# 

[**9.1: CNN Architectures I**](#_77ktfbebpfz2) **2**

[9.1.1: Setting the context](#_rhjoiqtvn152) 2

[9.1.2: The Imagenet Challenge](#_vntoe9p0218) 4

[9.1.3: Understanding the first layer of AlexNet](#_equkttswngjc) 5

[9.1.4: Understanding all the layers of AlexNet](#_u5z3holgpt95) 6

[9.1.5: ZFNet](#_nwbzdgk7lls) 9

[9.1.6: VGGNet](#_fjbubpa90axj) 11

[9.1.7: Summary](#_o19jfxppxhzy) 12

[**9.2: CNN Architectures II**](#_ojp1poy47422) **13**

[9.2.1: Setting the context](#_wnxshvg19bqm) 13

[9.2.2: Number of computations in a convolution layer](#_bs24evnqklp7) 14

[9.2.3: 1x1 Convolutions](#_2hqcvmfkv0op) 15

[9.2.4: The Intuition behind GoogLeNet](#_a47bcwebluck) 16

[9.2.5: The Inception Module](#_xs6nlceetj78) 18

[9.2.6: The GoogLeNet Architecture](#_irmbq89bbol7) 19

[9.2.7: Average Pooling](#_ggensdeg2ndm) 21

[9.2.8: Auxiliary Loss for training a deep network](#_ctrdtso2dpb9) 22

[9.2.9: ResNet](#_ulalugdq2uxg) 23

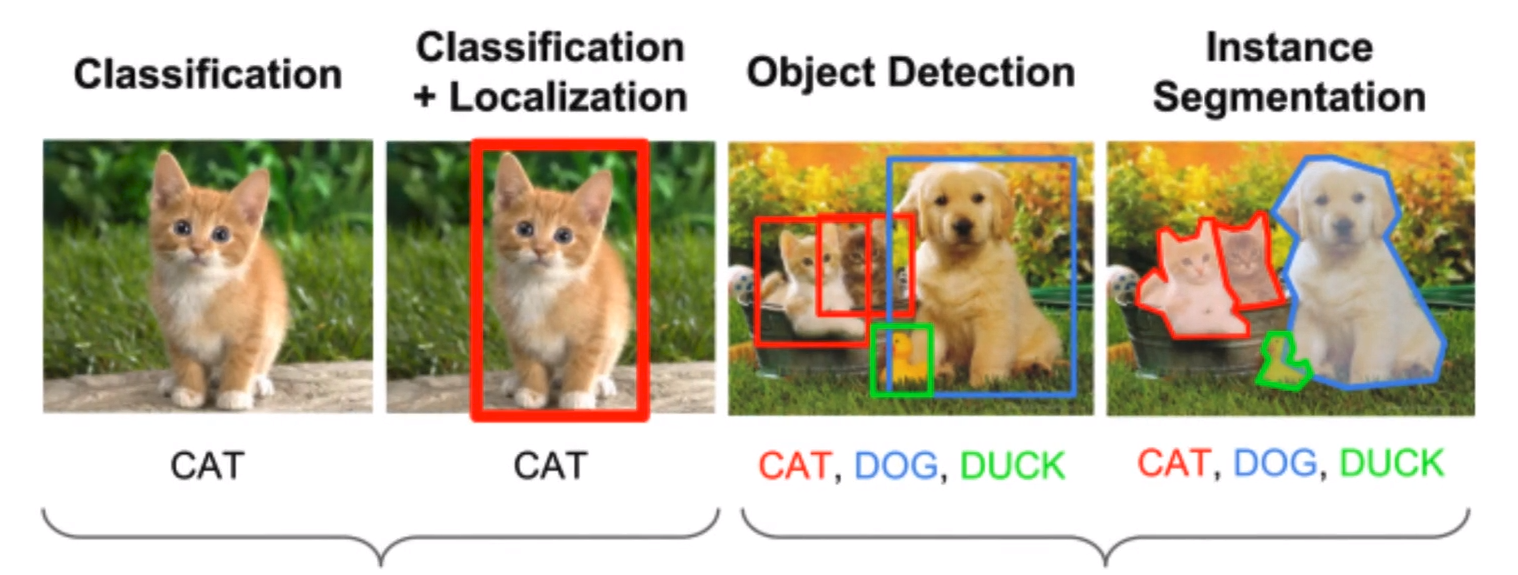
# 

# 9.1: CNN Architectures I

## 9.1.1: Setting the context

What are CNNs used for?

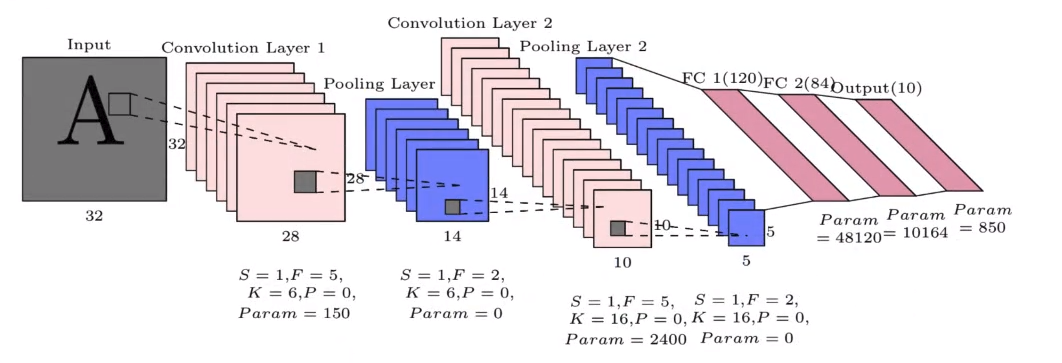
1. First, let’s see **what kind of tasks** are CNNs used for
   1. Consider the following image for an overview of the tasks that CNNs are used for



Single Object

Multiple Objects

* 1. Using popular datasets like Imagenet, which contains 1000 classes of objects, ranging from vehicles and sceneries, up to the different dog & cat breeds.
  2. **Classification**: In the first image, the task at hand is to correctly predict the class to which the object belongs to. This example uses an “Iconic Photo”, i.e. where the class-object of the photo occupies most of the photograph area.
  3. **Classification + Localization**: In the second image, we are predicting the class of the object and also precisely where it is located in the image. Given a starting point outside the object, we detect the width and height of the bounding box enclosing the object. This is both a classification problem and a regression problem.
  4. **Object Detection**: In the third image, we have multiple objects. Therefore, we must correctly detect each object and classify them respectively. It involves multiple Classification + Localization operations on the same image.
  5. **Instance Segmentation**: In the fourth image, we are moving a step further from object detection. Here, we are identifying the precise bounding area around each of the objects present in the image. This is commonly performed in autonomous driving etc, where we have to detect the presence of multiple objects at any given point in time.
  6. These are the four applications that commonly use CNNs. In fact, even our capstone-project of Character-detection and recognition also requires the above mentioned processes.

1. A typical recipe for solving image-related tasks is as follows
   1. First passing the input images through as series of convolutional layer
   2. We obtain a 3D tensor which we flatten to a single dimensional vector
   3. We pass the vector through a number of fully connected layers, which culminate in output prediction
   4. In these cases, we either perform Classification or Regression
2. Now, if we are going to use CNNs for these tasks, we have to make a few choices with regards to the design of the CNN layout.
3. What are some of the decisions that need to be taken?
   1. Let’s look at a sample CNN architecture
   2. Some of the factors under our control are as follows:
   3. Number of layers
   4. Number of filters in each layer
   5. Filter Size
   6. Max pooling
4. With the amount of choice we have, designing a CNN architecture could become a very messy process.
5. So a **standard practice is to use tried and tested architectures**.
6. In this chapter, we will be looking at the most popular CNN architectures.

## 

## 9.1.2: The Imagenet Challenge

The Imagenet challenge over the years

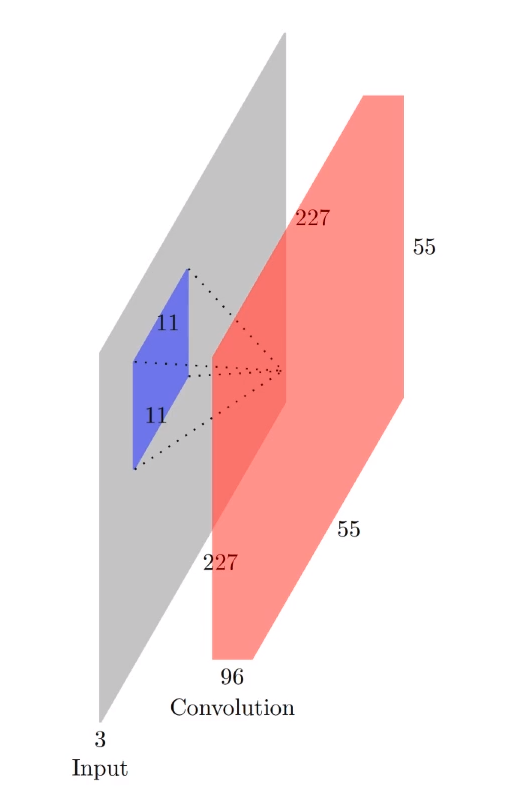
1. The Imagenet dataset is a 1000-class, 1,000,000-image dataset (1000 images per class)
2. The **I**magenet **L**arge **S**cale **V**isual **R**ecognition **C**hallenge (**ILSVRC**) or the **Imagenet Challenge** is an annual contest for contender’s models to correctly classify the images in the dataset
3. Let us look at the Challenge results between 2010-2015
4. Let us analyse the graph briefly
   1. The metric used to measure performance was Top-5 accuracy. I.e. If any of the top-five predicted class probabilities matched the true class, it was considered correct.
   2. **2010-2011**: This was the pre-DL era, and the Machine Learning models that were submitted had an error between 25-28 %
   3. **2012**: The AlexNet (CNN) architecture smashed existing records with 16.4% error. It kickstarted the Deep Learning Era.
   4. **2013-2014**: Significant improvements were made using models such as ZFNet, VGG and GoogLeNet. Error was brought down to ~6.7%
   5. **2015**: Microsoft's ResNet successfully brought the error down to 3.57% which is lower than the error scored by humans!
   6. One of the reasons for beating the human-error was because some of the classes in the Imagenet Dataset were very fine-grained, i.e. distinguishing between the different dog breeds.
   7. Another interesting point to note is the consistently increasing depth of the Networks used. From shallow networks in the ML era right up to 152 layers in ResNet

