NPTEL ONLINE CERTIFICATION COURSES

Course Name: Deep Learning

Faculty Name: Prof. P. K. Biswas
Department : E & ECE, IIT Kharagpur

Topic

Lecture 36: CNN Architecture

CONCEPTS COVERED

Concepts Covered:

- CNN
  - CNN Architecture
  - Convolution Layer
  - Receptive Field
  - Nonlinearity
  - Pooling

Convolution

1 D Convolution

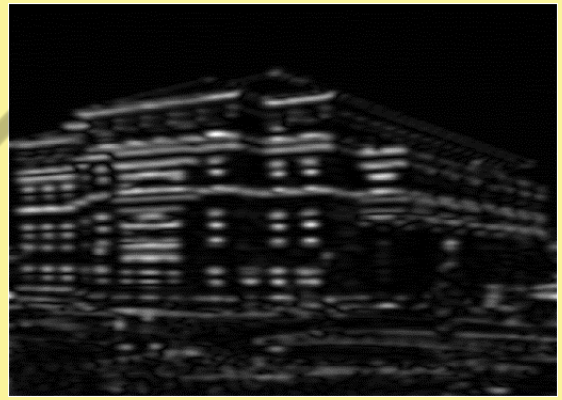
$$y(n) = \sum_{p=0}^{\infty} x(p)h(n-p) \qquad y(t) = \int_{0}^{\infty} x(\tau)h(t-\tau)d\tau$$

2 D Convolution

$$y(m,n) = \sum_{p=0}^{\infty}\sum_{q=0}^{\infty} x(p,q)h(m-p,n-q)$$

Finite Convolution Kernel

Feature at a point is local in nature



Convolution Kernel

1 D → 2A+1

$$y(n) = \sum_{p=-A}^{A} w(p)x(n-p)$$

2 D → (2A+1)x(2A+1)

$$y(m,n) = \sum_{p=-A}^{A} \sum_{q=-A}^{A} w(p,q)x(m-p,n-q)$$

Finite Convolution Kernel

0	0	X(0)	X(1)	X(2)	X(3)	.	X(n-2)	X(n-1)	X(n)	X(n+1)	X(n+2)	.
W(2)	W(1)	W(0)	W(-1)	W(-2)								
		Y(0)										

Finite Convolution Kernel

0	0	X(0)	X(1)	X(2)	X(3)	.	X(n-2)	X(n-1)	X(n)	X(n+1)	X(n+2)	.
	W(2)	W(1)	W(0)	W(-1)	W(-2)							
		Y(0)	Y(1)									

Finite Convolution Kernel

0	0	X(0)	X(1)	X(2)	X(3)	.	X(n-2)	X(n-1)	X(n)	X(n+1)	X(n+2)	.
		W(2)	W(1)	W(0)	W(-1)	W(-2)						
		Y(0)	Y(1)	Y(2)								

Finite Convolution Kernel

0	0	X(0)	X(1)	X(2)	X(3)	.	X(n-2)	X(n-1)	X(n)	X(n+1)	X(n+2)	.
			W(2)	W(1)	W(0)	W(-1)	W(-2)					
		Y(0)	Y(1)	Y(2)	Y(3)							

Finite Convolution Kernel

0	0	X(0)	X(1)	X(2)	X(3)	.	X(n-2)	X(n-1)	X(n)	X(n+1)	X(n+2)	.
						W(2)	W(1)	W(0)	W(-1)	W(-2)		
		Y(0)	Y(1)	Y(2)	Y(3)			Y(n-1)				

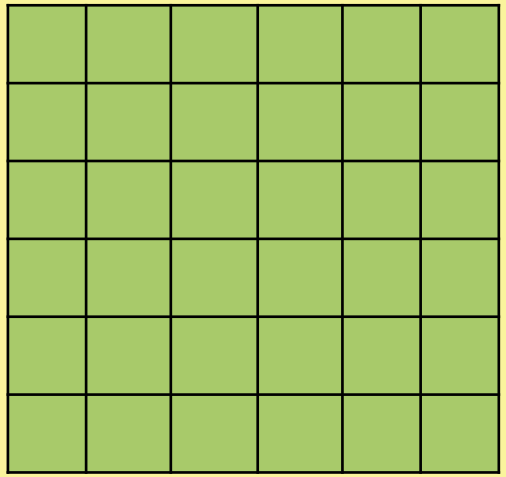
Finite Convolution Kernel

0	0	X(0)	X(1)	X(2)	X(3)	.	X(n-2)	X(n-1)	X(n)	X(n+1)	X(n+2)	.
							W(2)	W(1)	W(0)	W(-1)	W(-2)	
		Y(0)	Y(1)	Y(2)	Y(3)			Y(n-1)	Y(n)			

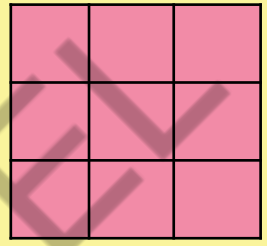
Finite Convolution Kernel

0	0	X(0)	X(1)	X(2)	X(3)	.	X(n-2)	X(n-1)	X(n)	X(n+1)	X(n+2)	.
								W(2)	W(1)	W(0)	W(-1)	W(-2)
		Y(0)	Y(1)	Y(2)	Y(3)			Y(n-1)	Y(n)	Y(n+1)		

