DL

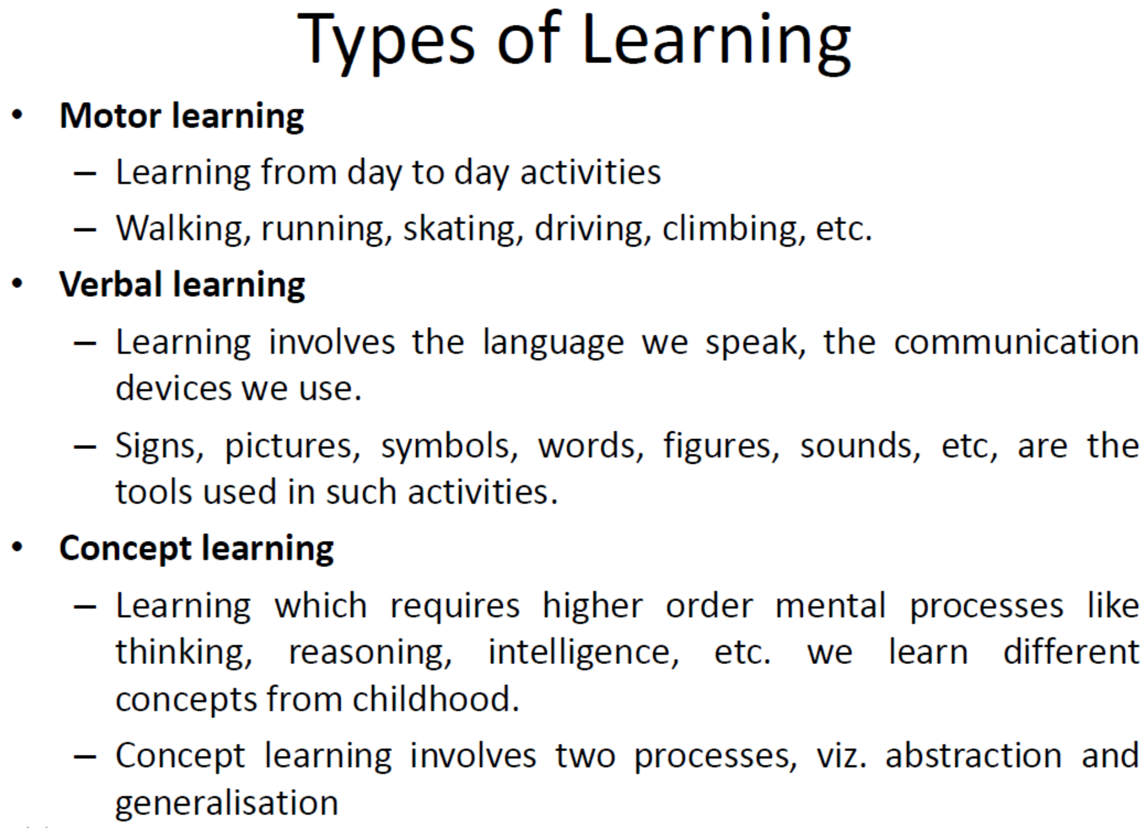
MID SEMESTER EXAM ANSWERS

Date: 7th March, 2023

Q1.) Answer the following (Any Two)

A.) Consider the following Scenario: A child was asked to sing a song in the school annual function. He started with good pace, however due to nervousness he was unable to continue; after a minute of his pause, some children among the audience cheered him up and finally he was able to perform very well in every such kind of program. Above scenario is an example of which type of Human learning to that child? Write briefly about different types of human learning.

Answer: **Operant Conditioning**



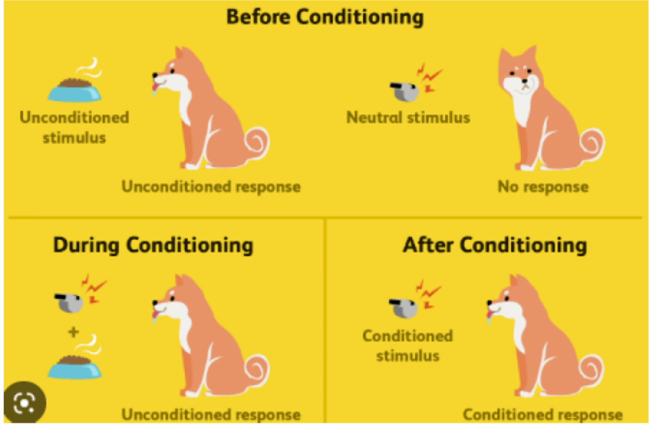
Types of Learning

- **Motor learning**
  - Learning from day to day activities
  - Walking, running, skating, driving, climbing, etc.
- **Verbal learning**
  - Learning involves the language we speak, the communication devices we use.
  - Signs, pictures, symbols, words, figures, sounds, etc, are the tools used in such activities.
- **Concept learning**
  - Learning which requires higher order mental processes like thinking, reasoning, intelligence, etc. we learn different concepts from childhood.
  - Concept learning involves two processes, viz. abstraction and generalisation