## 

## 9.1.3: Understanding the first layer of AlexNet

Let’s break down the first layer of the AlexNet architecture

1. AlexNet was the winning architecture for the 2012 Imagenet Challenge
2. Let us look at the convolutional layer



**Input**:

WI = 227

HI = 227

DI = 3

**Conv1**:

Filter Size (F) = 11

No. of Filters (K) = 96

Stride (S) = 4

Padding (P) = 0

Parameters = (11x11x3)x96 = 34848 parameters

**Output**:

WO = 55

HO = 55

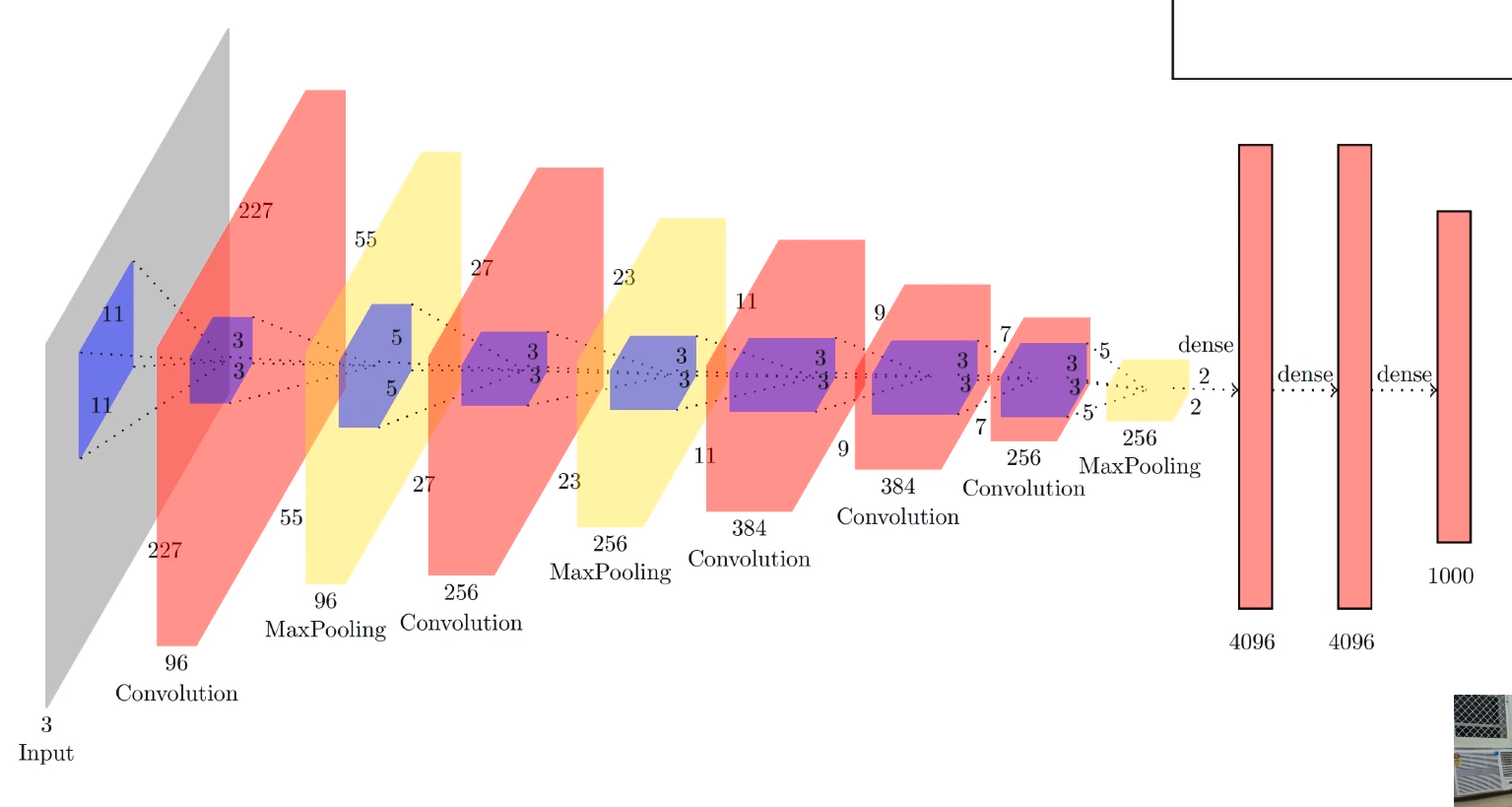
DO = K = 96

1. The details are as follows
   1. **Input images**: 227x227x3 (colour images of 227x227 Width x Height)
      1. WI = 227
      2. HI = 227
      3. DI = 3
   2. **Filter/Conv1** layer:
      1. Filter Size (F) = 11 (i.e. FxFxDI or 11x11x3)
      2. No. of Filters (K) = 96
      3. Stride (S) = 4
      4. Padding (P) = 0
      5. Parameters = (11x11x3) x 96 = 34,848
      6. These values were determined through extensive experimentation
   3. **Output**:
      1. WO = = 55
      2. HO = = 55
      3. DO = K = 96
2. This was a standard architecture and can be used for a variety of tasks.

## 9.1.4: Understanding all the layers of AlexNet

Let’s now look at the entire AlexNet

1. Before moving into the AlexNet architecture, let us understand why we have decided to use a Convolutional layer instead of a Fully-connected layer
   1. In the convolutional layer, the number of parameters is (11x11x3) x 96 = **34,848**
   2. However, in a FC layer, the number of parameters would be
      1. Input:(227x227x3) x Output:(55x55x96) = **4.49 x 1010**
   3. Thus, because of sparse-connectivity and weight sharing, we are able to achieve a similar degree of complexity with a significantly smaller number of parameters with a Convolutional layer.
2. Now, let us **break down the entire AlexNet architecture**



**Input Image**

**Conv1**

**MP1**

**Conv2**

**MP2**

**Conv3**

**Conv4**

**Conv5**

**MP3**

**FC1**

**FC2**

**FC3**

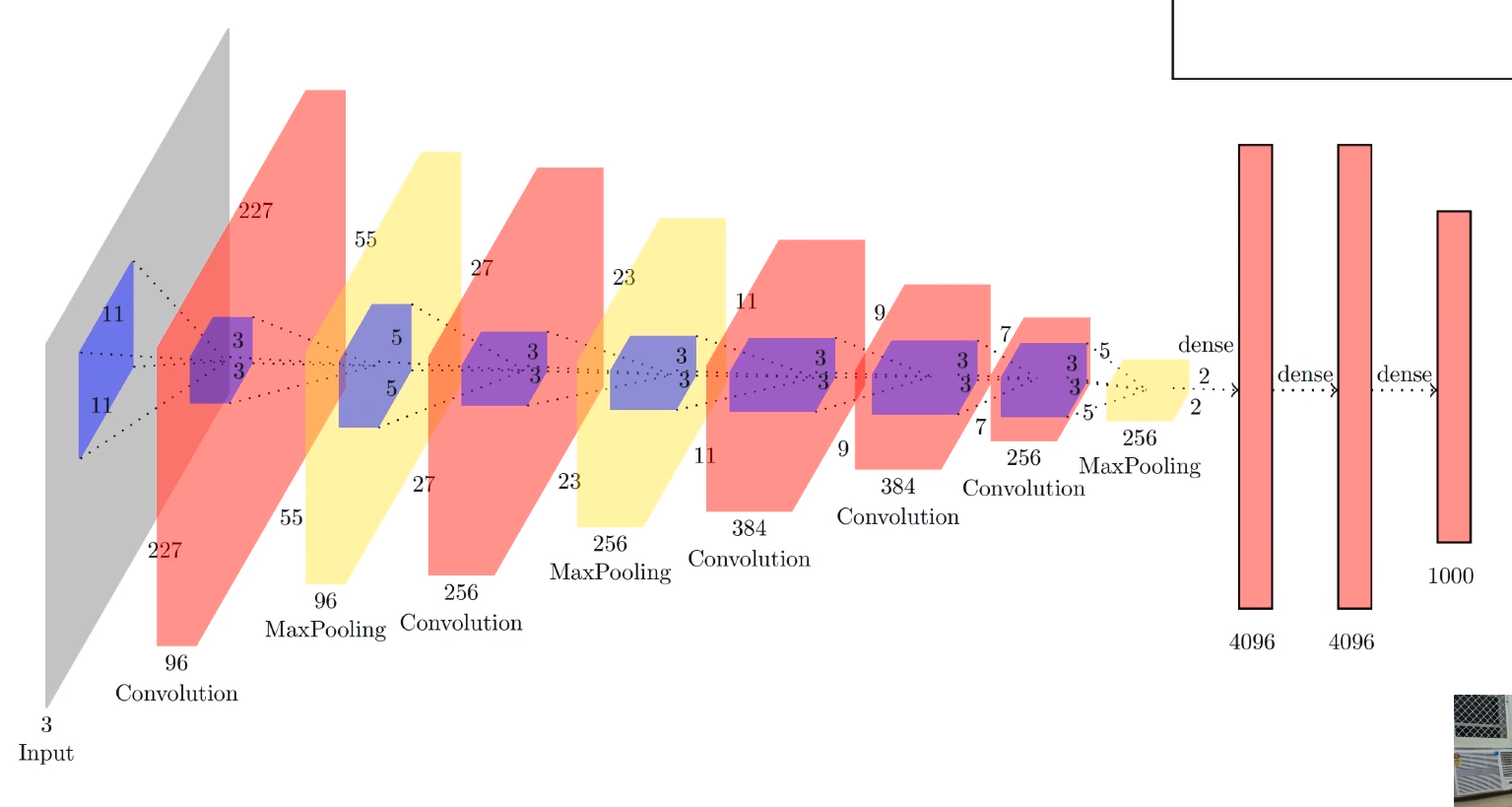
1. Let’s look at each of the layers in depth
2. **Input Layer**: 227x227x3 (colour images of 227x227 Width x Height)
   1. WIn = 227
   2. HIn = 227
   3. DIn = 3
3. **Convolutional Layer 1**: Input is 227x227x3
   1. Filter Size (**F**) = 11 (11x11x3)
   2. No. of Filters (**K**) = 96
   3. Stride (**S**) = 4
   4. Padding (**P**) = 0
   5. **Parameters** = (11x11x3) x 96 = 34,848
   6. **W1** = 55
   7. **H1** = 55
   8. **D1** = K = 96
   9. **ReLU** Non-linearity function is applied to every 2D area in the output volume.
4. **Max-Pooling Layer 1**: Input is 55x55x96
   1. Filter Size (**F**) = 3 (i.e. 3x3x96)
   2. Stride (**S**) = 4
   3. **Parameters** = 0 (no parameters in max pooling)
   4. **W1m** = 27
   5. **H1m** = 27
   6. **D1m** = 96
5. **Convolutional Layer 2**: Input is 27x27x96
   1. Filter Size (**F**) = 5 (5x5x96)
   2. No. of Filters (**K**) = 256
   3. Stride (**S**) = 1
   4. Padding (**P**) = 0
   5. **Parameters** = (5x5x96) x 256 = 614,400
   6. **W2** = 23
   7. **H2** = 23
   8. **D2** = K = 256
   9. **ReLU** Non-linearity function is applied.
6. **Max-Pooling Layer 2**: input is 23x23x256
   1. Filter Size (**F**) = 3 (3x3x256)
   2. Stride (**S**) = 2
   3. **Parameters** = 0
   4. **W2m** = 11
   5. **H2m** = 11
   6. **D2m** = 256
7. **Convolutional Layer 3**: input is 11x11x256
   1. Filter Size (**F**) = 3 (3x3x256)
   2. No. of Filters (**K**) = 384
   3. Stride (**S**) = 1
   4. Padding (**P**) = 0
   5. **Parameters** = (3x3x256) x 384 = 884,736
   6. **W3** = 9
   7. **H3** = 9
   8. **D3** = K = 384
   9. **ReLU** Non-linearity function is applied.
8. **Convolutional Layer 4**: input is 9x9x384
   1. Filter Size (**F**) = 3 (3x3x384)
   2. No. of Filters (**K**) = 384
   3. Stride (**S**) = 1
   4. Padding (**P**) = 0
   5. **Parameters** = (3x3x384) x 384 = 1,327,104
   6. **W4** = 7
   7. **H4** = 7
   8. **D4** = K = 384
   9. **ReLU** Non-linearity function is applied.
9. **Convolutional Layer 5**: input is 7x7x384
   1. Filter Size (**F**) = 3 (3x3x384)
   2. No. of Filters (**K**) = 256
   3. Stride (**S**) = 1
   4. Padding (**P**) = 0
   5. **Parameters** = (3x3x384) x 256 = 884,736
   6. **W5** = 5
   7. **H5** = 5
   8. **D5** = K = 256
   9. **ReLU** Non-linearity function is applied.
10. **Max-Pooling Layer 3**: input is 5x5x256
    1. Filter Size (**F**) = 3 (3x3x256)
    2. Stride (**S**) = 2
    3. **Parameters** = 0
    4. **W5m** = 2
    5. **H5m** = 2
    6. **D5m** = 256
11. **Fully Connected Layer 1**: input is 2x2x256 = 1024
    1. Number of Neurons = 4096
    2. Parameters = (2x2x256) x 4096 = 4,194,304
12. **Fully Connected Layer 2**: input = 4096
    1. Number of Neurons = 4096
    2. Parameters = 4096 x 4096 = 16,777,216
13. **Fully Connected Layer 3**: input = 4096
    1. Number of Neurons/output-classes = 1000
    2. Parameters = 4096 x 1000 = 4,096,000
14. Totally, there are around 27.55 Million parameters, out of which roughly 25 Million parameters were in the last 3 Fully-connected layers.
15. When counting the total number of layers, we do not include the max-pooling layers as they do not carry weights. Thus, we say that **AlexNet has 8 layers**