2 D Convolution

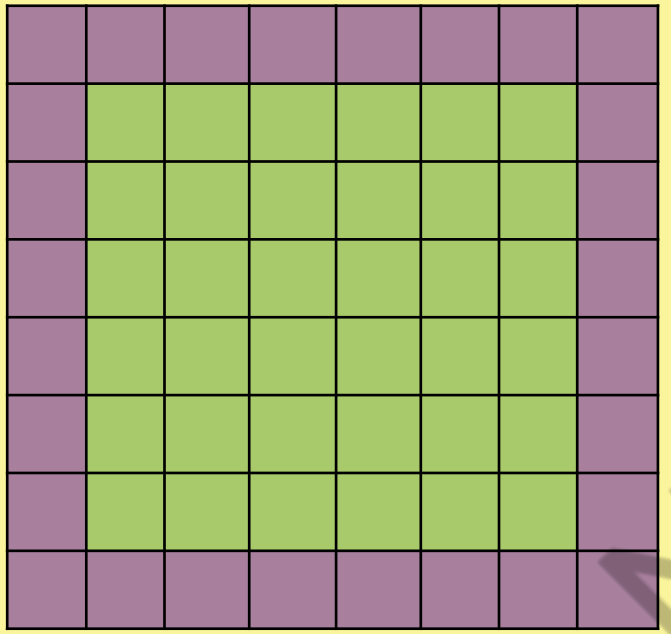


6 x 6 Image

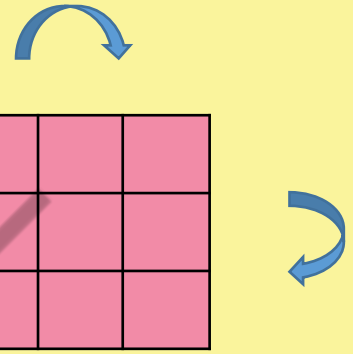


3 x 3 Kernel

2 D Convolution

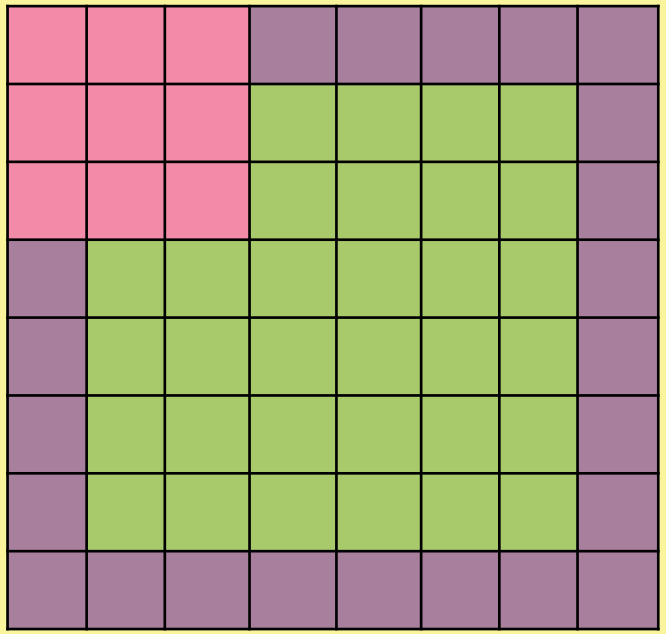


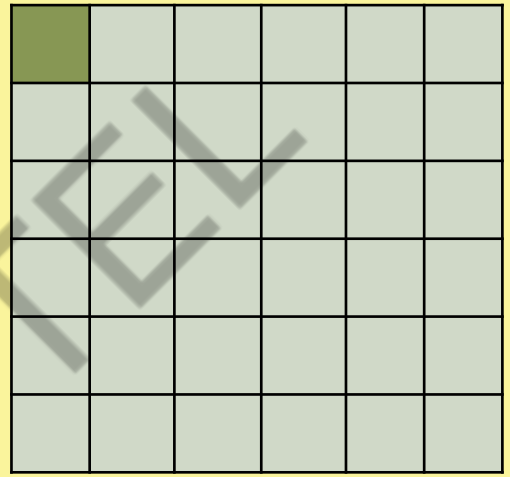
0 Padding



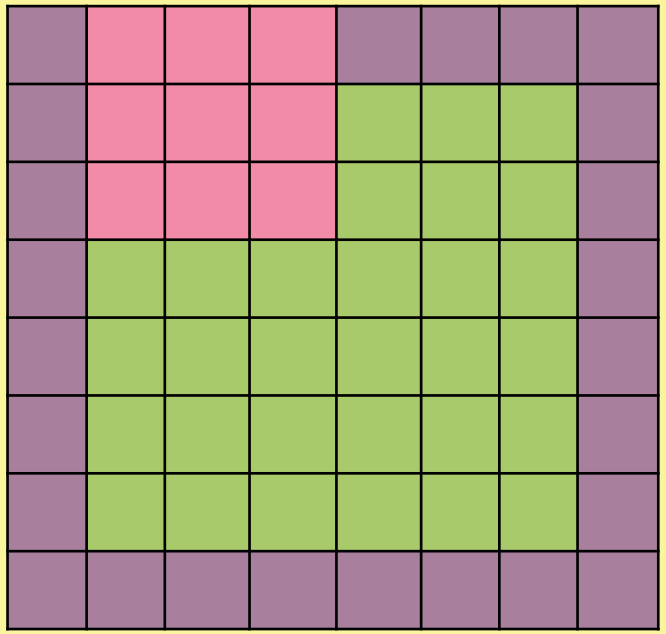
Flipping

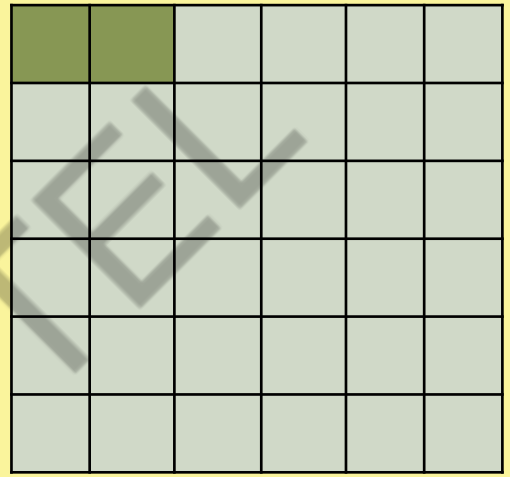
2 D Convolution



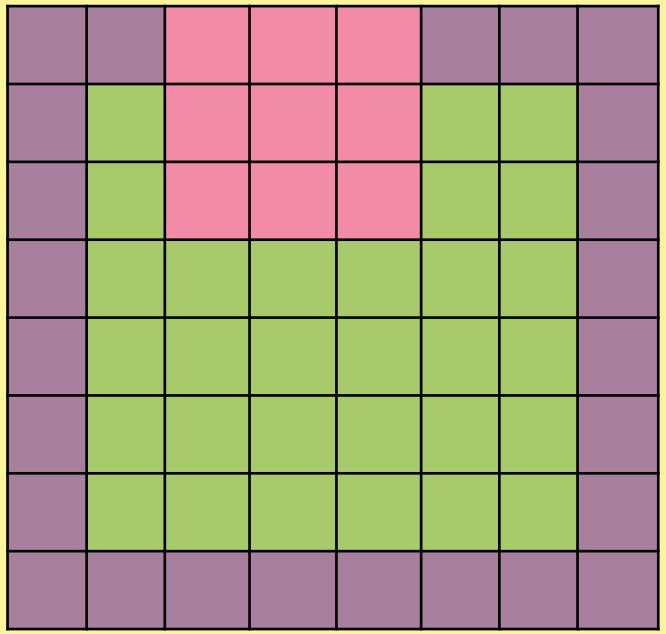


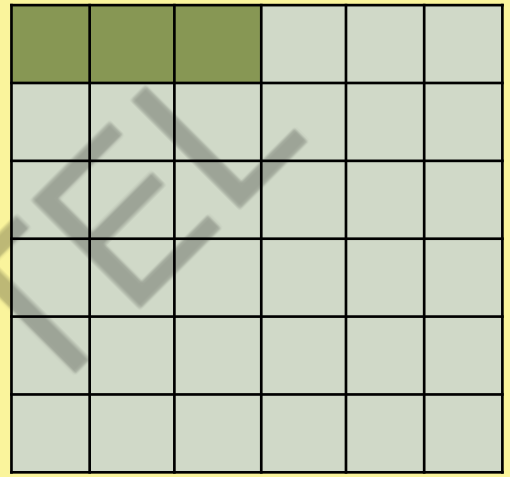
2 D Convolution



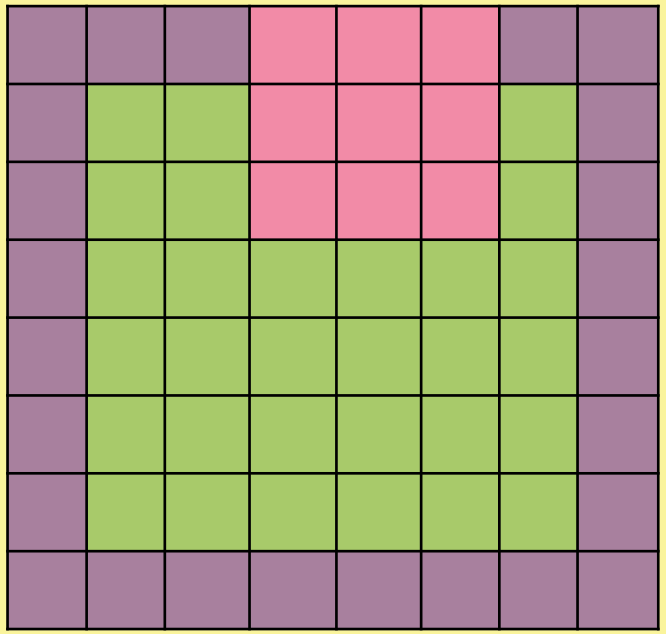


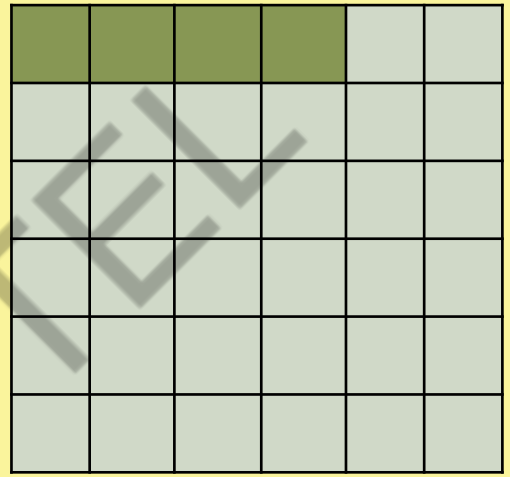
2 D Convolution



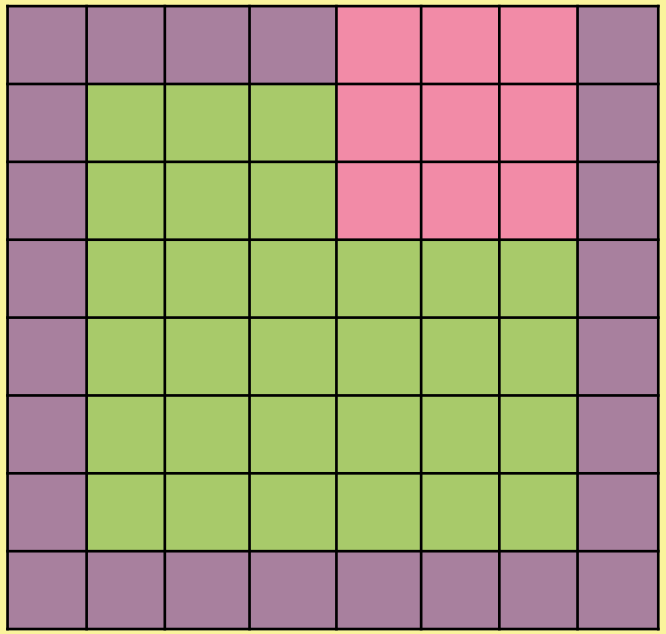


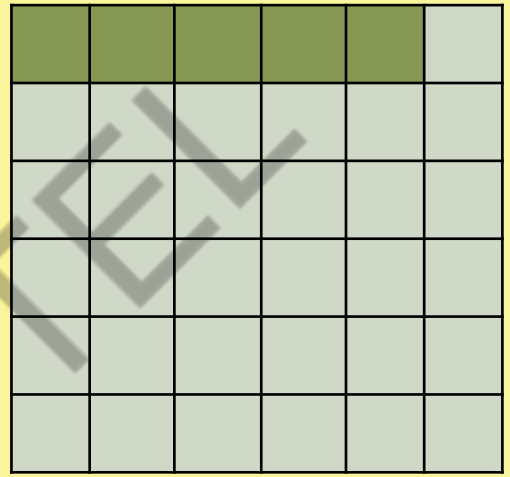
2 D Convolution



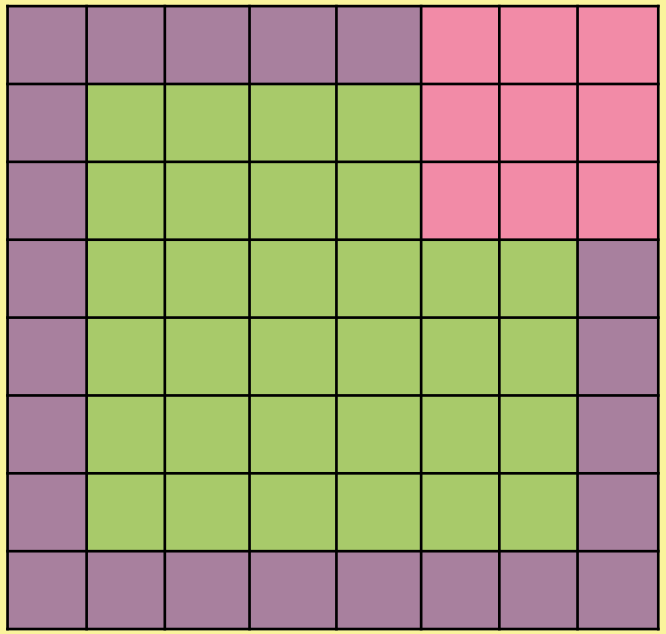


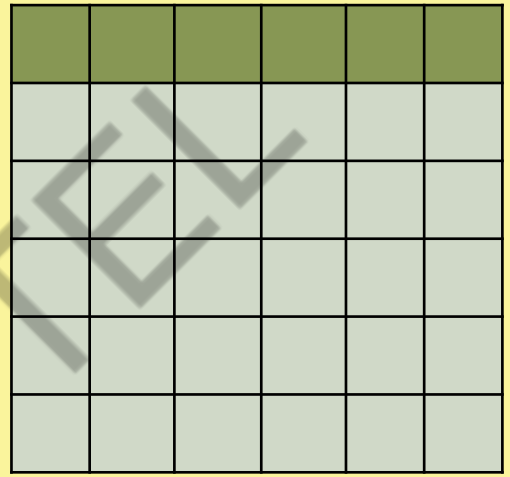
2 D Convolution



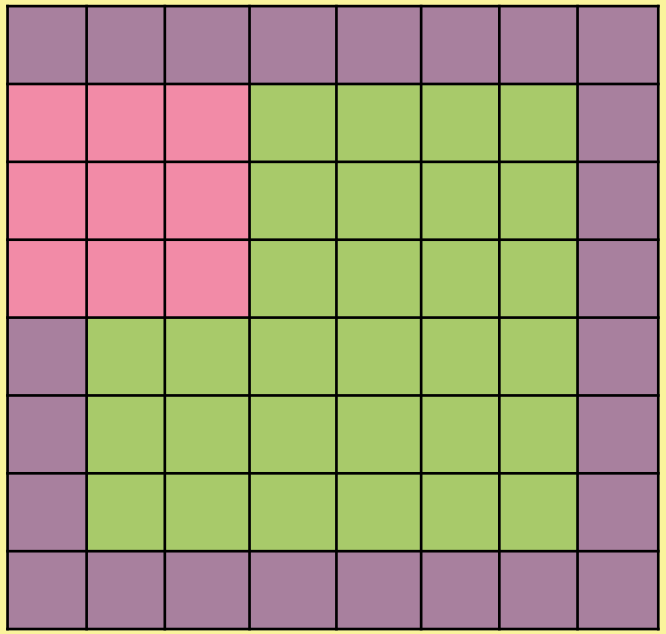


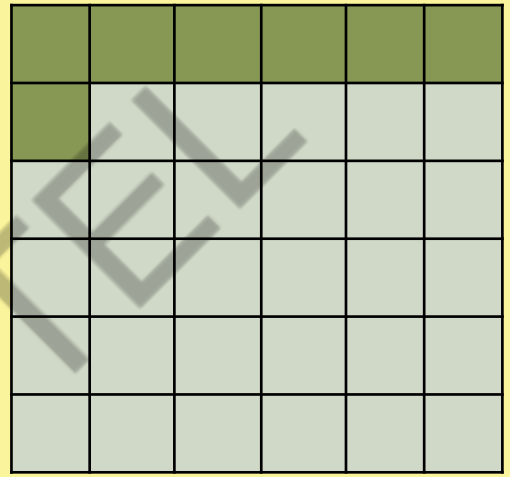
2 D Convolution



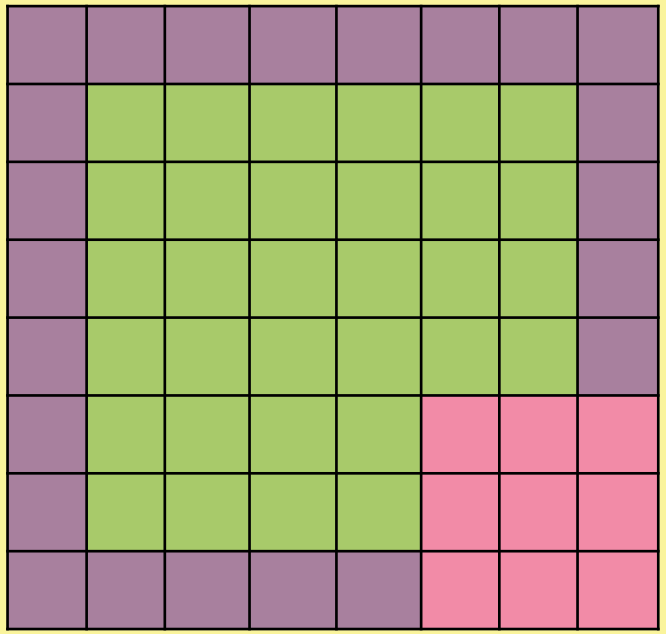


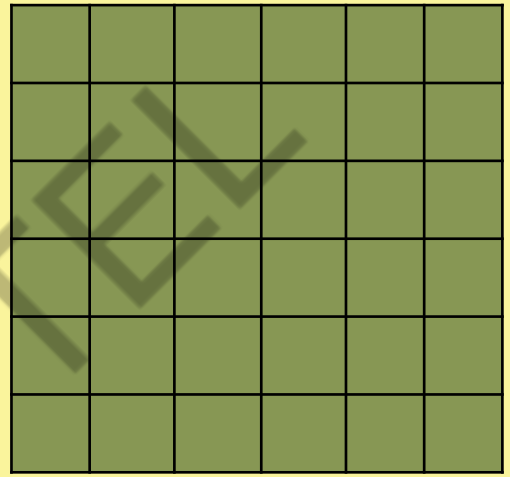
2 D Convolution





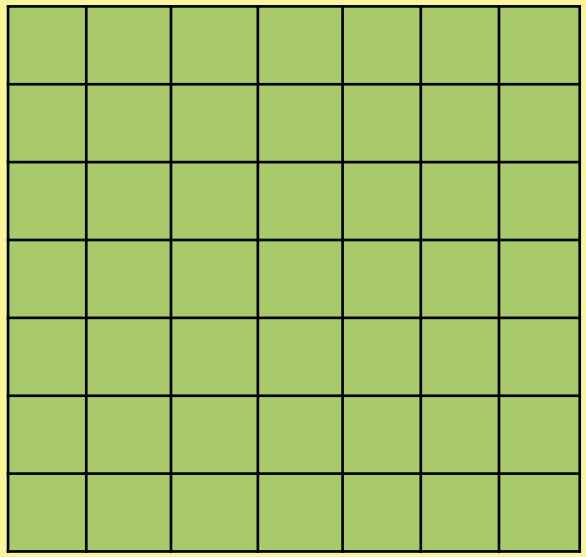
2 D Convolution



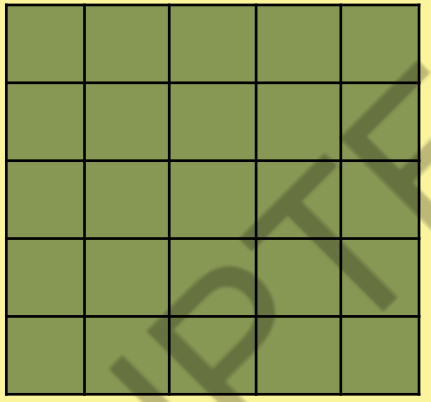


Stride

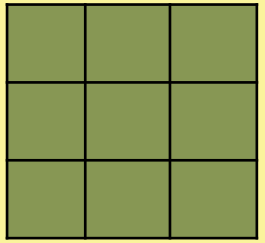
No. of steps the kernel is moved during convolution



7 x 7 Input Image
3 x 3 Kernel

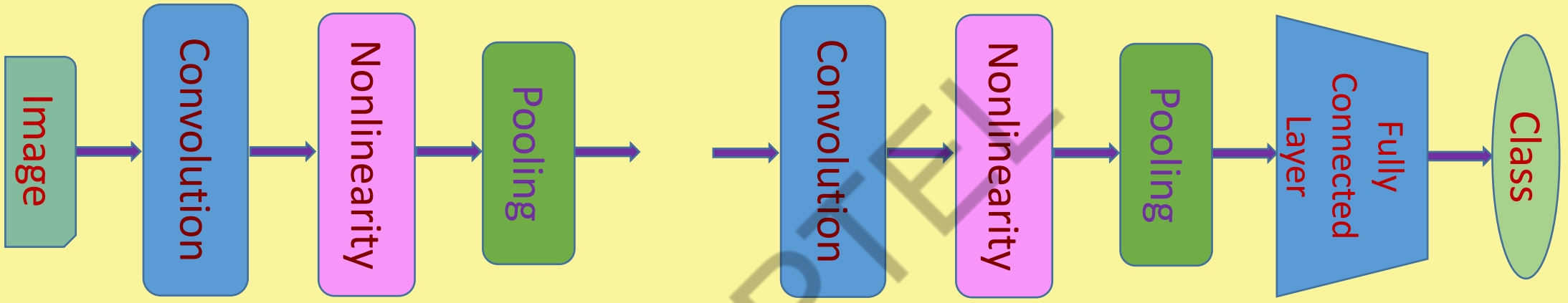


Stride =1



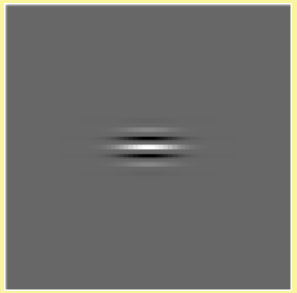
Stride =2

CNN Architecture



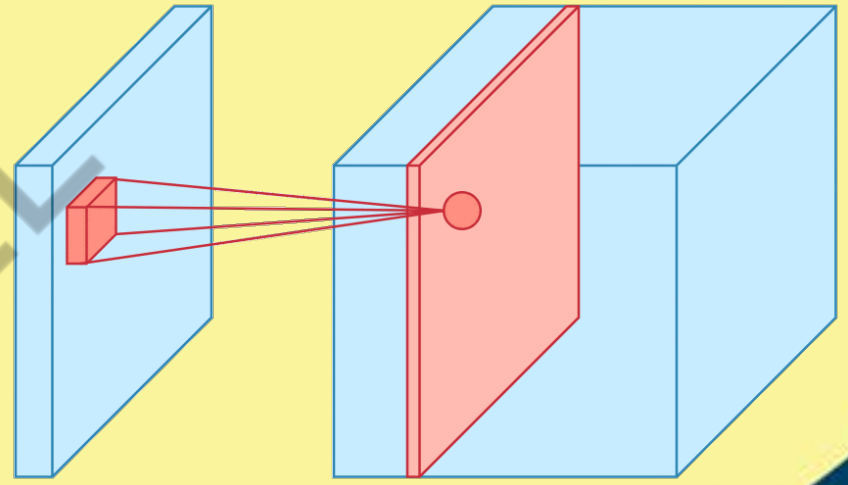
Convolution Layer: 3 D Convolution

- Color image has 3 dimensions: height, width and depth (depth is the color channels i.e RGB)
- Filter or kernels that will be convolved with the RGB image could also be 3D
- For multiple Kernels: All feature maps obtained from distinct kernels are stacked to get the final output of that layer



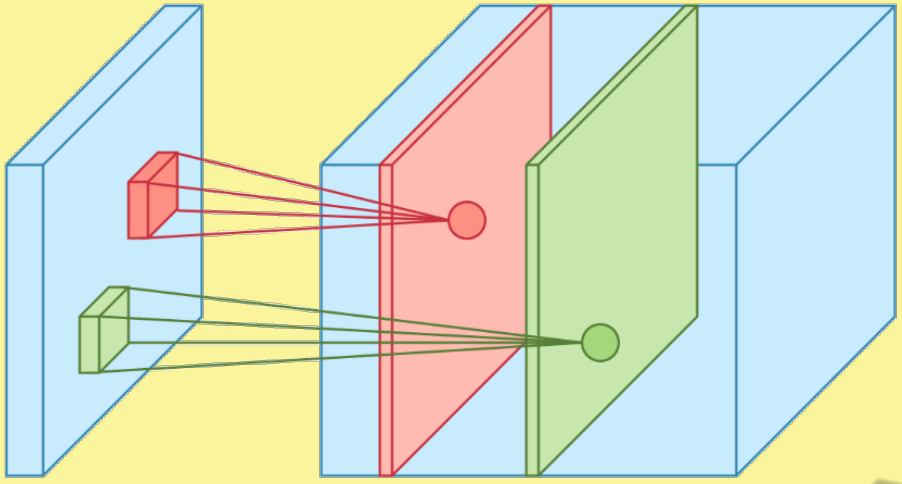
3 D Convolution- Visualization

- The kernel strides over the input Image.
- At each location $I(m,n)$ compute $f(m,n) = \sum\sum w(p,q)I(p-m,q-n)$ collect them in the feature map.
- The animation shows the sliding operation at 4 locations, but in reality it is performed over the entire input.



Animation:- Arden Dertat
https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2

3 D Convolution- Visualization



- Red and green boxes are two different featured maps obtained by convolving the same input with two different kernels. The feature maps are stacked along the depth dimension as shown.

Figure: Arden Dertat
https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2

3 D Convolution- Visualization

- An RGB Image of size 32X32X3
- 10 Kernels of size 5x5x3
- Output featuremap of size 32x32x10

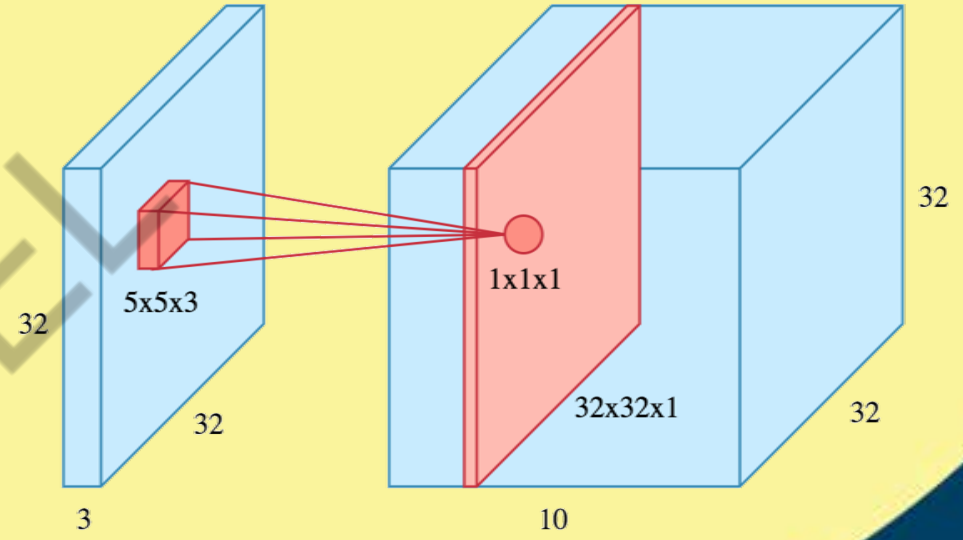


Figure: Arden Dertat
https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2

Nonlinearity

- ReLU is an element wise operation (applied per pixel) and replaces all negative pixel values in the feature map by zero

