श्रद्धावान् लभते ज्ञानम्

AMRITA
VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY

19CSE437
DEEP LEARNING FOR COMPUTER VISION
L-T-P-C: 2-0-3-3

Convolutional Neural Network Architectures

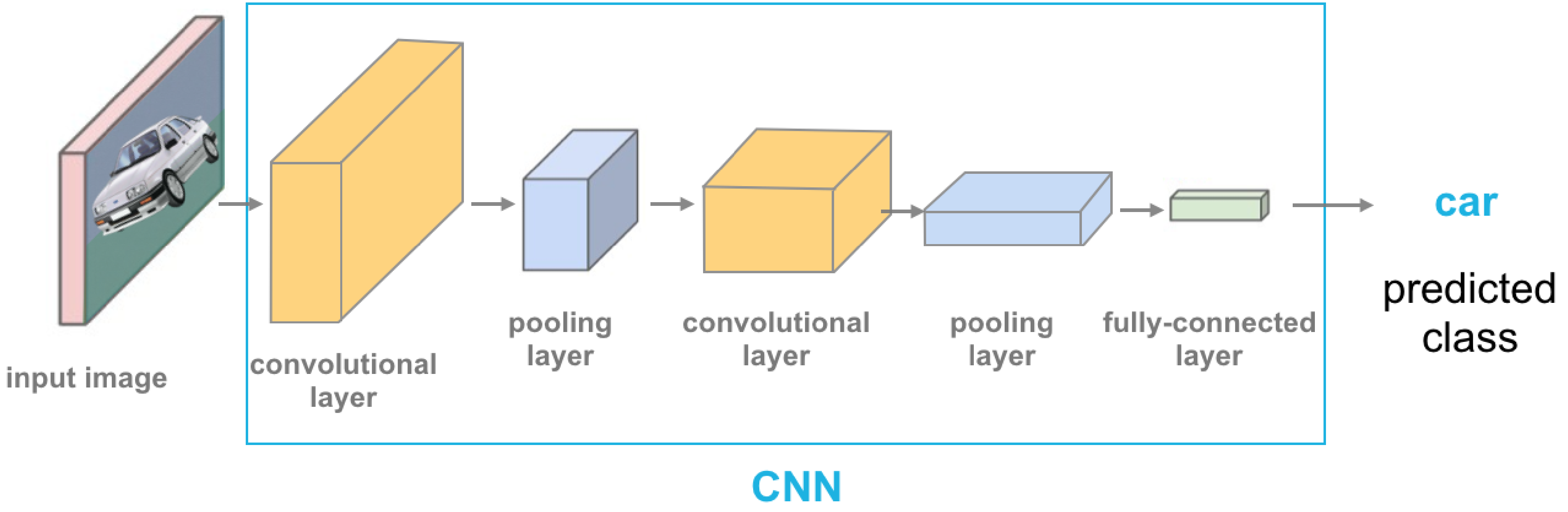
ZFNet(2013)

