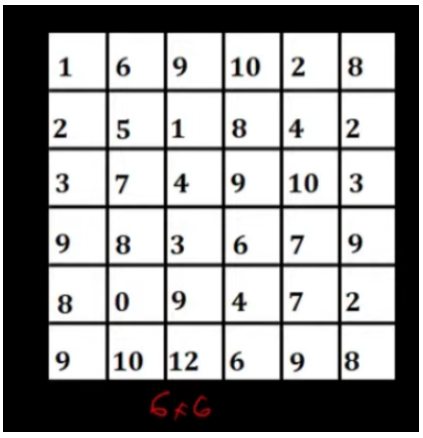
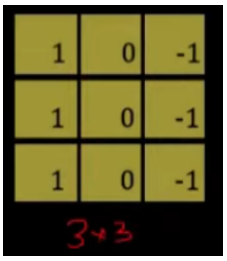
I. What is CNN in DL
   A. The main idea is to use a filter which is a sliding window that is responsible for detecting patterns in an image.
   B. We use it to detect edges in an image which is passed on to the next layer, which might detect features associated with the image and

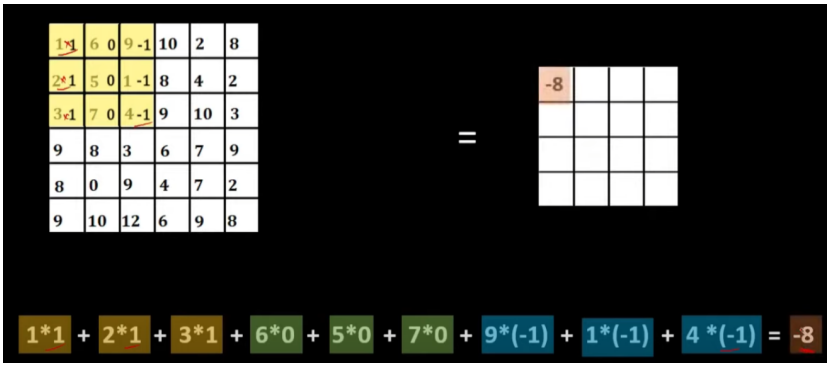
II. Convolution Operations in CNN
   A. Responsible for detecting edges and features in image
   B. 

=Gray Scale image

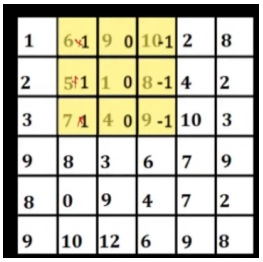


=Filter

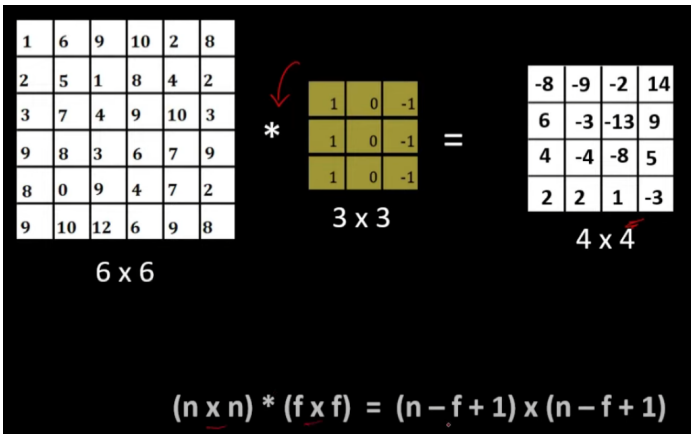
Convolution between filter and Gray scale image is



After the steps above we shift the filter one pixel to the right and re-run

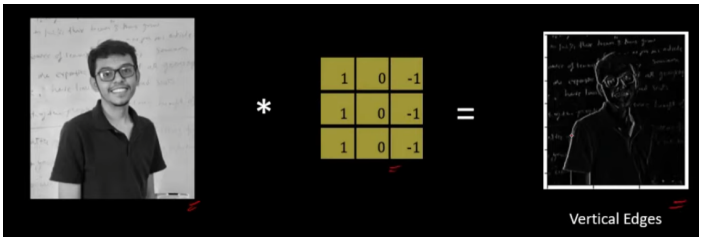


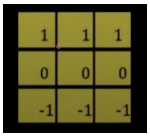
C. Using the above to create a formula for what convolution multiplication will finally look like

1. 

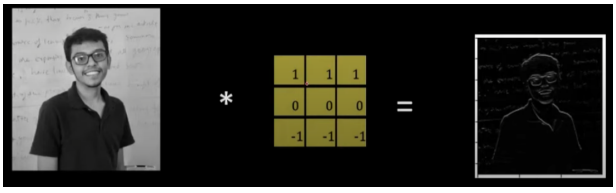
this

D. By using the filter above we are able to find the vertical edges of an image

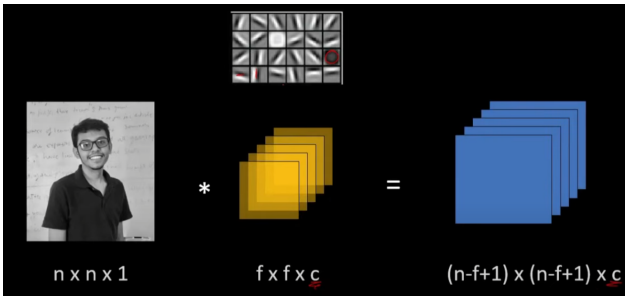
1. 

E. 

This filter provides you with the horizontal edges

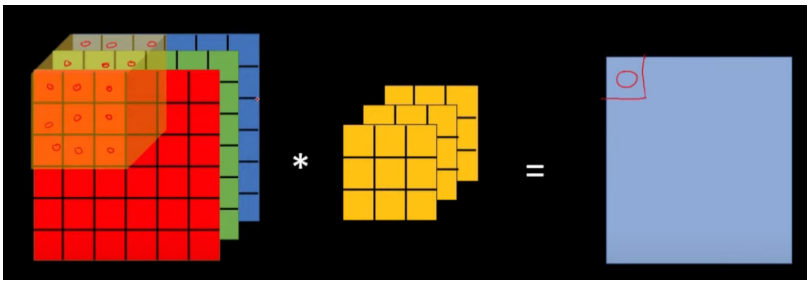
1. 

