ImageNet Classification with Deep Convolutional Neural Networks

Group 13

D.M. Gamage	E/15/106	e15106@eng.pdn.ac.lk
H.P.S.P. Hewapathirana	E/15/129	e15129@eng.pdn.ac.lk
K.A.D.T.R.S. Perera	E/15/260	e15260@eng.pdn.ac.lk
W.W.R. Sachinthaka	E/15/310	e15310@eng.pdn.ac.lk

19/02/2021

Research Background

- Their approach was to find most appropriate machine learning technique which can be used to recognize the objects.
- Simply saying, a module was trained with set of large number of labeled images (as animals, trees, food etc.) to predict which category the future images should belongs.
- But the prevailing datasets of labeled images (NORB, CIFAR-10/100) were relatively small with low resolution.
- It is necessary to choose much larger training set with high-resolution images to predict the category of a given image with higher accuracy.

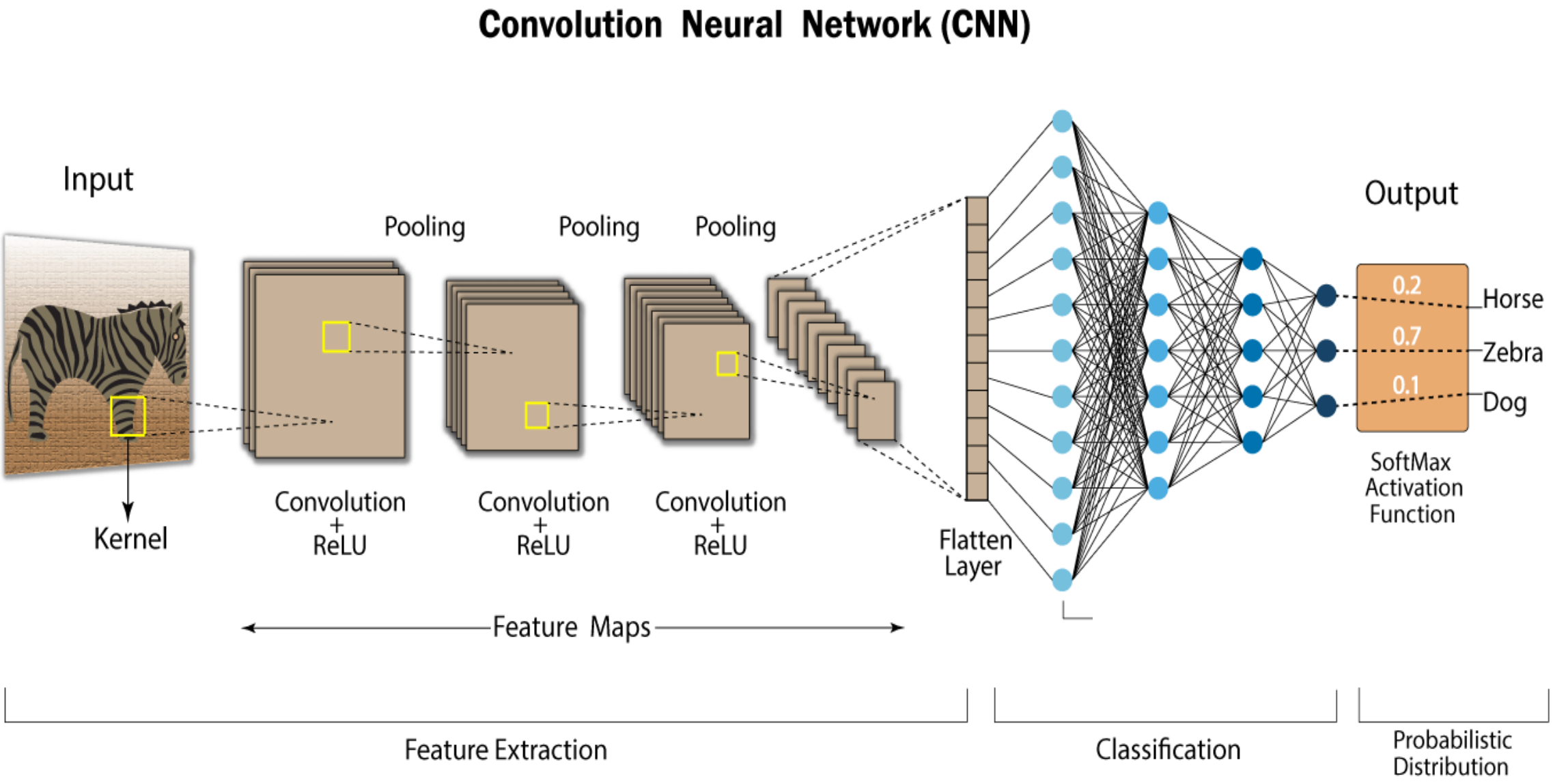
INTRODUCTION

- This was published by Alex Krizhevsky, Ilya Sutskever & Geoffrey E. Hinton in 2012.
- Used deep Convolutional Neural Network (CNN) to classify high-resolution images in the ImageNet platform.
- This is a supervised learning technique used for training modules.
- Used 2 Graphical processing Units (GPUs) to train a larger set of data very efficiently.
- Became one of the revolutional approach at the time when deep learning methods are concerned.
- To reduce overfitting while training data, recently-developed regularization method called “dropout” was used.

Why CNNs were used?

- Their capacity can be controlled by varying their depth and breadth.
- They also make strong and mostly correct assumptions about the nature of images
- Compared to standard feedforward neural networks with similarly-sized layers, CNNs have much fewer connections and parameters and so they are easier to train.

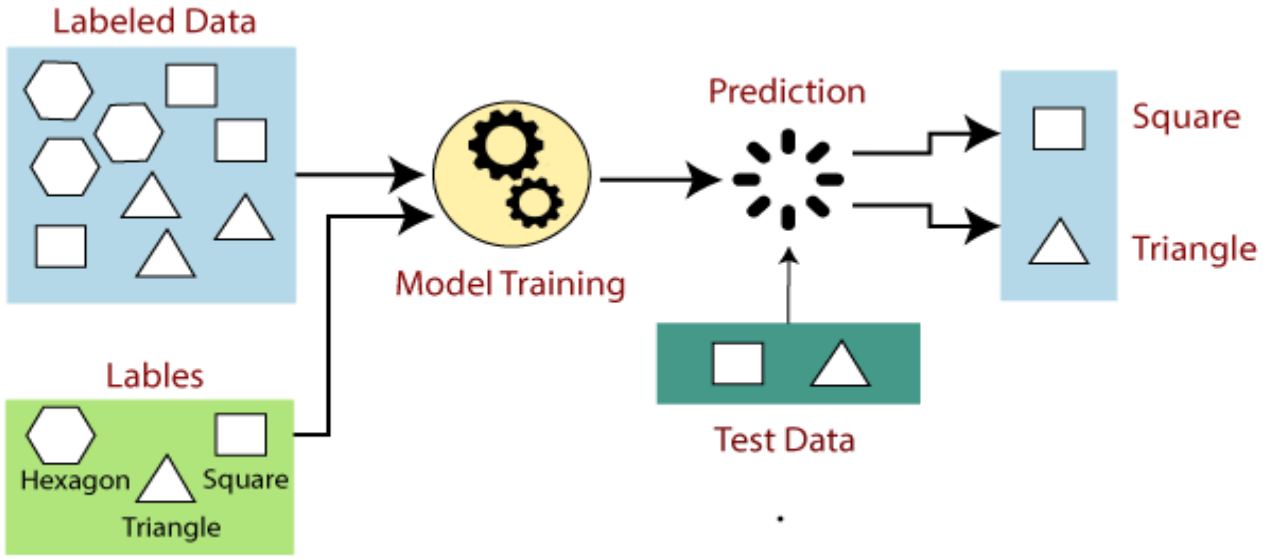
E/15/129 Pasindu



E/15/129 Pasindu

Training Data Set

- A larger set of data to be selected for higher accuracy predictions.
- NORB, CIFAR-10/100 are relatively small datasets of labeled images prevailed at the time with the resolution of 32×32 pixels.
- Instead of above, they have used data set in ImageNet which consists of over 15 million labeled high-resolution images of 256×256 pixels over 22,000 categories.
- Trained the network related to RGB values of the pixels.