Convolutional Neural Network

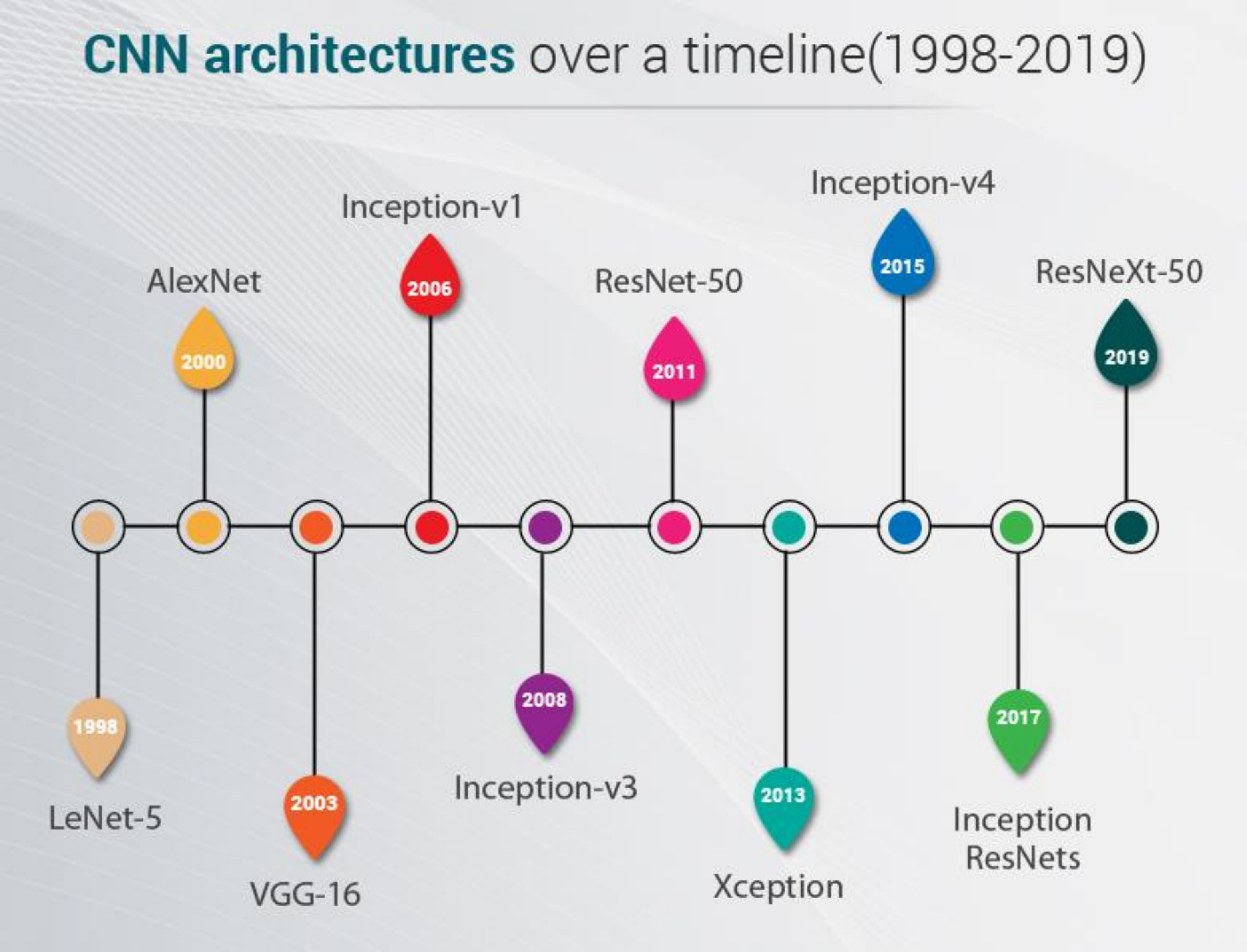
- In a convolutional network (ConvNet), there are basically three types of layers:
  1. Convolution layer
  2. Pooling layer
  3. Fully connected layer

After each convolution there is non-linearity applied using activation functions- Relu/Leaky Relu. H1=g(a1)

.