## 

## 9.1.5: ZFNet

Let’s now look at the entire AlexNet

1. ZFNet is another 8-layer CNN architecture. Let’s understand it better with a side-by-side comparison with AlexNet.



**Input Image**

**Conv1**

**MP1**

**Conv2**

**MP2**

**Conv3**

**Conv4**

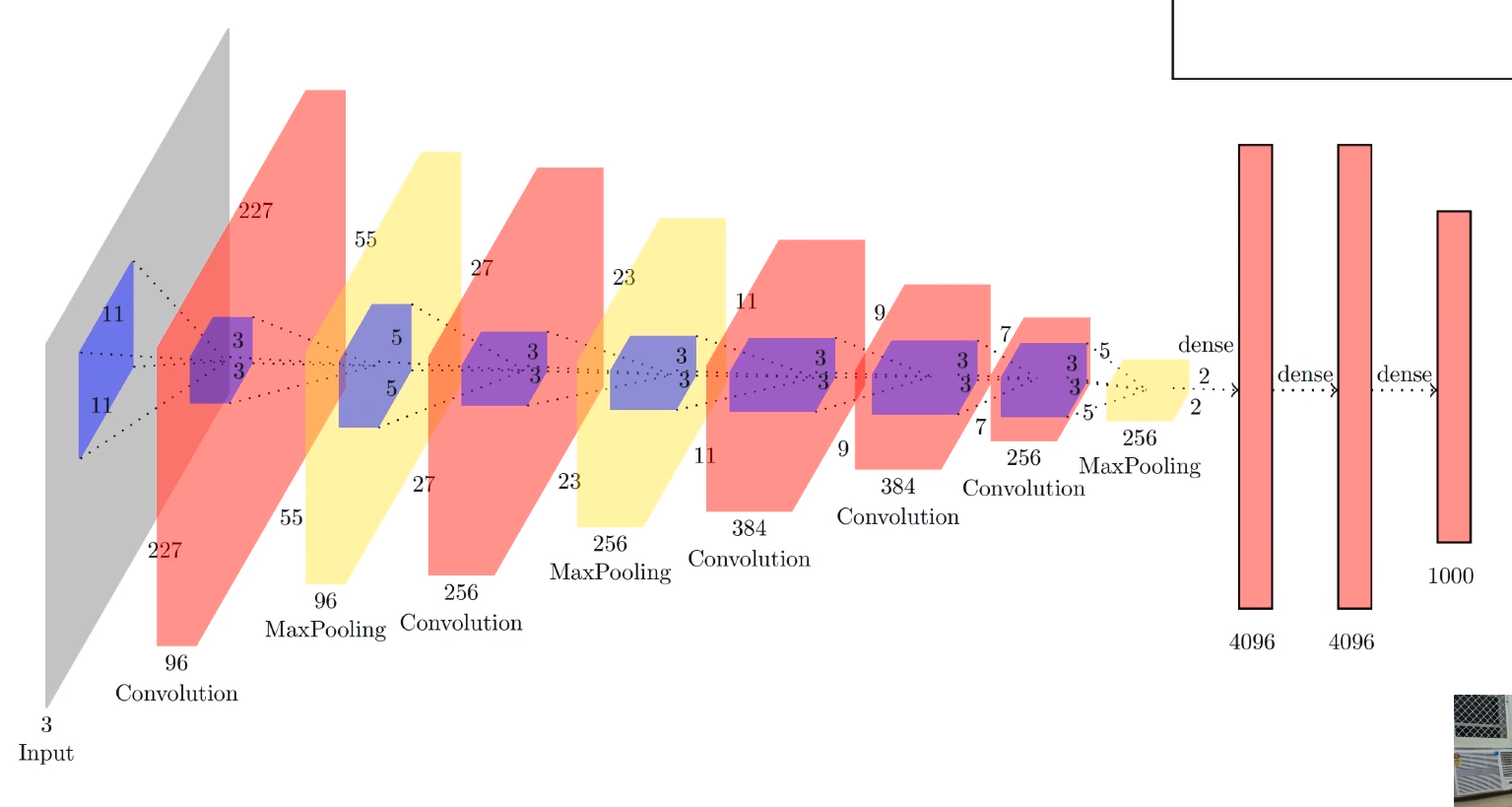
**Conv5**

**MP3**

**FC1**

**FC2**

**FC3**



**Input Image**

**Conv1**

**MP1**

**Conv2**

**MP2**

**Conv3**

**Conv4**

**Conv5**

**MP3**

**FC1**

**FC2**

**FC3**

**7**

**7**

**512**

**1024**

**512**

**512**

**\* Indicates differences in ZFNet from AlexNet**

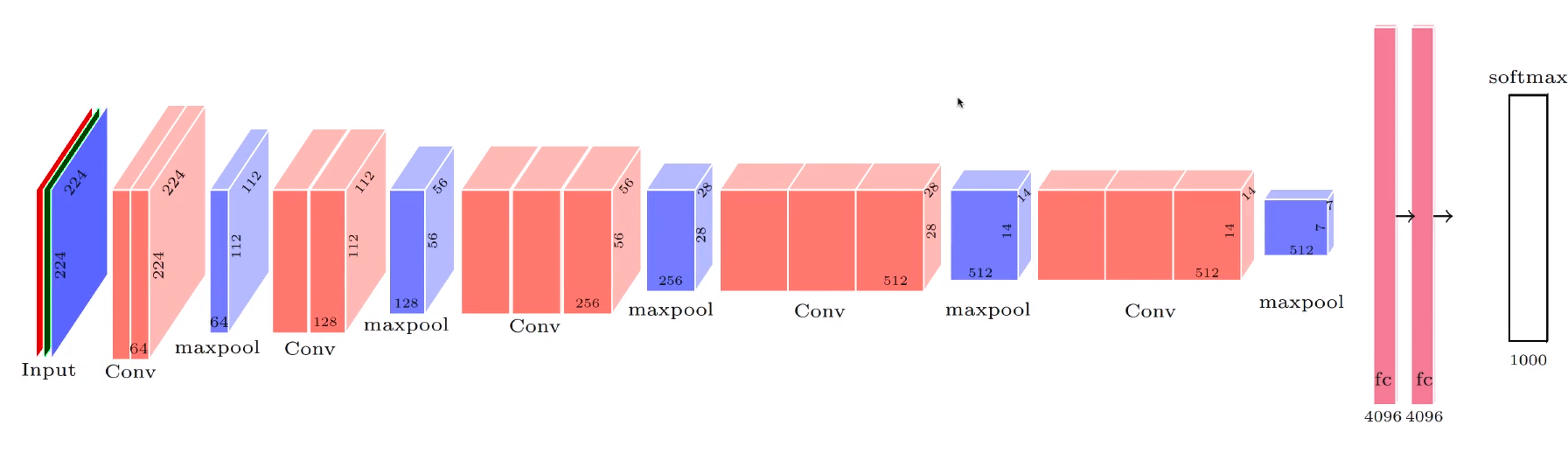
1. ZFNet is largely similar to AlexNet, with the exception of a few of the layers. Let us highlight those differences.
2. **Convolutional Layer 1**: Input is 227x227x3
   1. Filter Size (**F**) = 7 (7x7x3)
   2. No. of Filters (**K**) = 96
   3. Stride (**S**) = 4
   4. Padding (**P**) = 0
   5. **Parameters** = (7x7x3) x 96 = 14,112
   6. **W1** = 55
   7. **H1** = 55
   8. **D1** = K = 96
   9. **ReLU** Non-linearity function is applied to every 2D area in the output volume.
3. **Convolutional Layer 3**: input is 11x11x256
   1. Filter Size (**F**) = 3 (3x3x256)
   2. No. of Filters (**K**) = 512
   3. Stride (**S**) = 1
   4. Padding (**P**) = 0
   5. **Parameters** = (3x3x256) x 512 = 1,179,648
   6. **W3** = 9
   7. **H3** = 9
   8. **D3** = K = 512
   9. **ReLU** Non-linearity function is applied.
4. **Convolutional Layer 4**: input is 9x9x512
   1. Filter Size (**F**) = 3 (3x3x512)
   2. No. of Filters (**K**) = 1024
   3. Stride (**S**) = 1
   4. Padding (**P**) = 0
   5. **Parameters** = (3x3x512) x 1024 = 4,718,592
   6. **W4** = 7
   7. **H4** = 7
   8. **D4** = K = 1024
   9. **ReLU** Non-linearity function is applied.
5. **Convolutional Layer 5**: input is 7x7x1024
   1. Filter Size (**F**) = 3 (3x3x1024)
   2. No. of Filters (**K**) = 512
   3. Stride (**S**) = 1
   4. Padding (**P**) = 0
   5. **Parameters** = (3x3x1024) x 512 = 4,718,592
   6. **W4** = 5
   7. **H4** = 5
   8. **D4** = K = 512
   9. **ReLU** Non-linearity function is applied.
6. **Max-Pooling Layer 3**: input is 5x5x512
   1. Filter Size (**F**) = 3 (3x3x512)
   2. Stride (**S**) = 2
   3. **Parameters** = 0
   4. **W2m** = 2
   5. **H2m** = 2
   6. **D1m** = 512
7. **Fully Connected Layer 1**: input is 2x2x512 = 2048
   1. Number of Neurons = 4096
   2. Parameters = (2x2x512) x 4096 = 8,388,608
8. The **total difference in the number of parameters** ZFNet - AlexNet = 1.45 Million
9. There are other variants of ZFNet where we use a stride of 2 in the first convolutional layer, thereby changing the subsequent layer dimensions.