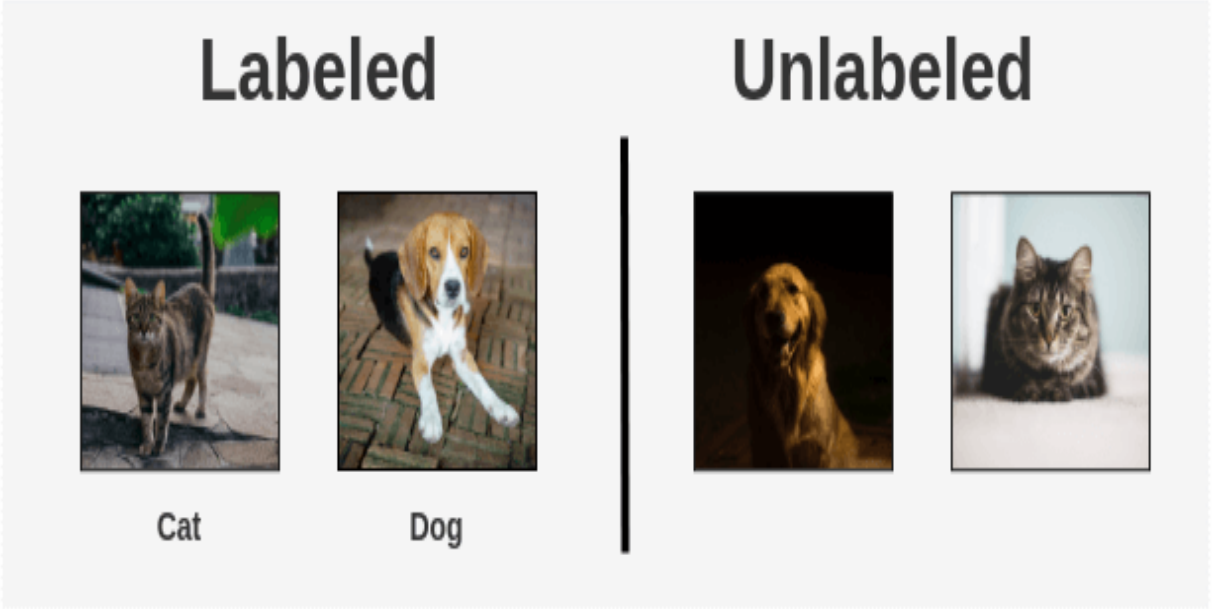


E/15/129 Pasindu

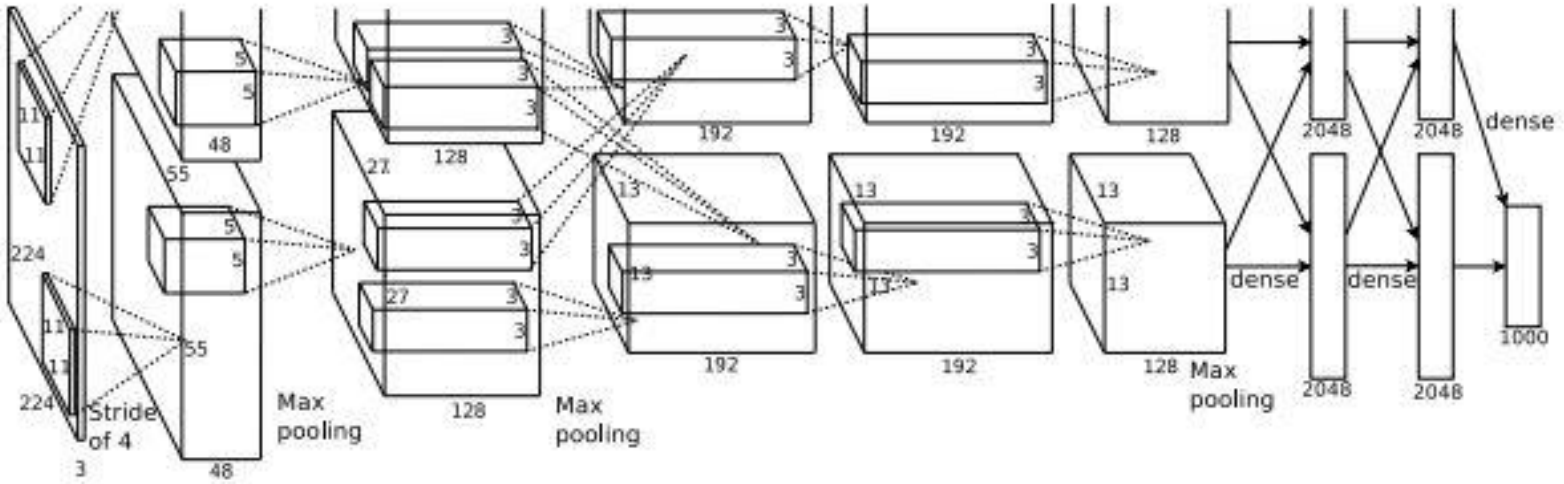
Their suggestions on drawbacks

Results can be improved simply by;