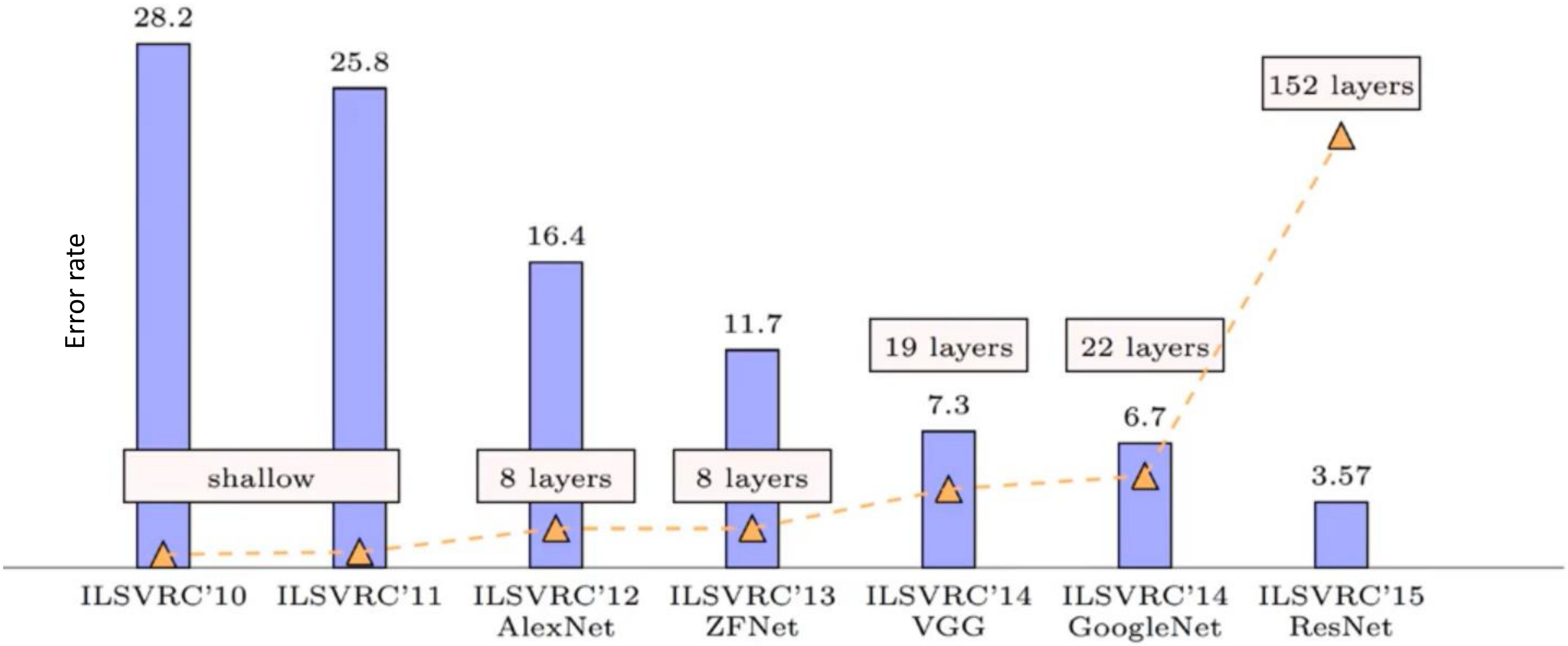
Different CNN Architectures



- **No: of Layers :**How many convolutional, max pooling fully connected layers?
- **No: of Filters in each layer**
- **Filter Size**
- **Max pooling:** What arrangement? 2 convolutional and then maxpooling or alternate convolutional and maxpooling?

Use standard tried and tested architectures!

ImageNet Large Scale Visual Recognition Challenge (ILSVRC)



https://image-net.org/challenges/LSVRC/index.php

ZFNet (2013)

- The **ImageNet** project is a large visual database designed for use in visual object recognition software research. The ImageNet project runs an annual software contest, the **ImageNet Large Scale Visual Recognition Challenge (ILSVRC)**, where software programs compete to correctly classify and detect objects and scenes.

- ILSVRC 2013 winner was a CNN which became known as ZFNet. It achieved a top-5 error rate of 14.8%
- It was mostly an achievement by tweaking the hyper-parameters of AlexNet while maintaining the same structure with additional Deep Learning elements
- Compared to AlexNet, the filter sizes are reduced and the stride of the convolutions are reduced.

The Top-5 error rate is the percentage of test examples for which the correct class was not in the top 5 predicted classes.

