Ain Shams University

Faculty of Engineering

Computer and Systems Department

CSE 616: Neural Networks and their applications

Assignment 2

Report

Name	Diaa Ahmed Abdelzaher Abdelaziz
Code	1902558

1- The shape of weight matrix is: 100 x 750000 and the shape of bias 100 x 1.
2- No of parameters equals = 5 x 5 x 3 x 10 + 10 x 1 = 760
3- The filter used is Sobel filter. The Sobel operator performs a 2-D spatial gradient measurement on an image and so emphasizes regions of high spatial frequency that correspond to edges. Typically, it is used to find the approximate absolute gradient magnitude at each point in an input grayscale image. The values of these filters respectively are:

-1	0	1
-2	0	2
-1	0	1

1	2	1
0	0	0
-1	-2	-1

4-

Batch Normalization

5-

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: γ, β

Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma^2_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma^2_{\mathcal{B}} + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma\widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

Some advantages of the Batch Normalization are:

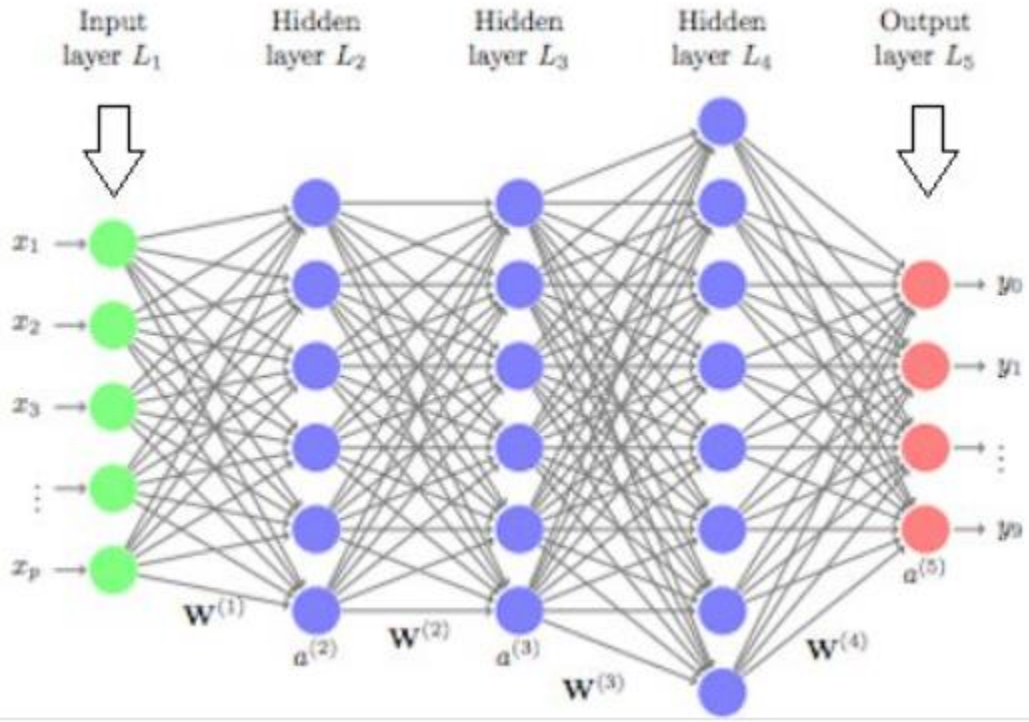
- Speed Up the Training by Normalizing the hidden layer activation the Batch normalization speeds up the training process.
- Acts as a form of regularization.
- Handles internal covariate shift, It solves the problem of internal covariate shift. Through this, we ensure that the input for every layer is distributed around the same mean and standard deviation.

6- The Size of receptive field = s1 * (k1 – 1) + s1 * (k2 - 1) + 1 = 5

7- The dimension of the output is 128 x 64 x 64.

8- With normal dropout at test time, you have to scale activations by dropout rate p, with inverted dropout, scaling is applied at the training time, but inversely. First, dropout all activations by dropout factor p, and second, scale them by inverse dropout factor 1/p. The advantage of inverted dropout is that you don't have to do anything at test time, which makes test operation faster.

9- In the case of deep neural networks each neuron in each layer is fully connected to all the neurons in the previous layer as you can see in the below figure.



Because of these large number of connections, the number of parameters to be learned increases. As the number of parameters increases the network becomes more complex. This more complexity of the network leads to overfitting.

Especially, in the case of Image data being pixel values of the images as features, the number of input features would be of large dimension. And most of the pixel portions of the images may not contribute in predicting the output.