- Increase the amount of labeled data to train the module.
- Use powerful GPU

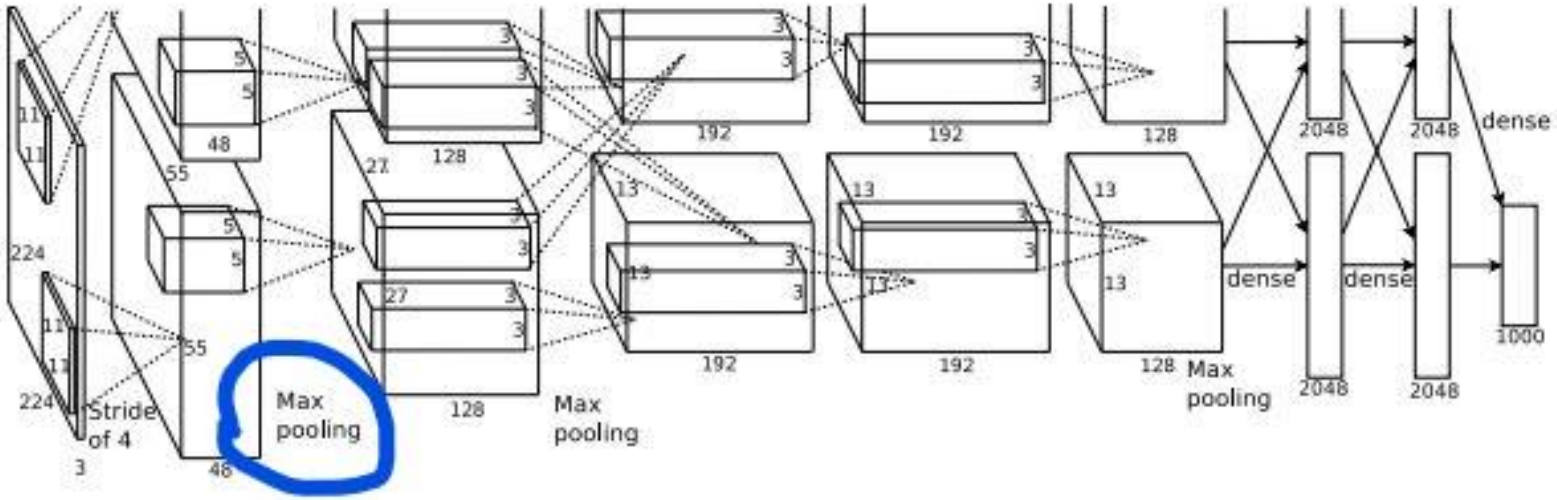


The Architecture



AlexNet Architecture

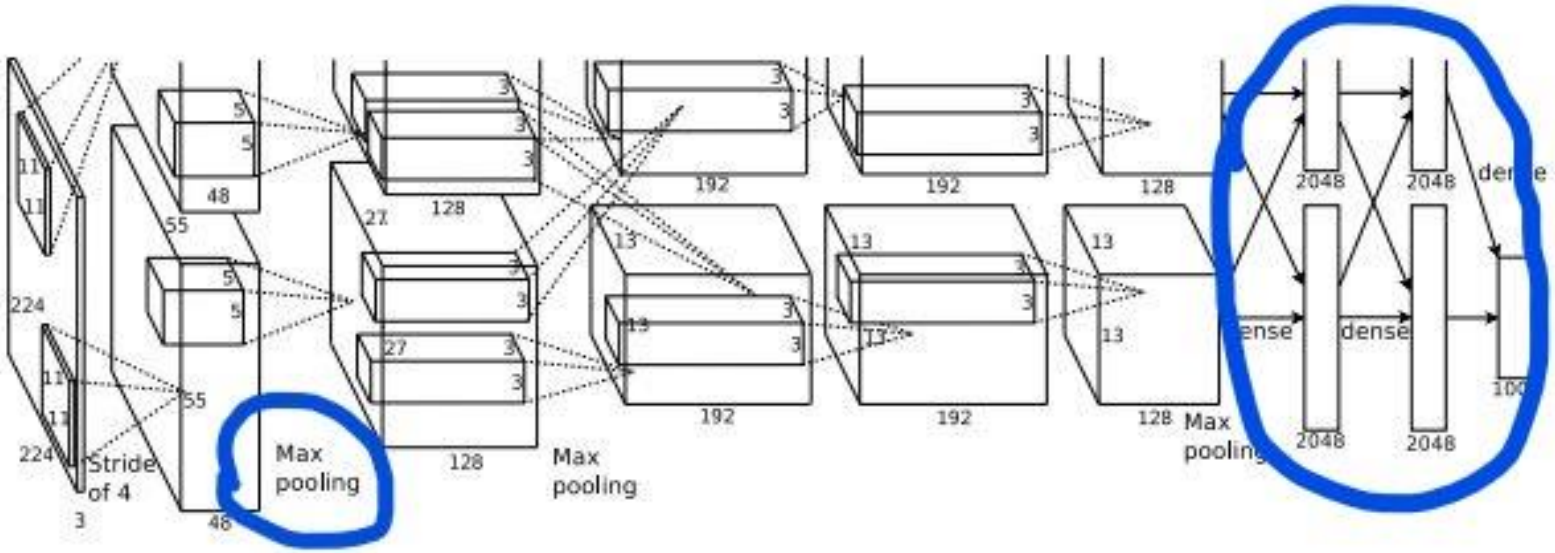
The Architecture



AlexNet Architecture

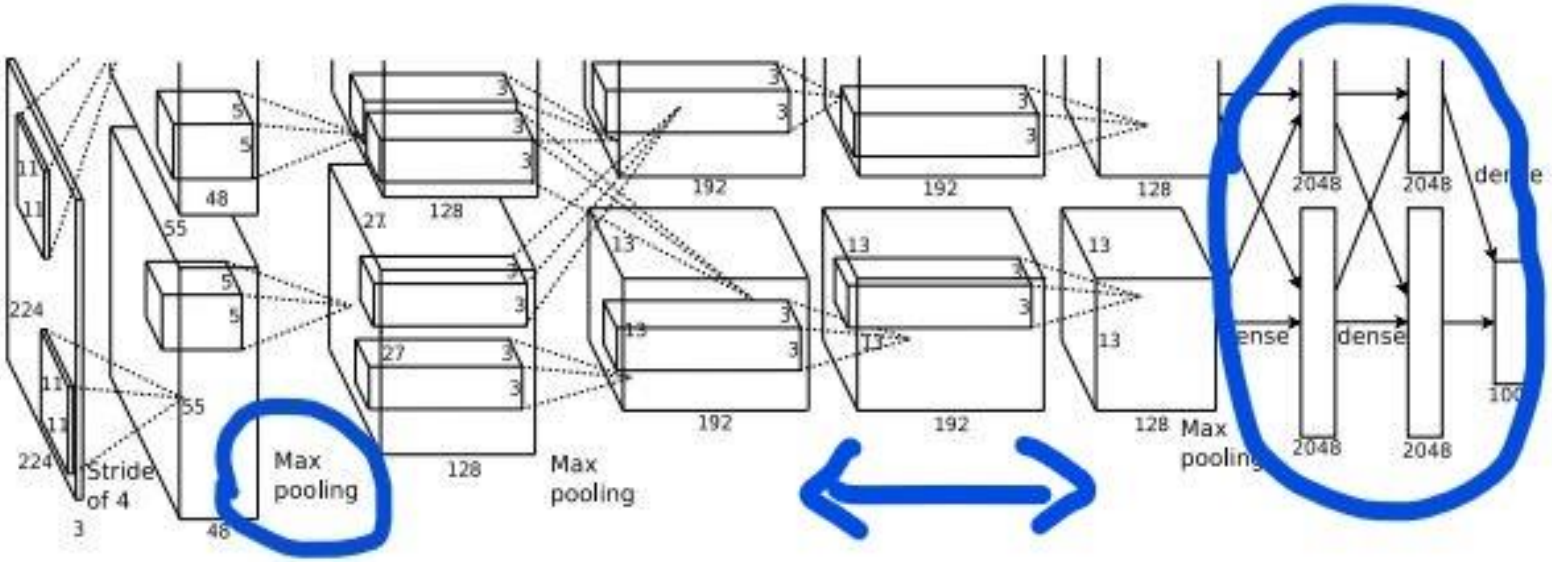
E/15/260 Thilina

The Architecture



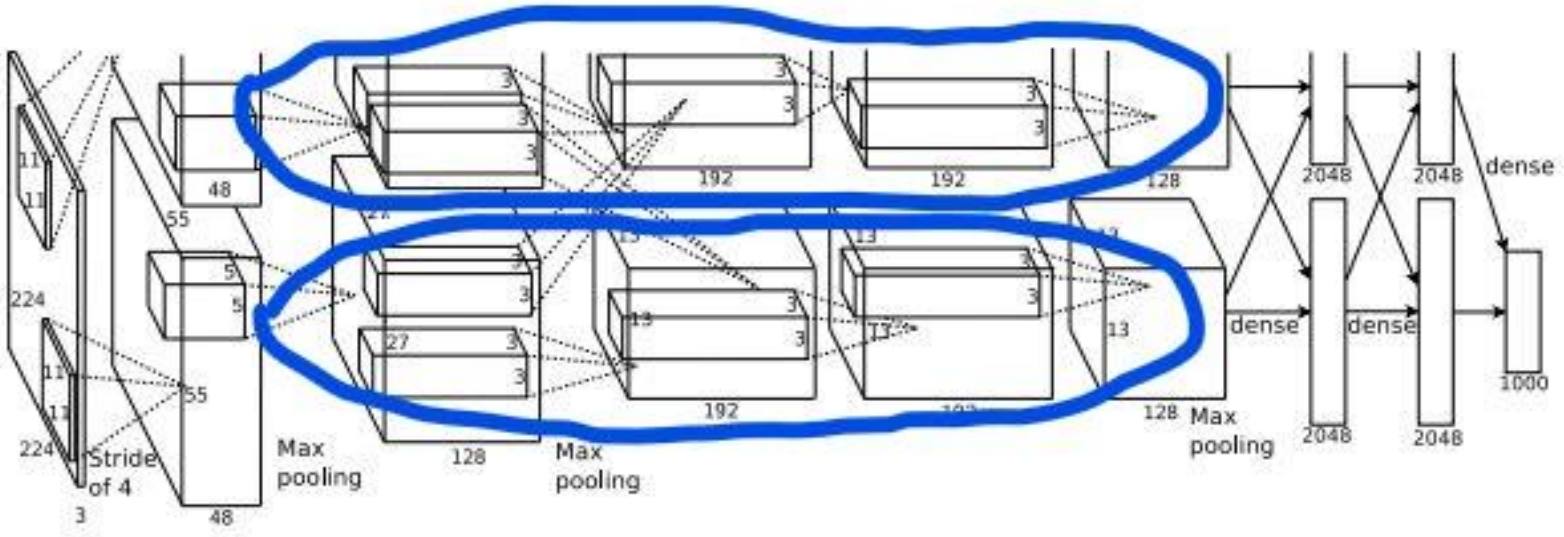
AlexNet Architecture

The Architecture



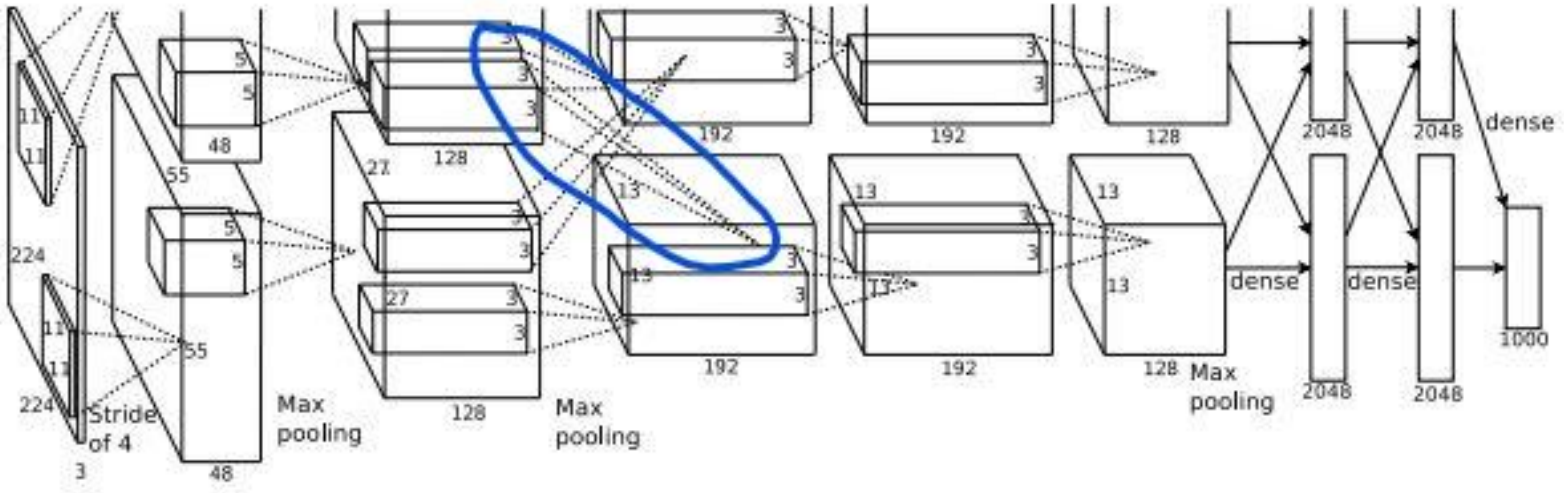