ZFNet (2013)

The ILSVRC 2013 winner was a CNN from Matthew Zeiler and Rob Fergus. It became known as **ZFNet**.
It improved on AlexNet by tweaking the architecture hyperparameters, in particular by expanding the size of the middle convolutional layers and making the stride and filter size on the first layer smaller, going from 11 x 11 stride 4 in AlexNet to 7 x 7 stride 2 in ZFNet.

The intuition behind this was that by using **bigger filters we were losing a lot of pixel information** and a smaller filter size in the first convolution layer helps to retain a lot of the original pixel information.

The number of filters increase as we go deeper. This network also used **ReLUs** for their activation and trained using **batch stochastic gradient descent.**

Also, AlexNet was trained on 15 million images, while ZFNet was trained on only **1.3 million images**:
obtain a test error of **14.8%**, on ImageNet2012 dataset

BY Matthew D. Zeiler zeiler, Rob Fergus
Dept. of Computer Science, Courant Institute, New York University

https://arxiv.org/pdf/1311.2901v3.pdf
Source paper

ZFNet (2013)

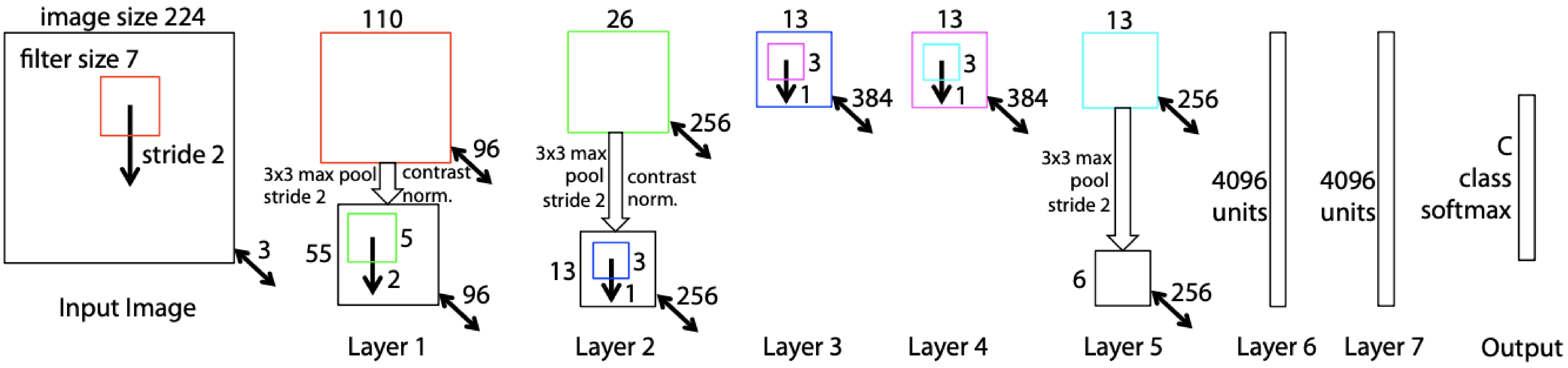
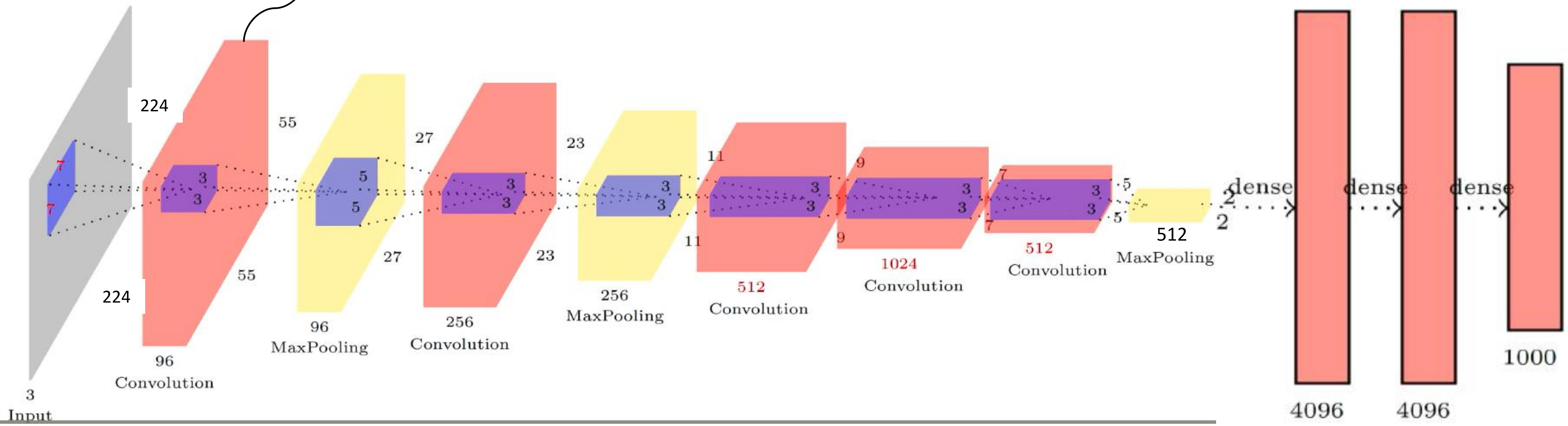