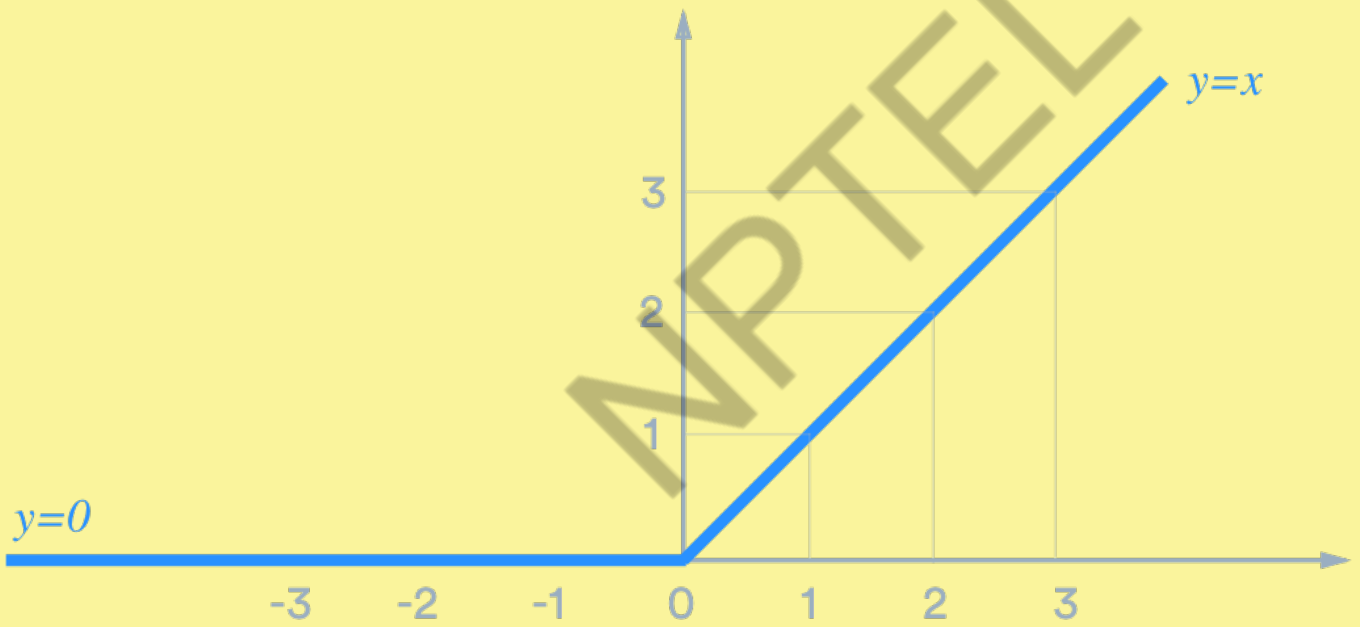


Figure: Arden Dertat
https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2

Poolin g

- Replaces the output of a node at certain locations with a summary statistic of nearby locations.
- Spatial Pooling can be of different types: Max, Average, Sum etc.
- Max Pooling report the maximum output within a rectangular neighborhood.
- Pooling helps to make the output approximately invariant to small translation.
- Pooling layers down sample each feature map independently, reducing the height and width, keeping the depth intact.
- In pooling layer stride and window size needs to be specified

Poolin g

- Figure below is the result of max pooling using a 2x2 window and stride 2. Each color denotes a different window. Since both the window size and stride are 2, the windows are not overlapping

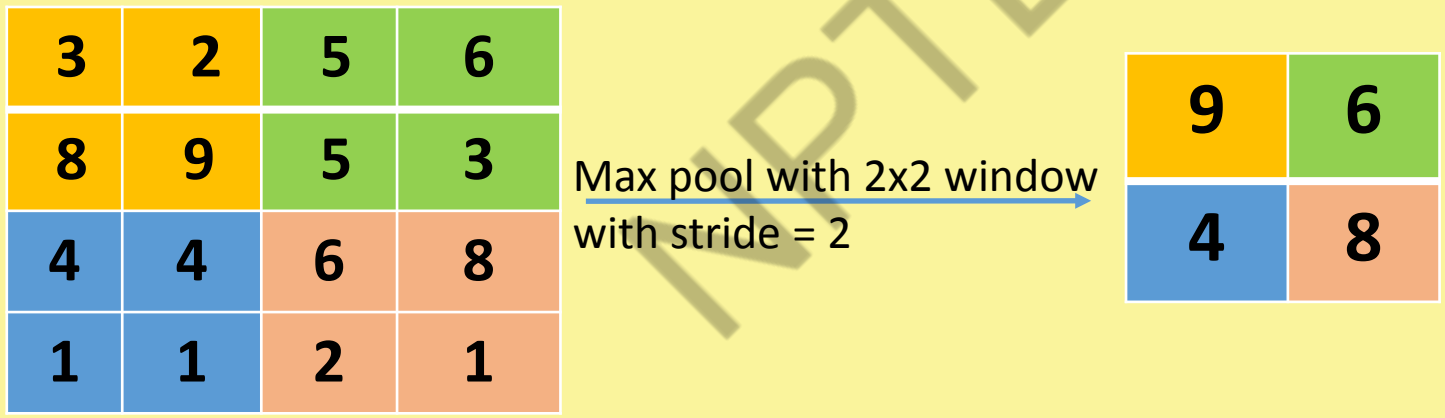


Figure: Arden Dertat
https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2

Poolin g

- Pooling reduces the height and the width of the feature map, but the depth remains unchanged as shown in figure
- Pooling operation is independently carried out across each depth

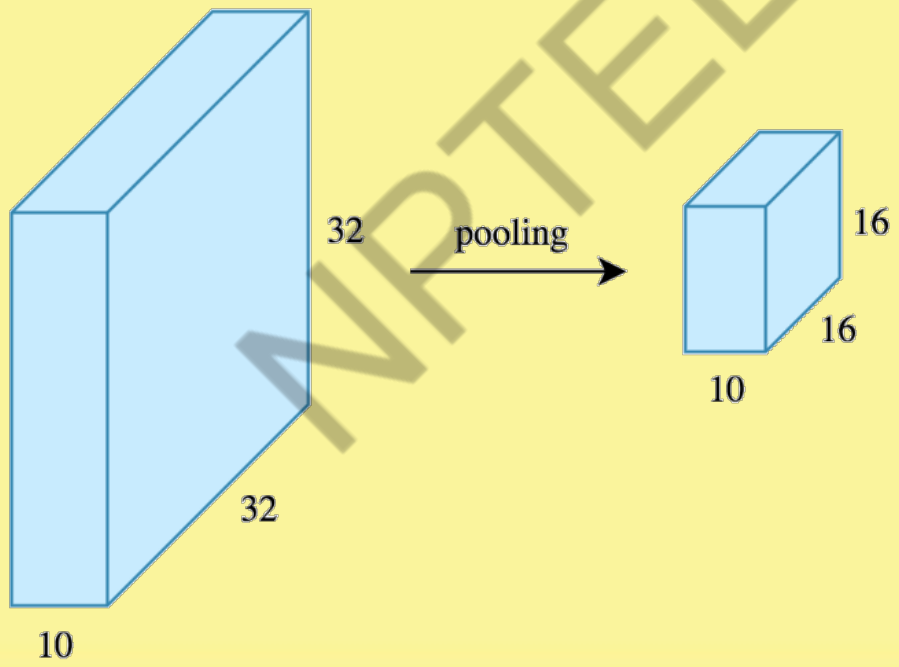
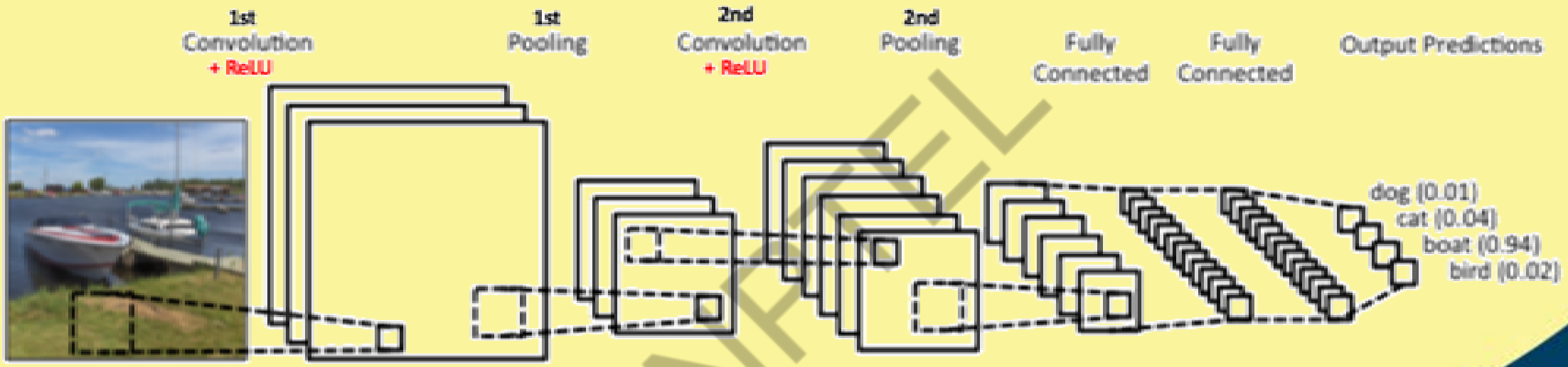


Figure: Arden Dertat
https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2

CNN Architecture



NPTEL ONLINE CERTIFICATION COURSES

Thank you

NPTEL ONLINE CERTIFICATION COURSES

Course Name: Deep Learning

Faculty Name: Prof. P. K. Biswas
Department : E & ECE, IIT Kharagpur

Topic

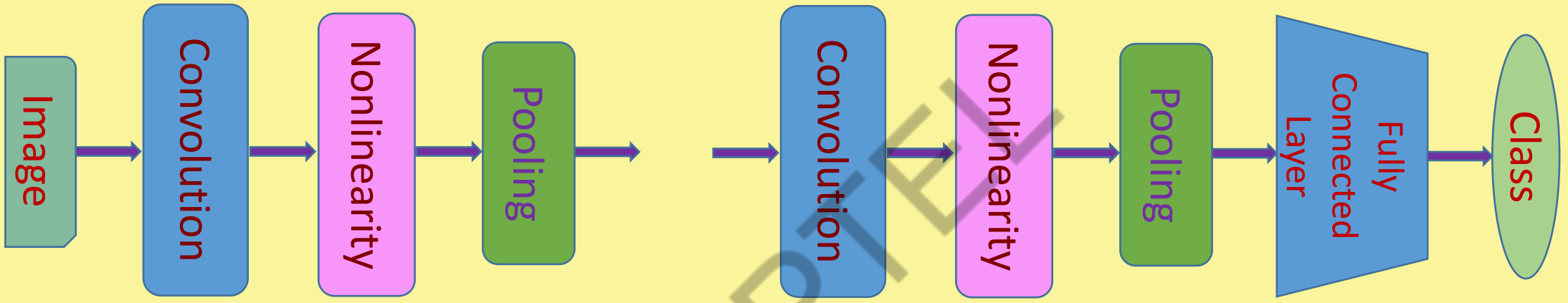
Lecture 37: Popular CNN Models

CONCEPTS COVERED

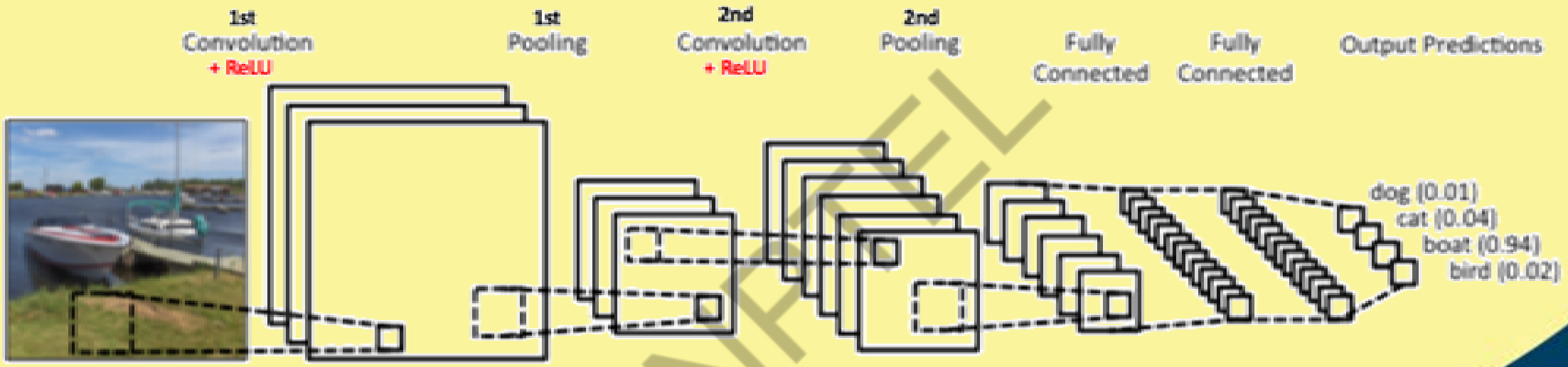
Concepts Covered:

- CNN
  - LeNet
  - AlexNet
  - VGG Net
  - GoogLeNet
  - etc.

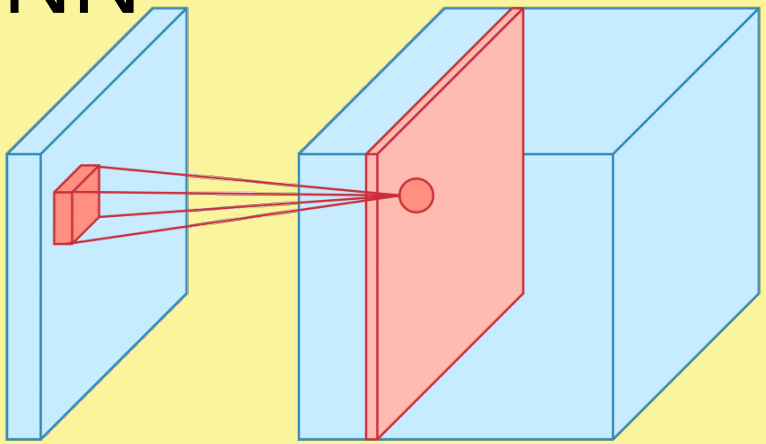
CNN Architecture

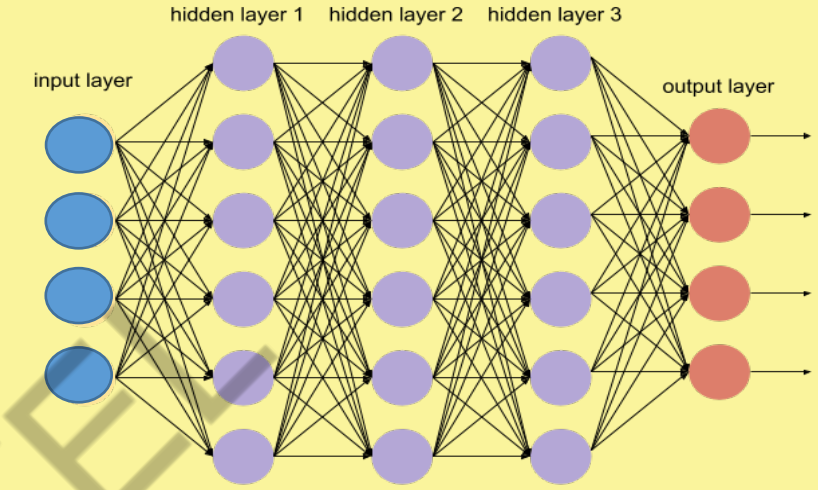


CNN Architecture



MLP vs CNN





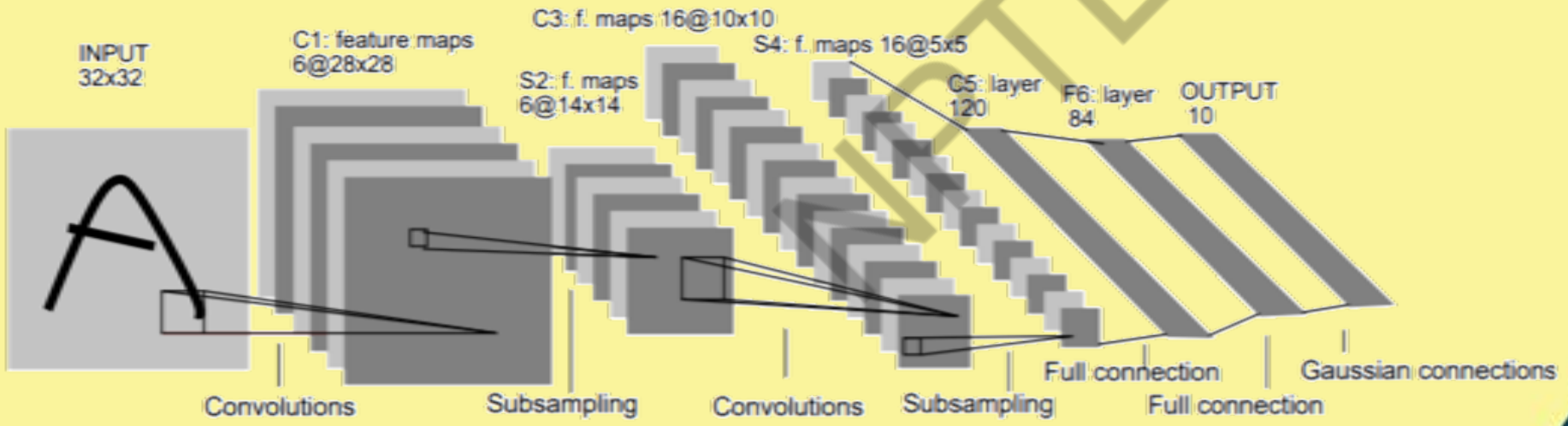
- Sparse Connectivity: Every node in the Convolution Layer receives input from a small number of nodes in the previous layer (Receptive Field), needing smaller number of parameters.
- Parameter Sharing: Each member of the Convolution Kernel is used at every position of the input, dramatically reducing the number of parameters.
- This makes CNN much more efficient than MLP.