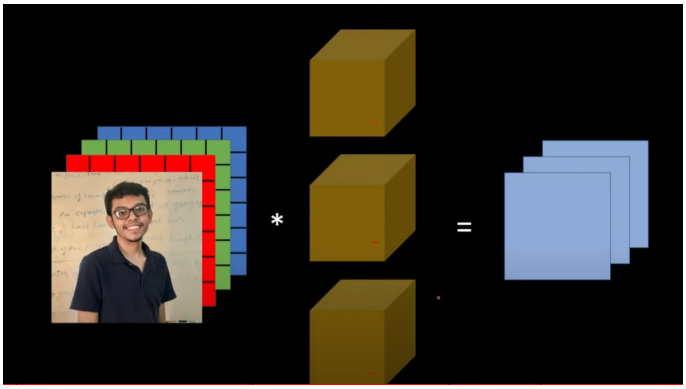
F. In a Single layer we will be using many filters to detect many different edges

1. 

    a) By using C filters to find C different type of edges our output has C different images outputed after this layer

G. A convolution on a colored image results in a 3 dimensional filter

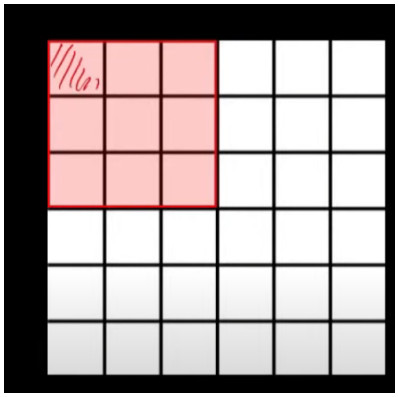
1.  you have to multiply by the filter then add all values up
2. MultiFilter example with a color image

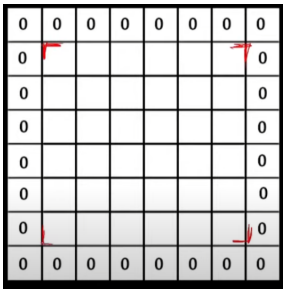
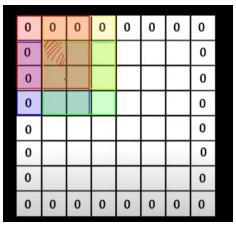
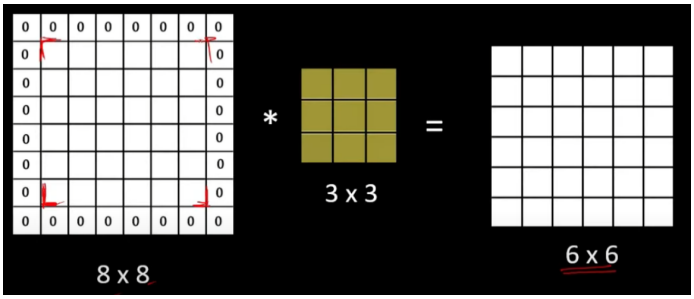


III. Padding CNN

A. Problems with Convolution

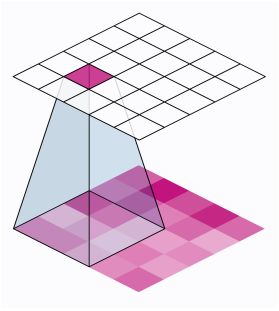
1. Firstly the size of the image is heavily reduced as we go from one layer to the next resulting in a significant loss of information
2. The top left pixel will only get exposed to the filter once as seen below as the filter moves one pixel length the border pixels are only seen once or

twice compared to a pixel closer to the center of the image which will be seen many more times 

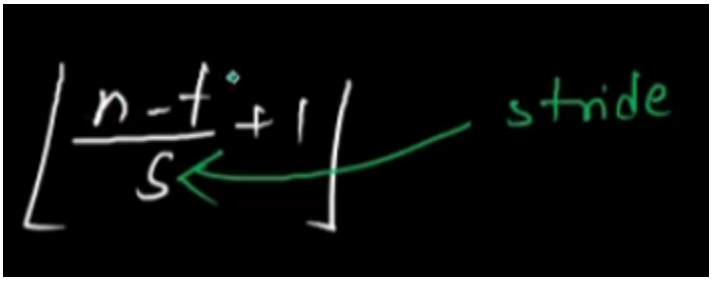
B. How to fix problem above - Padding
    1. Add a border of 0's all around the image
        a)  Here we pad with one pixel
    2.  proof that problem is fixed because the pixel gets exposed to the filter a number of times
C. Two ways to do padding
    1. Valid Convolution: Perform convolution with no padding at all
    2. Same Convolution: After performing the convolution operation the output image should be the same size as the original image minus the padding
        a) 

IV. Stride in Convolutional Neural Network

A. Strides are either to the right or down with a reset all the way to the left

1.  Look at this example to understand the down-by-one shift where the stride is of length 1
2. We can have different stride lengths

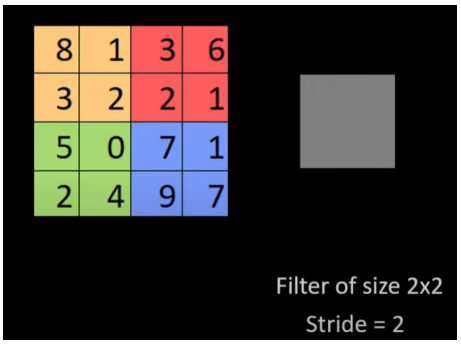
B. How to find the size of the image after convolving with the filter with stride S

1. 