Figure 3. Architecture of our 8 layer convnet model. A 224 by 224 crop of an image (with 3 color planes) is presented as the input. This is convolved with 96 different 1st layer filters (red), each of size 7 by 7, using a stride of 2 in both x and y. The resulting feature maps are then: (i) passed through a rectified linear function (not shown), (ii) pooled (max within 3x3 regions, using stride 2) and (iii) contrast normalized across feature maps to give 96 different 55 by 55 element feature maps. Similar operations are repeated in layers 2,3,4,5. The last two layers are fully connected, taking features from the top convolutional layer as input in vector form ($6 \cdot 6 \cdot 256 = 9216$ dimensions). The final layer is a C-way softmax function, C being the number of classes. All filters and feature maps are square in shape.

ZFNet Architecture

After convolution (110x110) and pooling



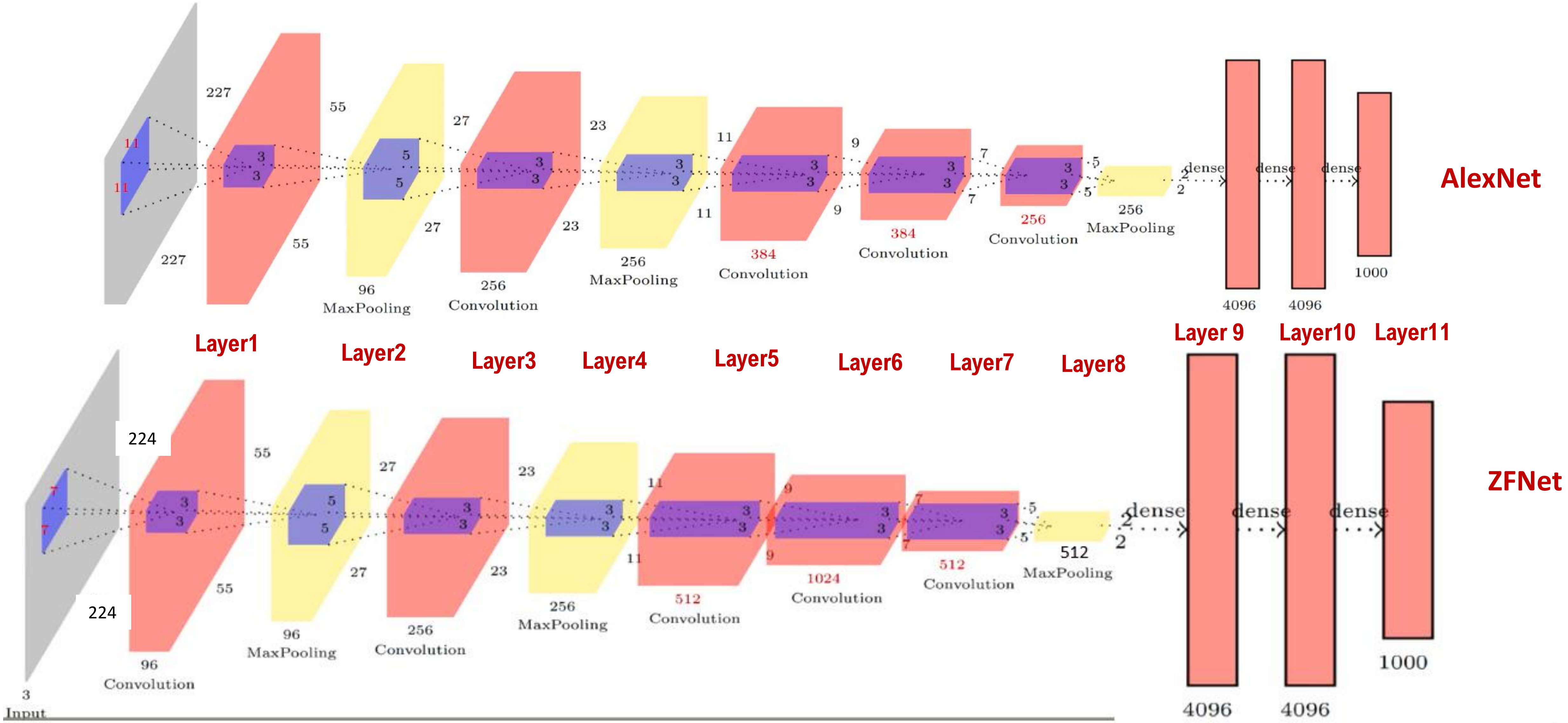
Layer I : 224x224 (3 channel) convolved with 96 filters of 7x7 size and stride =2, results in 96, 110x110 image

Wo=(Wi+2P-F) /S+1=(224+0-7)/2+1=110

This is then subjected to maxpooling resulting in 55x55

AlexNet-ZFNet Difference

Maxpooling also counted as layer for understanding purpose.

