Q.1 Red channel

1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1

Green channel

2	2	2	2
2	2	2	2
2	2	2	2
2	2	2	2

Blue channel

1	1	1	1
2	2	2	2
3	3	3	3
4	4	4	4

Convolution filter (3×3) :

1	1	1
1	1	1
1	1	1

Size of output image $= \frac{W-K+2P}{S} + 1$, $W=4, K=3, P=0, S=1$

$= \frac{4-3+0}{1} + 1 = 1+1 = 2 \Rightarrow$ (2×2) ↑ image.

first stride = (9 + 18 + (3+6+9)) = 27+18 = 45

Second stride = (9+18 + (3+6+9)) = 45

Third Stride = (9 + 18 + (6+9+12)) = 54

Fourth stride = (9+18 + (6+9+12)) = 54

Convolved image =

45	45
54	54

2) With zero padding

output image size $= \frac{W-K+2P}{S} + 1 = \frac{4-3+2}{1} + 1 = 3+1 = 4$

Image ⇒ 4×4

Red channel after zero padding

0	0	0	0	0	0
0	1	1	1	1	0
0	1	1	1	1	0
0	1	1	1	1	0
0	1	1	1	1	0
0	0	0	0	0	0

Green channel after padding

0	0	0	0	0	0
0	2	2	2	2	0
0	2	2	2	2	0
0	2	2	2	2	0
0	0	0	0	0	0
0	0	0	0	0	0

Blue channel after padding

0	0	0	0	0	0
0	1	1	1	1	0
0	2	2	2	2	0
0	3	3	3	3	0
0	4	4	4	4	0
0	0	0	0	0	0

filter

1	1	1
1	1	1
1	1	1

Output

18	27	27	18
30	45	45	30
36	54	54	36
26	37	37	26

Q.3 Red channel

0	0	0	0	0	0
0	1	1	1	1	0
0	1	1	1	1	0
0	1	1	1	1	0
0	1	1	1	1	0
0	0	0	0	0	0

Green channel

0	0	0	0	0	0
0	2	2	2	2	0
0	2	2	2	2	0
0	2	2	2	2	0
0	2	2	2	2	0
0	0	0	0	0	0

Blue channel

0	0	0	0	0	0
0	1	1	1	1	0
0	2	2	2	2	0
0	3	3	3	3	0
0	4	4	4	4	0
0	0	0	0	0	0

filter

1	1	1
1	1	1
1	1	1

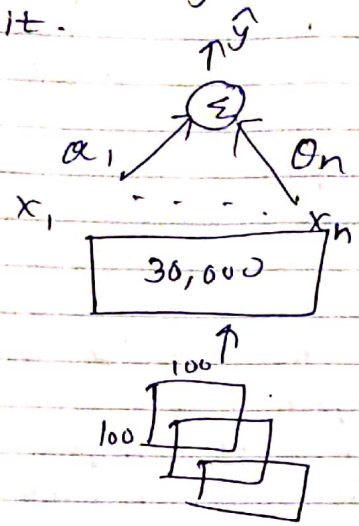
output Image

24	24
20	20

4) Template matching is the technique used to find small parts of an image which matches template image.
It slides the window across image that will provide a percent match with the template. When image is convoluted with the filter, it is expected to give portions of the image that match the particular template to give a higher response.

5) Multi-scale analysis can be achieved with a fixed window size by reducing an image size by continuously convoling it with filter. This will lead to a larger receptive field in the convoluted layer, better feature extraction and accounts for differently scaled objects.
ex. We start with 8x8 size image and fixed window size of 2x2. After first convolution, we get 4x4 output, running another convolution gives an image of 2x2. As the scale of image reduces the objects in the image get more receptive area and would actually lead to feature extraction.

Q.6 While flattening an image we loose the spatial context of it.



This problem is solved using convolution layer where we use filter of lower size which will lead to lesser number of parameters.

However this could lead to a loss of information as a result of decrease in the spatial resolutions.

This is compensated by adding the depth through each convolution layer, which is done by passing a higher number of filters.

