) turns into a flatten layer, and the shape is 4608x1 for each part and in total 9216x1. All 4608 neurons or units of each part of the flatten layer connected to all 2048 neurons or units of each part of the fully connected layer. In the fully-connected layer, the ReLU function used as an activation function. For better understanding, a picture of the process presented below.



Layer 10:

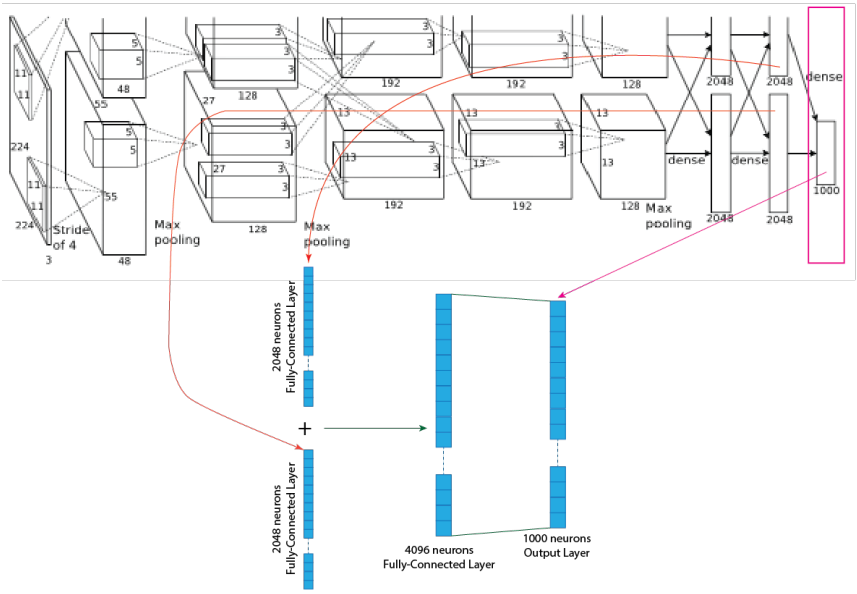
The tenth layer of the AlexNet architecture is a fully connected layer with 4096 feature maps and each of size 1x1. For each part, 2048 feature maps, and each of size 1x1. This layer fully connected with the previous fully connected layer with 4096 feature maps. In the fully-

connected layer, the ReLU function used as an activation function. For better understanding, a picture of the process presented below.



Output Layer:

The final layer of the AlexNet architecture is a fully connected output layer “y” shortly output layer with 1000 possible values where softmax function used as an activation function. Both fully connected layer of 2048 neurons connects together and make one fully connected layer of 4096 neurons, and it connected with the output layer. After summation, all the values of the units of the output layers must stay between 0-1. AlexNet architecture used a subset dataset of the ImageNet dataset, which has 1000 possible neurons of output layers corresponding to the digits from 0 to 999. For better understanding, a picture of the process presented below.



If this post helpfull for you than share it...

Twitter Facebook Google+

ALEXNET # INTRODUCTION TO ALEXNET # WHAT IS ALEXNET # ALEXNET CNN

We were unable to load Disqus. If you are a moderator please see our troubleshooting guide.

CONTACT ME