## 

## 9.1.6: VGGNet

Let’s look at the architecture of VGGNet

1. During the design of the VGGNet, it was found that alternating convolution & pooling layers were not required. So VGGnet uses multiple of Convolutional layers in sequence with pooling layers in between.
2. Let us break down the VGGNet architecture



Kernel Size is 3 x 3 throughout

Total parameters in non-FC layers = ~16M

Total parameters in FC layers = ~122M

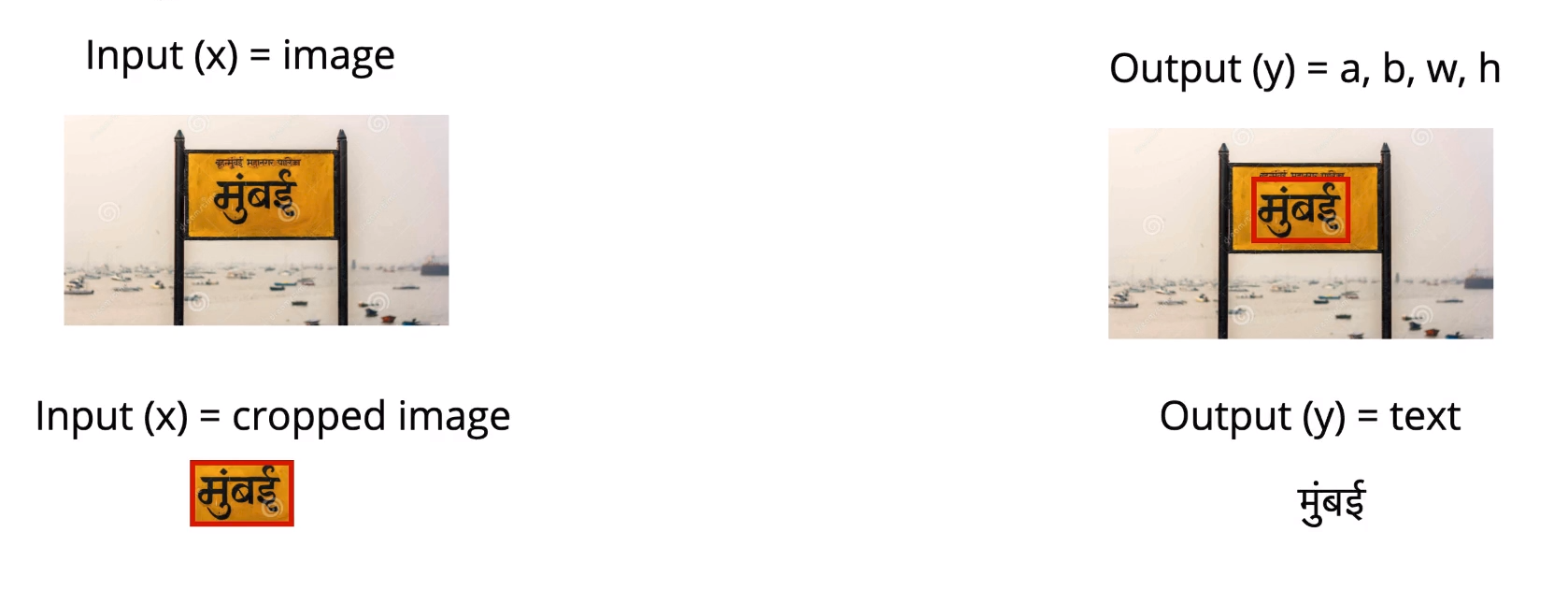
1. A few points to note
2. The kernel size 3x3 is maintained throughout the network, only the depth is changed between layers
3. Appropriate padding is provided to maintain the dimensions across the layers
4. **Convolutional Bundle 1**: There are **2 convolutional layers** of size **224x224x64**
5. **Max Pool Layer 1**: The size is **112x112x64**
6. **Convolutional Bundle 2**: There are **2 convolutional layers** of size **112x112x128**
7. **Max Pool Layer 2**: The size is **56x56x128**
8. **Convolutional Bundle 3**: There are **3 convolutional layers** of size **56x56x256**
9. **Max Pool Layer 3**: The size is **28x28x256**
10. **Convolutional Bundle 4**: There are **3 convolutional layers** of size **28x28x512**
11. **Max Pool Layer 4**: The size is **14x14x512**
12. **Convolutional Bundle 5**: There are **3 convolutional layers** of size **14x14x512**
13. **Max Pool Layer 5**: The size is **7x7x512**
14. The number of **parameters in the Non-FC layers is ~16 Million**
15. **FC Layer 1** has **4096** Neurons
16. **FC Layer 2** has **4096** Neurons
17. **FC Layer 3** is a softmax with **1000** Neurons/Output-classes
18. The number of **parameters in the FC layers is ~122 Million** (Most in FC Layer 1: ~102 Million)
19. Though the number of parameters in this network seems very large, it would have been exponentially larger if we had chosen an entirely Fully-Connected network.
20. The above shown VGGNet is a 16 layer network called VGG16. There are also other versions like the 19 layered VGG19

## 

## 9.1.7: Summary

Connecting this to the capstone project

1. Let’s see how CNNs come into play for our Signboard translation capstone project



**Regression Task: Finding bounding box**

CNN used to compute hidden-representation of input image, using a standard CNN architecture like VGGNet. Finally Add layer to perform regression and get the bounding box values.

Apply another CNN to obtain hidden representations before using RNNs and Encoder-Decoder models to get the text.

1. The above diagram shows how we use CNNs in our capstone project.

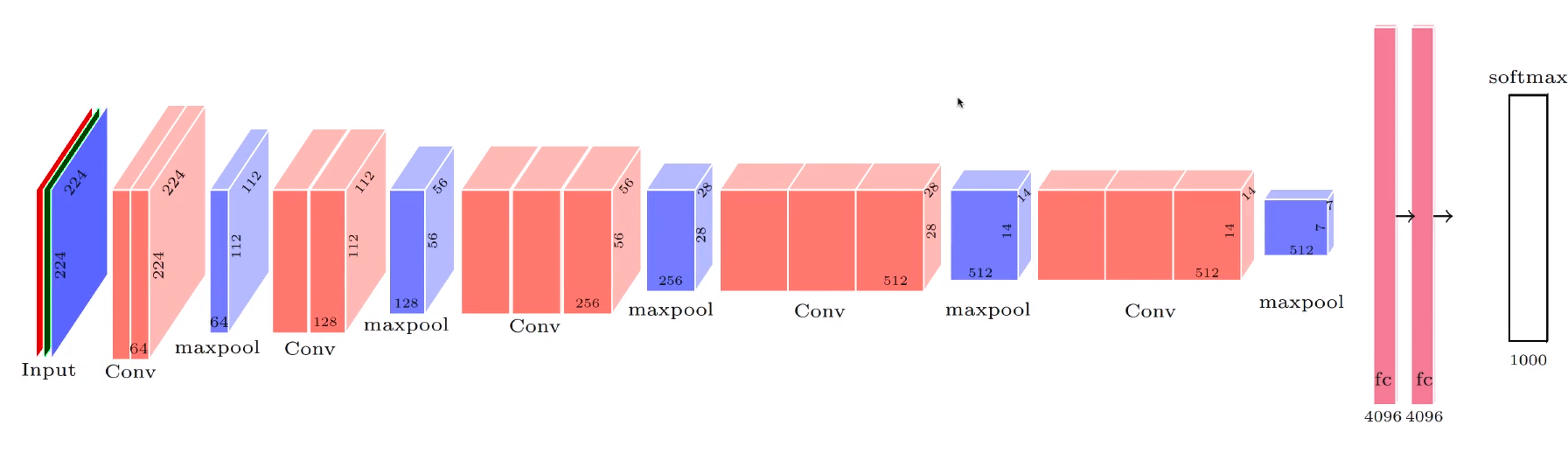
# 

# 9.2: CNN Architectures II

## 9.2.1: Setting the context

How do we address what’s been bothering us so far with CNNs

1. To pick up from VGGNet, can we do something to reduce the huge number of parameters incurred between the non-FC Layer and the FC Layer.