Q.7 $128 \times 128 \times 32 \rightarrow$ tensor

16 conv. filter of size $3 \times 3 \times 32$

$p = 0,\ w = 128,\ k = 3,\ s = 1$

size of output tensor $= \frac{w - k + 2p}{s} + 1$

$= \frac{128 - 3 + 0}{1} + 1$

$= 126.$

Resulting tensor dimension $= 126 \times 126 \times 16$

Q.8 $S = 2$

Size of output tensor $= \frac{W - K + 2P}{S} + 1$

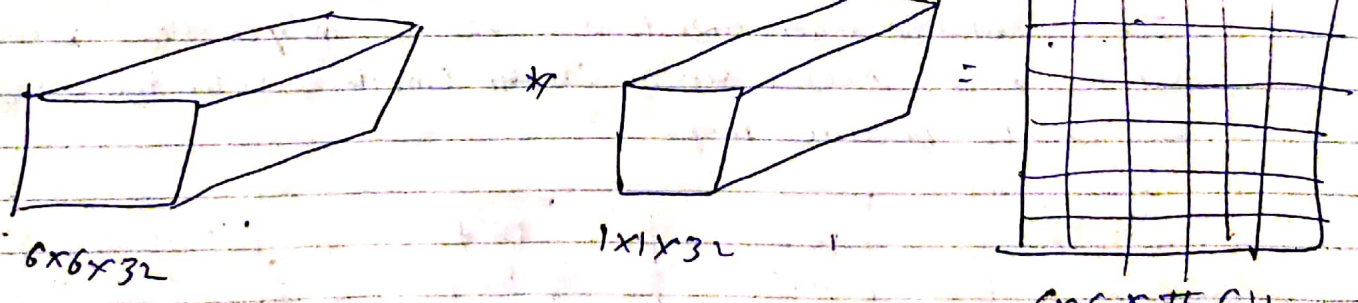
$= \frac{128 - 3 + 2\times0}{2} + 1$

$= \frac{125}{2} + 1$

$= 63.$

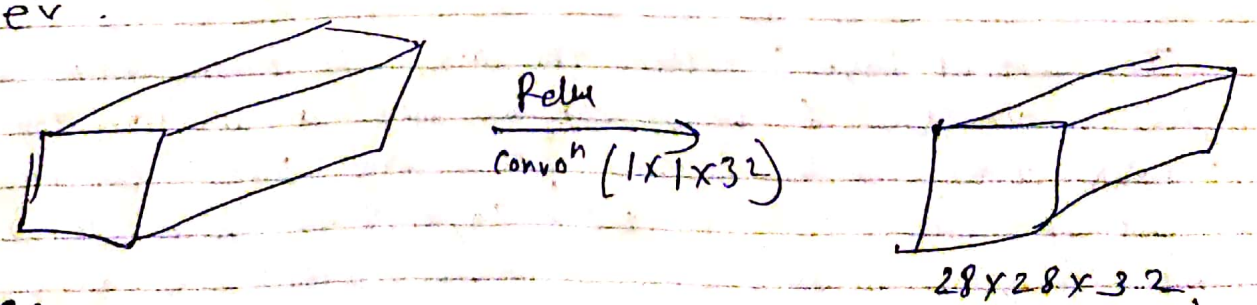
Resulting tensor dimension $= 63 \times 63 \times 16$

Q.9 Deep convolution N/w problem → No. of feature maps often increase with the depth of the network. This problem can result in a dramatic increase in the number of parameters and computation required when large filters are used. Such as 5x5 and 7x7. It is also called as network in network.