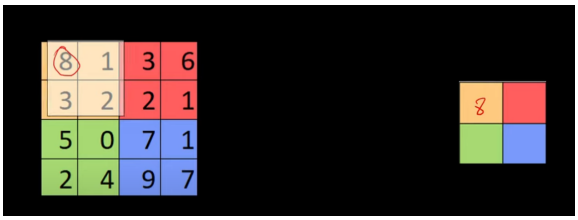
   a) N and F are the size of the original image and filter respectively
   b) Floor of the value above to make it a whole number

V. Max Pooling in CNN

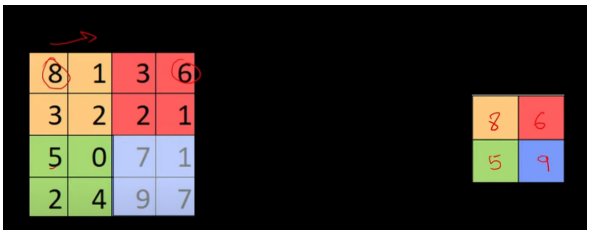
A. Steps for Max pooling

1. 

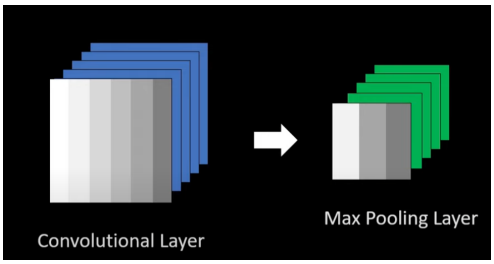
usually the Stride = length of the image
2. Start at the top left and choose the largest value in the filter

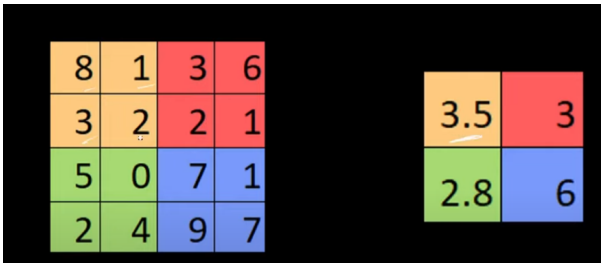


3. Slide by the stride amount and continue doing the same operation of finding the largest number in the filter

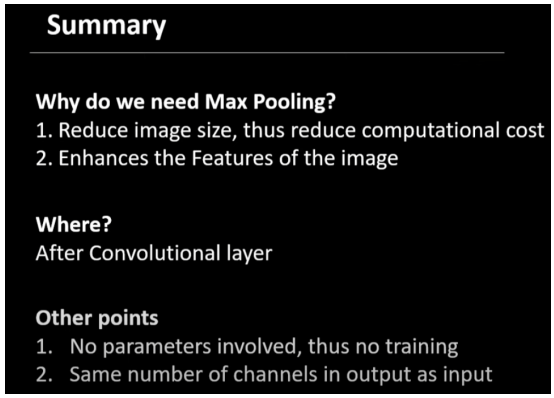


B. Why do we use max pooling
   1. Reduce image size
   2. Reduce computational cost
   3. Preserves features of an image - Enhances them because preserves features of the image by using the maximum value
C. Where is it applied in the workflow
   1. It is applied as the next step after the convolution



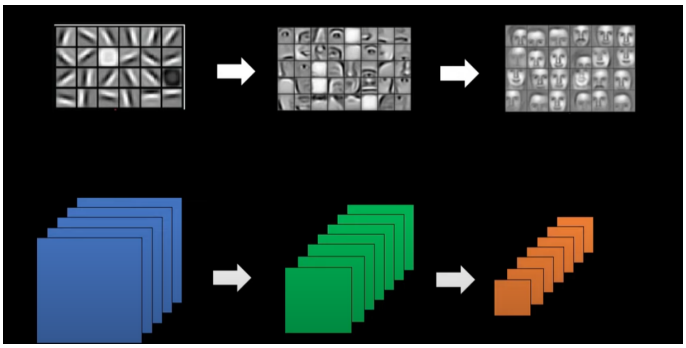
   2. It is applied here so that we can reduce the dimension of the convolution output and enhance the features
D. Average Pooling
   1. Like max pooling however now you take the average of the numbers in the filter rather than returning back the largest number in the filter
   2. 

ex 3.5 is avg(8+1+2+3)
E. Summary

1. 

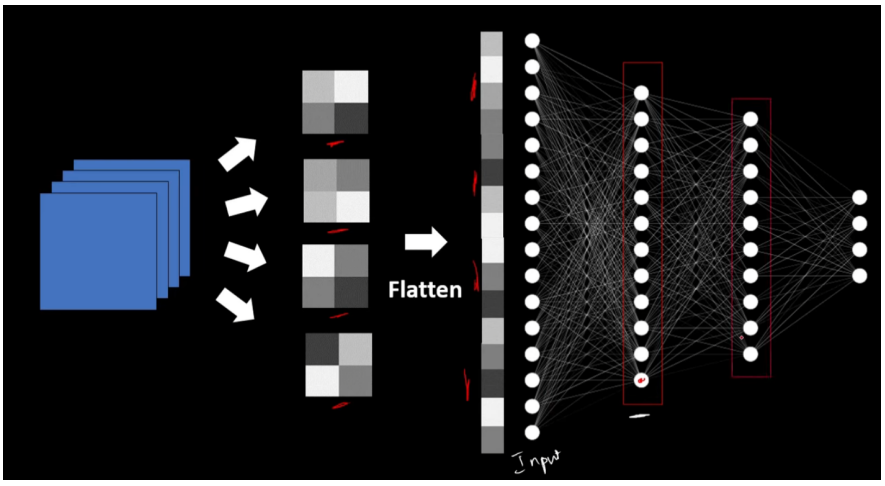
VI. Fully Connected Layers in CNN

A. Convolutional layers are used to extract the many features

1. 