102 Million

Parameters here

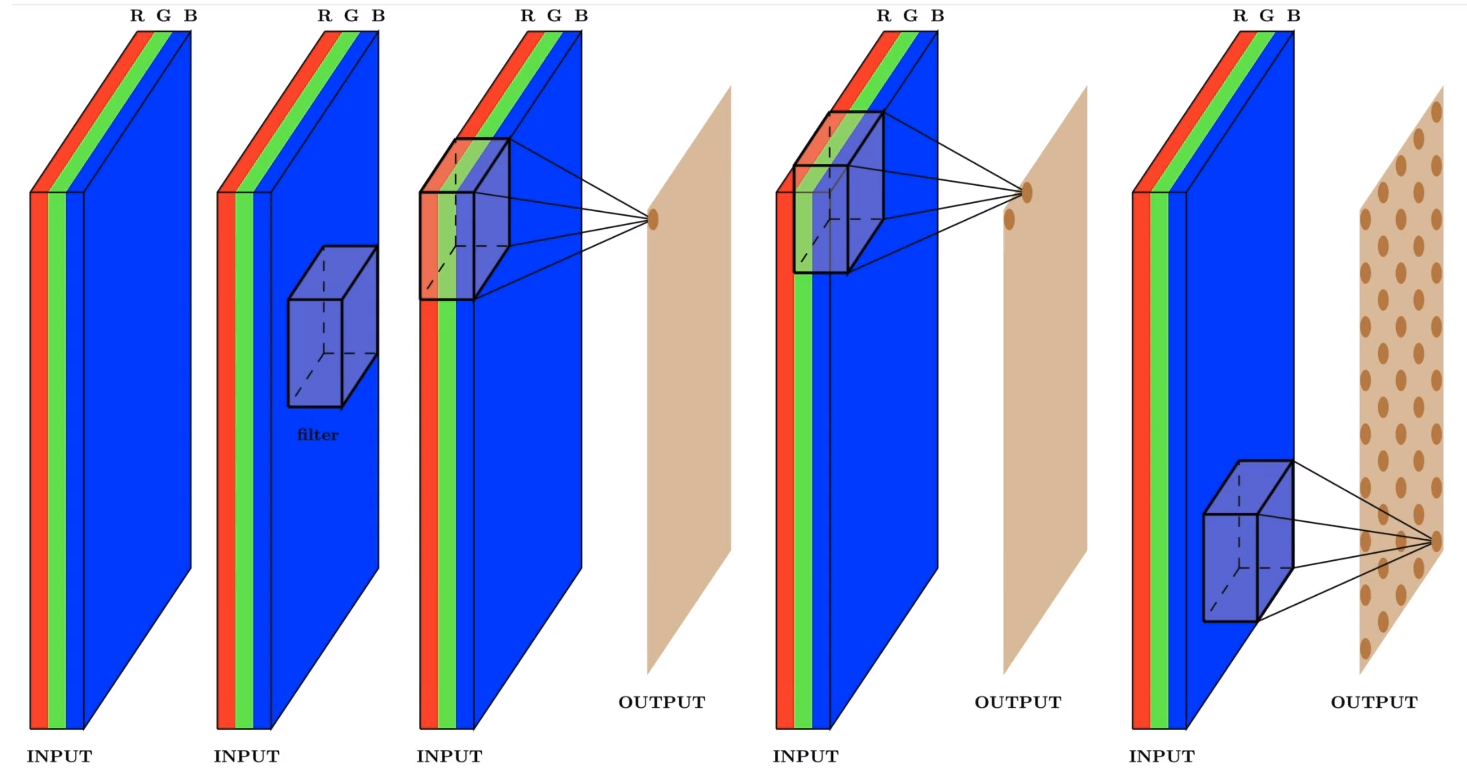
1. Another question we can ask is why we must stop at 16 layers, can we not go deeper.
2. So by combining the above two points, we can make the network deeper to reduce the number of parameters at the interface between the non-FC and the FC layers.
3. We must also make sure that the number of computations (Sliding a filter across the input) is not very high.
4. Our problem points can be summarised as follows
   1. Increase choice of filters
   2. Reduce number of parameters
   3. Reduce number of computations
   4. Make a deeper network

## 

## 9.2.2: Number of computations in a convolution layer

Let’s see how many computations are needed in a CNN.

1. We will be looking at the GoogLeNet architecture as an improvement to the VGGNet based on the points discussed in the previous section.
2. However, before that, we must look at two key concepts in the GoogLeNet layout: **1x1 convolution** and an **interesting way to perform max-pooling**.
3. To approach these two, we first need to see how many computations are needed in one convolutional layer .



Filter

**Dimensions**

**WI x HI x D**

**Dimensions**

**WO x HO**

**Filter**

**FxFxD**

**Assume S = 1 and we have used appropriate padding so that WO = WI = W and HO = HI = H**

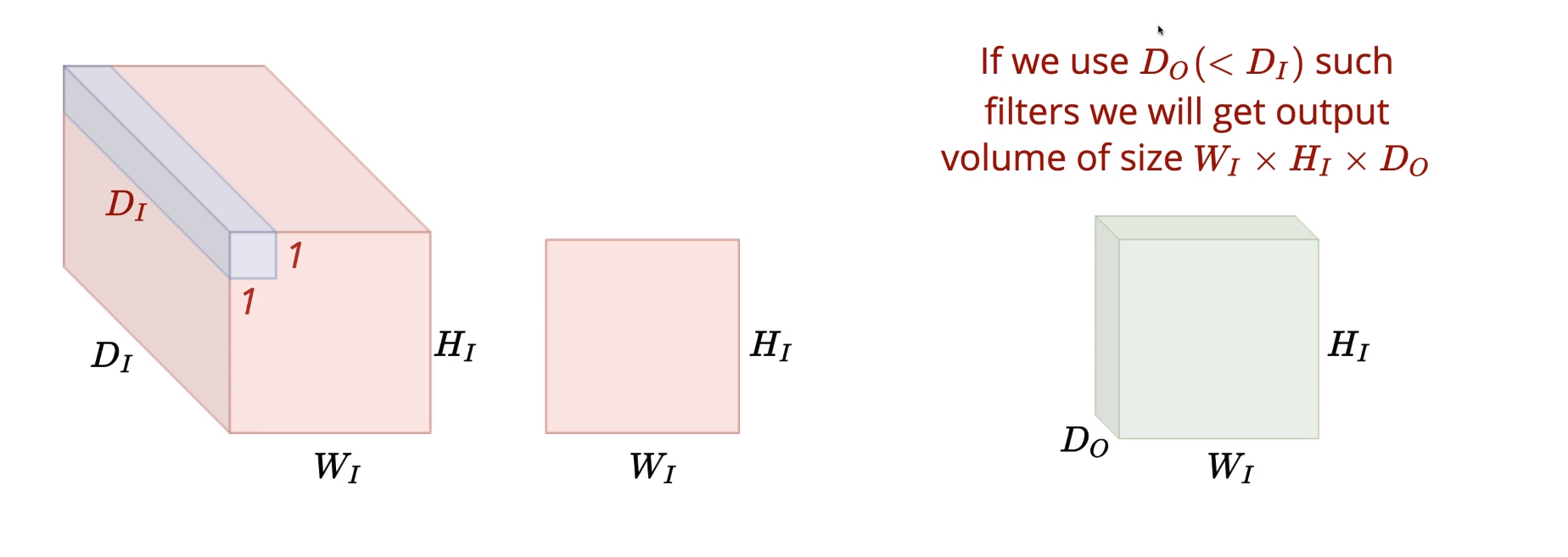
* 1. **Input dimensions**: WI x HI x D
  2. **Filter size**: F x F x DI
  3. **Output dimensions**: WO x HO
  4. Stride = 1 and appropriate padding so that**WO = WI = W and HO = HI = H**

1. To calculate the number of computations:
   1. For every pixel of interest, for D layers, we perform FxFxD computations
   2. So for an output area of WxH, we perform (WxH) x (FxFxD) computations
   3. From the previous point, we can observe that the Depth of the output layer will be very large if there is a large number of filters applied on the input layer, as each filter generates a 2D area of unit depth.
   4. So if we use a **large number of filters**, the **output volume will be very deep**, subsequently **increasing the number of computations in the next layer’s calculation** (Due to high D value).
   5. We can also try controlling W and H, but they can be more easily regulated using max-pooling. However, depth is directly related to the number of filters used.

## 9.2.3: 1x1 Convolutions

What is a 1x1 convolution used for?

1. We’ve mostly worked with 3x3 convolutions so far, i.e. a grid containing 3 rows, 3 columns with 9 cells in total.
2. The result of a single 3x3 operation is the weighted average of all the points in the grid, applied to our selected pixel.
3. Now, let us look at a 1x1 convolution

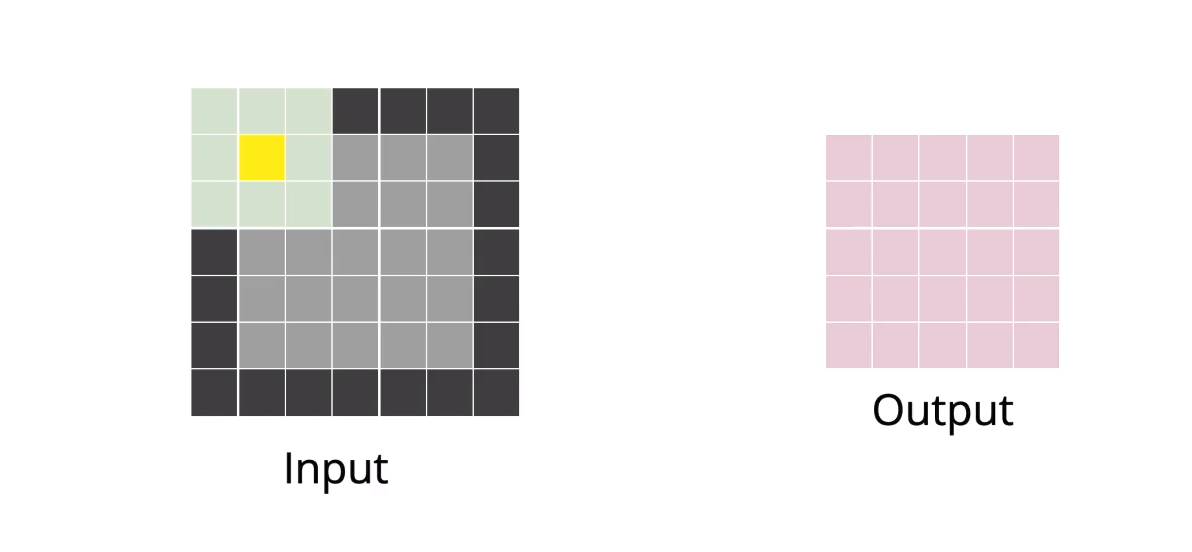


**1x1 convolutions**

Essentially **1x1xDI**

They average the values across the depth and create 2D output

1. A 1x1 kernel takes a neighborhood of 1 row and 1 column, which is essentially the pixel itself. Since the kernel is of size 1x1xDI, it computes the weighted average of all the pixels across the input depth DI
2. Here, the Input is 3D, the filter is 3D but the operation is 2D, as we are only moving horizontally and vertically. The 3D volume is compressed to a 2D area.
3. If we were to use DO number of filters, where (DO < DI), we will get an output of WI x HI x DO. Each of the 1x1 kernels will give one 2D output and DO such kernels gives us an output volume of the dimensions WI x HI x DO.
4. If DO  is much smaller thanDI, we effectively shrink the input volume while still effectively retaining the depth information (Due to averaging across depth).
5. Now, this output behaves as an input to the next layer, resulting in a much smaller number of computations due to smaller depth.
6. In a nutshell, 1x1 filters are used to compress input volumes across their depth to get a smaller output volume of same Width and Height.
7. Another operation we need to look at is Max Pooling. We usually perform Max-pooling with a Stride=2, resulting in halving the input dimensions. However, we can also perform it with a stride of 1. With S = 1 and appropriate padding, we can preserve input dimensions.