AlexNet Architecture

The Architecture



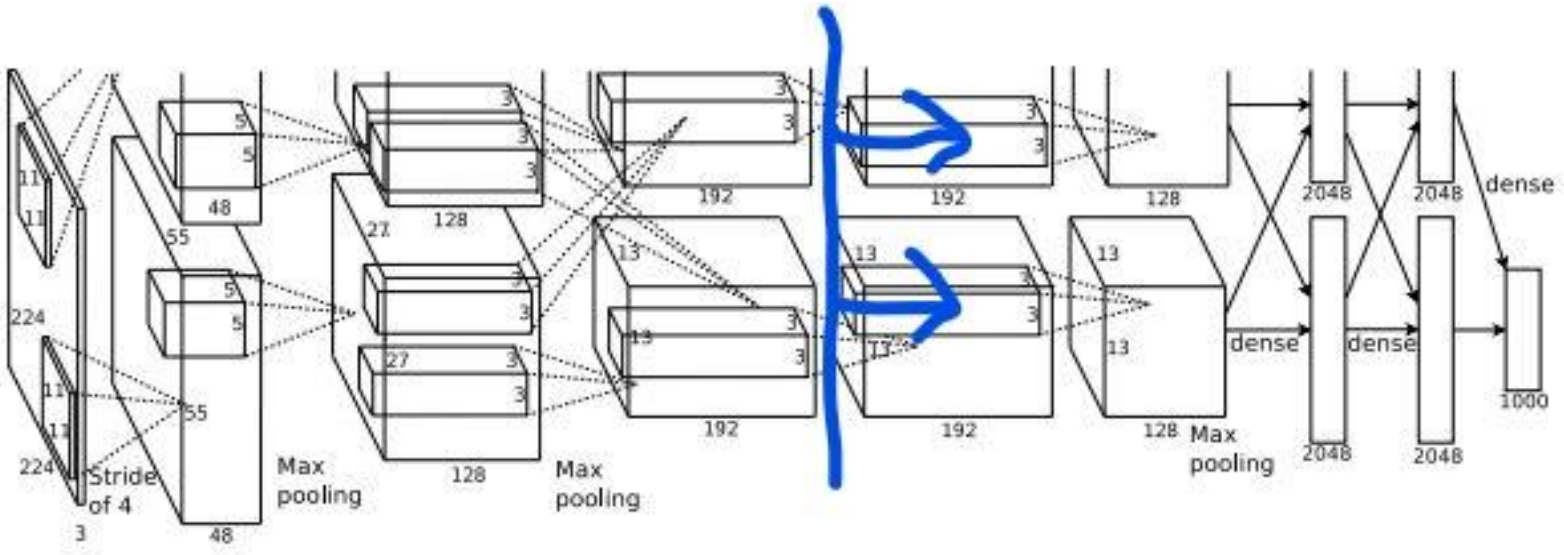
AlexNet Architecture

The Architecture



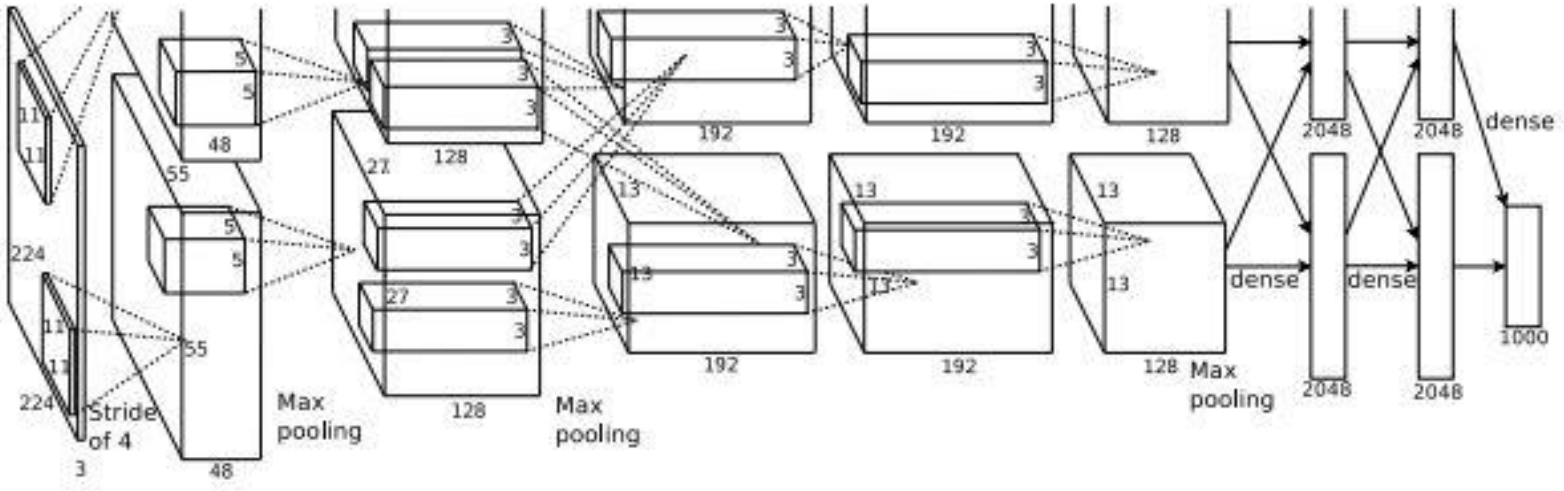
AlexNet Architecture

The Architecture



AlexNet Architecture

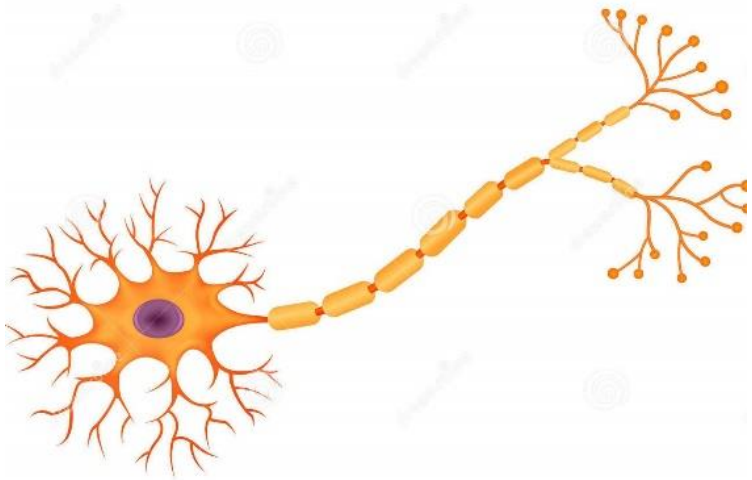
The Architecture



AlexNet Architecture

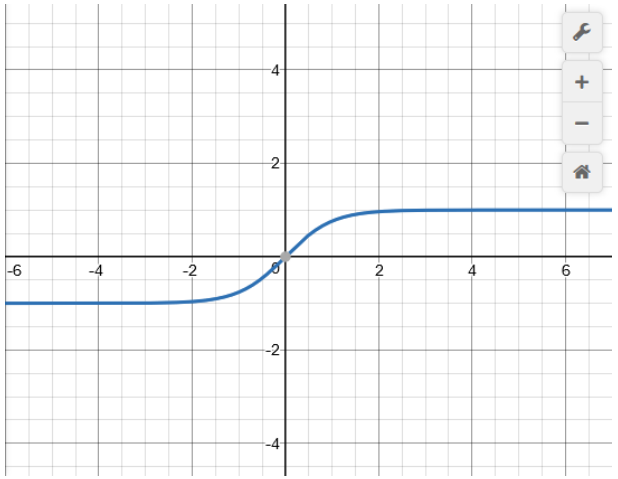
ReLU Non-linearity

- ReLU - Rectified Linear Unit
- It is a activation function.