B. How do we classify after the last convolutional layer above We use fully Connected layers?

C. Fully Connected layers every neuron in this layer is connected to every neuron in the next layer

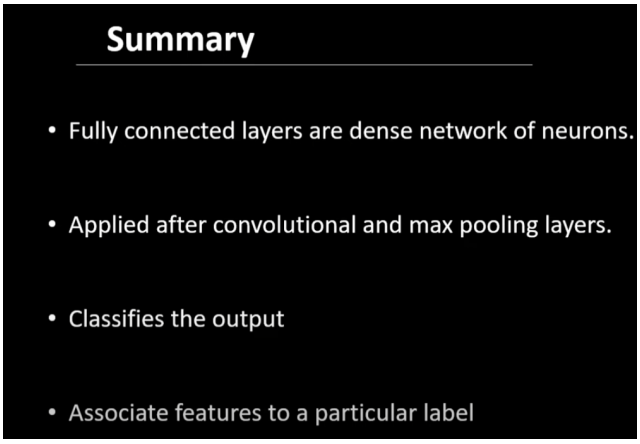
D. 

1. After we get the four images above we then give the images to the fully connected neural network by flattening the images into a single array
2. Number of neurons in the last fully connected layer =’s the different classifications
    a) So above it is 4 different classifications that can be made using SOFTMAX because we are trying to predict from 4 different things

b) If there were two elements we would SIGMOID activation function

3. Connections between neurons are all the weights(trainable parameter) and things that the parameter has to learn
    a) Used so that certain features are associated better with certain categories

E. Summary

1. 

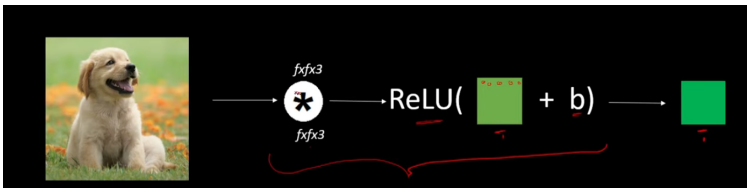
2. All above creates forward propagation

VII. CNN architecture

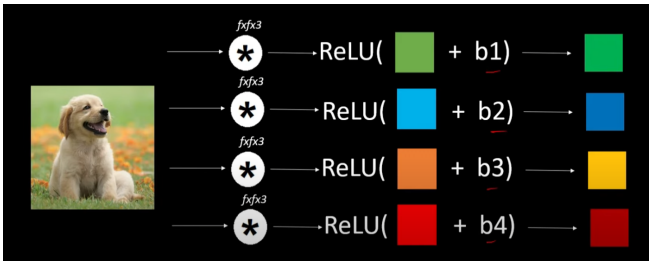
A. Convolution Layer

1. How its created
    a) Convolution between image and filter to create output
    b) Scale output by adding a bias and using a nonlinear function such as ReLU or Tanh on output+bias

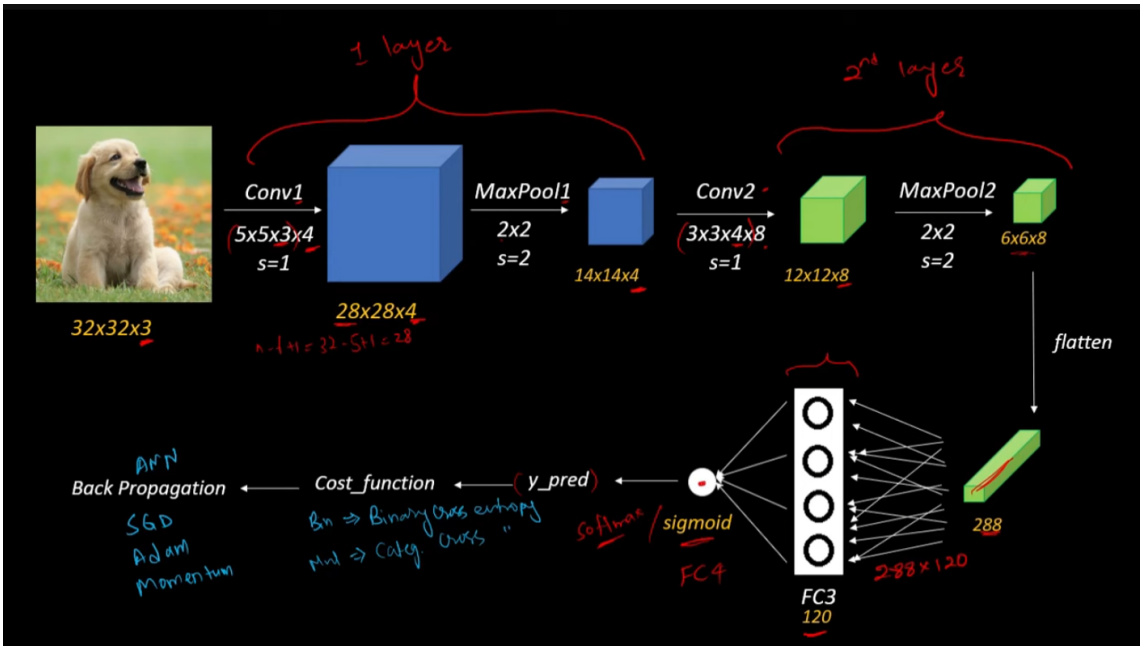
        (1) Output of the scaling is the same size as the original output

2. 

B. Example of finding the convolutional layer of 4 filters on the same input image Creating a 4 Channel output

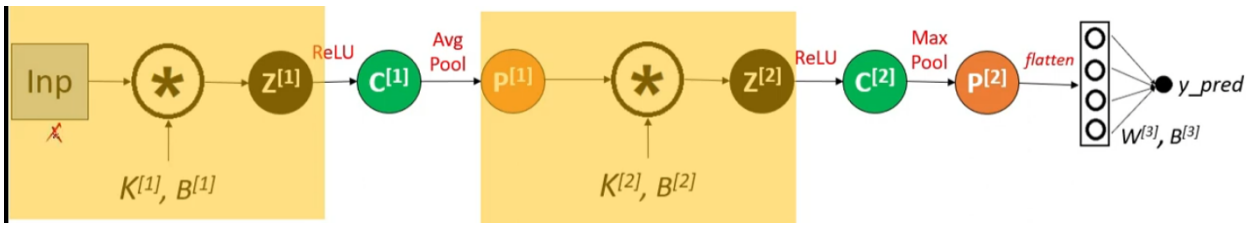
1. 