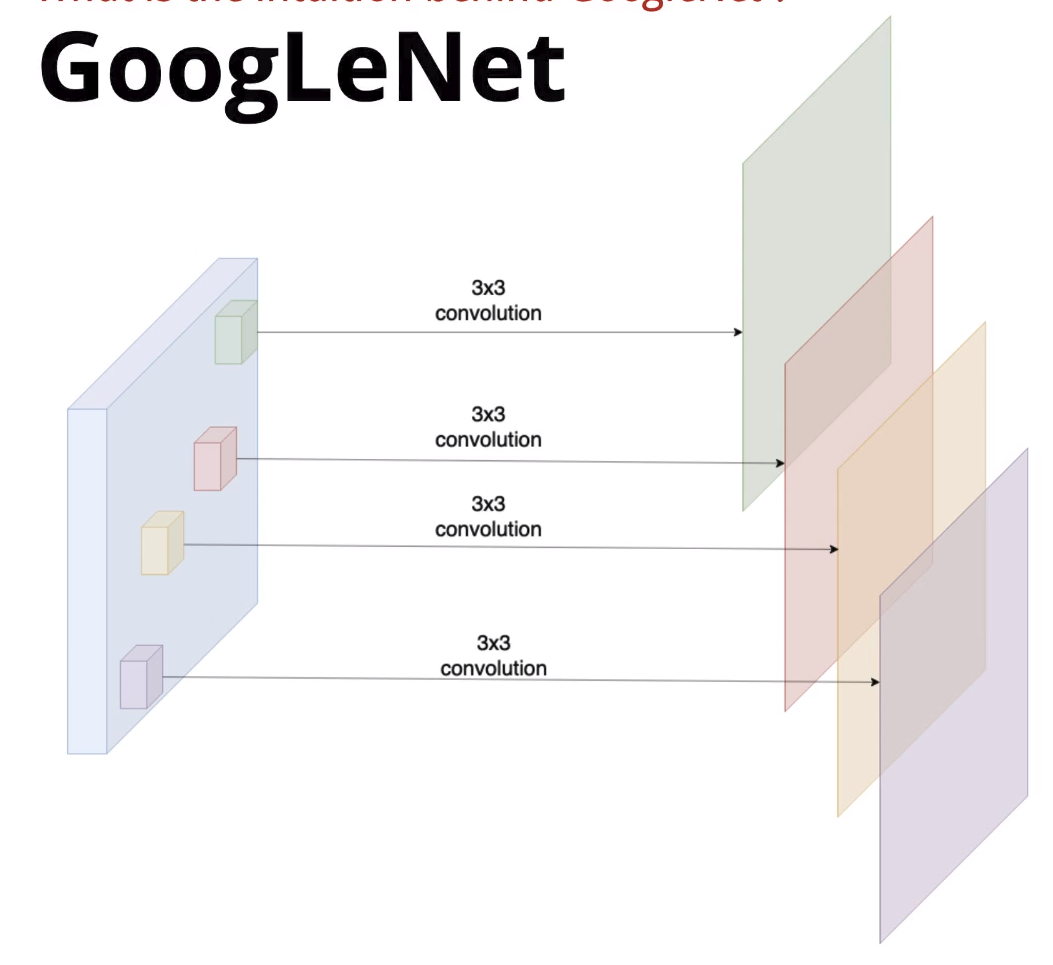
**S = 1, F = 3, P = 1**

**Output dimensions same as the input using S = 1 and P = 1 for F = 3**

## 9.2.4: The Intuition behind GoogLeNet

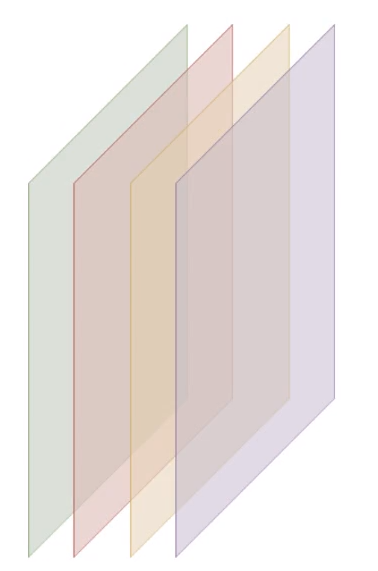
What is the intuition behind GoogLeNet?

1. One point to note in the architectures used thus far, is that we must always make a choice of a particular filter size for any given layer. For eg, in VGGNet, all filters used were 3x3
2. Another point is the interspersing of Max-pooling layers between convolutional layers. How do we decide the arrangement to follow?



Same filter size used for all operations in a layer

1. In GoogLeNet, the choice was eliminated, instead we are able to apply all our operations whatever combination of filter size and max-pooling/non that we’d like.



**Different filters and operations performed on a single layer**

**Stride and padding adjusted to maintain size b/w input and output**

D4 filters

D1 filters

D2 filters

D3 filters

D4

D1

D2

D3

Combined output depth: **D1 + D2 + D3 + D4**

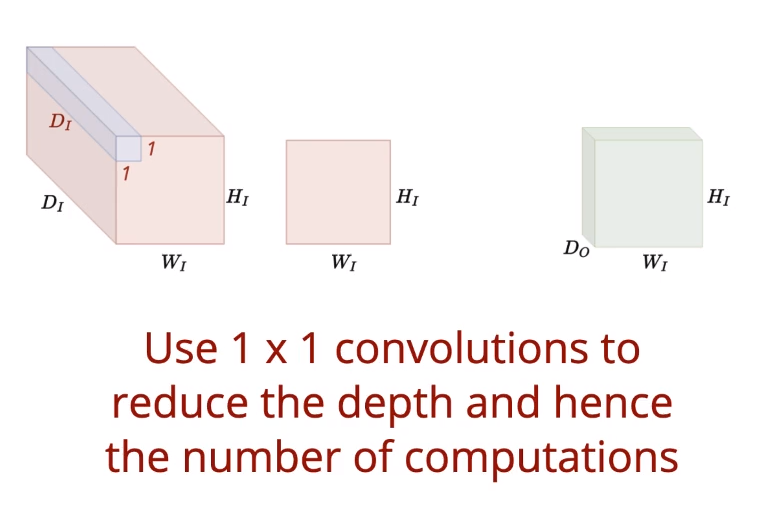
W

H

**W**

**H**

1. In GoogLeNet, we can apply multiple filters of varying size to perform either convolutional or max-pooling operations
2. Through experimentation with the older architectures, we have found that when there are multiple convolutional operations in a layer, we needn’t use larger filter sizes.
3. Therefore, 5x5 filters are usually the upper limit of size.
4. One constraint is that for each operation in a layer, appropriate padding and stride must be taken so as to preserve the width and height between the input and the output
5. The Number of filter for each operation can vary, (D0 to D3 etc)
6. In the output volume, the total depth is the combined depth of the individual volumes from each of the operations. D = D0 + D1 + D2 + D3
7. The problem with combining the depths is that it has the potential to become very large, thus drastically increasing the number of computations.
8. We can mitigate this problem by performing 1x1 convolutions



**Different filters and operations performed on a single layer**

**Stride and padding adjusted to maintain size b/w input and output**

D4 filters

D1 filters

D2 filters

D3 filters

W

H

Where

DI = D1+D2+D3 + D4

1. By performing 1x1 convolutions, we can reduce the depth of the output volume, thereby reducing the number of computations to be performed in the subsequent layers.