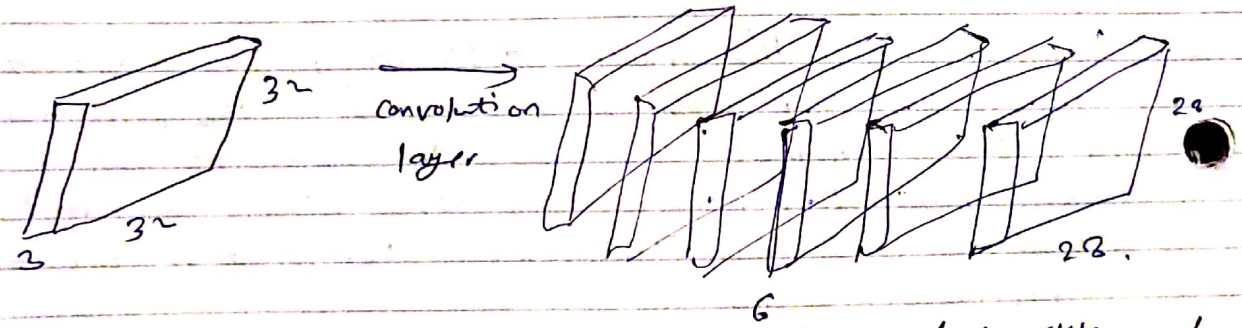
32 → no. of filters.

ex.



It is useful in shrinking the number of channels. Use 32 filters that are (1x1) → dimensions of each filter would be (1x1x192) → no. of channels is the filter should match the value in the input layer. Output of this results in (28x28x32). Pooling is used to shrink height and width of the image. No. of channels can be reduced by using 1x1 convolution.

Q.10 The convolution layer is the main building block of a convolution neural network.



The convolution layer consist of set of independent filter and each filter is independently convolved with the image and we end up feature maps.

The initial layers filters the image to become blobs of colored pieces and edges. As we go deeper in the convolution layers which takes smaller colored pieces and making larger pieces out of them. So the later layer learn higher level, more complex features of the image, like eyes, ears in the picture of human being.

Q.12

A pooling layer is another building block of a CNN. It's function is to successively ~~reduce~~ reduce the spatial size of the representation to reduce the amount of parameters and computations in the network.

Pooling layer operates on each feature map independently. The most common approach used in ~~the~~ pooling is max-pooling.

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

~~6 8~~
~~10~~ max pool with 2×2 filter with stride 2.

6	8
14	16

Q.11

1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1

Red channel

2	2	2	2
2	2	2	2
2	2	2	2
2	2	2	2

Green channel

1	1	1	1
2	2	2	2
3	3	3	3
4	4	4	4

Blue channel.

max pool with 2×2 filter and stride of size 2

1	1
1	1

2	2
2	2

2	2
4	4

Q.13 Data Augmentation is the strategy that helps to increase the diversity of the data available, without actually gathering more data points.

Techniques like cropping, padding, horizontal flipping are used to augment data. When we have less data, data augmentation is used to increase data set size.

It helps in reducing overfitting of the data when dataset is small and helps in better generalisation.

Q.14 Transfer learning is the idea of overcoming the isolated learning paradigm and utilizing knowledge acquired for one task to solve related problems.