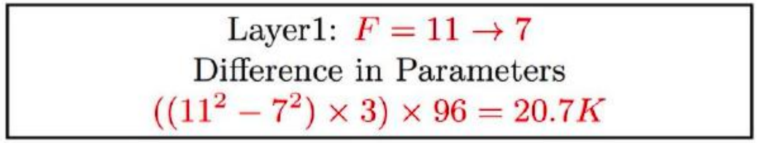


AlexNet

Layer1 Layer2 Layer3 Layer4 Layer5 Layer6 Layer7 Layer8 Layer 9 Layer10 Layer11

ZFNet

AlexNet-ZFNet Difference in parameters

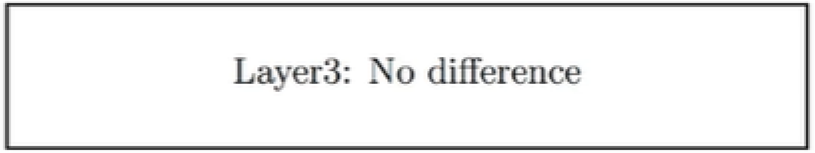


Layer1: $F = 11 \to 7$

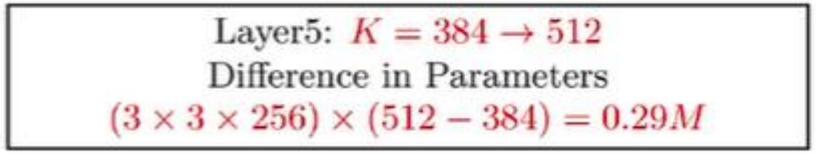
Difference in Parameters

$((11^2 - 7^2) \times 3) \times 96 = 20.7K$

11x11x3x96-7x7x3x96



Layer3: No difference

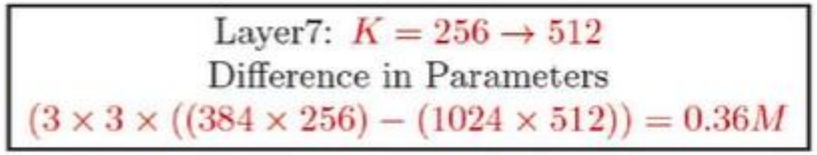


Layer5: $K = 384 \to 512$

Difference in Parameters

$(3 \times 3 \times 256) \times (512 - 384) = 0.29M$

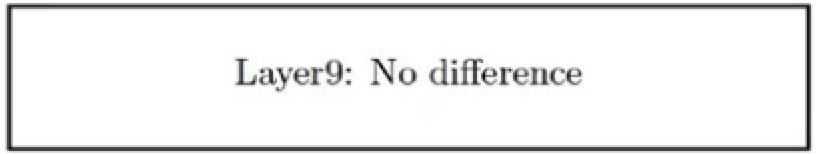