- **Discrimination learning**
  - Learning to differentiate between stimuli and showing an appropriate response to these stimuli is called discrimination learning. Example, sound horns of different vehicles like bus, car, ambulance, etc.
- **Learning of principles**
  - Individuals learn certain principles related to science, mathematics, grammar, etc. in order to manage their work effectively.
  - Example: formulae, laws, associations, correlations, etc

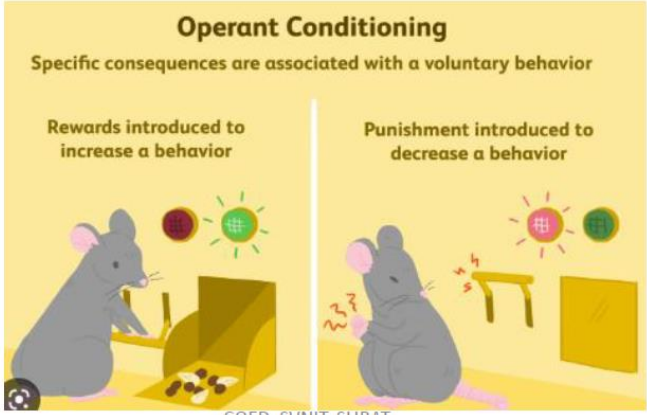
Learning through association - Classical Conditioning

If a neutral stimulus (a stimulus that at first elicits no response) is paired with a stimulus that already evokes a reflex response, then eventually the new stimulus will by itself evoke a similar response.



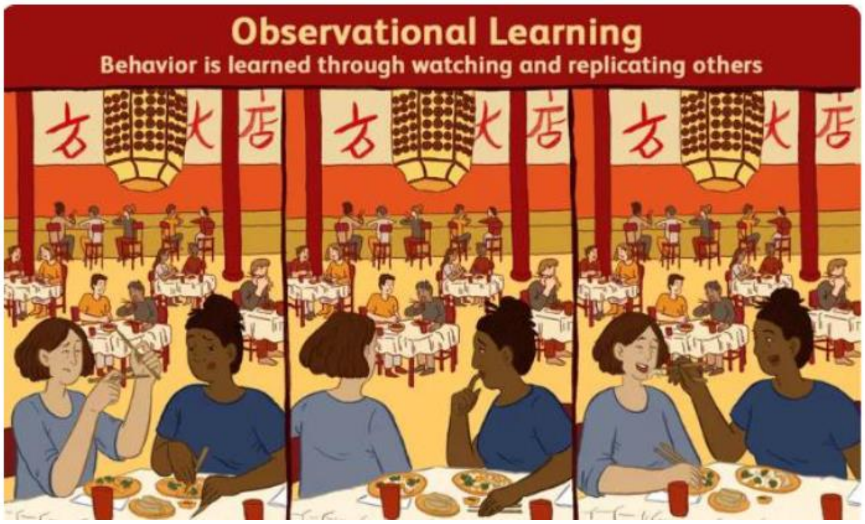
Learning through consequences – Operant Conditioning

- The organism operates on its environment in some way; the behavior in which it engages are instrumental to achieving some outcome.


Operant Conditioning
Specific consequences are associated with a voluntary behavior
Rewards introduced to increase a behavior
Punishment introduced to decrease a behavior