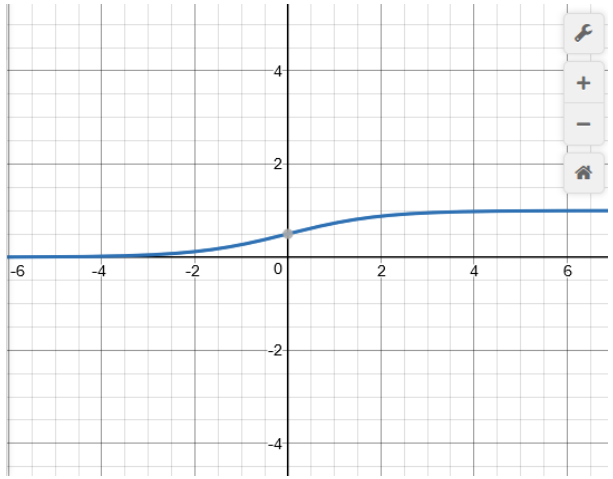
Hyperbolic function

$$f(x) = \tanh(x)$$



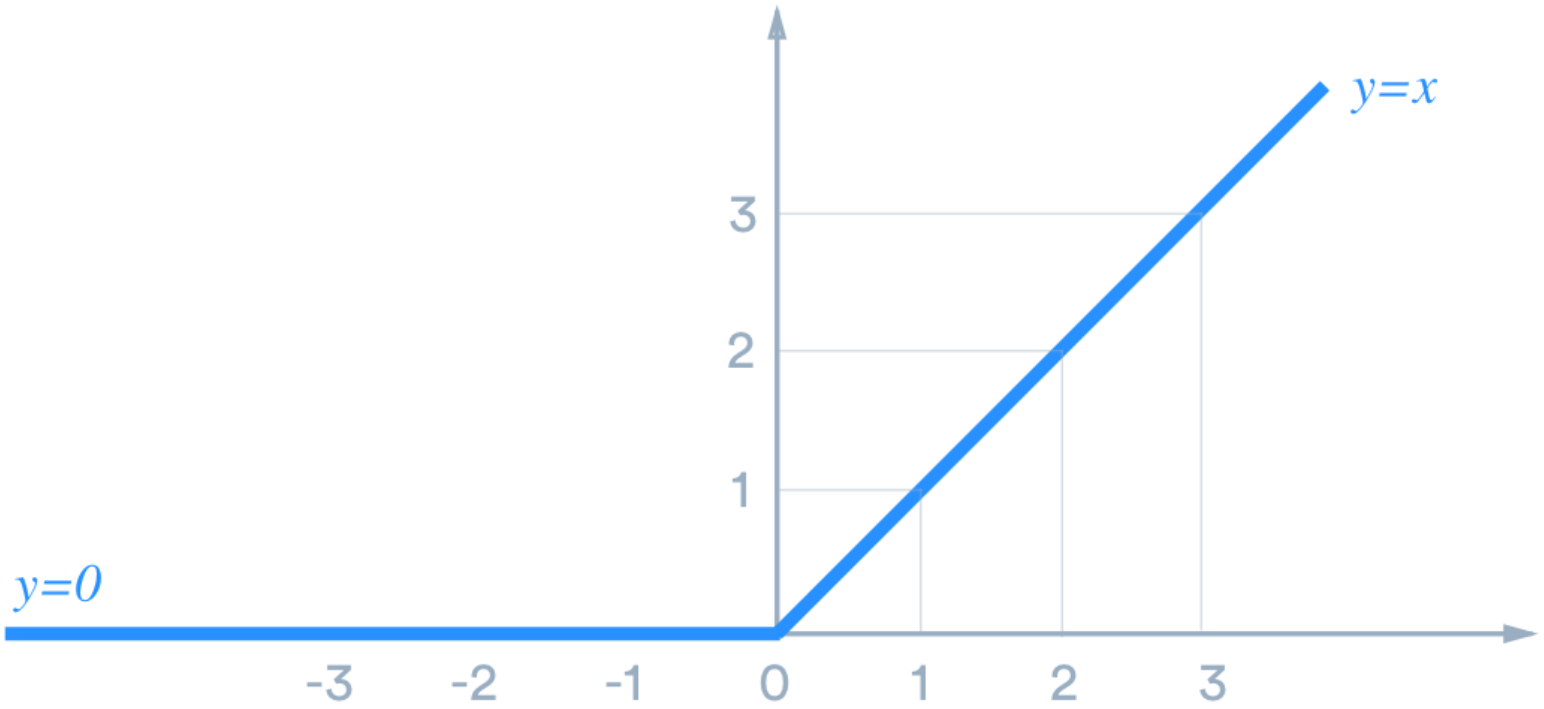
Sigmoid function

$$f(x) = (1 + e^{-x})^{-1}$$

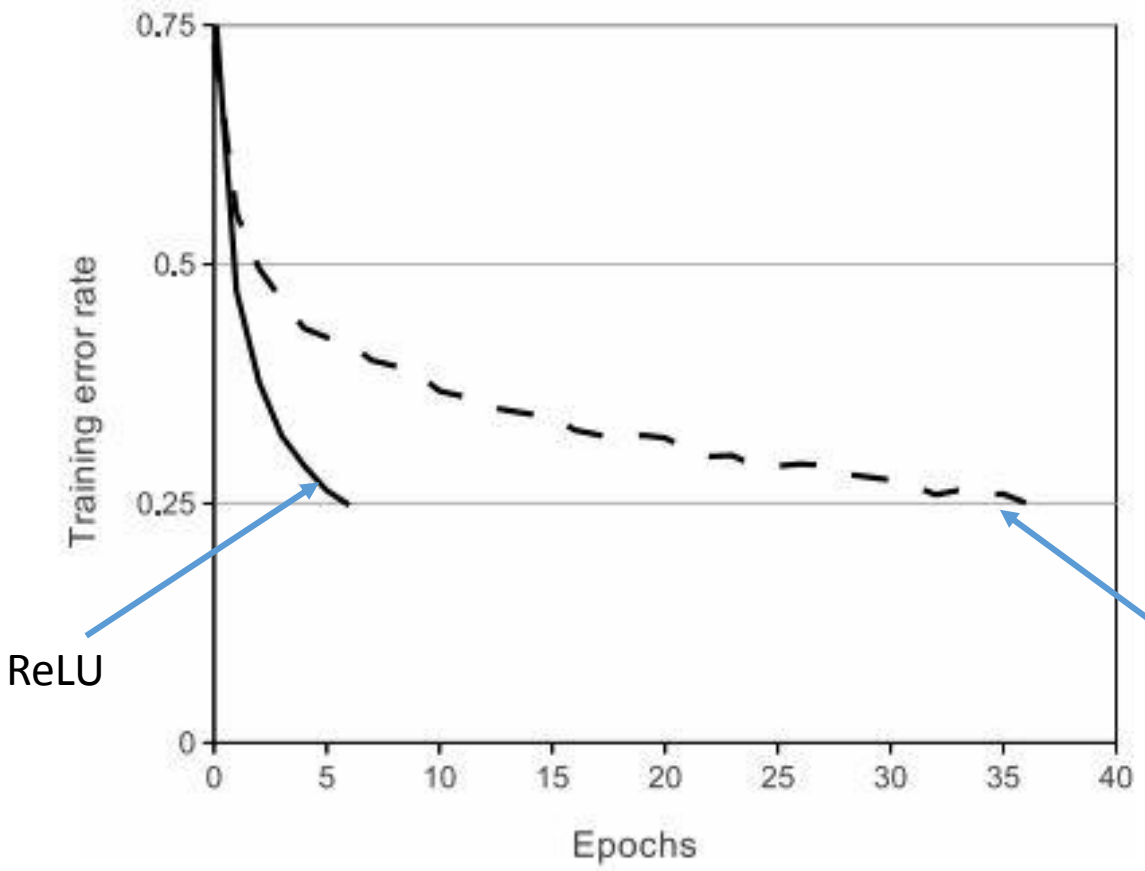


E/15/260 Thilina

ReLU Non-linearity



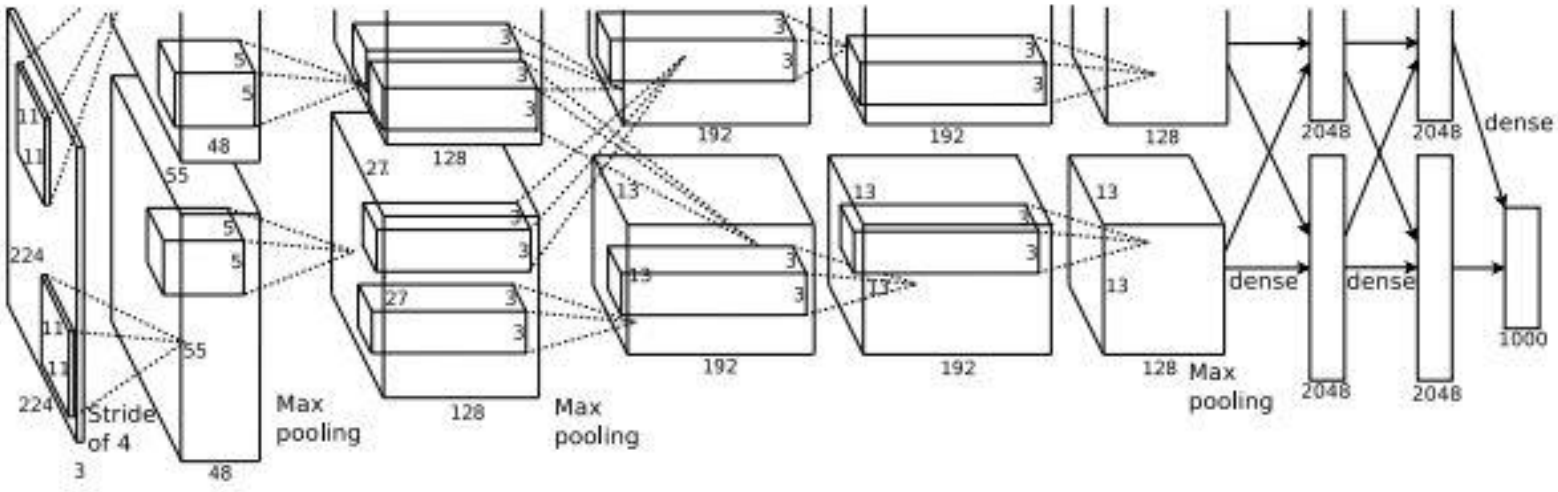
ReLU Non-linearity



6 times faster than hyperbolic tangent

Training on Multiple GPUs

- They had GTX 580 GPU with 3GB memory
- 1.2 million training examples too big to fit on one GPU
- they spread the net across two GPUs



GPUs communicate only in certain layers

Classification Results

- top-1 error rates by 1.7%
- top-5 error rates by 1.2%

Model	Top-1	Top-5
Sparse coding [2]	*47.1%*	*28.2%*
SIFT + FVs [24]	*45.7%*	*25.7%*
CNN	**37.5%**	**17.0%**

Error rate comparison

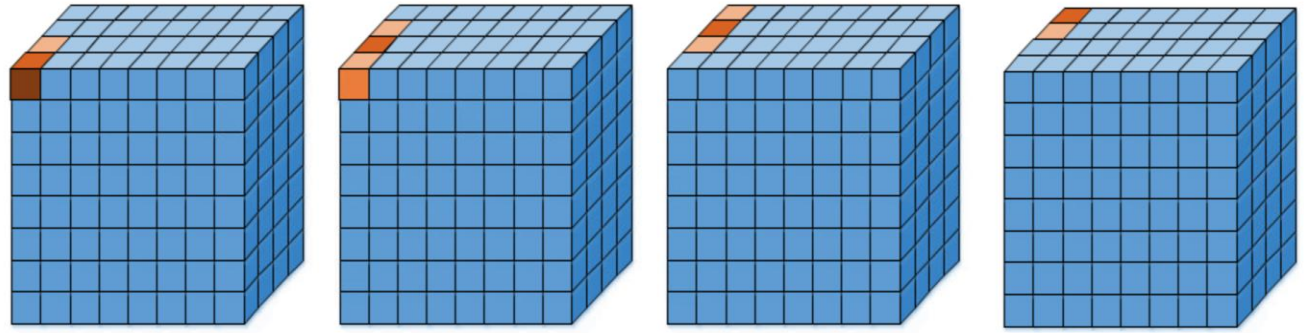
Local Response Normalization

Local Response Normalization (LRN) was first introduced in AlexNet architecture where the activation function used was ReLU as opposed to the more common tanh and sigmoid at that time. The reason for using LRN was to encourage lateral inhibition.

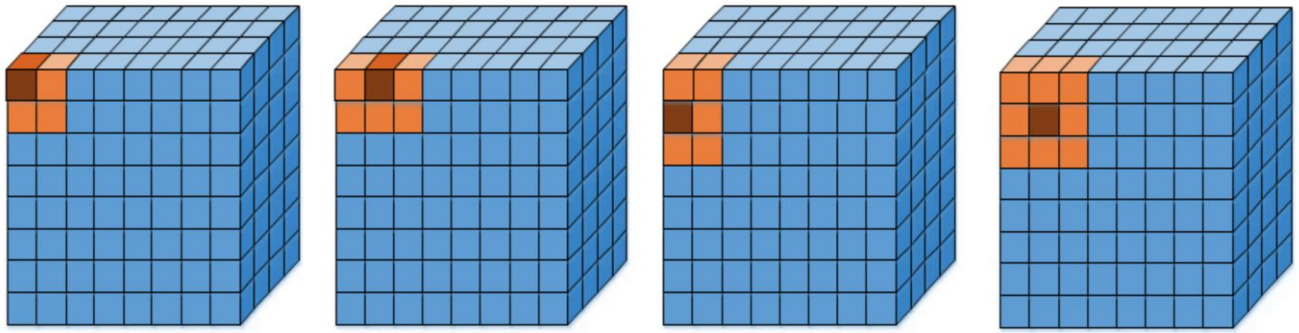
E/15/106 Dhanushka

Types of LRN

- LRN is a non-trainable layer that square-normalizes the pixel values in a feature map within a local neighborhood. There are two types of LRN based on the neighborhood defined.



a) Inter-Channel LRN (n=2)



b) Intra-Channel LRN (n=2)

Inter-Channel LRN:

- This is originally what the AlexNet paper used. The neighborhood defined is across the channel. For each (x,y) position, the normalization is carried out in the depth dimension and is given by the following formula

$$b^i_{x,y} = a^i_{x,y} / \left(k + \alpha \sum_{j=\max(0, i-n/2)}^{\min(N-1, i+n/2)} (a^j_{x,y})^2 \right)^{\beta}$$

E/15/106 Dhanushka

Intra-Channel LRN

- In Intra-channel LRN, the neighborhood is extended within the same channel only as can be seen in the figure above. The formula is given by

$$b^k_{x,y} = a^k_{x,y} / \left(k + \alpha \sum_{i=max(0,x-n/2)}^{min(W,x+n/2)} \sum_{j=max(0,y-n/2)}^{min(H,y+n/2)} (a^k_{i,j})^2 \right)^{\beta}$$

E/15/106 Dhanushka

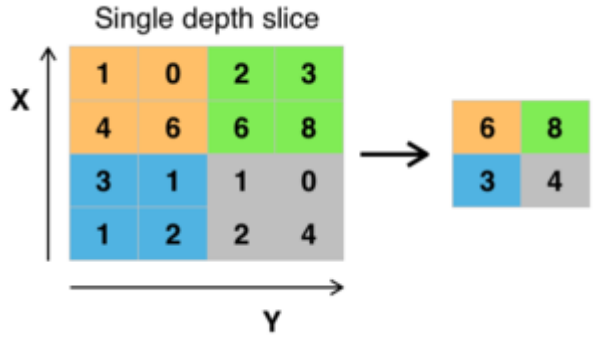
Overlapping Pooling

- Max Pooling layers are usually used to downsample the width and height of the tensors, keeping the depth same. Overlapping Max Pool layers are similar to the Max Pool layers, except the adjacent windows over which the max is computed overlap each other.

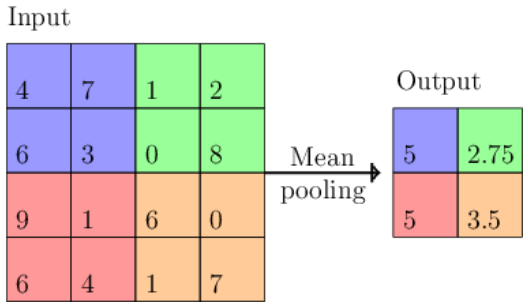
- The authors used pooling windows of size 3×3 with a stride of 2 between the adjacent windows. This overlapping nature of pooling helped reduce the top-1 error rate by 0.4% and top-5 error rate by 0.3% respectively when compared to using non-overlapping pooling windows of size 2×2 with a stride of 2 that would give same output dimensions.

There are many pooling techniques. They are as follows

- Mean pooling , where we take mean of the pixel values of a segment.