Some popular CNN Models

LeNet

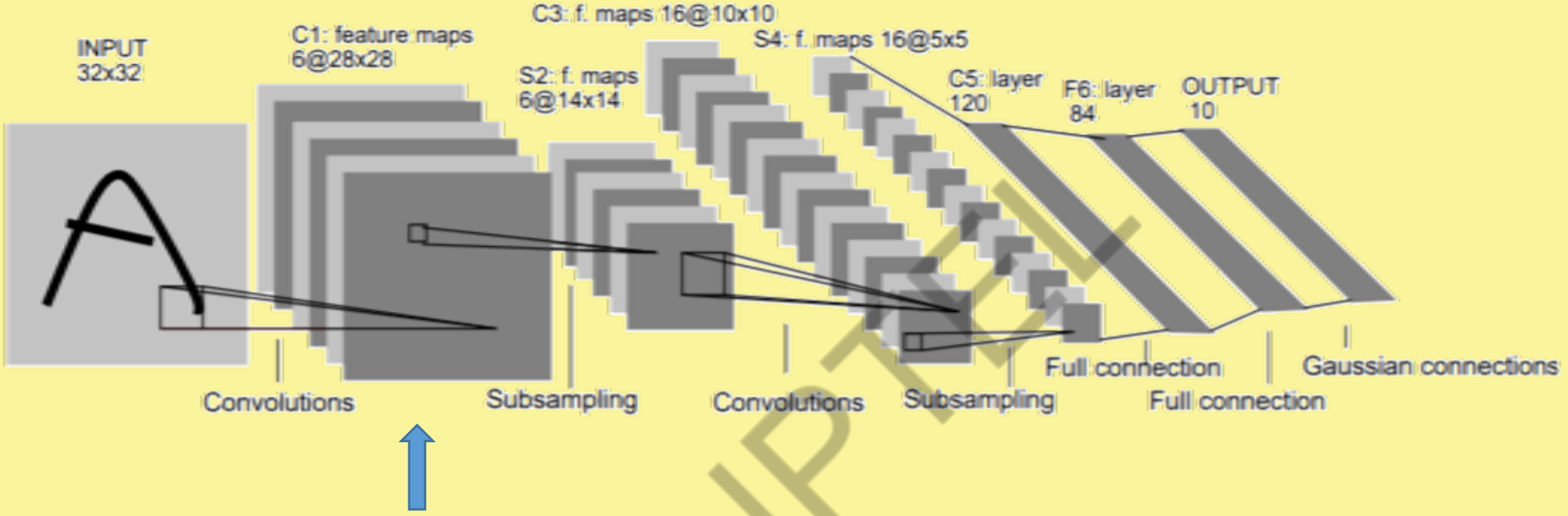
LeNet 5

- Proposed by Yann LeCun, Leon Bottou, Yosuha Bengio and Patrick Haffner for handwritten and machine-printed character recognition.
- Used by many Banks for recognition of hand written numbers on cheques.
- This architecture achieves an error rate as low as **0.95%** on test data



Yann LeCun, Leon Bottou, Yosuha Bengio and Patrick Haffner, “Gradient –Based Learning Applied to Document Recognition”, Proc. IEEE, Nov. 1998

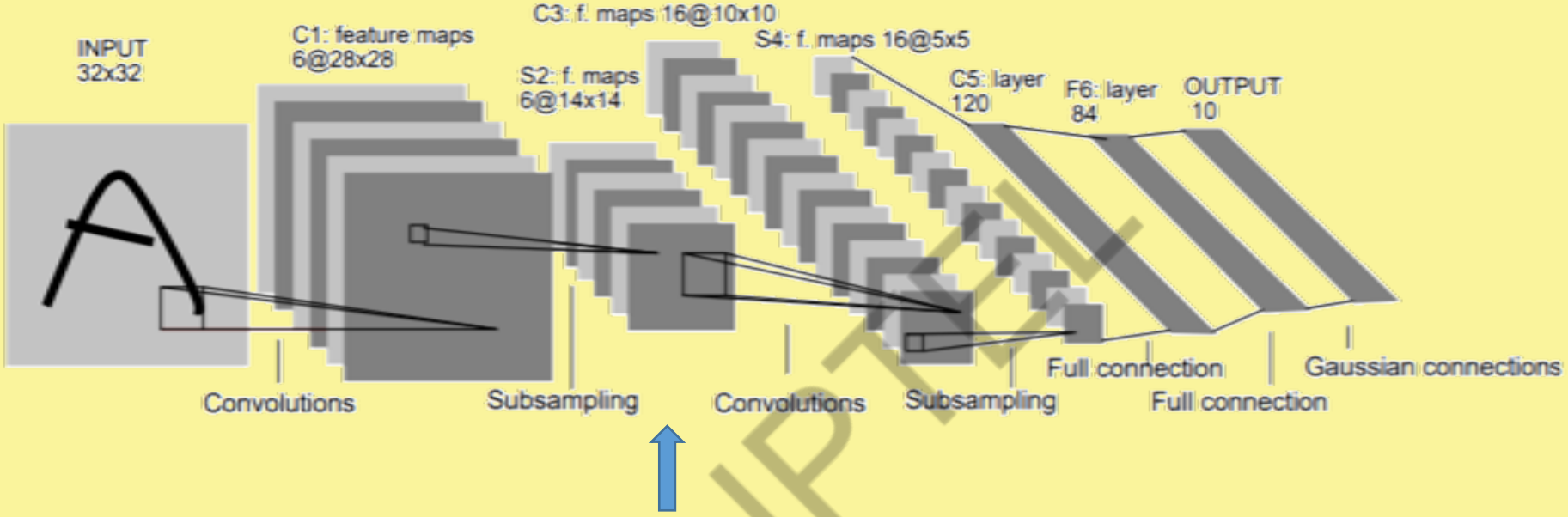
LeNet 5



No. of Kernels- 6
Kernel Size- 5 x 5
Stride- 1

Yann LeCun, Leon Bottou, Yosuha Bengio and Patrick Haffner, “Gradient –Based Learning Applied to Document Recognition”, Proc. IEEE, Nov. 1998

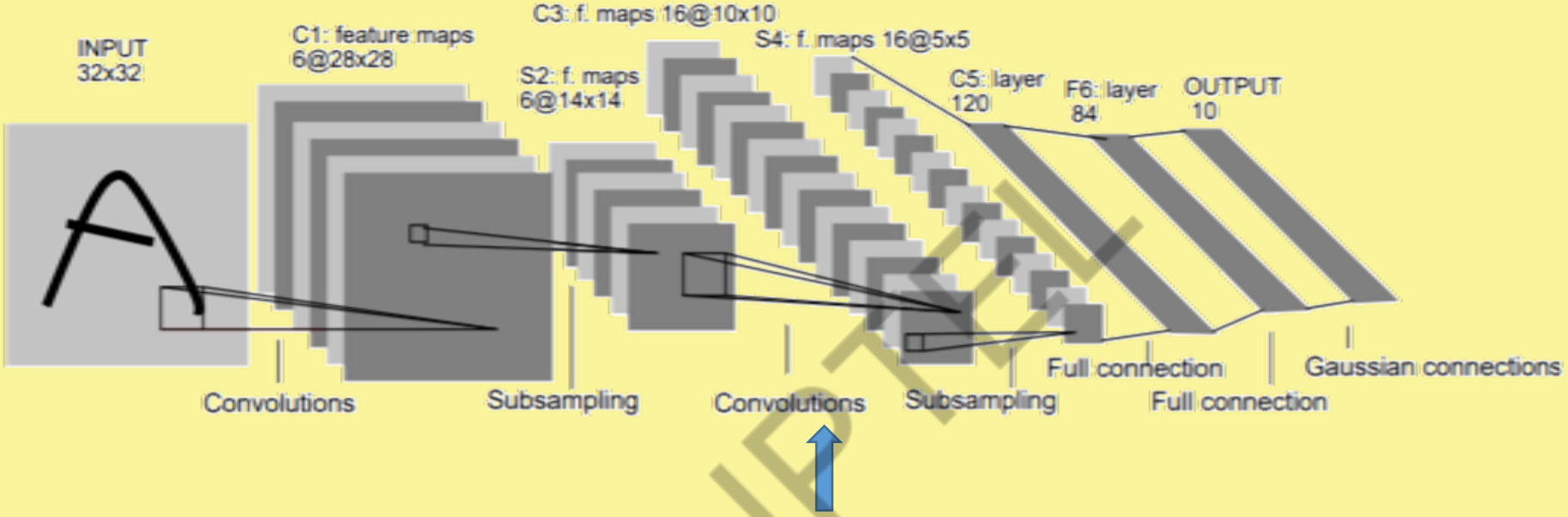
LeNet 5



Average Pooling
Window Size- 2 x 2
Stride- 2

Yann LeCun, Leon Bottou, Yosuha Bengio and Patrick Haffner, “Gradient –Based Learning Applied to Document Recognition”, Proc. IEEE, Nov. 1998

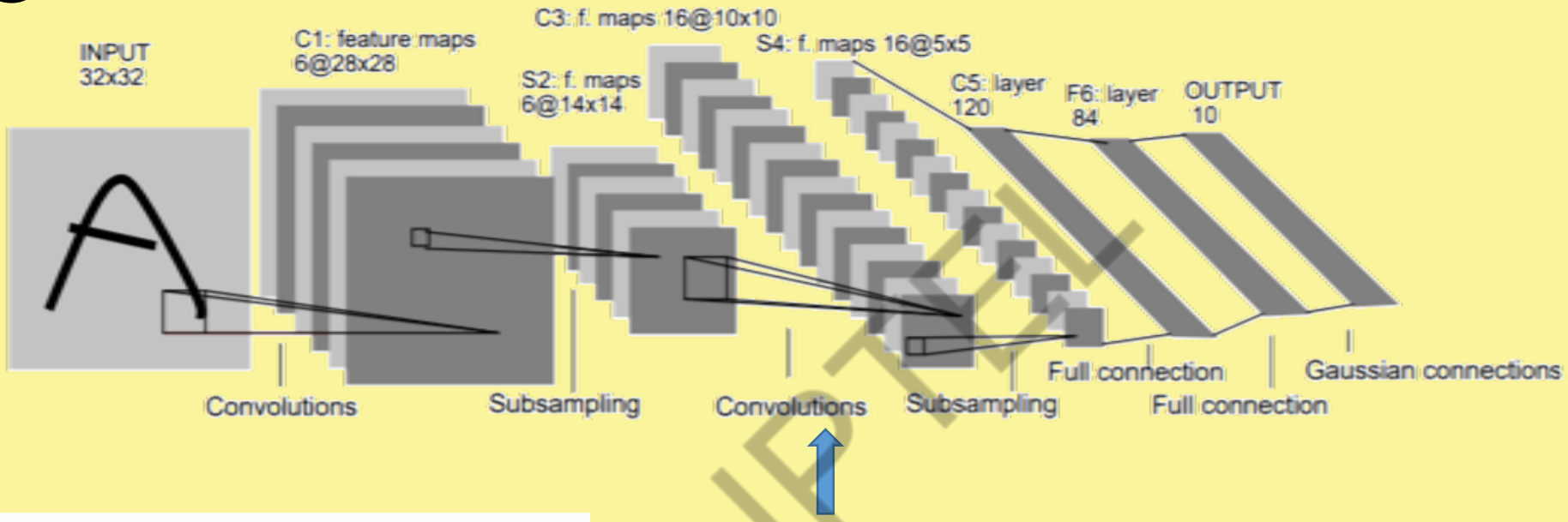
LeNet 5



No. of Kernels- 16
Kernel Size- 5 x 5
Stride- 1

Yann LeCun, Leon Bottou, Yosuha Bengio and Patrick Haffner, “Gradient –Based Learning Applied to Document Recognition”, Proc. IEEE, Nov. 1998

LeNet 5

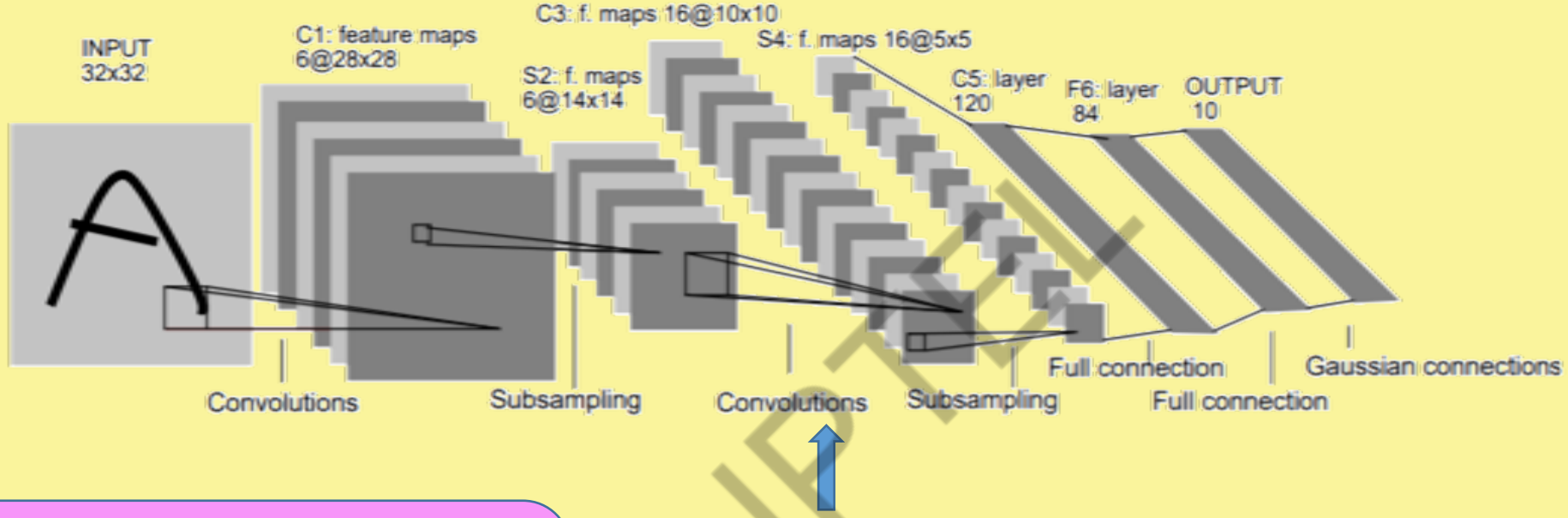


	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	X				X	X	X			X	X	X	X		X	X
1	X	X				X	X	X			X	X	X	X		X
2	X	X	X				X	X	X			X		X	X	X
3		X	X	X			X	X	X	X			X		X	X
4			X	X	X			X	X	X	X		X	X		X
5				X	X	X			X	X	X	X		X	X	X

No. of Kernels- 16
Kernel Size- 5 x 5
Stride- 1

Yann LeCun, Leon Bottou, Yosuha Bengio and Patrick Haffner, “Gradient –Based Learning Applied to Document Recognition”, Proc. IEEE, Nov. 1998

LeNet 5

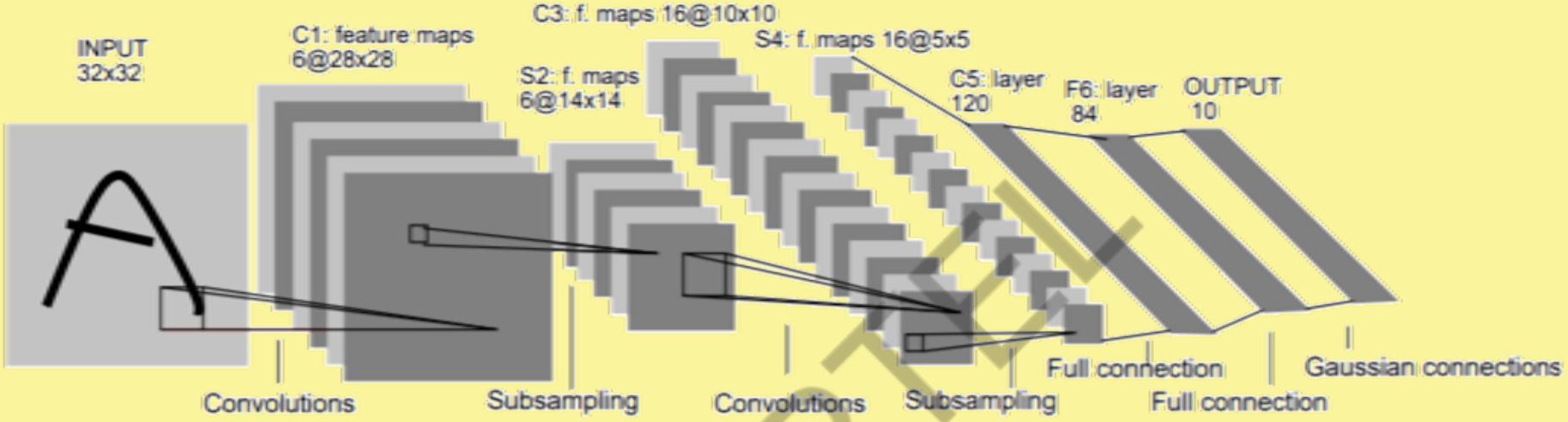


- *Break the symmetry in the network*
- *Keep number of connections within reasonable bounds.*

No. of Kernels- 16
Kernel Size- 5 x 5
Stride- 1

Yann LeCun, Leon Bottou, Yosuha Bengio and Patrick Haffner, “Gradient –Based Learning Applied to Document Recognition”, Proc. IEEE, Nov. 1998

LeNet 5



Average Pooling
Window Size- 2 x 2
Stride- 2

Yann LeCun, Leon Bottou, Yosuha Bengio and Patrick Haffner, “Gradient –Based Learning Applied to Document Recognition”, Proc. IEEE, Nov. 1998

LeNet 5: Summary

Layer		Feature Map	Size	Kernel Size	Stride	Activation
Input	Image	1	32x32	-	-	-
1	Convolution	6	28x28	5x5	1	tanh
2	Average Pooling	6	14x14	2x2	2	tanh
3	Convolution	16	10x10	5x5	1	tanh
4	Average Pooling	16	5x5	2x2	2	tanh
5	Convolution	120	1x1	5x5	1	tanh
6	FC	-	84	-	-	tanh
Output	FC	-	10	-	-	softmax

IMAGENET Large Scale Visual Recognition Challenge (ILSVRC)

https://engmrk.com/lenet-5-a-classic-cnn-architecture/

ILSVRC

- IMAGENET Large Scale Visual Recognition Challenge.
- Evaluates algorithms for Object Detection and Image Classification on large image database.
- Helps researchers to review state of the art Machine Learning techniques for object detection across a wider variety of objects.
- Monitor the progress of computer vision for large scale image indexing for retrieval and annotation.
- Database contains large number of Images from 1000 categories.
- More than 1000 images in every category.