C. Architecture(Workflow)

1. 

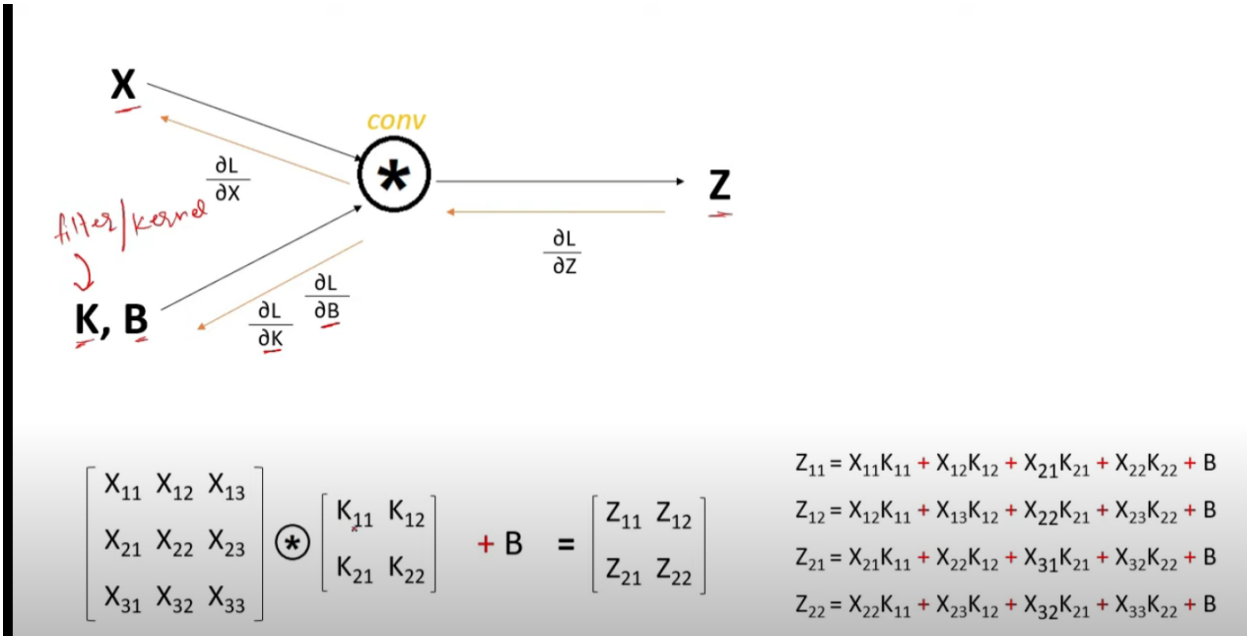
   a) Keep in mind max pooling layers can be skipped if we are trying to create a large CNN because we want to preserve all features
      (1) Pooling is better for smaller architectures
   b) Each number in the flattened array is its own neuron and therefore to FC3 there are weights from each number to neuron connection
   c) Sigmoid or Softmax is applied in the final layer
      (1) Softmax = Multi Classification Model
         (a) Number of neurons in final layer = Number of Classifications
      (2) Sigmoid = Dual Classification model
   d) Cost function is just for training
   e) Back Propagation is how we minimize the cost function

VIII. Back Propagation for the convolution



   a) We find how to do the back propagation on yellow part above

B.



$$\begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \\ X_{31} & X_{32} & X_{33} \end{bmatrix} \circledast \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} + B = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix}$$

$$Z_{11} = X_{11}K_{11} + X_{12}K_{12} + X_{21}K_{21} + X_{22}K_{22} + B$$
$$Z_{12} = X_{12}K_{11} + X_{13}K_{12} + X_{22}K_{21} + X_{23}K_{22} + B$$
$$Z_{21} = X_{21}K_{11} + X_{22}K_{12} + X_{31}K_{21} + X_{32}K_{22} + B$$
$$Z_{22} = X_{22}K_{11} + X_{23}K_{12} + X_{32}K_{21} + X_{33}K_{22} + B$$

1. K = Kernal which is the filter
2. B is Bias
3. Use partial derivative $\frac{\partial L}{\partial Z}$ to find partial derivatives $\frac{\partial L}{\partial K}\ \frac{\partial L}{\partial B}\ \frac{\partial L}{\partial X}$
4. We use $\frac{\partial L}{\partial K}\ \frac{\partial L}{\partial B}$ to update the k Parameters using the formula

$$K = K - \alpha \cdot \frac{\partial L}{\partial K}$$

$$B = B - \alpha \cdot \frac{\partial L}{\partial B}$$

C. dL/dK

1. Given $$\frac{\partial L}{\partial K_{mn}} = \sum^{1} \frac{\partial L}{\partial Z_{ij}} * \frac{\partial Z_{ij}}{\partial K_{mn}}$$ Fill in using

$$Z_{11} = X_{11}K_{11} + X_{12}K_{12} + X_{21}K_{21} + X_{22}K_{22} + B$$
$$Z_{12} = X_{12}K_{11} + X_{13}K_{12} + X_{22}K_{21} + X_{23}K_{22} + B$$
$$Z_{21} = X_{21}K_{11} + X_{22}K_{12} + X_{31}K_{21} + X_{32}K_{22} + B$$
$$Z_{22} = X_{22}K_{11} + X_{23}K_{12} + X_{32}K_{21} + X_{33}K_{22} + B$$