3x3x256x512 – 3x3x256x384



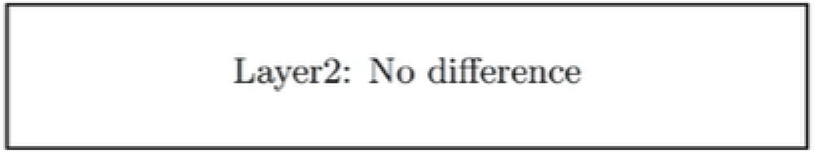
Layer7: $K = 256 \to 512$

Difference in Parameters

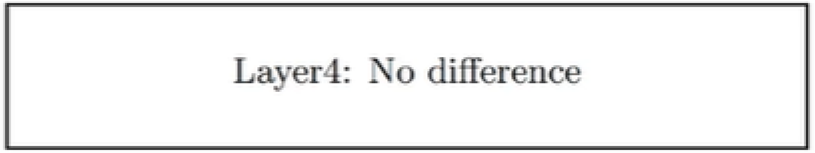
$(3 \times 3 \times ((384 \times 256) - (1024 \times 512)) = 0.36M$



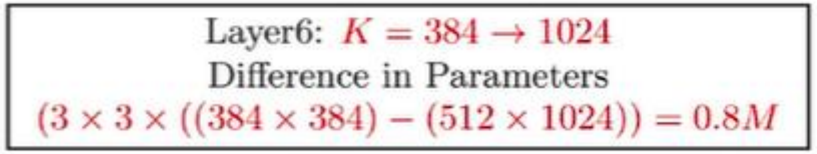
Layer9: No difference



Layer2: No difference



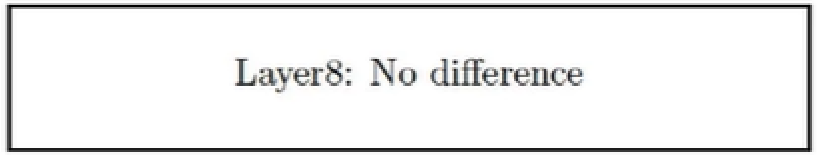
Layer4: No difference



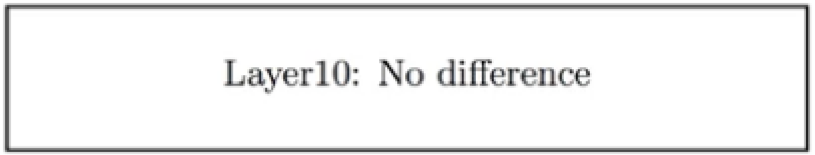
Layer6: $K = 384 \to 1024$

Difference in Parameters

$(3 \times 3 \times ((384 \times 384) - (512 \times 1024)) = 0.8M$



Layer8: No difference



Layer10: No difference

Namah Shivaya