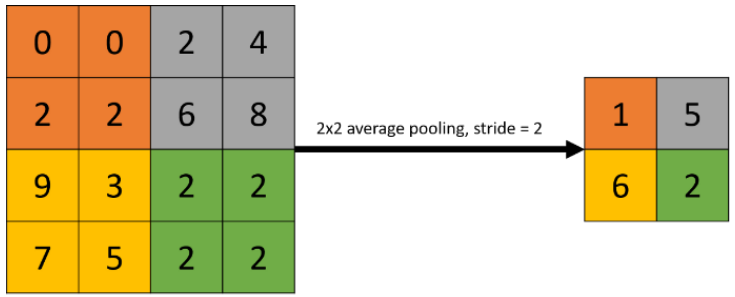


- Max pooling where we take largest of the pixel values of a segment.



E/15/106 Dhanushka

- Average pooling where we take average values of a segment.



We generally observe during training that models with overlapping pooling find it slightly more difficult to overfit.

Reducing Overfitting

What is overfitting?

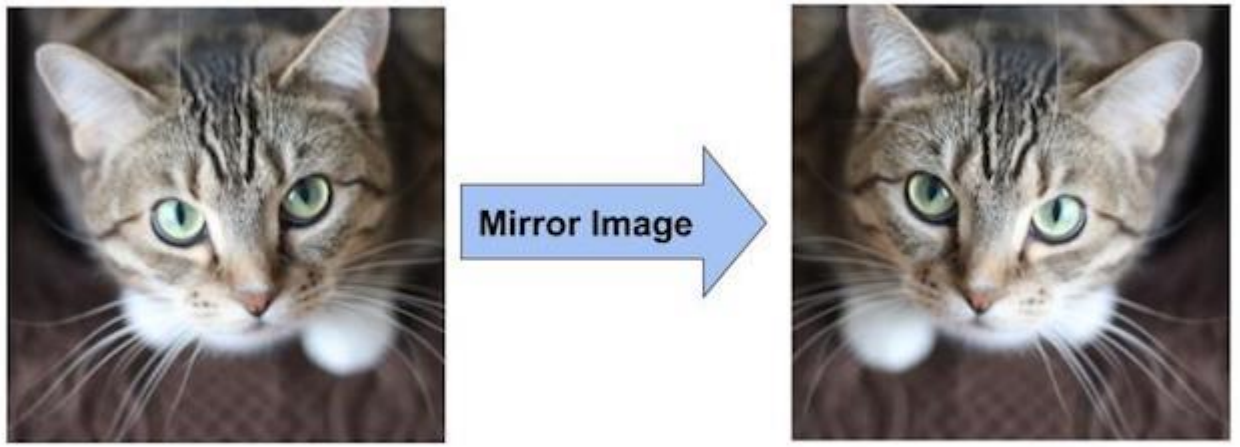
- The size of the Neural Network is its capacity to learn, but if we are not careful, it will try to memorize the examples in the training data without understanding the concept. As a result, the Neural Network will work exceptionally well on the training data, but they fail to learn the real concept. It will fail to work well on new and unseen test data. This is called overfitting.

- The authors of AlexNet reduced overfitting using a couple of different methods.

Data Augmentation

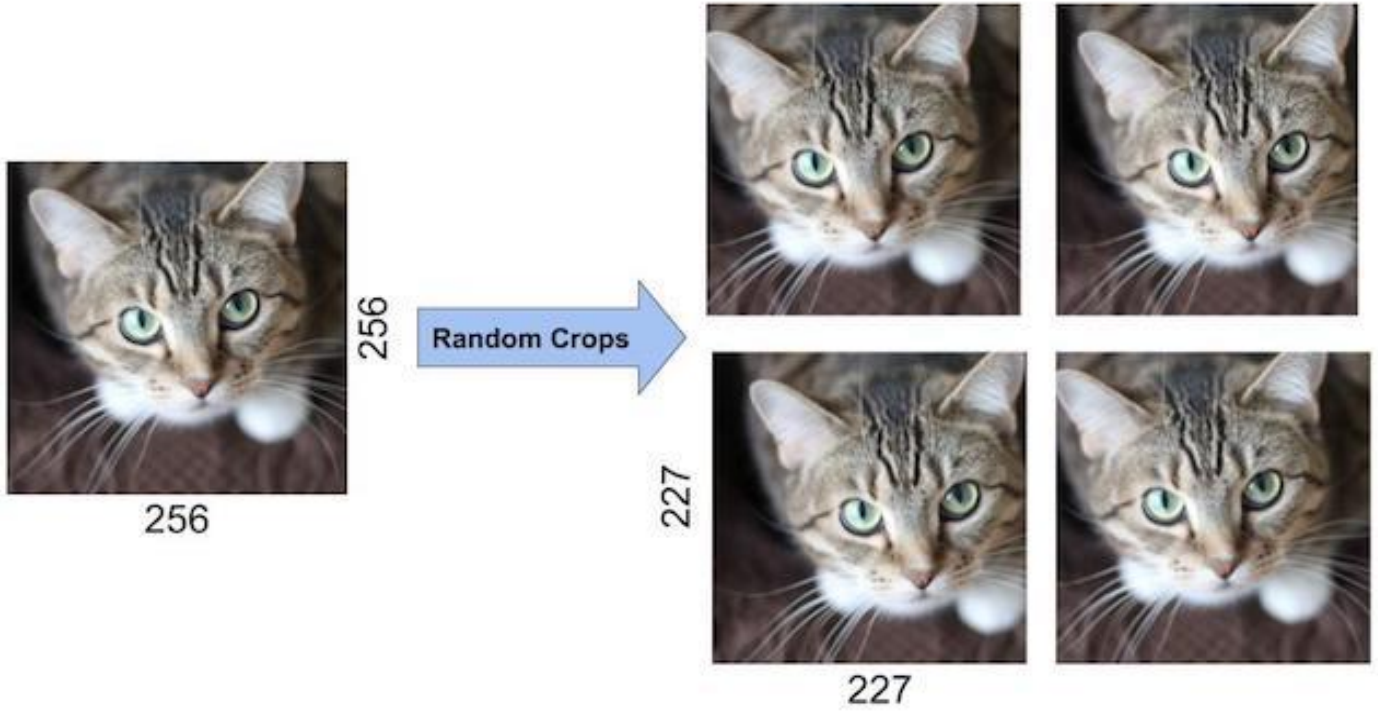
- Showing a Neural Net different variation of the same image helps prevent overfitting. We are forcing it to not memorize Often it is possible to generate additional data from existing data for free. Here are few tricks used by the AlexNet team.

1) Data Augmentation by Mirroring



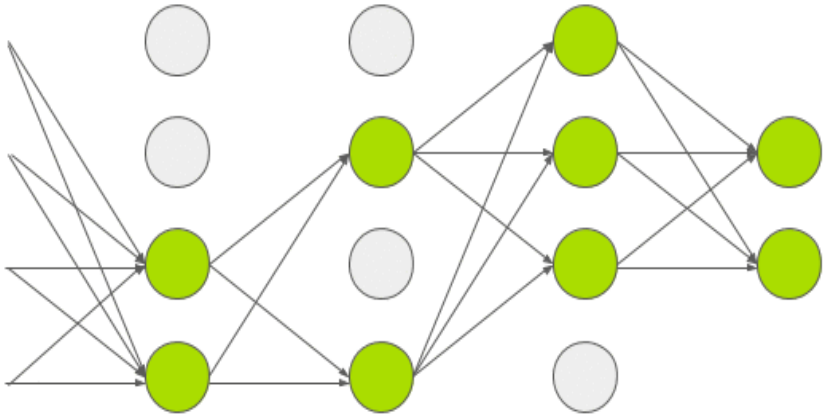
2) Data Augmentation by Random Crops

In addition, cropping the original image randomly will also lead to additional data that is just a shifted version of the original data.



Dropout

- In dropout, a neuron is dropped from the network with a probability of 0.5.
- When a neuron is dropped, it does not contribute to either forward or backward propagation.
- Combining the predictions of many different models ⟶ reduce the test errors.
- Without dropout network exhibits substantial overfitting.
- Dropout roughly double the number of iterations required to converge.