ILSVRC

- Every year of the challenge the forum also organizes a workshop at one of the premier computer vision conferences.
- The purpose of the workshop is to disseminate the new findings of the challenge.
- Contestants with the most successful and innovative techniques are invited to present their work.

NPTEL ONLINE CERTIFICATION COURSES

Thank you

NPTEL ONLINE CERTIFICATION COURSES

Course Name: Deep Learning

Faculty Name: Prof. P. K. Biswas
Department : E & ECE, IIT Kharagpur

Topic

Lecture 38: Popular CNN Models II

CONCEPTS COVERED

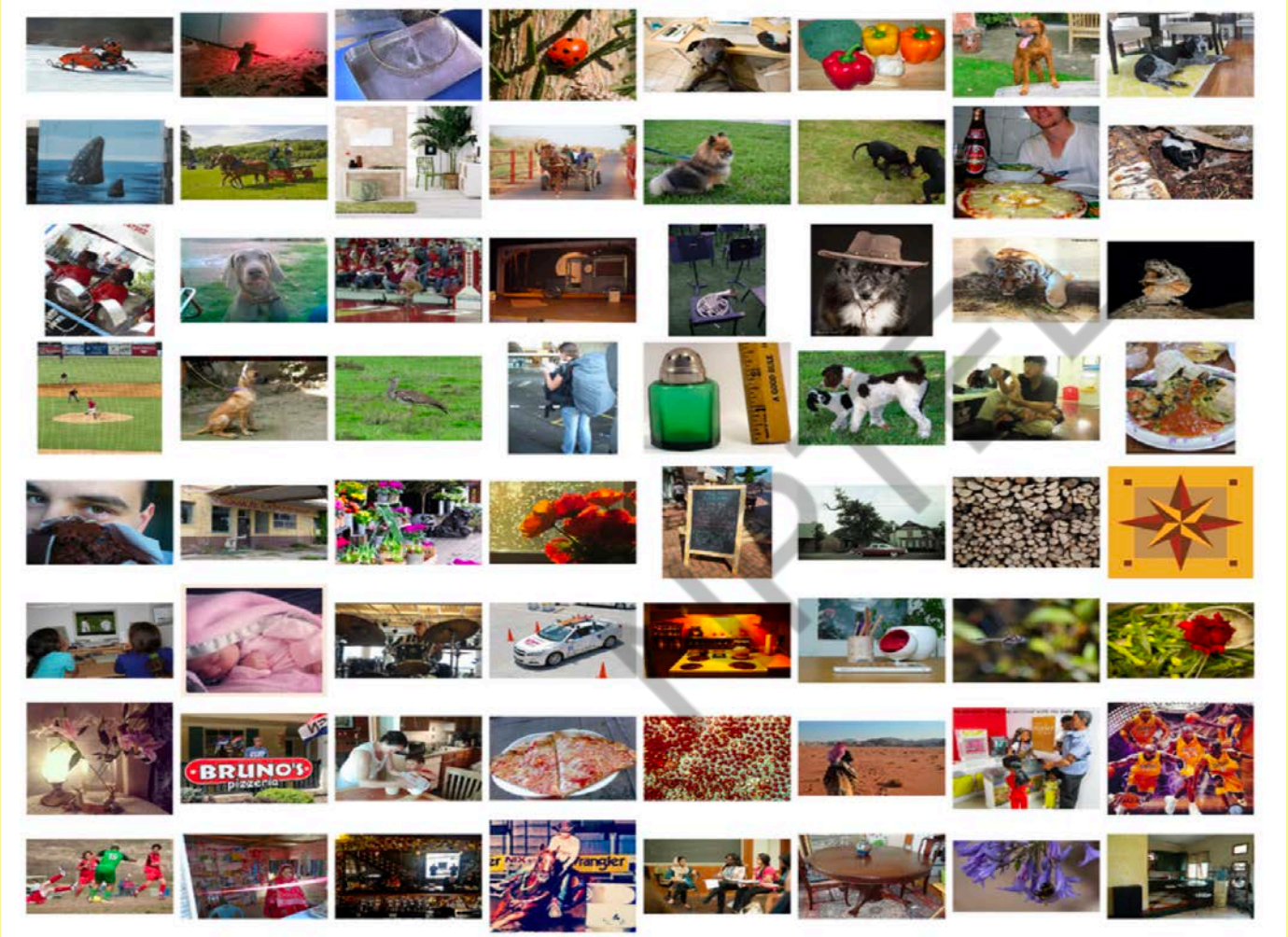
Concepts Covered:

- CNN
  - LeNet
  - ILSVRC
  - AlexNet
  - VGG Net
  - GoogLeNet
  - etc.

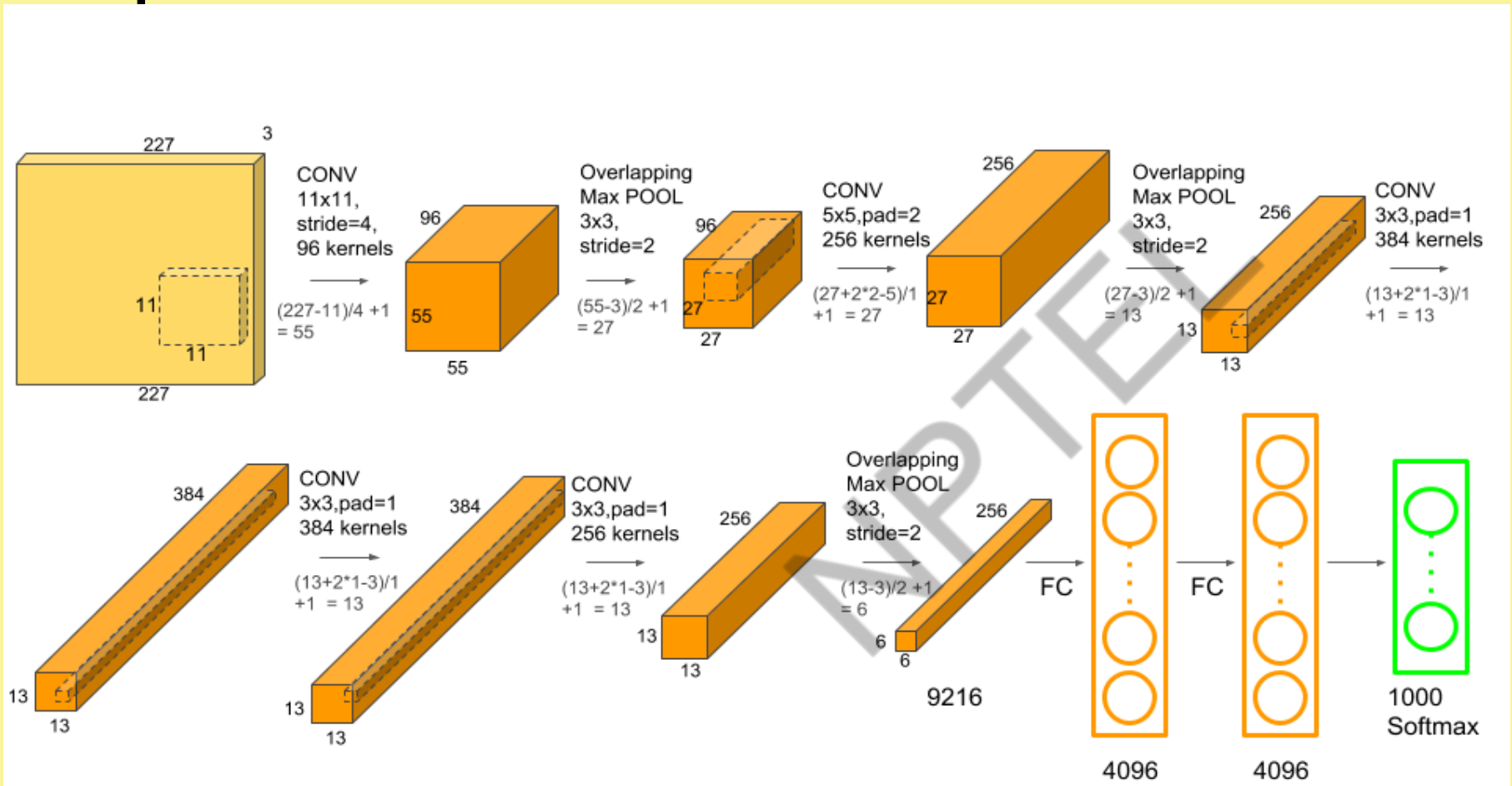
AlexNet
ILSVRC 2012 Winer

Krizhevsky Alex, Ilya Sutskever and Geoffrey E. Hilton, “Imagenet Classification with deep convolutional neural networks”, Advances in Neural Information Processing Systems, 2012

Sample Images from ImageNet Dataset



AlexNe
t

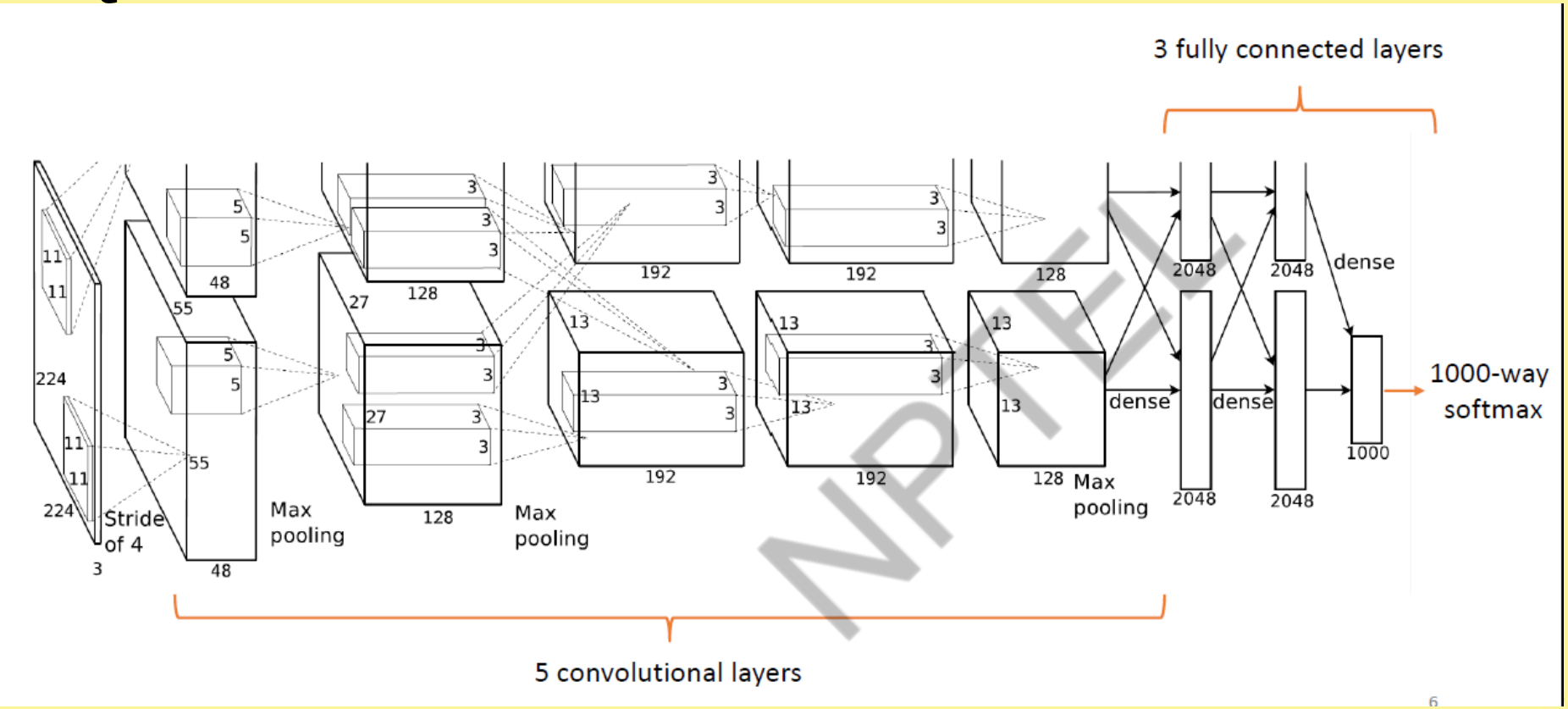


https://www.learnopencv.com/understanding-alexnet/

ILSVRC 2012
Winner

https://www.learnopencv.com/understanding-alexnet/

AlexNet



Krizhevsky Alex, Ilya Sutskever and Geoffrey E. Hilton, "Imagenet Classification with deep convolutional neural networks", Advances in Neural Information Processing Systems, 2012

AlexNet

- 60 Million parameters and 650000 neurons.
- The network is split into two pipelines and was trained on two GPU.
- Input Image size 256 x 256 RGB.
- Grey scale images to be replicated to obtain 3-Channel RGB
- Random crops of size 227 x 227 are fed to the input layer of AlexNet.
- Stochastic Gradient Descent with Momentum Optimizer.
- Top-5 error rate 15.3%.

Krizhevsky Alex, Ilya Sutskever and Geoffrey E. Hilton, “Imagenet Classification with deep convolutional neural networks”, Advances in Neural Information Processing Systems, 2012

Vanishing Gradient Problem

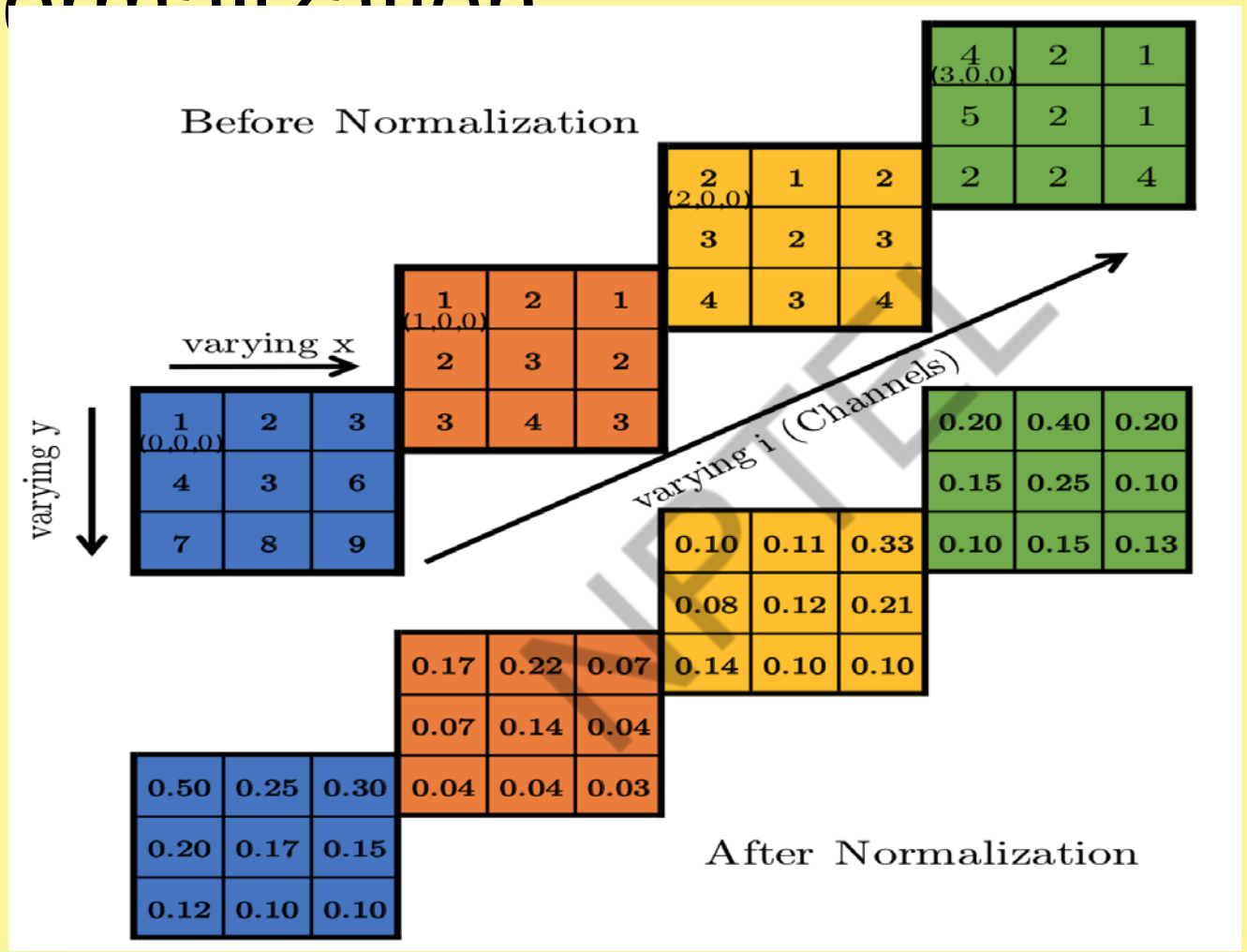
- Uses ReLU activation instead of sigmoidal function.
- ReLU output is unbounded- uses Local Response Normalization (LRN).
- LRN carries out a normalization amplifying the excited neuron while dampening the surrounding neurons at the same time in a local neighbourhood.
- Encourage *Lateral Inhibition*: concept in neuro biology that indicates capacity of a neuron to reduce activity of its neighbours.