Learning through observation – Modeling/Observational Learning

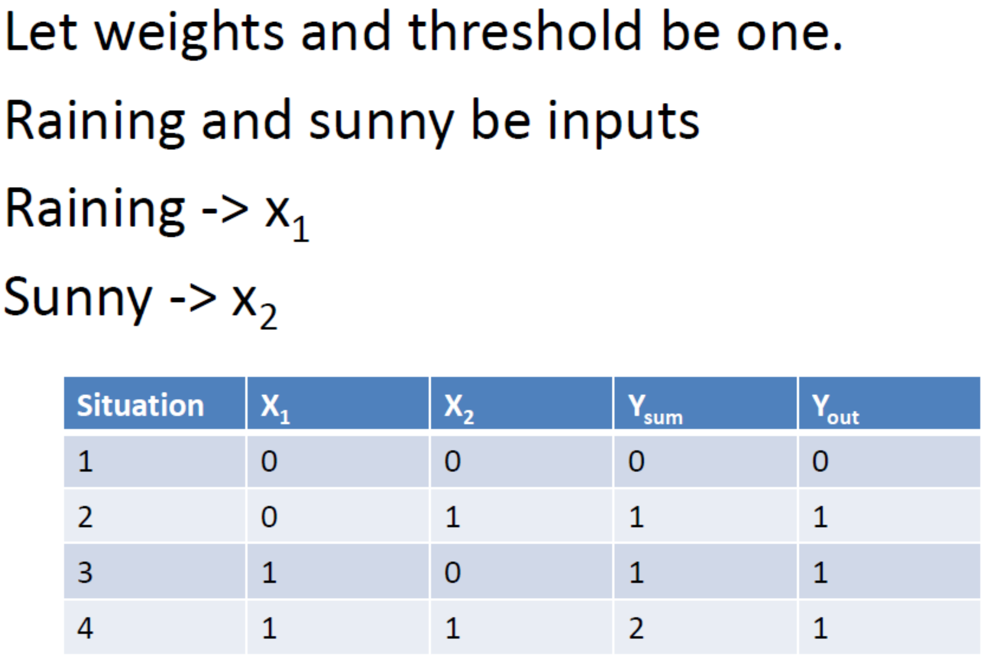
- The process of learning by watching others, retaining the information, and then later replicating the behaviors that were observed.



Observational Learning

Behavior is learned through watching and replicating others

B.) Given the following Customized AC System: System works on the fact that Customized AC will be turned OFF if the Room door is open and the temperature is low. Given the following situations: S1: Room door is closed, temperature is high; S2: Room is closed, temperature is low; S3: Room is open, temperature is high, S4: Room is open, temperature is low. Design an appropriate McCulloch O Pitts Neuron model for above system.



Let weights and threshold be one.

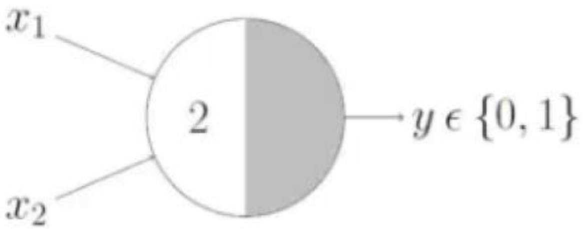
Raining and sunny be inputs

Raining -> x_1

Sunny -> x_2

Situation	X_1	X_2	Y_{sum}	Y_{out}
1	0	0	0	0
2	0	1	1	1
3	1	0	1	1
4	1	1	2	1

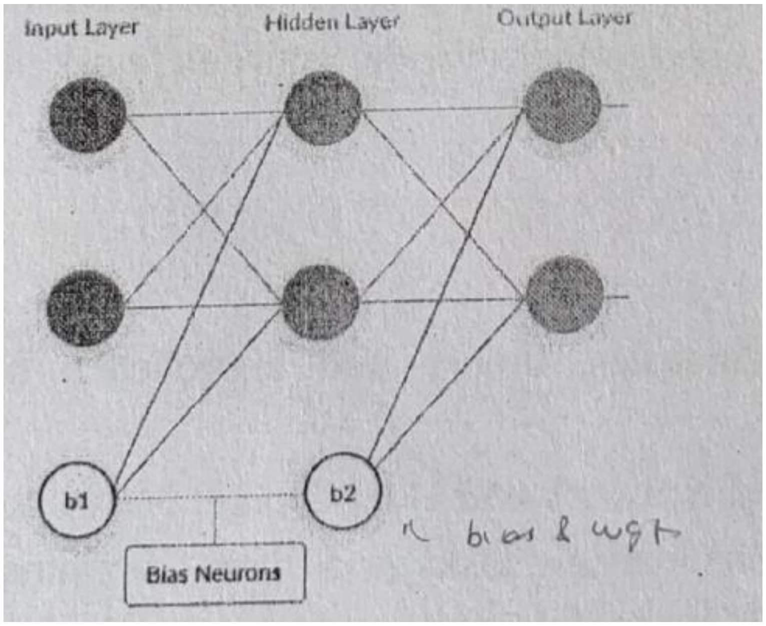
AND Function



AND function

$$x_1 + x_2 = \sum_{i=1}^{2} x_i \geq 2$$

C.) Refer to the neural network given below: List and calculate learnable parameters at each layer and for overall network. List and comment on hyper-parameters of the deep neural networks in brief.