Results

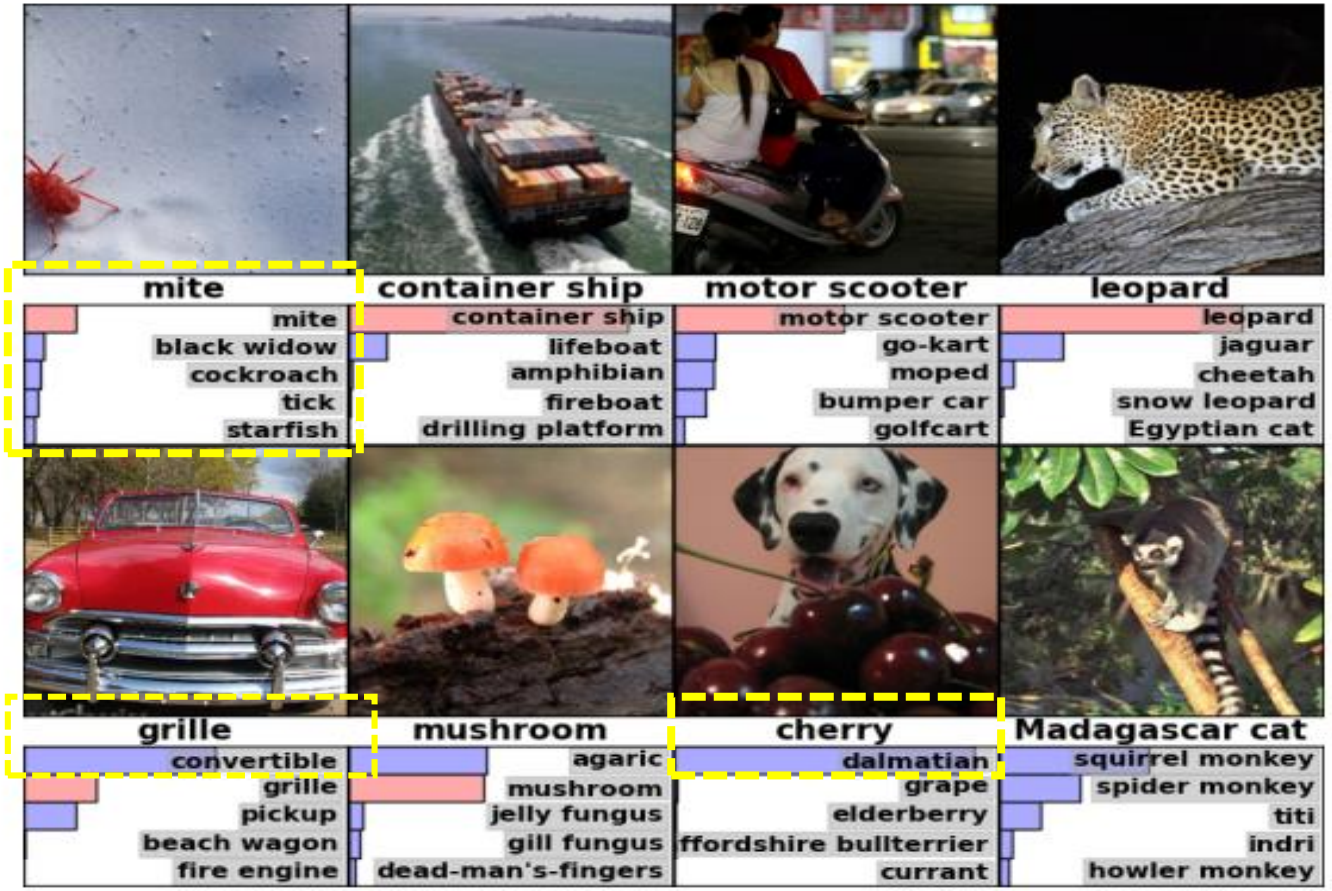
- Used momentum to train and do some qualitative analysis.
- Network achieves top-1 and top-5 test set error rates of 67.4% and 40.9% (ILSVRC-2009).
- Network achieves top-1 and top-5 test set error rates of 37.5% and 17.0% (ILSVRC-2010).
- Network is “fine-tuning ” with ILSVRC-2011 to get error rate on ILSVRC-2012.

Model	Top-1 (val)	Top-5 (val)	Top-5 (test)
SIFT + FVs [7]	—	—	*26.2%*
1 CNN	40.7%	18.2%	—
5 CNNs	38.1%	16.4%	**16.4%**
1 CNN*	39.0%	16.6%	—
7 CNNs*	36.7%	15.4%	**15.3%**

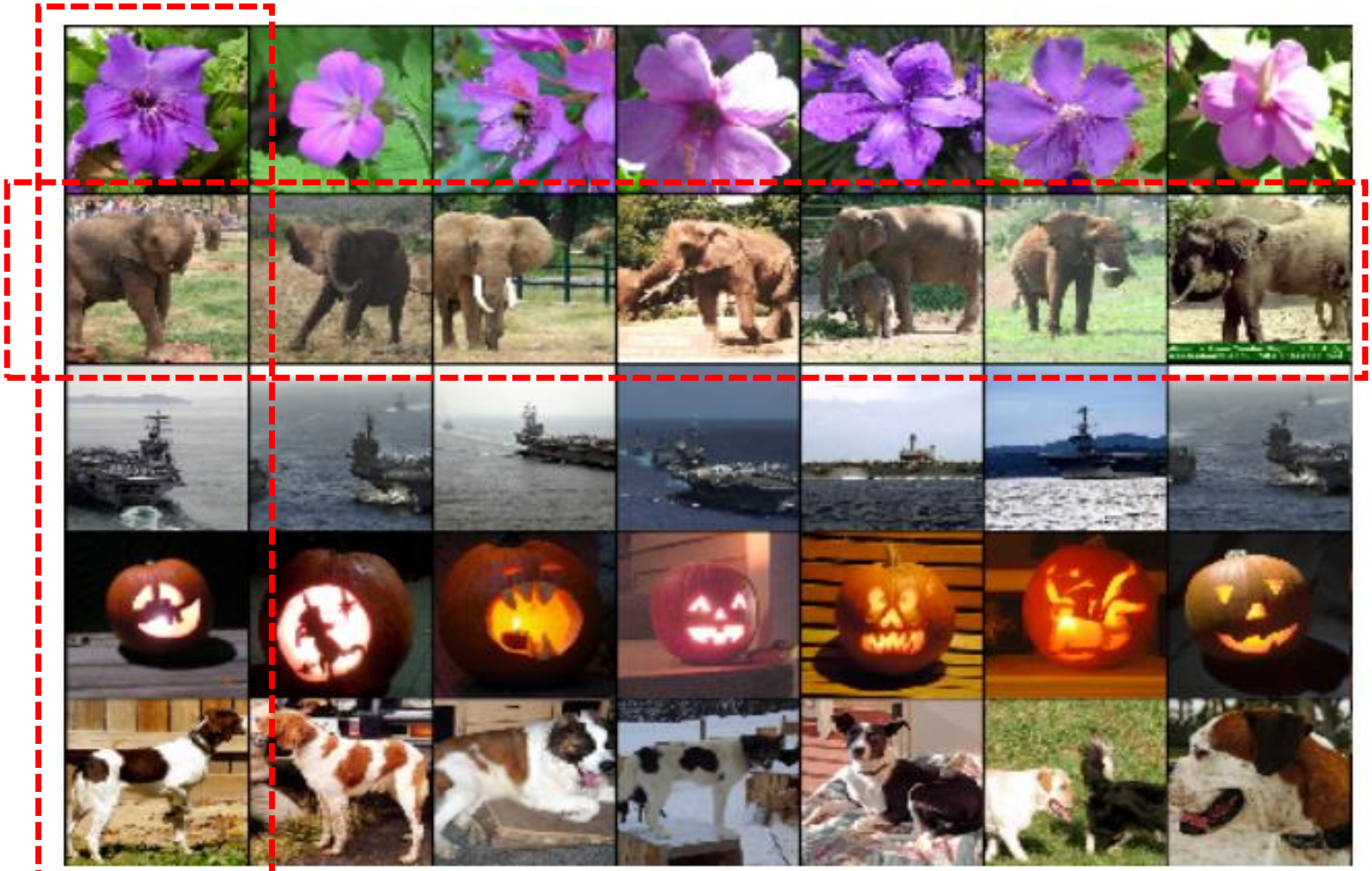
Table: Comparison of error rates on ILSVRC-2012 validation and test sets

Qualitative evaluation

- **RED** bar is the correct bar.



E/15/310 Ravindu



Training set image

E/15/310 Ravindu

Discussion

- A large, deep CNN is capable of achieving record-breaking results on a highly challenging dataset using purely **supervised learning.**

- Network performance degrade ⟶ if the single convolutional layer is removed.

  Example:- Removing any of the middle layer results in a loss of about 2% for the top-1 performance.

- Depth of the network is very important to achieve the results.

- Did not use any unsupervised pre-training.

- Further, like to use very large and deep convolutional nets on video sequences.


THANK
YOU

Q & A ?

1) What are the famous datasets/competitions for image classification?

2) Explain the words like training/validation/testing sets.

3) What are the metrics used in image classification tasks?

4) What are other image-related tasks (except for classification)?

Q & A ?

5) Compare this NN to the newer image classification networks. What are the differences you see between this and newer ones?

6) What is the purpose of using dropout layers in convolutional neural networks?

7) Are there any advantages when directly applying a pooling layer after each convolutional layer that is used in this paper rather than stacking multiple convolutional layers?

8) Pooling layers are used in CNNs architecture. What is the condition to occur overlapping pooling?


Any questions?