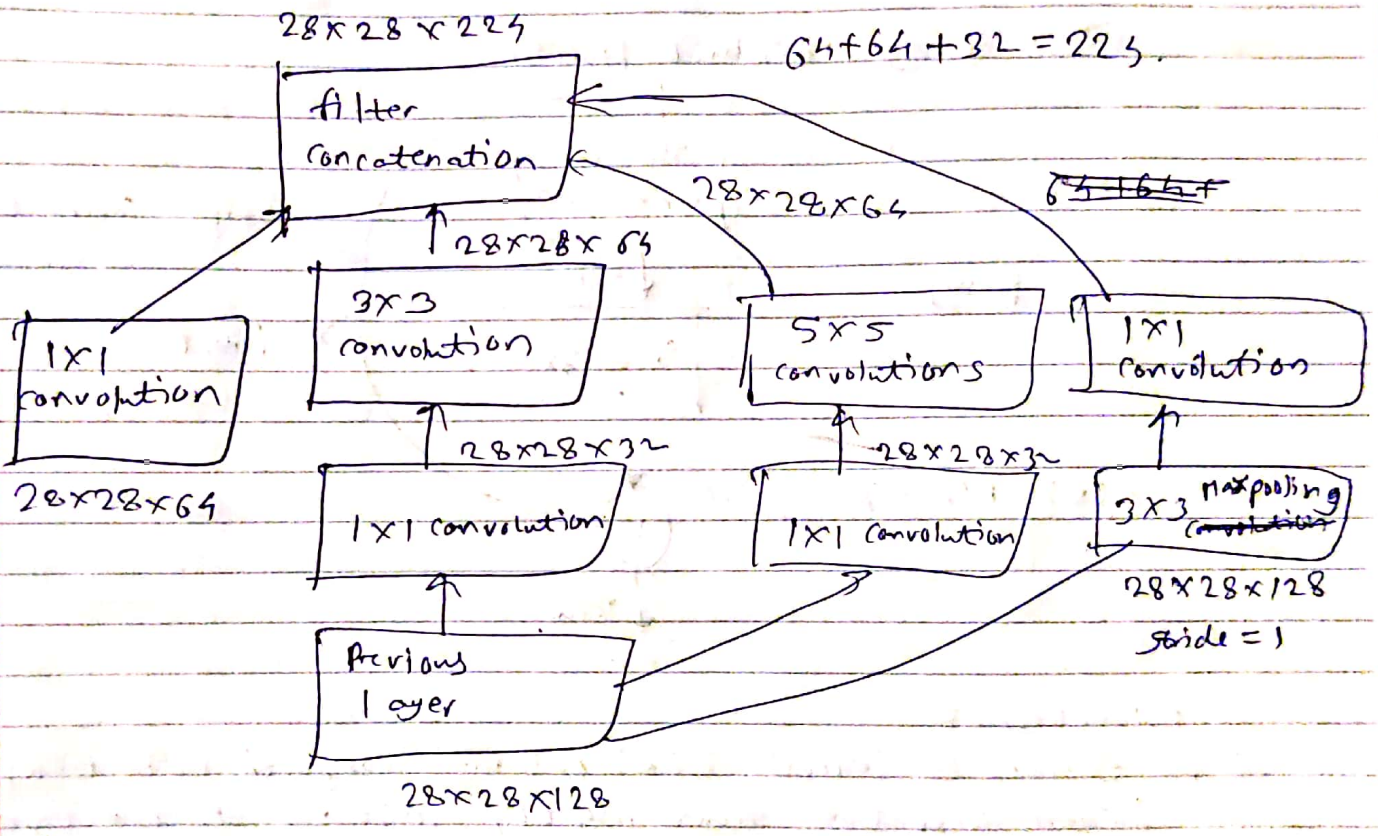
It can be conceptualized as the feature extraction mechanism in the problems when we have smaller dataset. A model that has been trained on a larger dataset can be used an initial layer of our model to perform feature extraction. These extracted feature will provide input for our custom classifier.

Q.15 Transfer learning uses a pretrained network like 'Imgenet object classifier', which is trained of 1.5 million image dataset with 1000 classes. The initial layers of the model comes from a pre-trained model whose weights have been already learnt. The last few layers of the model would be custom classifier layer whose weights are initialised to small values close to 0. During training we get a gradient value for that, for the classifier layers, would be small because the initial weights are small.

Q.16 In the fine tuning, we first freeze the pre-trained network and train the shallow network. Once the shallow network is trained we unfreeze the pretrained network and train the entire model so that the pretrained network parameters are fine tuned and the model performs better.

Q.17 Inception layer is a combination of all those layers (namely 1x1 convolution layer, 3x3 convolution layer, 5x5 convolution layer) with their output layer concatenated into a single output vector forming the input of next stage.

Eg:- Inception V1 Googlenet



Purpose of inception block:

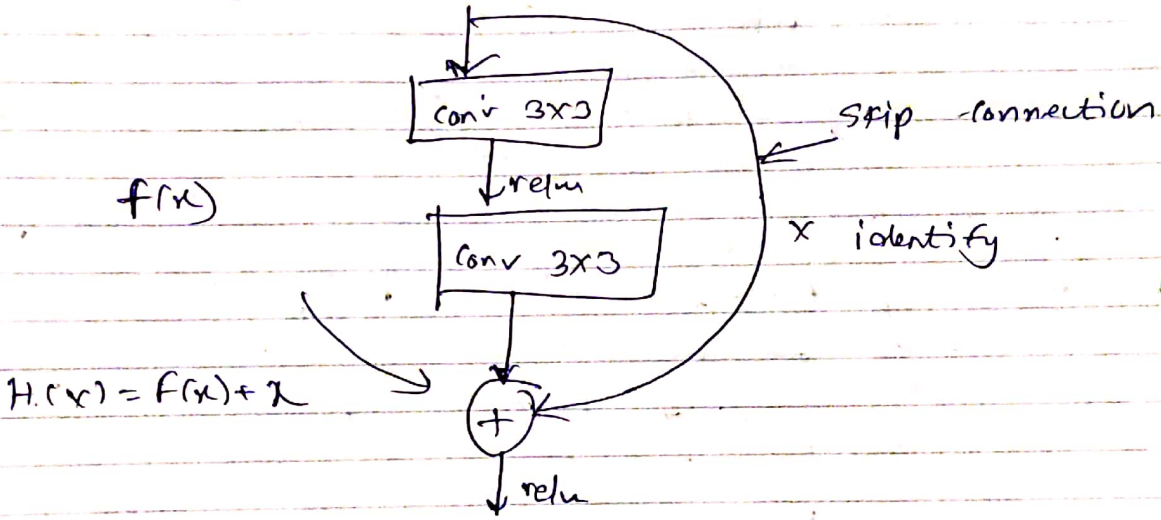
They allow for more efficient computation through dimensionality reduction with stack 1x1 convolutions.

The modules were designed to solve the problem of computational expense as well as overfitting and vanishing gradient problems.

Q.18) Residual blocks

ResNet allow the flow of the memory (or information) from initial layers to last layers. ~~Despite~~ Despite the absense of gates in their skip connections, residual network perform as good as any other highway network in practical scenario.

* Using residual blocks.



Advantages:

1. Easier to learn f(x) residual compared with H(x) (learn derivation from identity instead of function).

2. Skip connections helps to solve vanishing gradient problem.

Q.19 ~~Steps~~ to visualise intermediate activations of convolutional layers given an input.

1. An input is decomposed into the different filters learned by the network.
2. This gives a view into how each channel encodes relatively independent features so the proper way to visualise these features maps is by independently plotting the contents of every channel as a 2D image.

Visualising intermediate activation consists of displaying the feature maps that are output by various convolution and pooling layer in a network, given a certain input (the output of a layer is often called as an its activation, the output of the activation function).

Purpose :- Intermediate activations are useful for understanding how successive convnet layer transform their input and for getting a first idea of the meaning of individual convnet filters.

Q.20 Steps to visualise filter weight of convolution layer:-
1. Filters may be interpreted as the templates that being matched.
2. Using gradient ascent find an input that maximises the response of the filter. (i.e. correlated with the filter)
3. Visualize the input that will minimize loss (or maximize response).

Purpose :- These filter visualizations tell you about how convnet layer see the world: each layer in a convnet learns a collection of filter such that their inputs can be expressed as a combination of the filters.

Q. 21) The heatmap of the image is then generated using the probability values. The heatmap signifies which part of the image are more important.

Steps to visualize filter the heatmap of class activations for specific image and class :-

1. Feed an image to network.
2. Compute gradients of selected output node w.r.t. each channel of the target layer where activation is to be completed ("conv5").
3. Compute the average-gradients at eakh channel.
4. Add the activations of each channel weighted by their average gradient magnitude.
5. Superimpose activation on input image (heatmap).

Limitations of class activation :-

1. Need to have a GAP layer in architecture. If not we cant apply this method.
2. Can only be applied to visualise the final layer heatmap. cant be used if we wanted to visualize a previous layer in the network.

In Grad-CAM (Gradient Weighted Class Activation Mapping), the gradient of the output class value with regards to each channel in it representing the gradient of the corresponding channel in a feature map.

The gradient channel map obtained is then global average pooled, and the values obtained henceforths are the weights of importance of each channel in the feature map. The weighted feature map is then used as a heatmap, just like in class activation Mapping (CAM).

Purpose of Visualization :-

Since the gradient output can be calculated with regards to any layer, there's no restriction of only using a final layer — also there's no mention of network architecture anywhere, so any kind of architecture works.