2. Doing above returns

$$\frac{\partial L}{\partial K_{11}} = \left(\frac{\partial L}{\partial Z_{11}} * \frac{\partial Z_{11}}{\partial K_{11}}\right) + \left(\frac{\partial L}{\partial Z_{12}} * \frac{\partial Z_{12}}{\partial K_{11}}\right) + \left(\frac{\partial L}{\partial Z_{21}} * \frac{\partial Z_{21}}{\partial K_{11}}\right) + \left(\frac{\partial L}{\partial Z_{22}} * \frac{\partial Z_{22}}{\partial K_{11}}\right)$$
$$\frac{\partial L}{\partial K_{12}} = \frac{\partial L}{\partial Z_{11}} * \frac{\partial Z_{11}}{\partial K_{12}} + \frac{\partial L}{\partial Z_{12}} * \frac{\partial Z_{12}}{\partial K_{12}} + \frac{\partial L}{\partial Z_{21}} * \frac{\partial Z_{21}}{\partial K_{12}} + \frac{\partial L}{\partial Z_{22}} * \frac{\partial Z_{22}}{\partial K_{12}}$$
$$\frac{\partial L}{\partial K_{21}} = \frac{\partial L}{\partial Z_{11}} * \frac{\partial Z_{11}}{\partial K_{21}} + \frac{\partial L}{\partial Z_{12}} * \frac{\partial Z_{12}}{\partial K_{21}} + \frac{\partial L}{\partial Z_{21}} * \frac{\partial Z_{21}}{\partial K_{21}} + \frac{\partial L}{\partial Z_{22}} * \frac{\partial Z_{22}}{\partial K_{21}}$$
$$\frac{\partial L}{\partial K_{22}} = \frac{\partial L}{\partial Z_{11}} * \frac{\partial Z_{11}}{\partial K_{22}} + \frac{\partial L}{\partial Z_{12}} * \frac{\partial Z_{12}}{\partial K_{22}} + \frac{\partial L}{\partial Z_{21}} * \frac{\partial Z_{21}}{\partial K_{22}} + \frac{\partial L}{\partial Z_{22}} * \frac{\partial Z_{22}}{\partial K_{22}}$$

   a) Notice when taking the $\frac{\partial Z_{11}}{\partial K_{11}}$ derivatives for diff ij and mn terms we hold everything else constant except for where i = 1 and j=1 for the partial derivative $\frac{\partial Z_{11}}{\partial K_{11}}$ which = X_{11} therefore $\frac{\partial Z_{12}}{\partial K_{11}}$ = X_{12} and etc

3. Using the partial derivative explanation above we can simplify the above to get

$$\frac{\partial L}{\partial K_{11}} = \frac{\partial L}{\partial Z_{11}} * X_{11} + \frac{\partial L}{\partial Z_{12}} * X_{12} + \frac{\partial L}{\partial Z_{21}} * X_{21} + \frac{\partial L}{\partial Z_{22}} * X_{22}$$
$$\frac{\partial L}{\partial K_{12}} = \frac{\partial L}{\partial Z_{11}} * X_{12} + \frac{\partial L}{\partial Z_{12}} * X_{13} + \frac{\partial L}{\partial Z_{21}} * X_{22} + \frac{\partial L}{\partial Z_{22}} * X_{23}$$
$$\frac{\partial L}{\partial K_{11}} = \frac{\partial L}{\partial Z_{11}} * X_{21} + \frac{\partial L}{\partial Z_{12}} * X_{22} + \frac{\partial L}{\partial Z_{21}} * X_{31} + \frac{\partial L}{\partial Z_{22}} * X_{32}$$
$$\frac{\partial L}{\partial K_{11}} = \frac{\partial L}{\partial Z_{11}} * X_{22} + \frac{\partial L}{\partial Z_{12}} * X_{23} + \frac{\partial L}{\partial Z_{21}} * X_{32} + \frac{\partial L}{\partial Z_{22}} * X_{33}$$

4. Notice above can be simplified to $$\frac{\partial L}{\partial K} = \text{conv}(X, \frac{\partial L}{\partial Z})$$

D. dL/dB

$$\frac{\partial L}{\partial B} = \sum \frac{\partial L}{\partial Z_{ij}} * \frac{\partial Z_{ij}}{\partial B}$$

1. Given and

$$Z_{11} = X_{11}K_{11} + X_{12}K_{12} + X_{21}K_{21} + X_{22}K_{22} + B$$
$$Z_{12} = X_{12}K_{11} + X_{13}K_{12} + X_{22}K_{21} + X_{23}K_{22} + B$$
$$Z_{21} = X_{21}K_{11} + X_{22}K_{12} + X_{31}K_{21} + X_{32}K_{22} + B$$
$$Z_{22} = X_{22}K_{11} + X_{23}K_{12} + X_{32}K_{21} + X_{33}K_{22} + B$$

2. Using the above and plugging into the formula we get

$$\frac{\partial L}{\partial Z_{11}} * \frac{\partial Z_{11}}{\partial B} + \frac{\partial L}{\partial Z_{12}} * \frac{\partial Z_{12}}{\partial B} + \frac{\partial L}{\partial Z_{21}} * \frac{\partial Z_{21}}{\partial B} + \frac{\partial L}{\partial Z_{22}} * \frac{\partial Z_{22}}{\partial B}$$

   a) Notice the partial derivative of $\frac{\partial Z_{11}}{\partial B}$ is

$$\frac{\partial Z_{11}}{\partial B} = X_{11}K_{11} + X_{12}K_{12} + X_{21}K_{21} + X_{22}K_{22} + B$$

      (1) We hold everything as a constant except for B and partial derivative of B is 1
      (2) This statement holds true for rest of the partial derivatives

3. Therefore using the logic above we can simplify the question found above to $\frac{\partial L}{\partial B} = \frac{\partial L}{\partial Z_{11}} + \frac{\partial L}{\partial Z_{12}} + \frac{\partial L}{\partial Z_{21}} + \frac{\partial L}{\partial Z_{22}}$