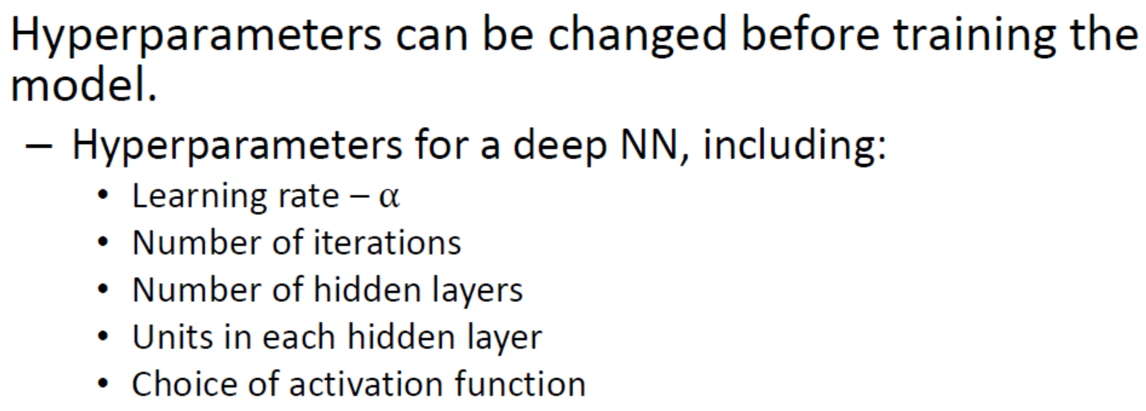


Answer: inputs * outputs + biases

Learnable Parameters: Weights & Biases

I/P & Hidden = 2*2 + 2 = 6

Hidden & O/P = 2*2 + 2 = 6

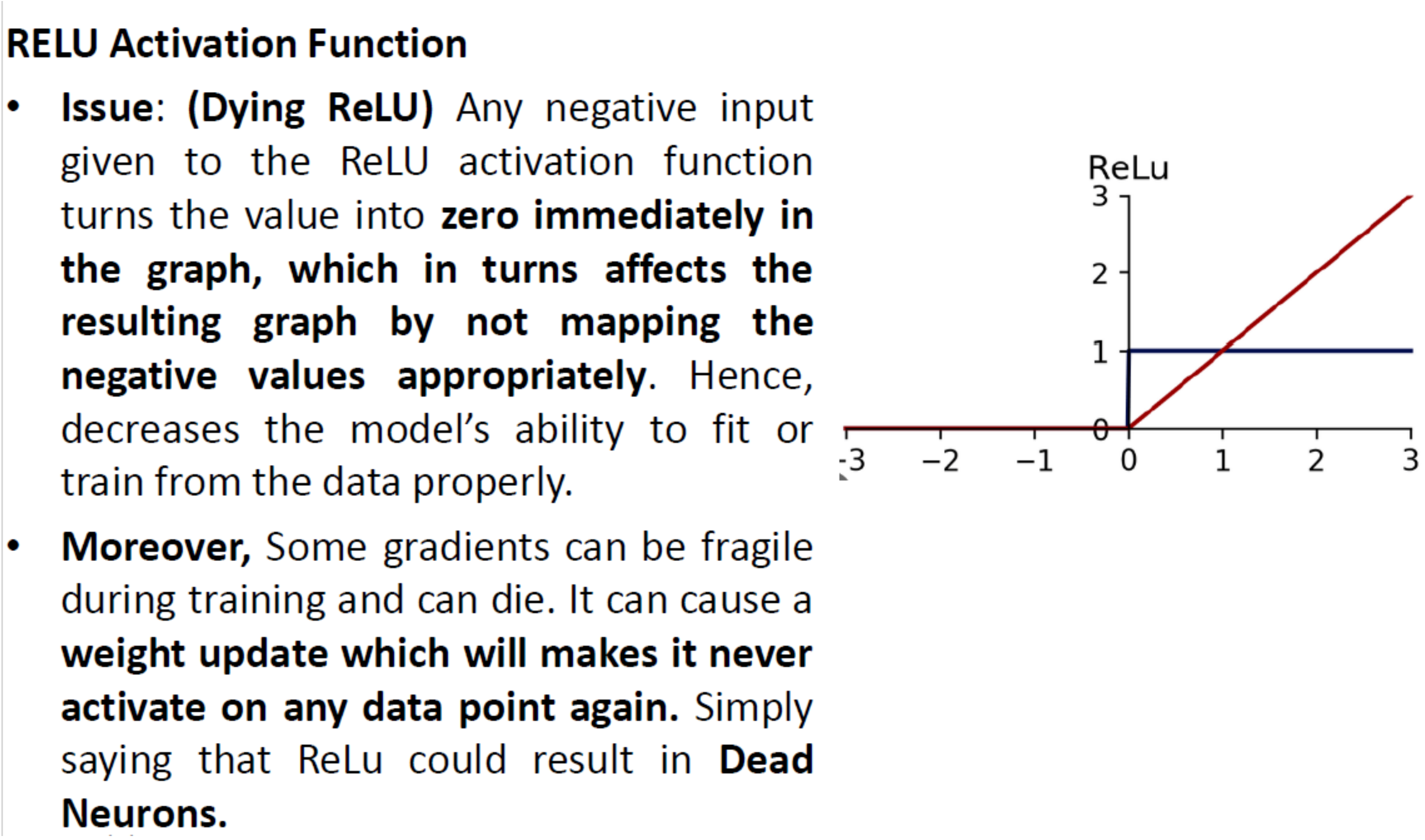


Hyperparameters can be changed before training the model.

- Hyperparameters for a deep NN, including:
  - Learning rate – α
  - Number of iterations
  - Number of hidden layers
  - Units in each hidden layer
  - Choice of activation function

Q2.) Answer the following.

A.) What do you mean by dead neurons in deep neural networks? By using which activation function/s issue of dead neurons can occur? How to overcome them?



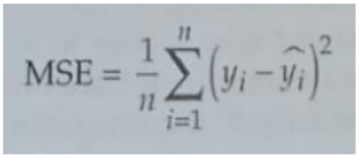
RELU Activation Function

- **Issue**: **(Dying ReLU)** Any negative input given to the ReLU activation function turns the value into **zero immediately in the graph, which in turns affects the resulting graph by not mapping the negative values appropriately**. Hence, decreases the model's ability to fit or train from the data properly.
- **Moreover,** Some gradients can be fragile during training and can die. It can cause a **weight update which will makes it never activate on any data point again.** Simply saying that ReLu could result in **Dead Neurons.**