Krizhevsky Alex, Ilya Sutskever and Geoffrey E. Hilton, "Imagenet Classification with deep convolutional neural networks", Advances in Neural Information Processing Systems, 2012

Local Response Normalization (Inter-Channel)

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

Local Response Normalization

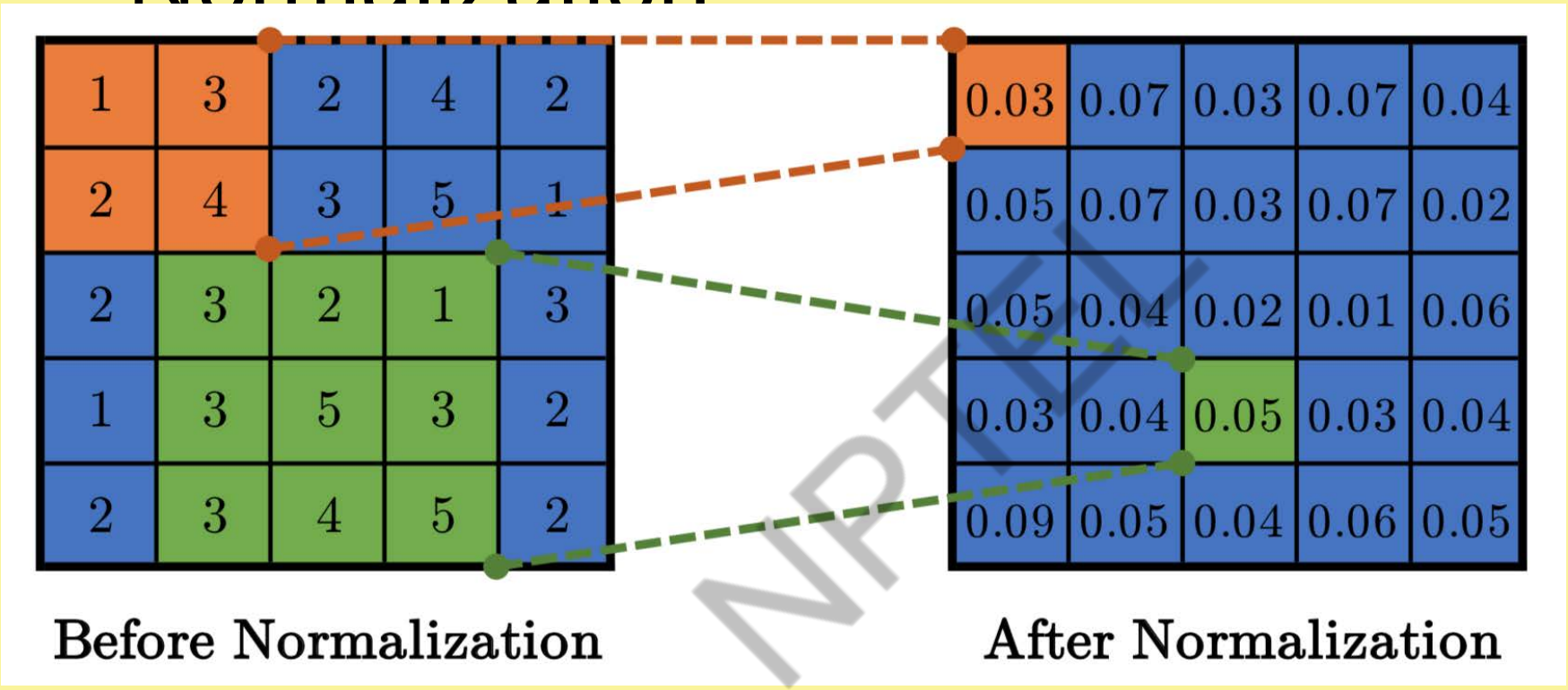


https://towardsdatascience.com/difference-between-local-response-normalization-and-batch-normalization-272308c034ac

Local Response Normalization (Intra-Channel)

$$b_{x,y}^{i} = \frac{a_{x,y}^{i}}{\left(k + \alpha \sum_{p=\max(0,x-n/2)}^{\max(W,x+n/2)} \sum_{q=\max(0,y-n/2)}^{\min(H,y+n/2)} \left(a_{p,q}^{i}\right)^2\right)^{\beta}}$$

Local Response Normalization



https://towardsdatascience.com/difference-between-local-response-normalization-and-batch-normalization-272308c034ac

Reducing Overfitting

- Train the network with different variants of the same image helps avoiding overfitting.
  - Generate additional data from existing data (Augmentation).
  - Data augmentation by mirroring.
  - Data Augmentation by random crops.
- Dropout Regularization.

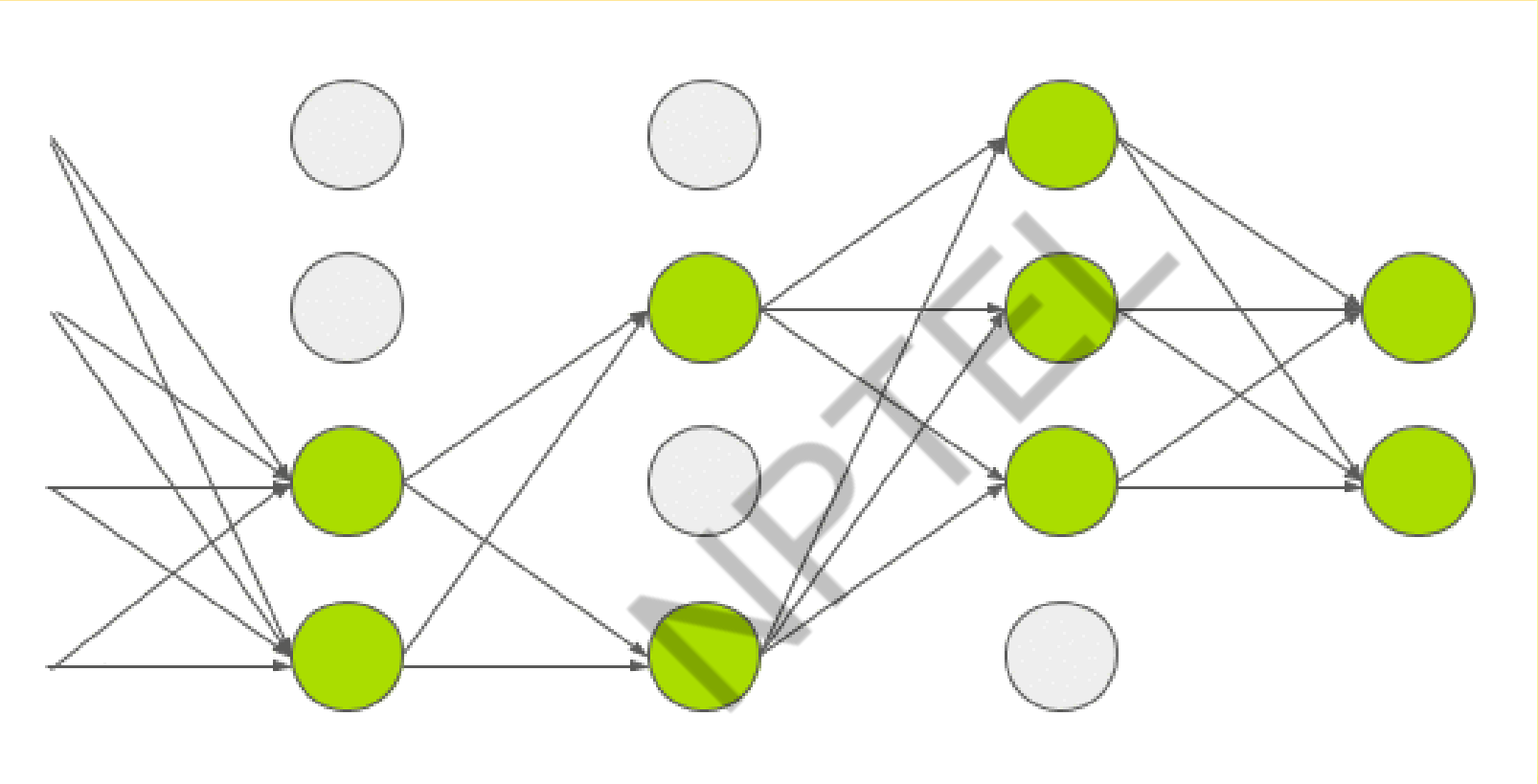
Krizhevsky Alex, Ilya Sutskever and Geoffrey E. Hilton, “Imagenet Classification with deep convolutional neural networks”, Advances in Neural Information Processing Systems, 2012

Dropout

- Regularization Technique proposed by Srivastava et. al. in 2014.
- During training randomly selected neurons are dropped from the network (with probability 0.5) temporarily .
- Their activations are not passed to the downstream neurons in the forward pass.
- In the backward pass weight updates are not applied to theses neurons.

Srivastava Nitish et. al. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting” Journal of Machine Learning Research 15 (2014), 1929-1958

Dropout



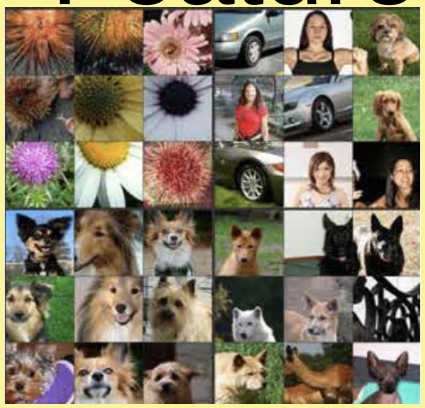
https://www.learnopencv.com/understanding-alexnet/

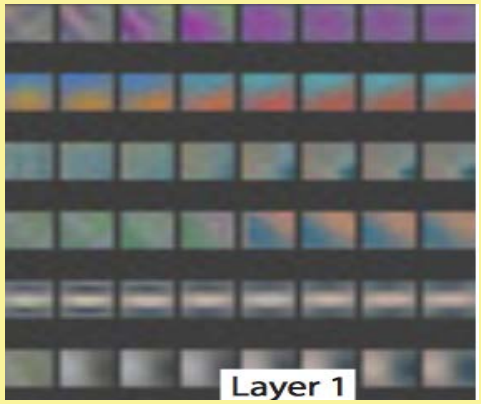
How does it help?

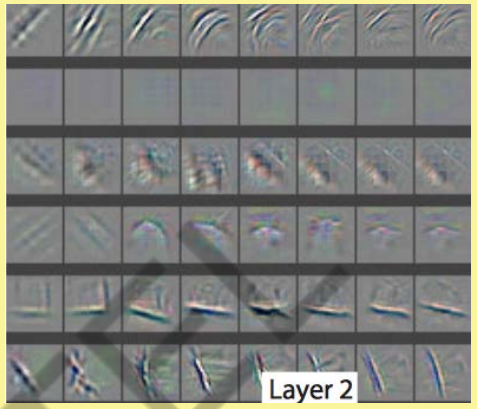
- While training weights of neurons are tuned for specific features that provides some sort of specialization.
- Neighbouring neurons starts relying on these specializations (co-adaptation).
- This leads to a neural network model too specialized to the training data.
- As neurons are randomly dropped other neurons have to step in to compensate.
- Thus the network learns multiple independent representations

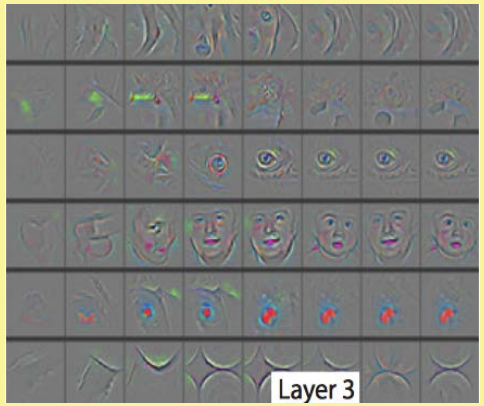
Srivastava Nitish et. al. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting” Journal of Machine Learning Research 15 (2014), 1929-1958

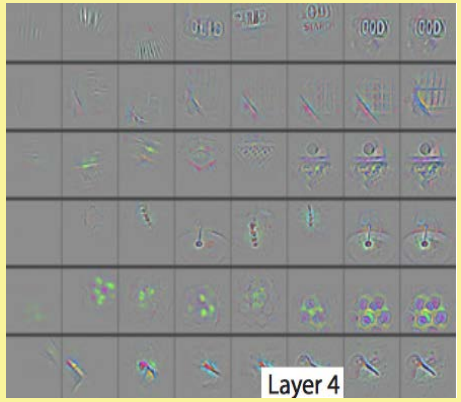
Learned Features

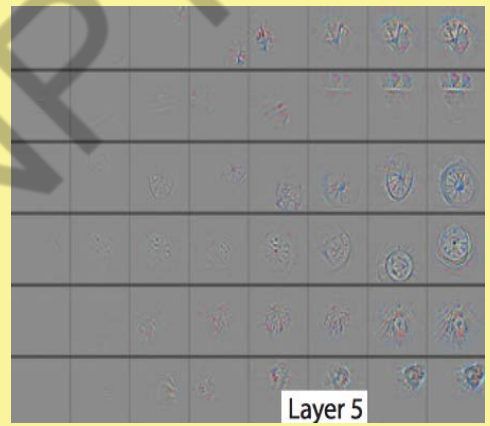












How does it help?

- This makes the network less sensitive to specific weights.
- Enhances the generalization capability of the network
- Less vulnerable to overfitting.
- The whole network is used during testing – there is no dropout.
- Dropout increases number of iterations for the network to converge.
- But helps avoid overfitting.

Srivastava Nitish et. al. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting” Journal of Machine Learning Research 15 (2014), 1929-1958

NPTEL ONLINE CERTIFICATION COURSES

Thank you

NPTEL ONLINE CERTIFICATION COURSES

Course Name: Deep Learning

Faculty Name: Prof. P. K. Biswas
Department : E & ECE, IIT Kharagpur

Topic

Lecture 39: Popular CNN Models III

CONCEPTS COVERED

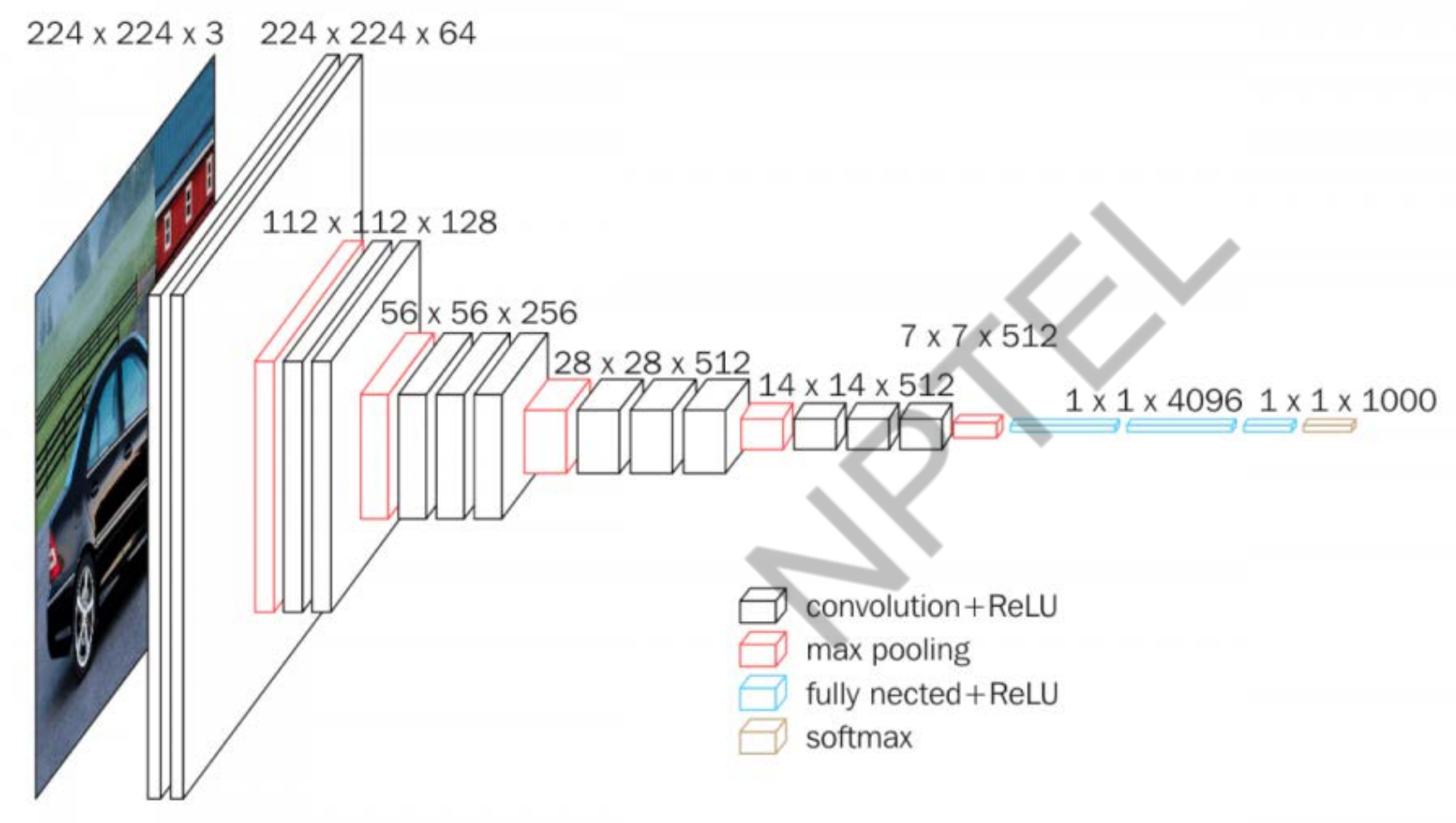
Concepts Covered:

- CNN
  - AlexNet
  - VGG Net
  - Transfer Learning
  - GoogLeNet
  - ResNet
  - etc.