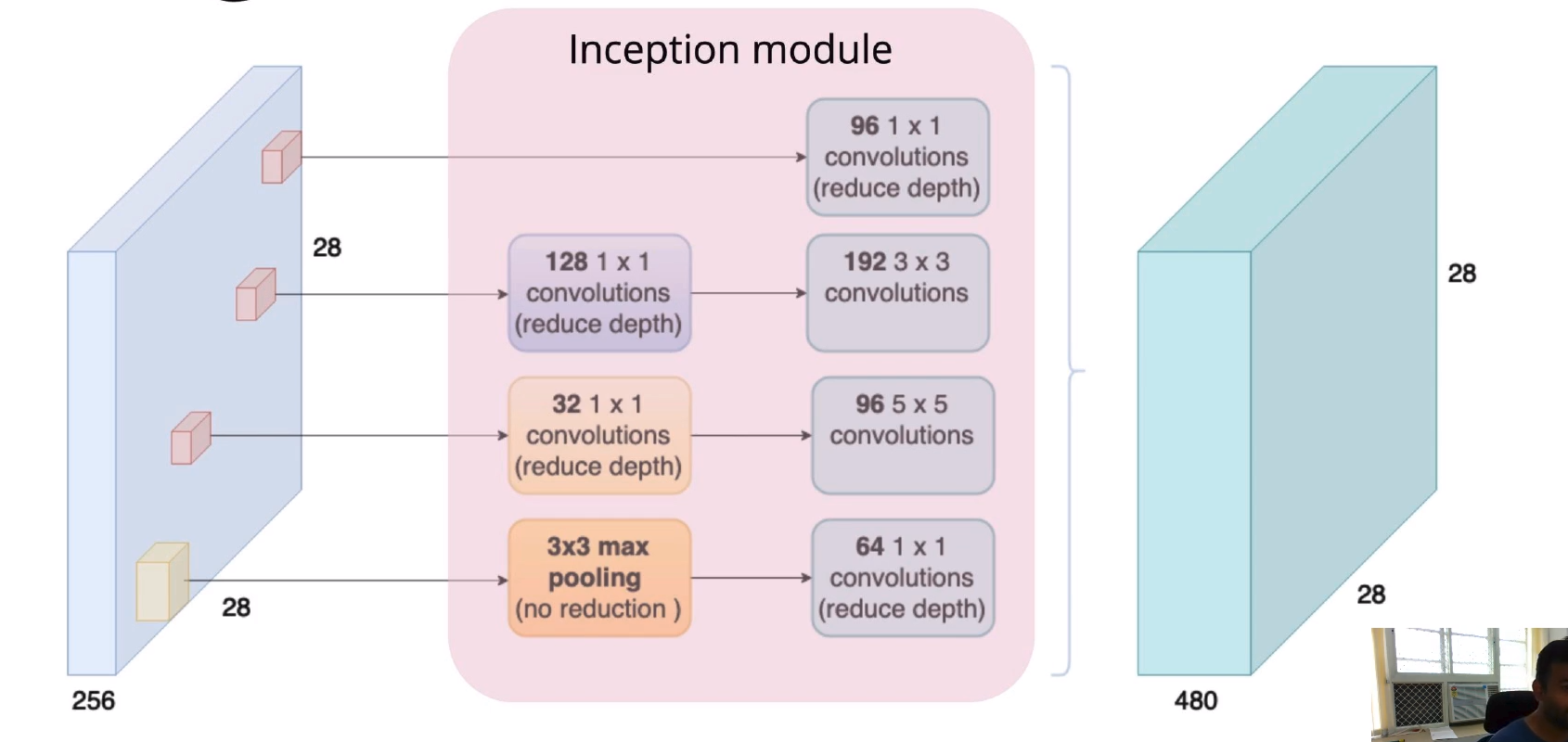
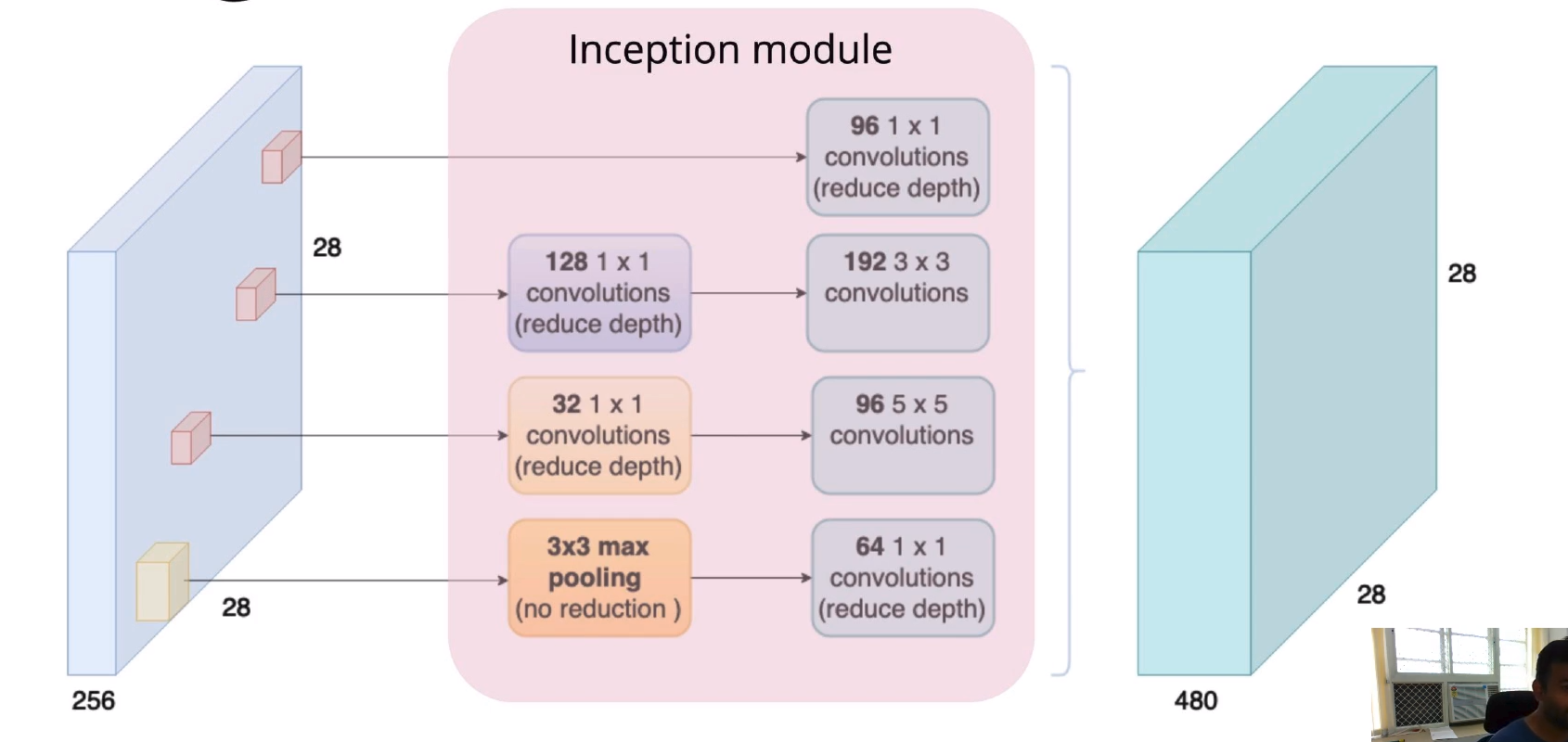
## 

## 

## 9.2.5: The Inception Module

What is the Inception Module?

1. GoogLeNet is also called Inception net. It is made up of multiple modular operation-blocks known as inception modules.
2. Let us break down a single inception module



**128x28x28**

**32x28x28**

**256x28x28**

**192x28x28**

**96x28x28**

**96x28x28**

**96x28x28**

**Combined Depth**

**480**

**96**

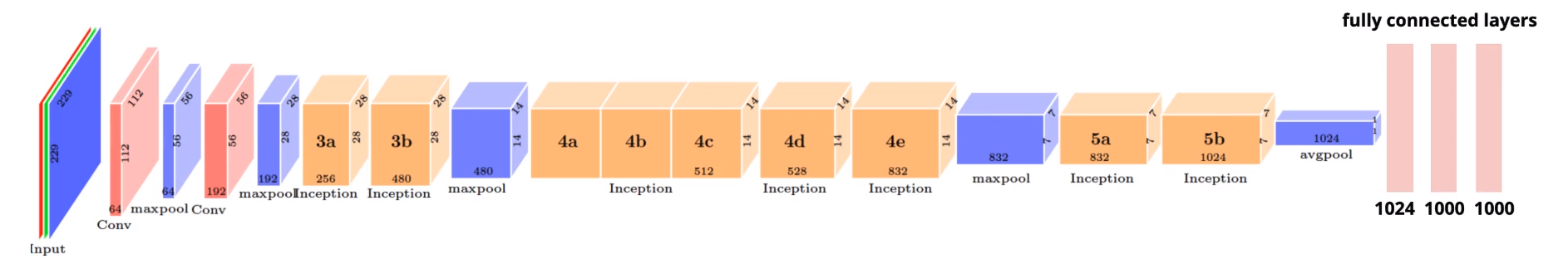
1. Here, we can see the sequence of operations performed in this inception module.
2. There are a few operations that are a blueprint of inception modules
3. Direct 1x1 convolutions
4. 1x1 convolutions followed by 3x3 and 5x5 convolutions
5. 3x3 convolutions followed by 1x1 convolutions
6. Here, The number of filters can change but the same blueprint repeats itself throughout the network.
7. Summing the output of all of these, we get the output volume as shown in the figure.

## 

## 9.2.6: The GoogLeNet Architecture

What does the full network look like?

1. Let’s take a look at the entire GoogLeNet architecture



**Input**

224x224x3

**Conv 1:**

F=3, K=64

S=2

112x112x64

**Max-pool 1**:

F=3, S=2

56x56x64

**Conv 2**:

F=3, K=192

S=1, P=1

56x56x192

**Max-pool 2**:

F=3, S=2

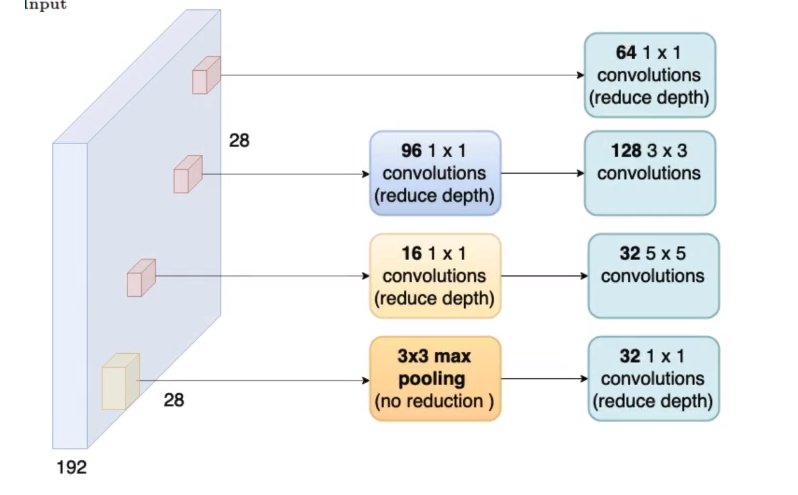
28x28x192

**Inception Modules**

**Max Pool Layers**

**Average Pool layer**

1. Up till the second max-pooling layer, the architecture is similar to what we’ve seen in earlier configurations. Post that, we begin moving into the Inception modules.
2. **Inception Module 1:**



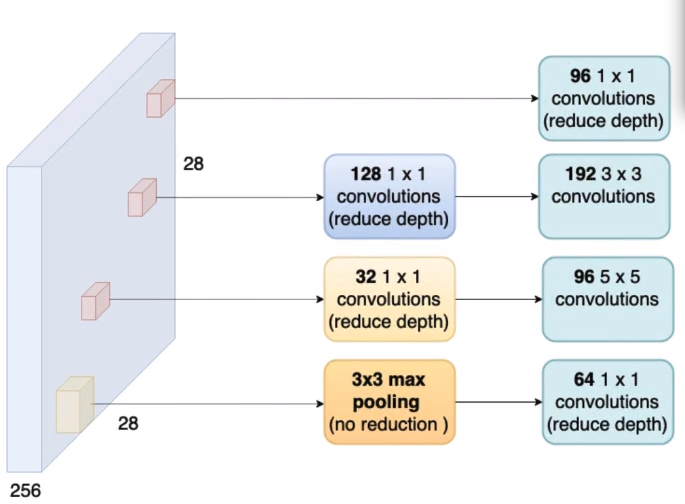
Input:

128x28x28

Output:

256x28x28

1. **Inception Module 2**:



**96**

Input:

256x28x28

Output:

480x28x28

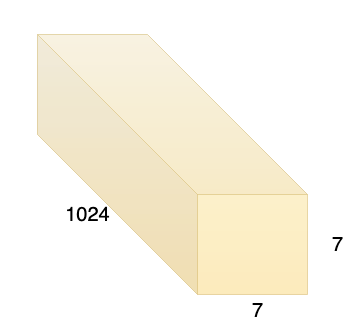
1. Then there is a **Max-pooling layer which reduces the Dimensions by Half (480x14x14)**. This Max-pooling layer is used because the Max-pooling layers in the Inception modules do not reduce the dimensions.
2. **Inception Module 3,4 & 5**:
   1. **Input at Inception Module 3** is: 480x14x14
   2. **Output at Inception Module 5** is: 512x14x14
3. **Inception Module 6**:
   1. **Input**: 512x14x14
   2. **Output**: 528x14x14
4. **Inception Module 7**:
   1. **Input**: 528x14x14
   2. **Output**: 832x14x14
5. Then there is a **Max-pooling layer which reduces the Dimensions by Half (832x7x7)**.
6. **Inception Module 8**:
   1. **Input**: 832x7x7
   2. **Output**: 832x7x7
7. **Inception Module 9**:
   1. **Input**: 832x7x7
   2. **Output**: 1024x7x7
8. **Average Pool layer is used to reduce the output from Inception Module 9**, thereby reducing the number of parameters between the non-FC and FC layer interface.
9. Each Inception Module counts as 2 layers, therefore we have more than 20 layers in GoogLeNet.