Overcome : Leaky ReLU & ELU

B.) Assume a dataset has multiple outliers and we need to design a deep neural network for the same; which loss function will be optimum for such a dataset? Explain the same in brief.

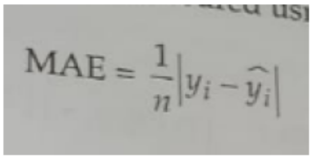
- **MSE/L2/ Quadratic Loss**
- Means Square error is one of the most commonly used Cost function methods.
- Because of the square of the difference, it avoids any possibility of negative error.
- The formula for calculating MSE is given below:



$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{y_i}\right)^2$$

- Mean squared error is also known as **L2 Loss**.
- In MSE, **each error is squared, and it helps in reducing a small deviation in prediction as compared to MAE.**
- But if the dataset has outliers that generate more prediction errors, then squaring of this error will further increase the error multiple times. **Hence, we can say MSE is less robust to outliers**

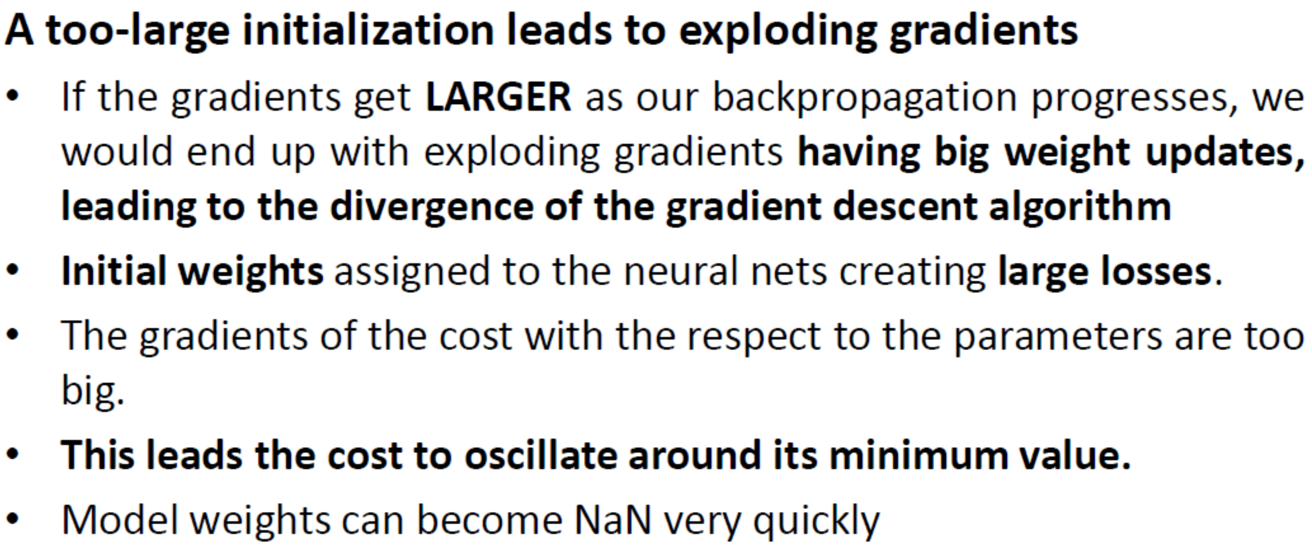
- **Mean Absolute Error (MAE)**
- Mean Absolute error also **overcome the issue of the Mean Square error cost function by taking the absolute difference between the actual value and predicted value**.
- The formula for calculating Mean Absolute Error is given below:



$$\mathrm{MAE} = \frac{1}{n}\left|y_i - \widehat{y_i}\right|$$

- This means the Absolute error cost function is also known as **L1 Loss**. It is not affected by noise or outliers, hence giving better results if the dataset has noise or outlier.

C.) "In Deep neural networks, as and when the weights increase, the gradient increases". Above mentioned statement gives the implication of which problem? List and brief possible way/s to mitigate the same?



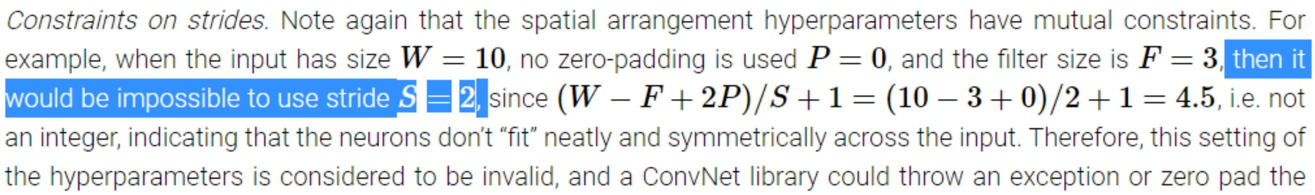
A too-large initialization leads to exploding gradients

- If the gradients get **LARGER** as our backpropagation progresses, we would end up with exploding gradients **having big weight updates, leading to the divergence of the gradient descent algorithm**
- **Initial weights** assigned to the neural nets creating **large losses**.
- The gradients of the cost with the respect to the parameters are too big.
- **This leads the cost to oscillate around its minimum value.**
- Model weights can become NaN very quickly

Mitigation: Gradient Clipping & Backpropagation through Time

Q3.) Answer the following

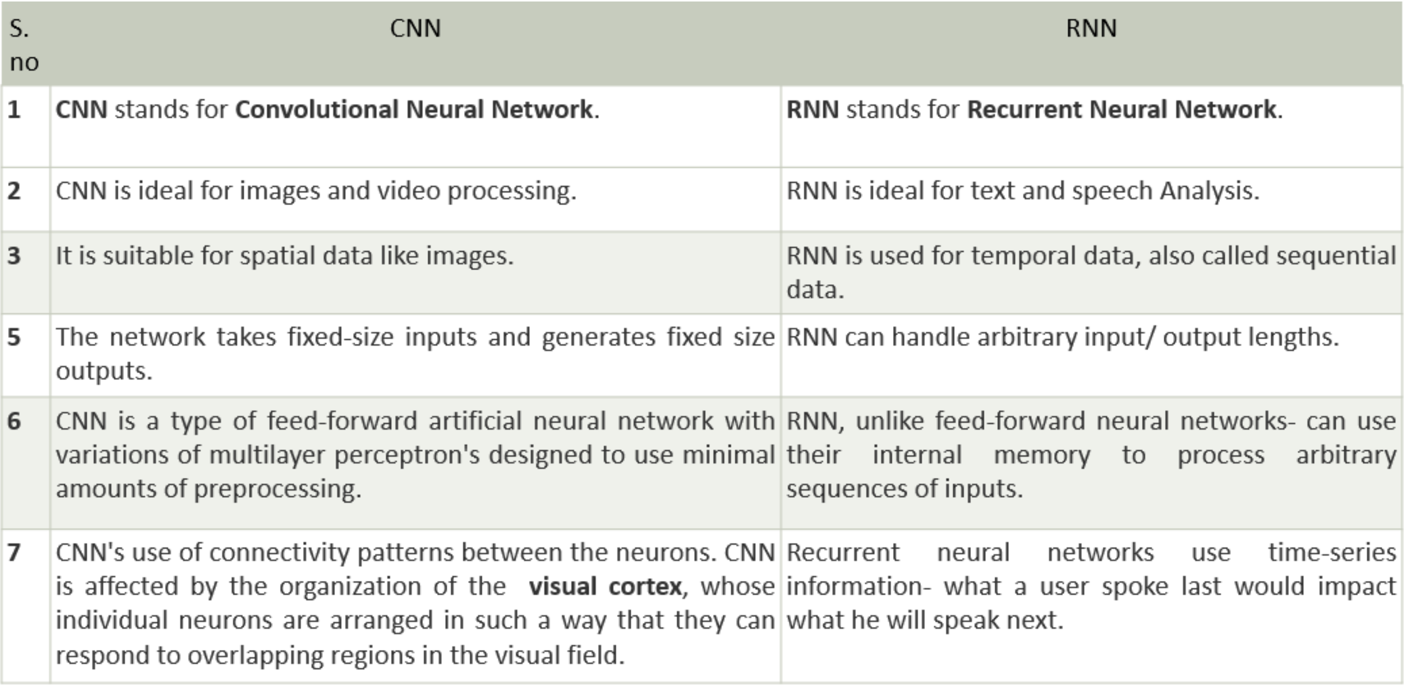
A.) Consider an input image of size 13x13 and 64 filters of size 3x3. Discuss whether it is possible to perform convolutions with strides 2, 3, 4, and 5. Justify your answer in each case.