4. Summing above returns $\frac{\partial L}{\partial B} = \text{sum}(\frac{\partial L}{\partial Z})$

E. Finding dl/dx

$$\frac{\partial L}{\partial X_{mn}} = \sum_{i,j} \frac{\partial L}{\partial Z_{ij}} * \frac{\partial Z_{ij}}{\partial X_{mn}}$$

$$Z_{11} = X_{11}K_{11} + X_{12}K_{12} + X_{21}K_{21} + X_{22}K_{22} + B$$
$$Z_{12} = X_{12}K_{11} + X_{13}K_{12} + X_{22}K_{21} + X_{23}K_{22} + B$$
$$Z_{21} = X_{21}K_{11} + X_{22}K_{12} + X_{31}K_{21} + X_{32}K_{22} + B$$
$$Z_{22} = X_{22}K_{11} + X_{23}K_{12} + X_{32}K_{21} + X_{33}K_{22} + B$$

$$\begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \\ X_{31} & X_{32} & X_{33} \end{bmatrix} \circledast \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} + B = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix}$$

$$\frac{\partial L}{\partial X_{11}} = \frac{\partial L}{\partial Z_{11}} * \left(\frac{\partial Z_{11}}{\partial X_{11}}\right) = \frac{\partial L}{\partial Z_{11}} * K_{11}$$

$$\frac{\partial L}{\partial X_{12}} = \left(\frac{\partial L}{\partial Z_{11}} * \frac{\partial Z_{11}}{\partial X_{12}}\right) + \left(\frac{\partial L}{\partial Z_{12}} * \frac{\partial Z_{12}}{\partial X_{12}}\right)$$

1.

   a) Notice to find all the different terms for derivative we first choose an X at m,n and we look at what values that x can affection output z based on the filters that are created to create the equations
      (1) EX: X_{11} can only affect Z_{11} because when doing convolution the only filter that takes into account X_{11} is the green filter that outputs Z_{11} whereas for X_{12} it is in both green and blue filter therefore affects both Z_{11} and Z_{12}
   b) You do the above for every single X

$$\frac{\partial L}{\partial X_{11}} = \frac{\partial L}{\partial Z_{11}} * K_{11}$$
$$\frac{\partial L}{\partial X_{12}} = \frac{\partial L}{\partial Z_{11}} * K_{12} + \frac{\partial L}{\partial Z_{12}} * K_{11}$$
$$\frac{\partial L}{\partial X_{13}} = \frac{\partial L}{\partial Z_{12}} * K_{12}$$
$$\frac{\partial L}{\partial X_{21}} = \frac{\partial L}{\partial Z_{11}} * K_{21} + \frac{\partial L}{\partial Z_{21}} * K_{11}$$
$$\frac{\partial L}{\partial X_{22}} = \frac{\partial L}{\partial Z_{11}} * K_{22} + \frac{\partial L}{\partial Z_{12}} * K_{21} + \frac{\partial L}{\partial Z_{21}} * K_{12} + \frac{\partial L}{\partial Z_{22}} * K_{11}$$
$$\frac{\partial L}{\partial X_{23}} = \frac{\partial L}{\partial Z_{12}} * K_{22} + \frac{\partial L}{\partial Z_{22}} * K_{12}$$
$$\frac{\partial L}{\partial X_{31}} = \frac{\partial L}{\partial Z_{21}} * K_{21}$$
$$\frac{\partial L}{\partial X_{32}} = \frac{\partial L}{\partial Z_{21}} * K_{22} + \frac{\partial L}{\partial Z_{22}} * K_{21}$$
$$\frac{\partial L}{\partial X_{33}} = \frac{\partial L}{\partial Z_{22}} * K_{22}$$

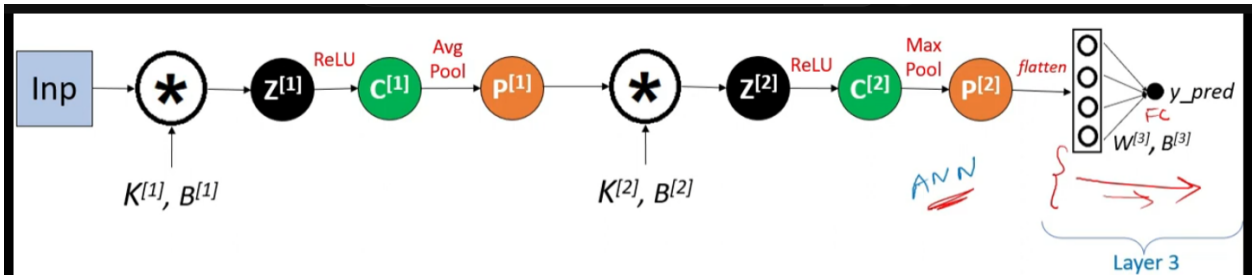
2. Above sums into these 9 Equations which can be simplified even more to look like the convolution of padded output and the invariant filter matrix which is just a 180^0 rotation of the kernel

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{\partial L}{\partial Z_{11}} & \frac{\partial L}{\partial Z_{12}} & 0 \\ 0 & \frac{\partial L}{\partial Z_{21}} & \frac{\partial L}{\partial Z_{22}} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \circledast \begin{bmatrix} K_{22} & K_{21} \\ K_{12} & K_{11} \end{bmatrix} = \begin{bmatrix} \frac{\partial L}{\partial X_{11}} & \frac{\partial L}{\partial X_{12}} & \frac{\partial L}{\partial X_{13}} \\ \frac{\partial L}{\partial X_{21}} & \frac{\partial L}{\partial X_{22}} & \frac{\partial L}{\partial X_{23}} \\ \frac{\partial L}{\partial X_{31}} & \frac{\partial L}{\partial X_{32}} & \frac{\partial L}{\partial X_{33}} \end{bmatrix}$$

   a) This can bee summed up into

$$\frac{\partial L}{\partial X} = \text{conv}\left(\text{padded}\left(\frac{\partial L}{\partial Z}\right), \text{180° rotated filter K}\right)$$

IX. Back Propagation for the Entire CNN Structure

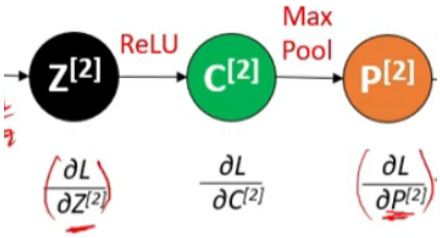


A. For Layer 3

1. $$dZ^{[3]} = \frac{\partial L}{\partial Z^{[3]}} = (y\_pred - y)$$
$$dW^{[3]} = \frac{\partial L}{\partial W^{[3]}} = dZ^{[3]} * f^{T}$$
$$dB^{[3]} = \frac{\partial L}{\partial B^{[3]}} = dZ^{[3]}$$

$$df = \frac{\partial L}{\partial f} = W^{[3]T} * dZ^{[3]}$$

   a) Y = real labels
   b) F = flatten layer
   c) Df is Gradient for Flatten Layer