## 

## 9.2.7: Average Pooling

How does average pooling reduce the output size?

1. At the final Inception Module, we have an output dimension of 1024x7x7. If this was to directly interface with a Fully-Connected layer with a 1000 Neurons, we would get ~50 Million parameters
2. To reduce this number, Google added another layer which performs Average-pooling



**Avg. of each 7x7 grid is taken across the depth of 1024**

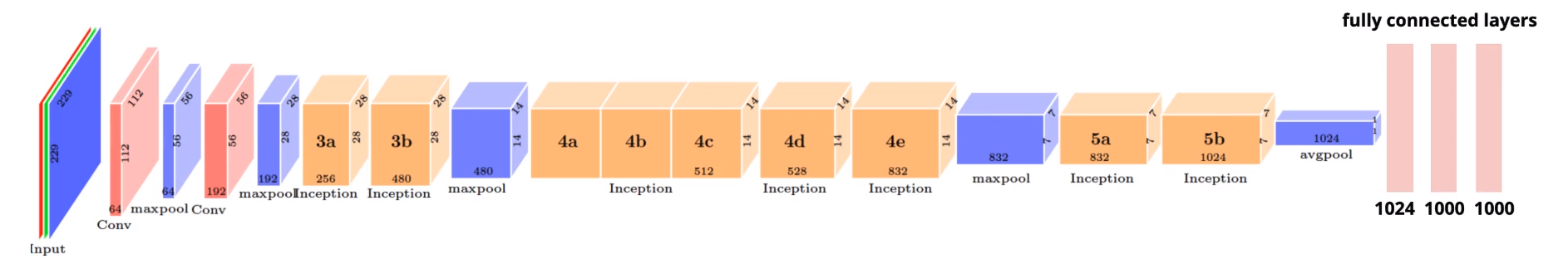
1. By taking the average value of each 7x7 slice to get a 1x1 value across the depth, we are left with a 1024x1x1 vector.
2. This vector of 1024 values interfaces with the Fully connected layer with a 1000 Neurons.
3. This gives us ~1 Million parameters, as opposed to the earlier seen 50 Million parameters.

## 

## 9.2.8: Auxiliary Loss for training a deep network

Can auxiliary loss help to train the network better?

1. Let’s look at how GoogLeNet responds to the 4 problem points from the previous CNN architectures
2. **Increase choice of filters**: Parallel convolutions/max-pooling
3. **Reduce number of parameters**: Average Pooling
4. **Reduce number of computations**: 1x1 convolutions
5. **Make a deeper network**: Has 22 layers as opposed to VGG19’s 19 layers.
6. Now, since it is a very deep network, there is a possibility for vanishing gradients to occur when backpropagating the Loss. This is mitigated using a technique called Auxiliary Loss



x

ŷ

ŷ

ŷ

Input

Output

Dummy output predictions

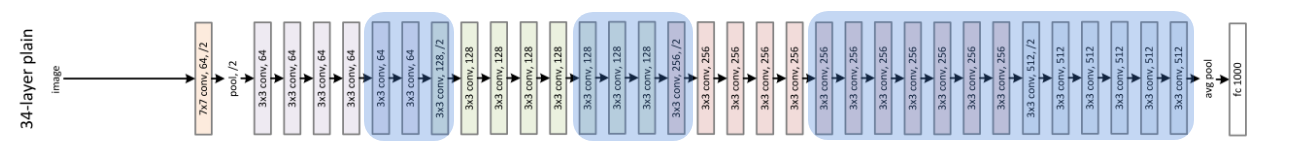
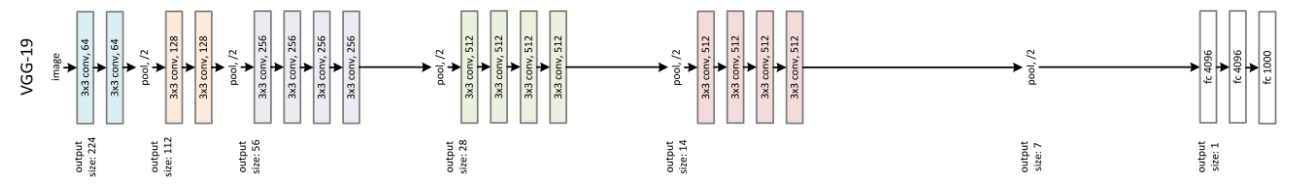
1. In addition to the final output prediction, we are also trying to make partial predictions from the above specified regions in the network.
2. We compute the loss at the final prediction and at both the dummy predictions.
3. Now we can backpropagate from the final loss or from the dummy-losses obtained, thereby shortening the effective depth of the network and lowering the chance of vanishing gradients occurring.
4. Some interesting points to note about GoogLeNet
   1. 12x less parameters than AlexNet
   2. 2x more computations than AlexNet
   3. Improved performance on ImageNet

## 

## 9.2.9: ResNet

What happens if you increase the depth of the network?

1. Let us consider the basic architecture of the VGG-19 network when compared to another 34 layer network architecture

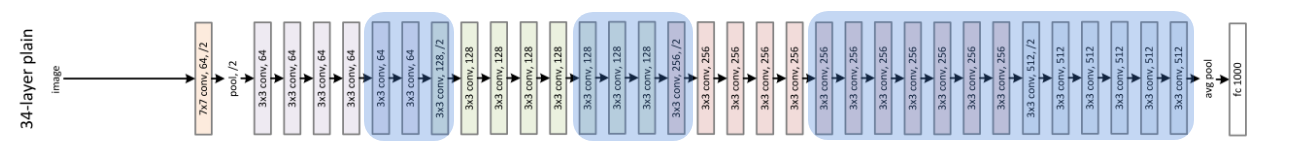
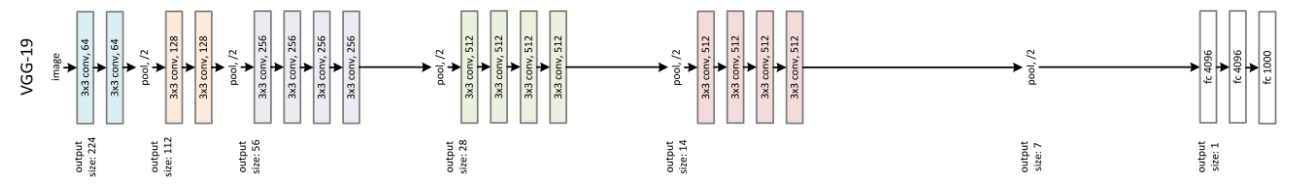


Additional layers added

1. The train/test error curves were plotted for some other 20-layer and 56-layer networks

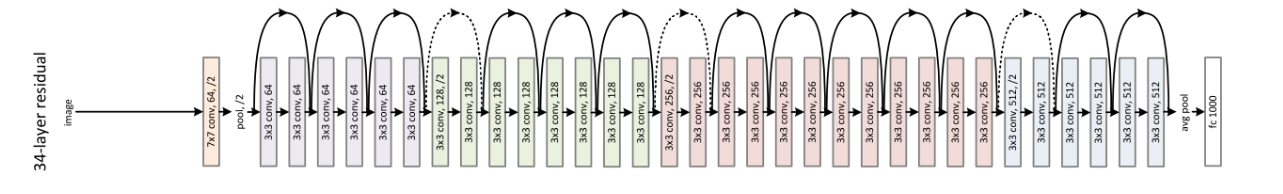
|  |  |
| --- | --- |
| **Training curves** | **Test Curves** |
|  |  |
| Unexpectedly, the 56-layer had a higher error than the 20-layer. It was expected to have overfit the training data, thereby having a lower training error.  It was hypothesised that the gradients were not able to flow well through this deeper network. | Here, as predicted, the 56-layer performed worse than the 20-layer due to overfitting. |

1. Now, when comparing the 19 and 34-layer networks, we see that at the very least, the 34-layer network should be able to match the performance of the 19-layer network.
2. Matching the performance could be done by bypassing the additional layers in the 34-layer network by using Identity Mapping.
3. Identity mapping refers to learning a filter values such that the output is preserved identical across a finite number of layers, till it reaches the target layer. Basically cloning the layer output till required.



Identity mapping

Essentially bypassing the highlighted layers

1. However, it wasn’t able to match the 19-layer’s error. This implies that the information from the input is getting highly morphed and by the time we reach the output, it is highly transformed.
2. A simple solution would be to keep passing the input information repeatedly in stages.
3. To attempt this, they tried the Residual Network or the ResNet
4. In the ResNet, **every two layers**, **we pass the input given to the first layer along with the output obtained at the second layer.**
   1. Input: x1
   2. Output: x2  = f(x1) + x1
   3. Output after two layers: x3 = f(x2) + x2
5. This helped the gradients to flow back better and the training to improve
6. It is called a Residual Network because at every stage, there is a residue of the input which is passed once again with the output.
7. Using this technology, they were able to train very deep Neural networks of up to 151 layers.
8. The ResNet showed remarkable performance among the various tasks
9. It was the winner among the 4 main tasks across the following datasets
   1. ImageNet Classification **(ResNet-151)**
   2. ImageNet Localization **(ResNet-101)**
   3. ImageNet Detection **(ResNet-101)**
   4. Coco Detection **(ResNet-101)**
   5. Coco Segmentation **(ResNet-101)**
10. Some of the popular ResNets are **(ResNet-51, ResNet-101 and ResNet-151)**