Constraints on strides. Note again that the spatial arrangement hyperparameters have mutual constraints. For example, when the input has size $W = 10$, no zero-padding is used $P = 0$, and the filter size is $F = 3$, then it would be impossible to use stride $S = 2$, since $(W - F + 2P)/S + 1 = (10 - 3 + 0)/2 + 1 = 4.5$, i.e. not an integer, indicating that the neurons don't "fit" neatly and symmetrically across the input. Therefore, this setting of the hyperparameters is considered to be invalid, and a ConvNet library could throw an exception or zero pad the

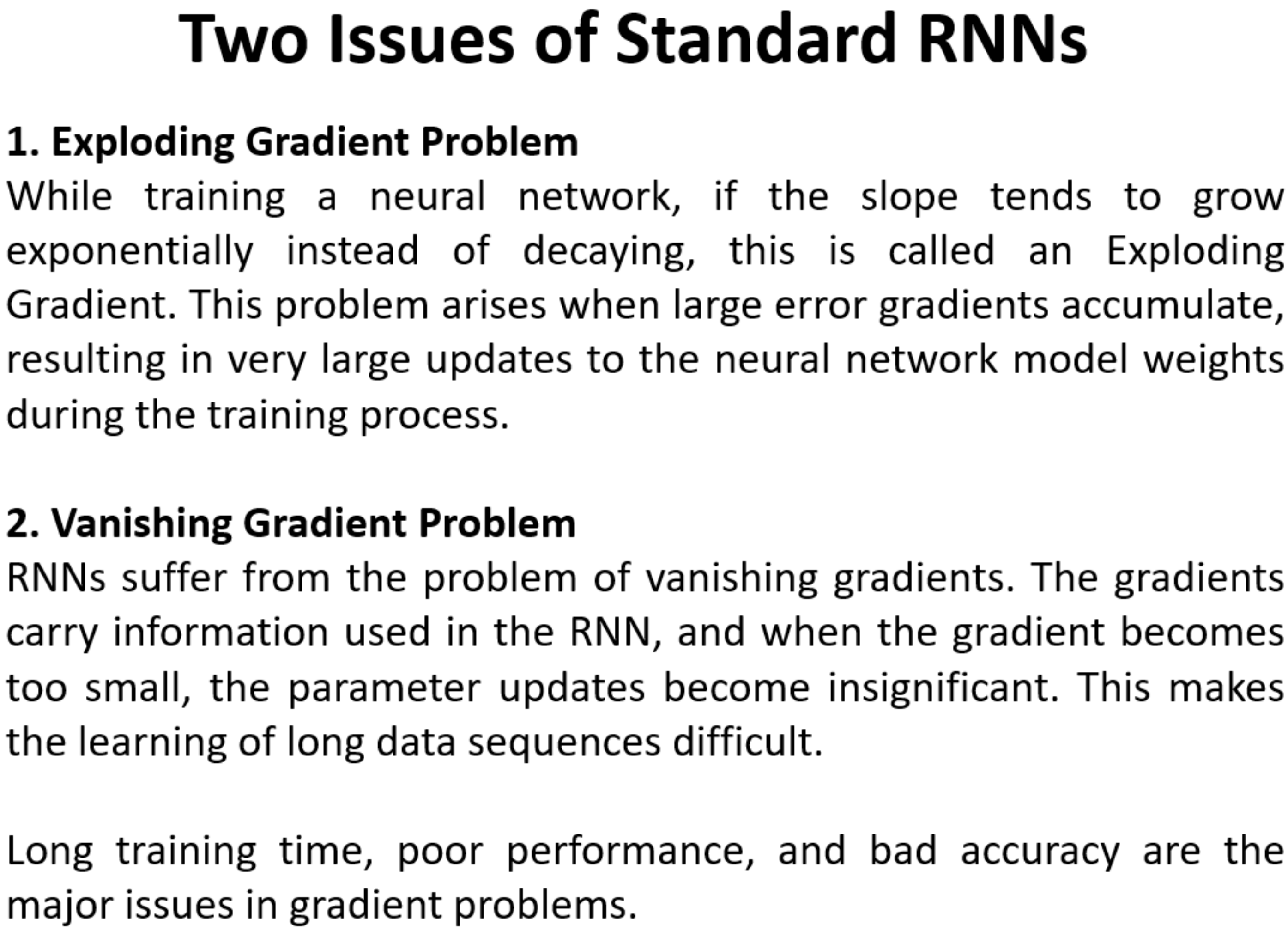
For 2,4,5 -> Possible, 3 -> Not Possible