VGG 16
ILSVRC 2014 1st Runner-Up

Visual Geometry Group
Oxford University

VGG 16



Very Deep Convolutional Networks for Large-Scale Image Recognition by Karen Simonyan and Andrew Zisserman

VGG 16

- Input to the architecture are color images of size 224x224.
- The image is passed through a stack of convolutional layers.
- Every convolution filter has very small receptive field: 3×3, Stride 1.
- Uses row and column padding to maintain spatial resolution after convolution.
- There are 13 Convolution Layers.
- There are 5 max-pool layers.
- Max pooling window size 2x2, stride 2.

VGG 16

- Not every convolution layer is followed by max-pool layer.
- 3 Fully connected layers.
- First two FC layers have 4096 channels each.
- Last FC layer has 1000 channels.
- Last layer is a softmax layer with 1000 channels, one for each category of images in ImageNet database.
- Hidden layers have ReLU as activation function.

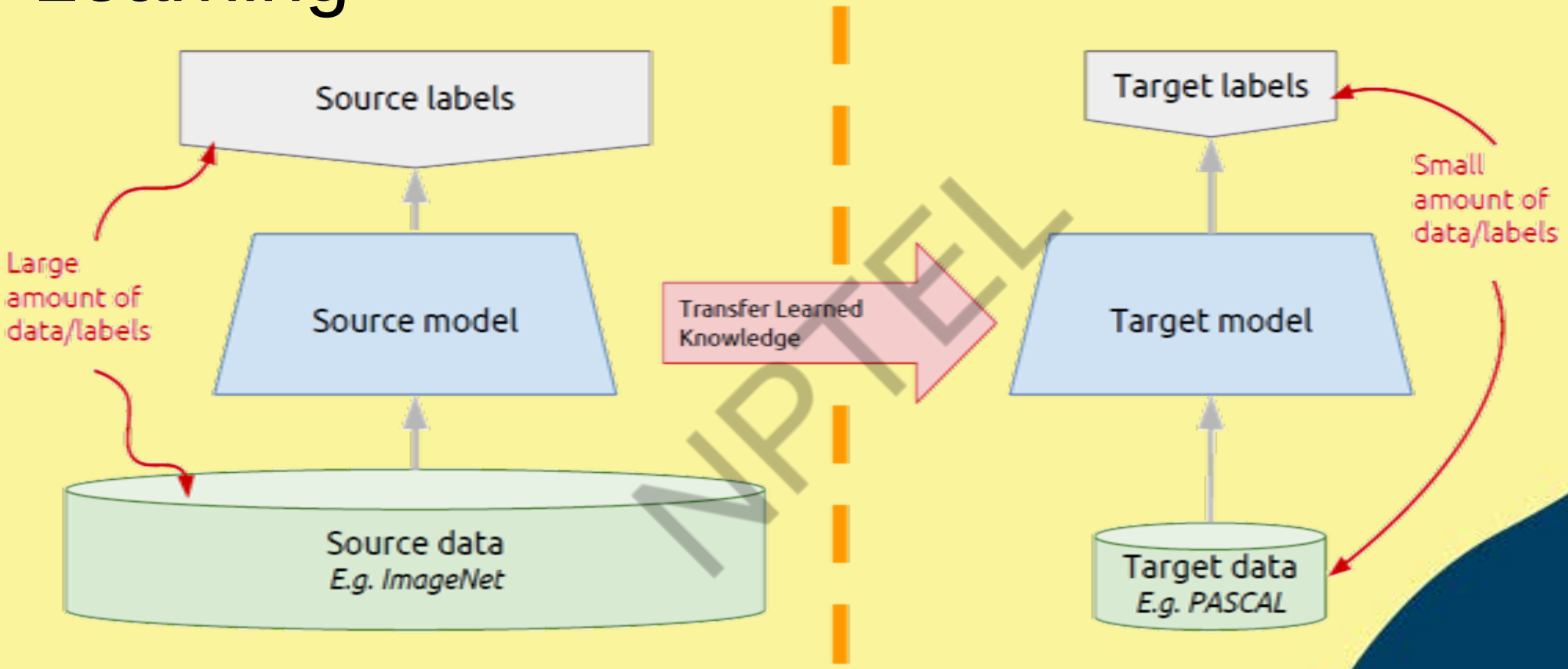

VGG 16

Striking difference from AlexNet

- All convolution kernels are of size 3x3 with stride 1.
- All maxpool kernels are of size 2x2 stride 2
- All variable size kernels as in AlexNet can be realised using multiple 3x3 kernels.
- This realisation is in terms of size of the receptive field covered by the kernels.
- Top-5 error rate ~ 7 %

Transfer Learning

Transfer Learning



Kevin McGuinness
https://www.slideshare.net/xavigiro/transfer-learning-d2l4-insightdcu-machine-learning-workshop-2017

Transfer Learning

CNN as Fixed Feature Extractor:

- Take a pre-trained CNN architecture trained on a large dataset (like ImageNet)
- Remove the last fully connected layer of this pre-trained network
- Remaining CNN acts as a fixed feature extractor for the new dataset

Transfer Learning

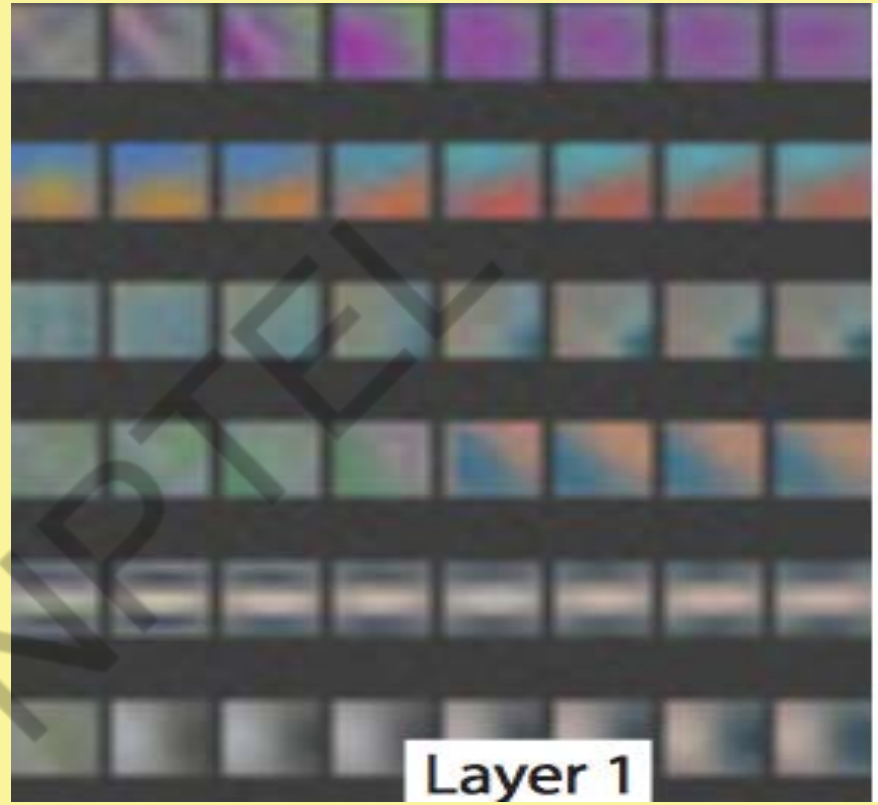


Image Source:-
https://becominghuman.ai/what-exactly-does-cnn-see-4d436d8e6e52

Transfer Learning

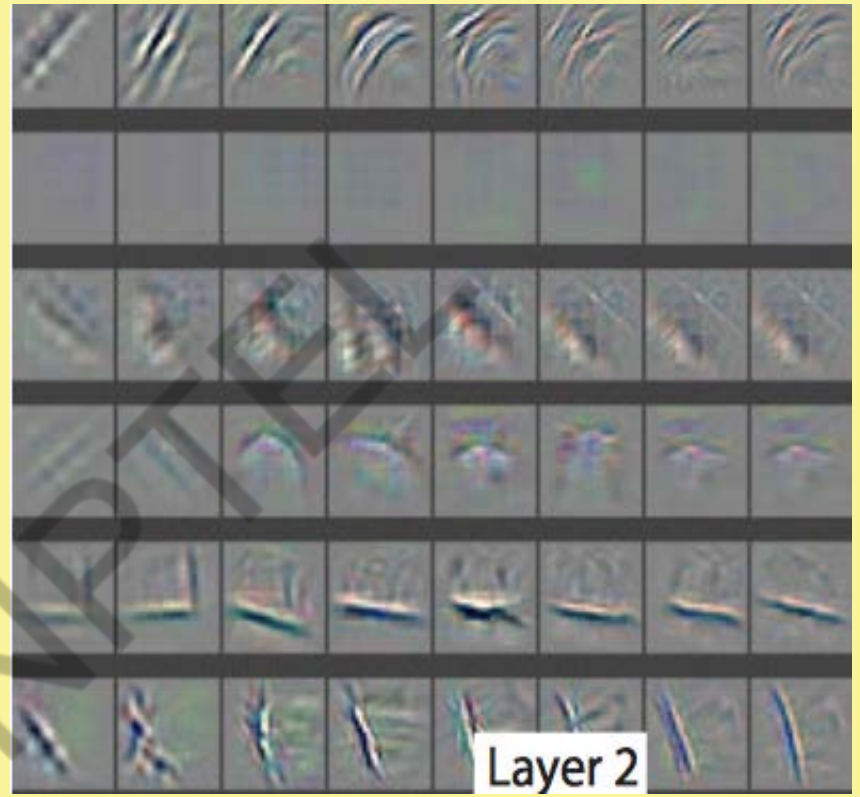
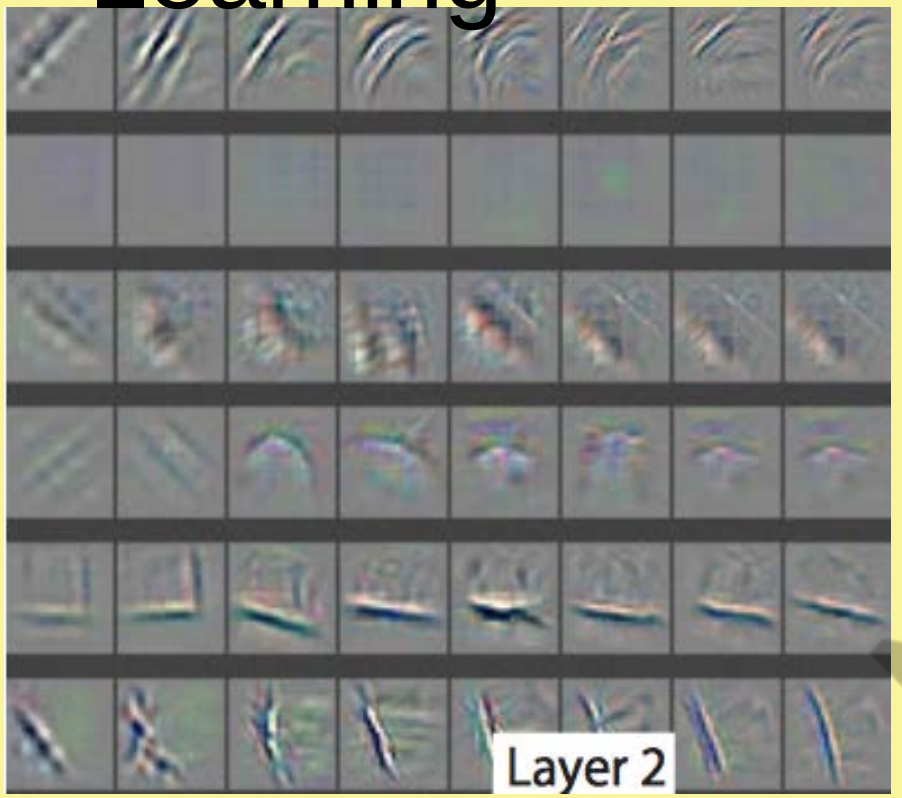


Image Source:-
https://becominghuman.ai/what-exactly-does-cnn-see-4d436d8e6e52

Transfer Learning



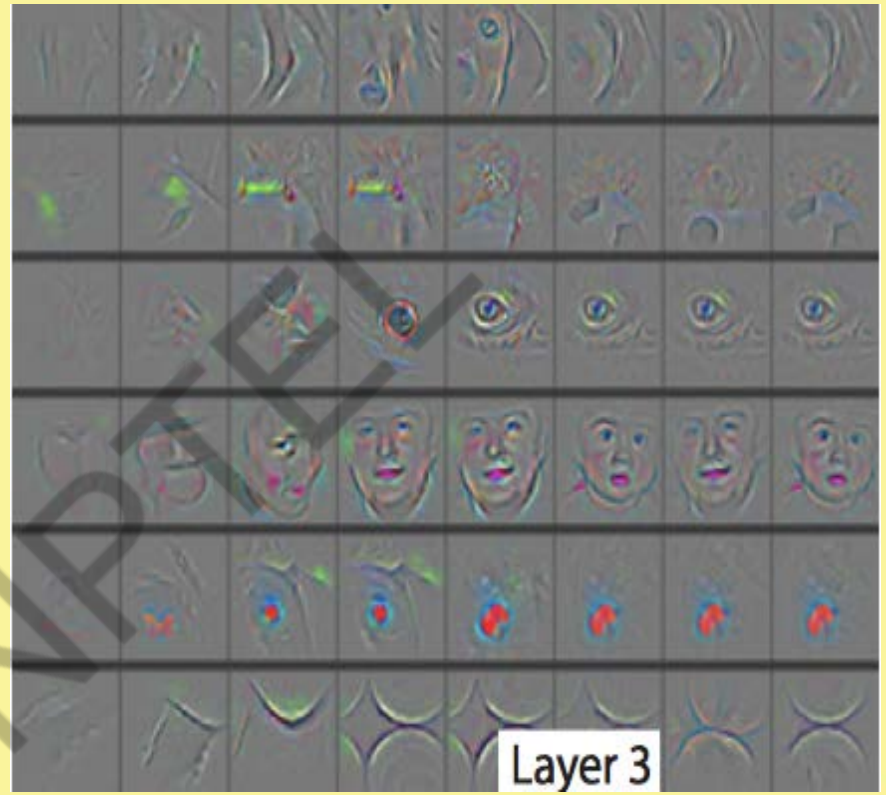
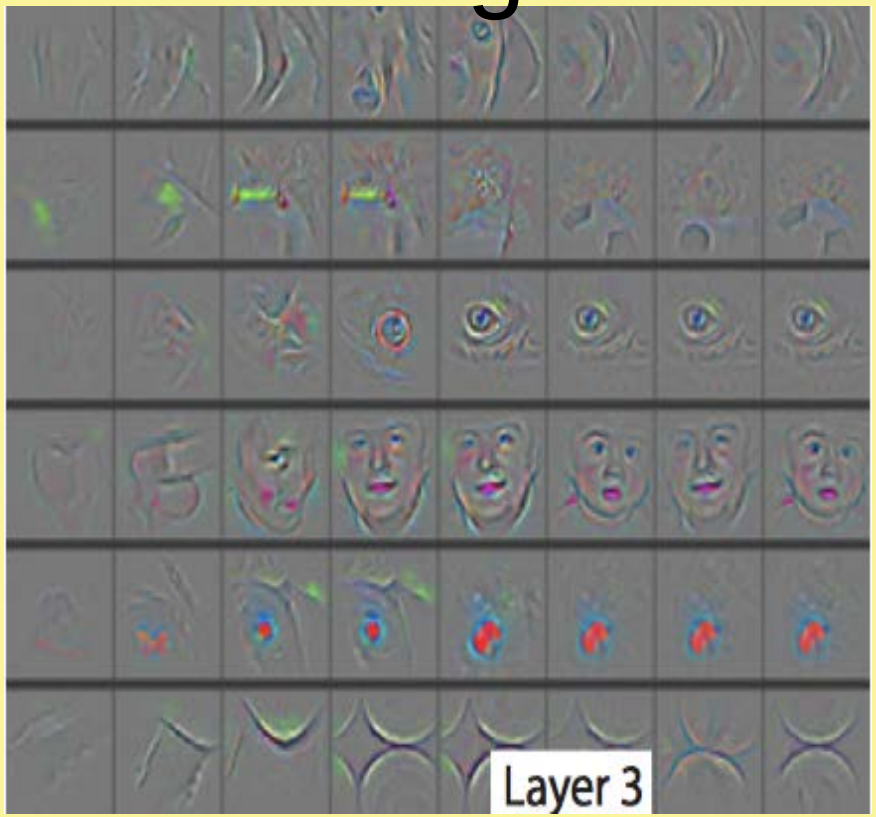