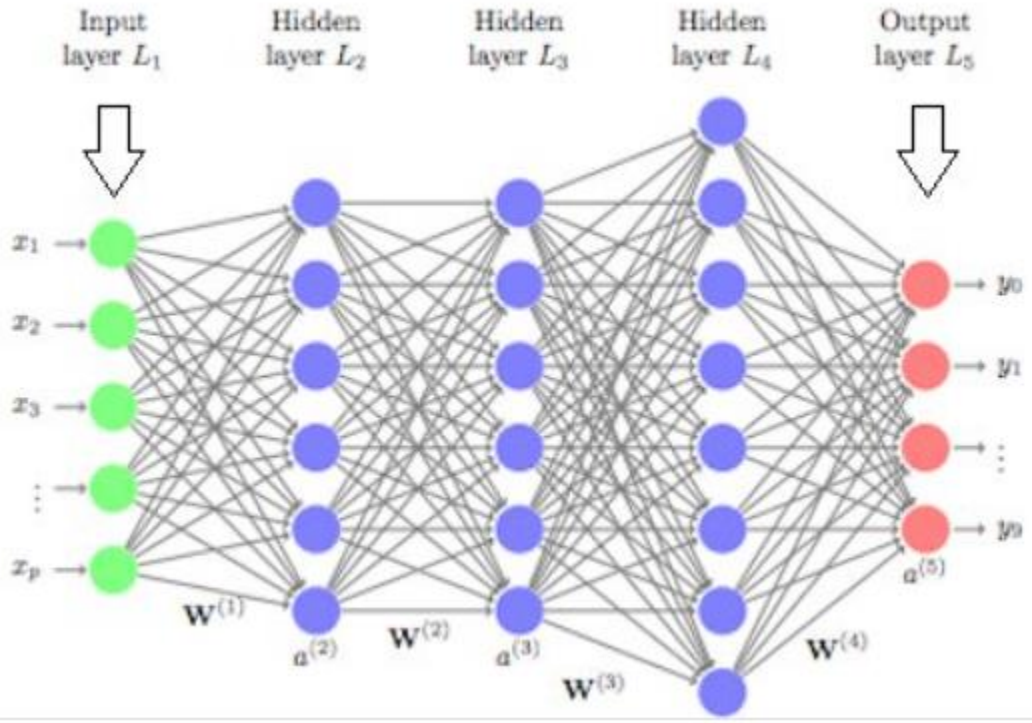
10- The steps are shown in the following table:

	4	1	-1	3		result
1	-2					-8
	1	-2				2
		1	-2			3
			1	-2		-7
				1	-2	3

a. So, the resulting array is [-8, 2, 3, -7, 3]

11- What happened is that there is a learning rate decay has been used at the mentioned epochs.

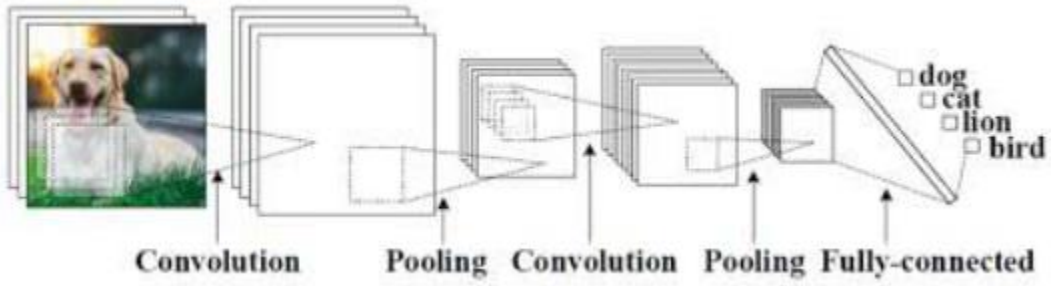
12- In the case of deep neural networks each neuron in each layer is fully connected to all the neurons in the previous layer as you can see in the below figure.



Because of these large number of connections, the number of parameters to be learned increases. As the number of parameters increases the network becomes more complex. This more complexity of the network leads to overfitting.

Especially, in the case of Image data being pixel values of the images as features, the number of input features would be of large dimension. And most of the pixel portions of the images may not contribute in predicting the output.

To overcome these challenges, the Convolution Neural Networks were discovered. In this, the input image data will be subjected to set of convolution operations such as filtration and max pooling. Then, the resultant data which will be of lesser dimension compared to the original image data will be subjected to Fully connected layers to predict output as shown in the below figure.



By performing the convolution operations, the dimensionality of the data shrinks significantly large. Hence, the number of parameters to be learned decreases. Hence, the network complexity decreases which leads to less chances of overfitting!

This is the reason why we use CNN's while in the case of Image classification.

13- Dropout refers to ignoring units (i.e., neurons) during the training phase of certain set of neurons which is chosen at random. Ignoring these units means that these units are not considered during a particular forward or backward pass.

More technically, at each training stage, individual nodes are either dropped out of the net with probability 1-p or kept with probability p, so that a reduced network is left, incoming and outgoing edges to a dropped-out node are also removed.

Training Phase:

Training Phase: For each hidden layer, for each training sample, for each iteration, ignore (zero out) a random fraction, p, of nodes (and corresponding activations).

Testing Phase:
Use all activations but reduce them by a factor p (to account for the missing activations during training).

14-

15- Update Rule for AdaGrad:

$$v_t^w = v_{t-1}^w + (\nabla w_t)^2$$

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{v_t^w + \epsilon}} * \nabla w_t$$

$$v_t^b = v_{t-1}^b + (\nabla b_t)^2$$

$$b_{t+1} = b_t - \frac{\eta}{\sqrt{v_t^b + \epsilon}} * \nabla b_t$$

It is clear from the update rule that history of the gradient is accumulated in v. The smaller the gradient accumulated, the smaller the v value will be, leading to a bigger learning rate (because v divides η).