B. For Layer 2

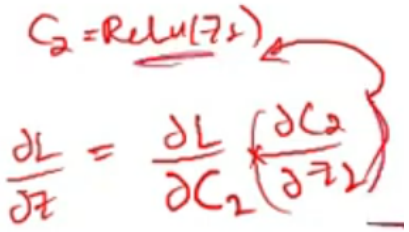
1. $$dP^{[2]} = \frac{\partial L}{\partial P^{[2]}} = df.reshape(P^{[2]}.shape)$$

   a) We can do this to find dp^2 because it is just the same derivative as flatten layer but reshaped because flatten layer is just flatten of matrix in p^2

2. $$dC_{mn}^{[2]} = \frac{\partial L}{\partial C_{mn}^{[2]}} = \begin{cases} \frac{\partial L}{\partial P_{xy}^{[2]}} & \text{, If } C_{mn} \text{ is the max element} \\ 0 & \text{, otherwise} \end{cases}$$

a) Definition for backpropagation of Max pooling exp[lained more below

(1) If $C^{[2]} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ $P^{[2]} = [4]$ this because max value is C = P and the max value is 4 therefore the bottom right number is the only thing we care for therefore if $\frac{\partial L}{\partial P^{[2]}} = [2]$ then $\frac{\partial L}{\partial C^{[2]}} = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix}$

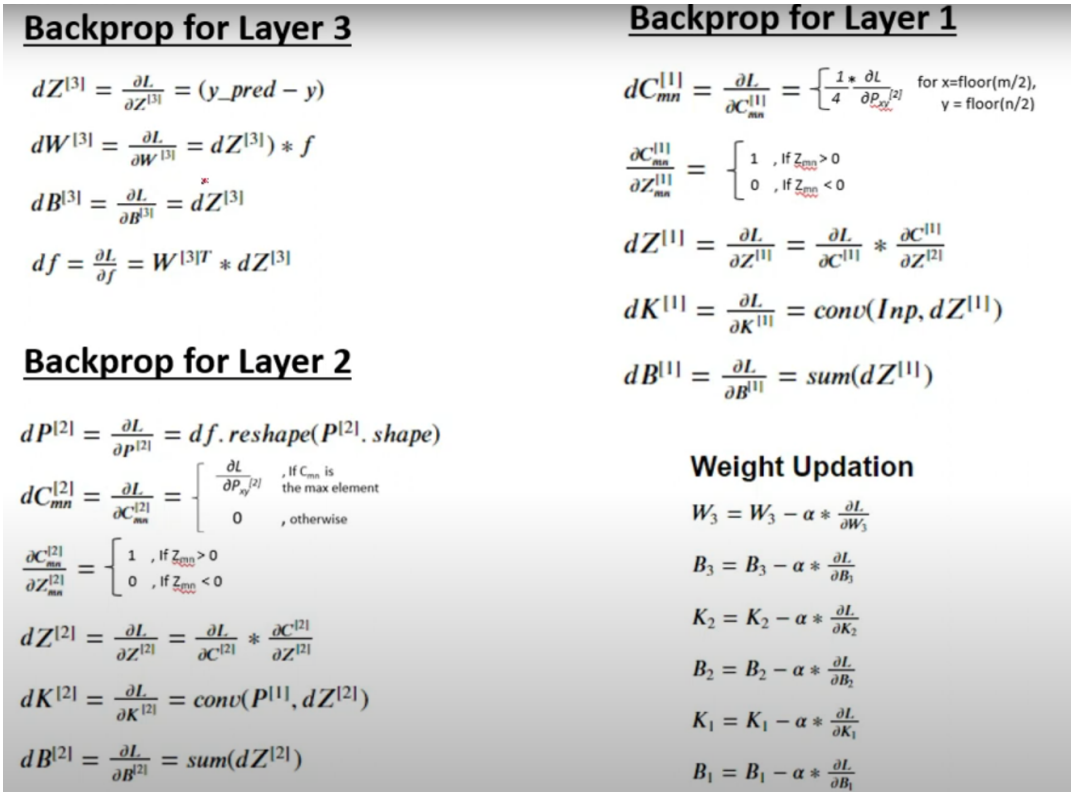


3. Now after We have solved for dL/dZ^2 We can plug that in to find the Back Propagation of the convolution using the quotations we found above

a)
$$dK^{[2]} = \frac{\partial L}{\partial K^{[2]}} = conv(P^{[1]}, dZ^{[2]})$$
$$dB^{[2]} = \frac{\partial L}{\partial B^{[2]}} = sum(dZ^{[2]})$$
$$\frac{\partial L}{\partial P^{[1]}} = conv(padded(dZ^{[2]}), 180^{o} rotated\ filter\ K^{[2]})$$

C. NOW repeat the steps above For Layer One all steps above describe how to do Layer 2 only difference is dL/dP^1 will be dX^2

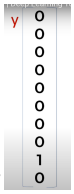
D. Summary

1. 

X. Tensor Flow s Keras
   A. Keras mainly used for making smal models for mock up not really for big model
   B. Tensorflow is for big models
      1. Work on GPU and TPU
      2. Allows us to run model on anything by allowing us to run model on any language so written in python deployed on java
   C.

1. Easy to code	1. Not so easy to code
2. Training the model is slow	2. Training the model is fast
3. Used for rapid prototyping	3. Used for bigger and high level applications
4. Lesser need to debug	4. Bit difficult to debug
5. Used for small dataset	5. Used for large dataset
6. Smaller community support	6. Bigger community support

XI. Categorical cross entropy vs Sparse categorical cross entropy

A. Use Categorical cross entropy if Y is a one-hot encoded vector 

B. Use Sparse categorical cross entropy if y is just a number $y = 8$