Image Source:- https://becominghuman.ai/what-exactly-does-cnn-see-4d436d8e6e52

Transfer Learning



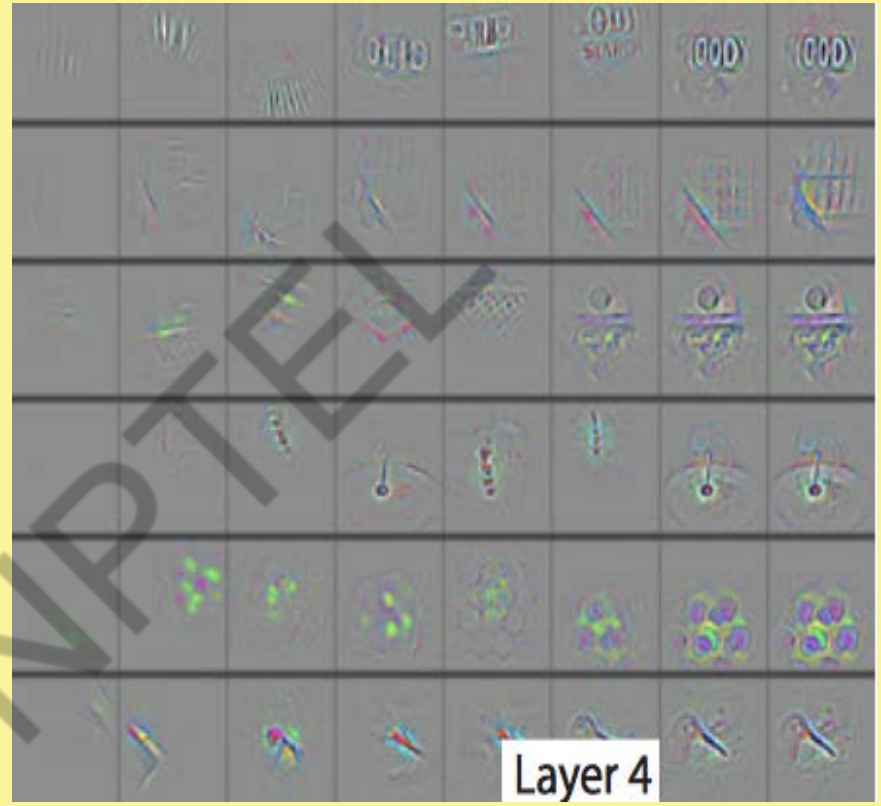
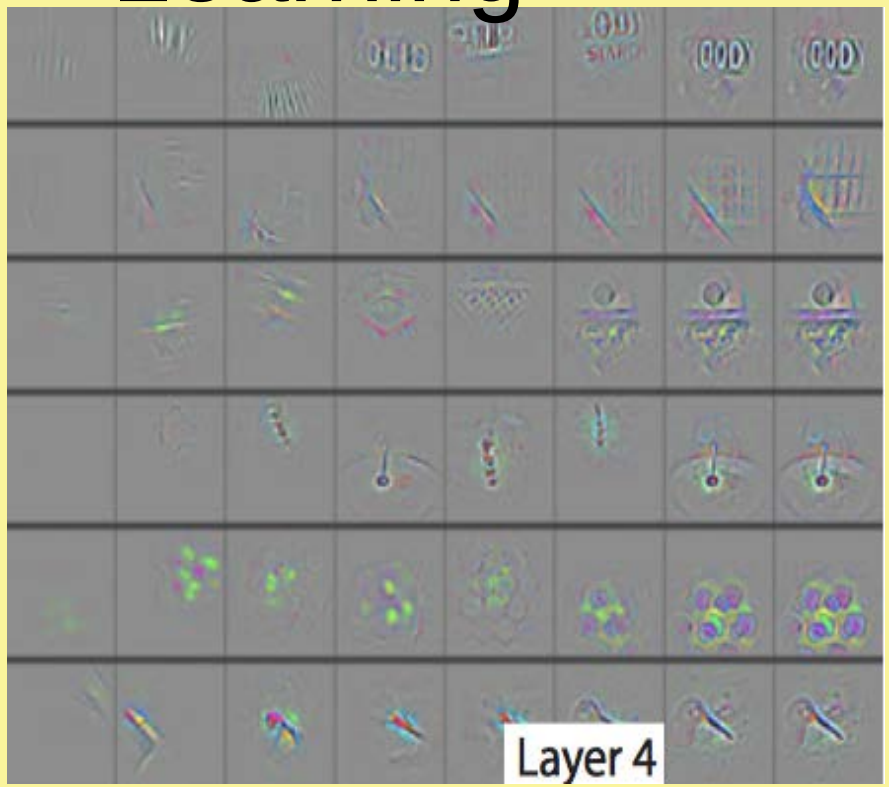


Image Source:- https://becominghuman.ai/what-exactly-does-cnn-see-4d436d8e6e52

Transfer Learning



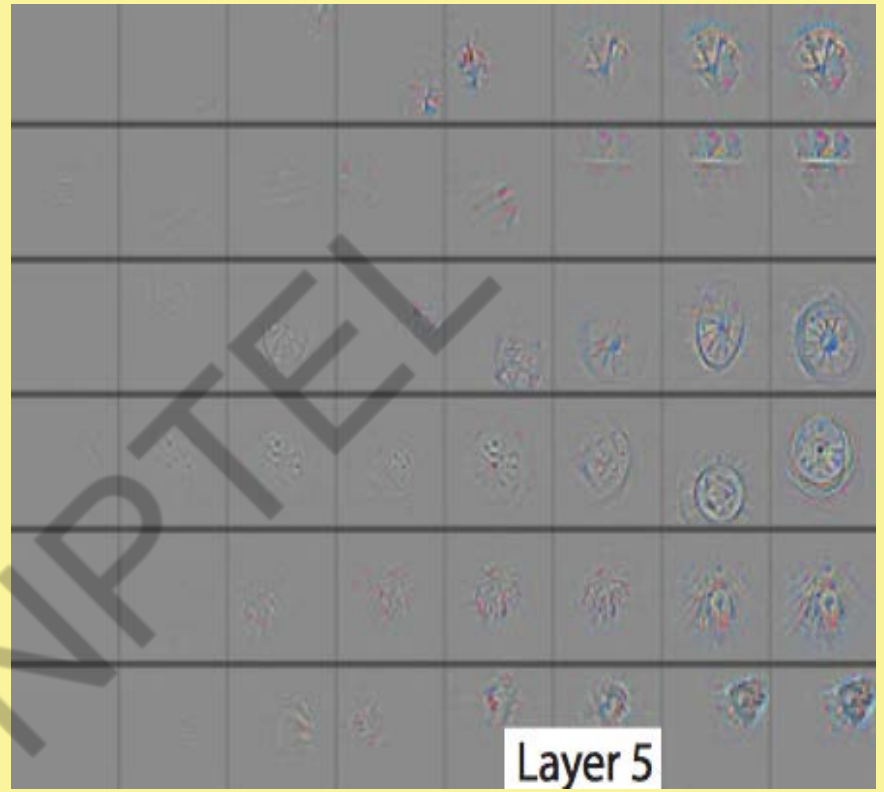


Image Source:-
https://becominghuman.ai/what-exactly-does-cnn-see-4d436d8e6e52

Transfer Learning

- Lower layers generate more general features:- knowledge transfers very well to other tasks.
- Higher layers are more task specific.
- Fine-tuning improves generalization when sufficient examples are available.
- Transfer learning and fine tuning often lead to better performance than training from scratch on the target dataset.
- Even features transferred from distant tasks often perform better than random initial weights.

Fine tuning

- Weights of the pre-trained CNN is fine-tuned for the new dataset by continuing the back propagation.
- Fine-tuning can be done for all layers.
- Due to overfitting concern, the earlier layers of the net may be fixed and fine tuning is done only on the higher layers.
- Earlier layers can be fixed as lower layers extract features that are more generic.
- Higher layers on the other hand are task specific.

NPTEL ONLINE CERTIFICATION COURSES

Thank you

NPTEL ONLINE CERTIFICATION COURSES

Course Name: Deep Learning

Faculty Name: Prof. P. K. Biswas
Department : E & ECE, IIT Kharagpur

Topic

Lecture 40: Popular CNN Models IV

CONCEPTS COVERED

Concepts Covered:

- CNN
  - AlexNet
  - VGG Net
  - Transfer Learning
  - Challenges in Deep Learning
  - GoogLeNet
  - ResNet
  - etc.

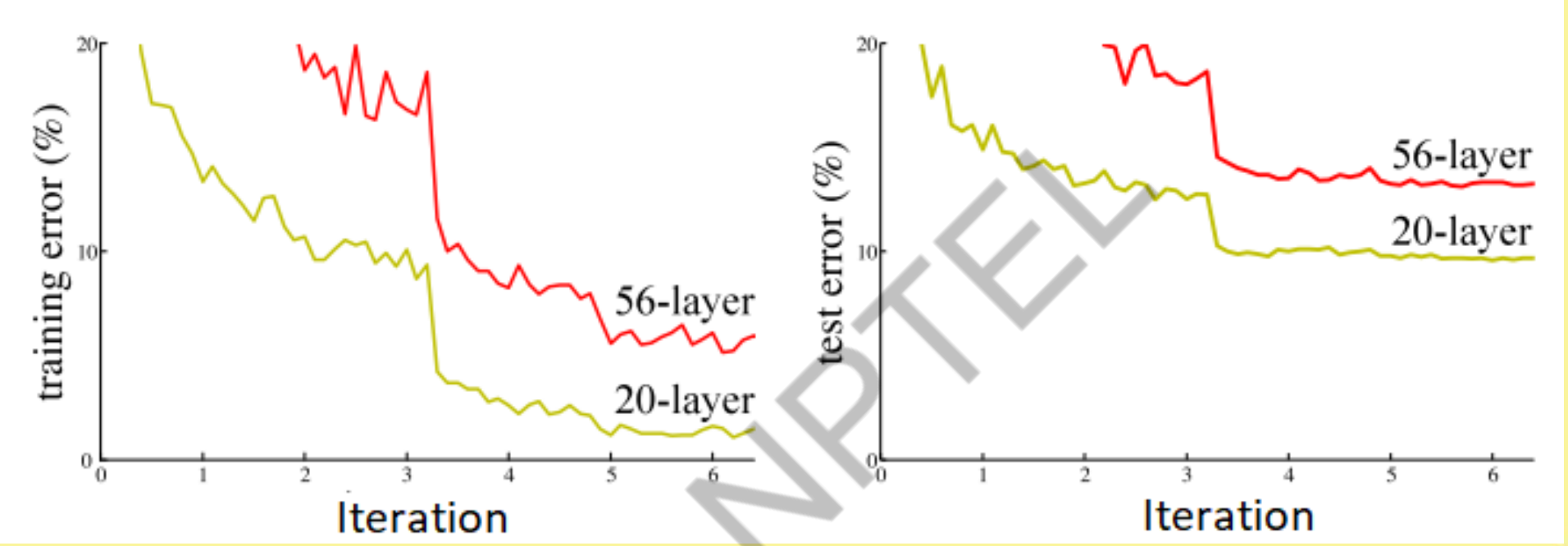
Deep Learning Challenges

Challenges

- Deep learning is data hungry.
- Overfitting or lack of generalization.
- Vanishing/Exploding Gradient Problem.
- Appropriate Learning Rate.
- Covariate Shift.
- Effective training.

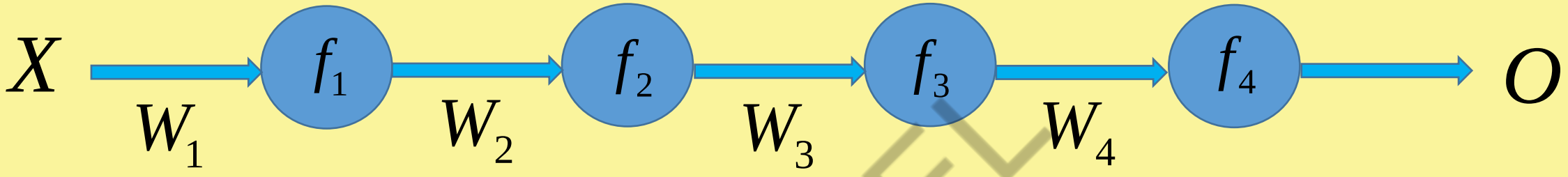
Vanishing Gradient

Vanishing Gradient Problem



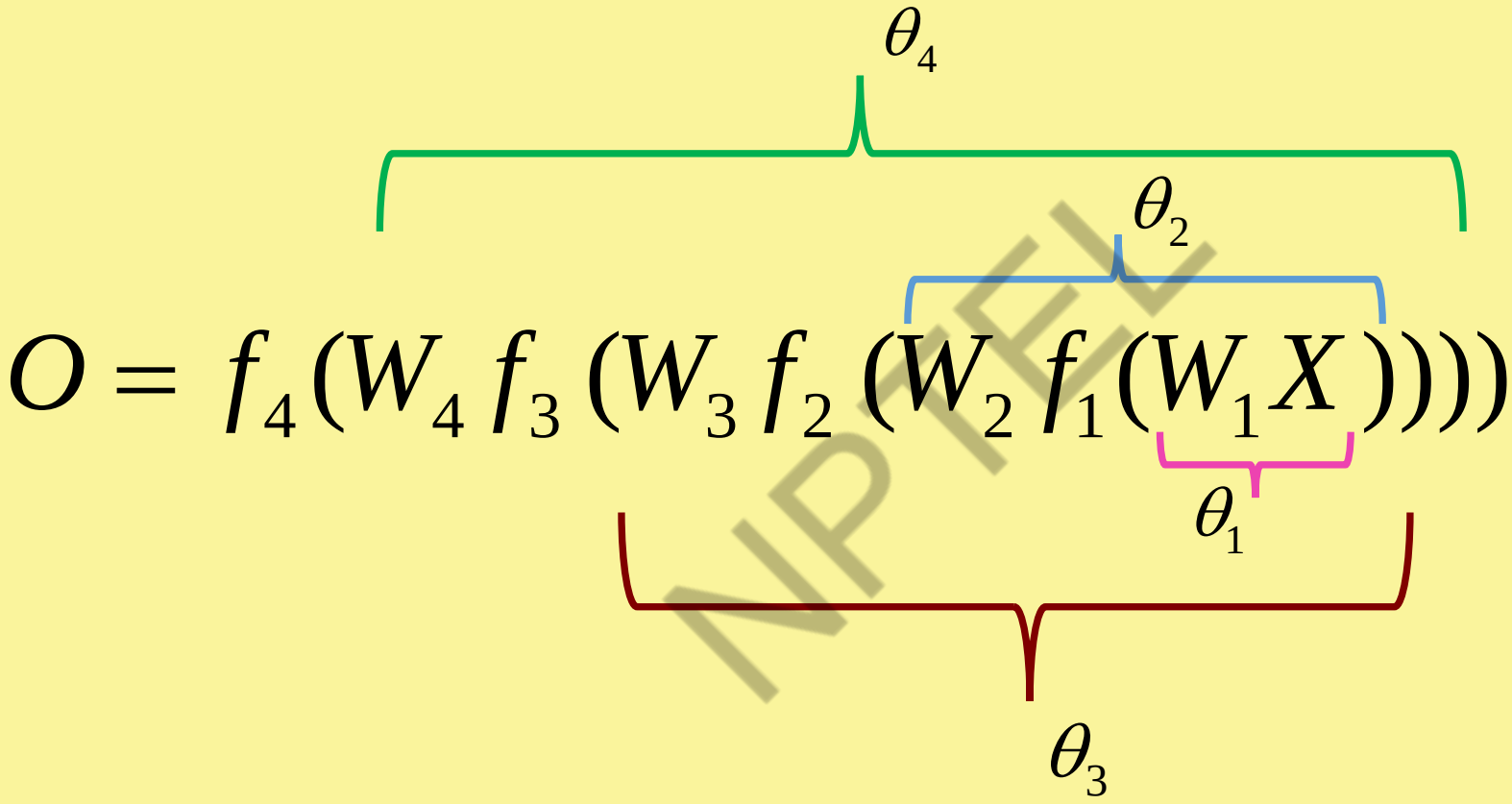
https://towardsdatascience.com/an-overview-of-resnet-and-its-variants-5281e2f56035

Vanishing Gradient Problem



$$O = f_4(W_4 f_3(W_3 f_2(W_2 f_1(W_1 X))))$$

Vanishing Gradient Problem



Vanishing Gradient Problem

$$O = f_4(\theta_4) \quad \theta_4 = W_4 f_3(\theta_3) \quad \theta_3 = W_3 f_2(\theta_2) \quad \theta_2 = W_2 f_1(\theta_1) \quad \theta_1 = W_1 X$$

$$\frac{\partial O}{\partial W_1} = \frac{\partial O}{\partial \theta_4} \cdot \frac{\partial \theta_4}{\partial f_3} \cdot \frac{\partial f_3}{\partial \theta_3} \cdot \frac{\partial \theta_3}{\partial f_2} \cdot \frac{\partial f_2}{\partial \theta_2} \cdot \frac{\partial \theta_2}{\partial f_1} \cdot \frac{\partial f_1}{\partial \theta_1} \cdot \frac{\partial \theta_1}{\partial W_1} = X \cdot f_1' \cdot W_2 \cdot f_2' \cdot W_3 \cdot f_3' \cdot W_4 \cdot \frac{\partial O}{\partial \theta_4}$$

$$\frac{\partial O}{\partial W_2} = \frac{\partial O}{\partial \theta_4} \cdot \frac{\partial \theta_4}{\partial f_3} \cdot \frac{\partial f_3}{\partial \theta_3} \cdot \frac{\partial \theta_3}{\partial f_2} \cdot \frac{\partial f_2}{\partial \theta_2} \cdot \frac{\partial \theta_2}{\partial W_2} = f_1 \cdot f_2' \cdot W_3 \cdot f_3' \cdot W_4 \cdot \frac{\partial O}{\partial \theta_4}$$

Vanishing Gradient Problem

- Choice of activation function: ReLU instead of Sigmoid.
- Appropriate initialization of weights.
- Intelligent Back Propagation Learning Algorithm.

NPTEL ONLINE CERTIFICATION COURSES

Thank you