B.) Differentiate between CNN and RNN with suitable examples.



S. no	CNN	RNN
1	**CNN** stands for **Convolutional Neural Network**.	**RNN** stands for **Recurrent Neural Network**.
2	CNN is ideal for images and video processing.	RNN is ideal for text and speech Analysis.
3	It is suitable for spatial data like images.	RNN is used for temporal data, also called sequential data.
5	The network takes fixed-size inputs and generates fixed size outputs.	RNN can handle arbitrary input/ output lengths.
6	CNN is a type of feed-forward artificial neural network with variations of multilayer perceptron's designed to use minimal amounts of preprocessing.	RNN, unlike feed-forward neural networks- can use their internal memory to process arbitrary sequences of inputs.
7	CNN's use of connectivity patterns between the neurons. CNN is affected by the organization of the **visual cortex**, whose individual neurons are arranged in such a way that they can respond to overlapping regions in the visual field.	Recurrent neural networks use time-series information- what a user spoke last would impact what he will speak next.

C.) What are the major issues in RNN? Discuss the methods to address these issues.



Two Issues of Standard RNNs

1. Exploding Gradient Problem

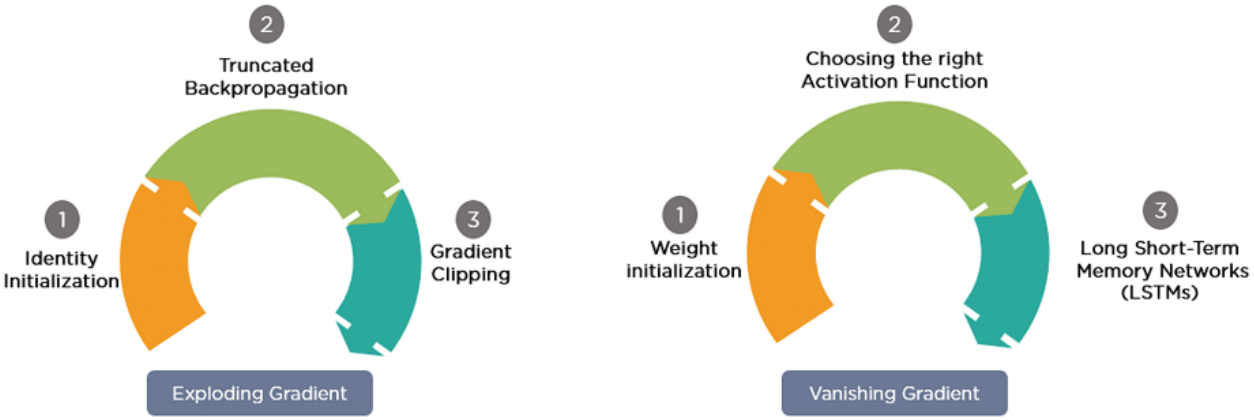
While training a neural network, if the slope tends to grow exponentially instead of decaying, this is called an Exploding Gradient. This problem arises when large error gradients accumulate, resulting in very large updates to the neural network model weights during the training process.

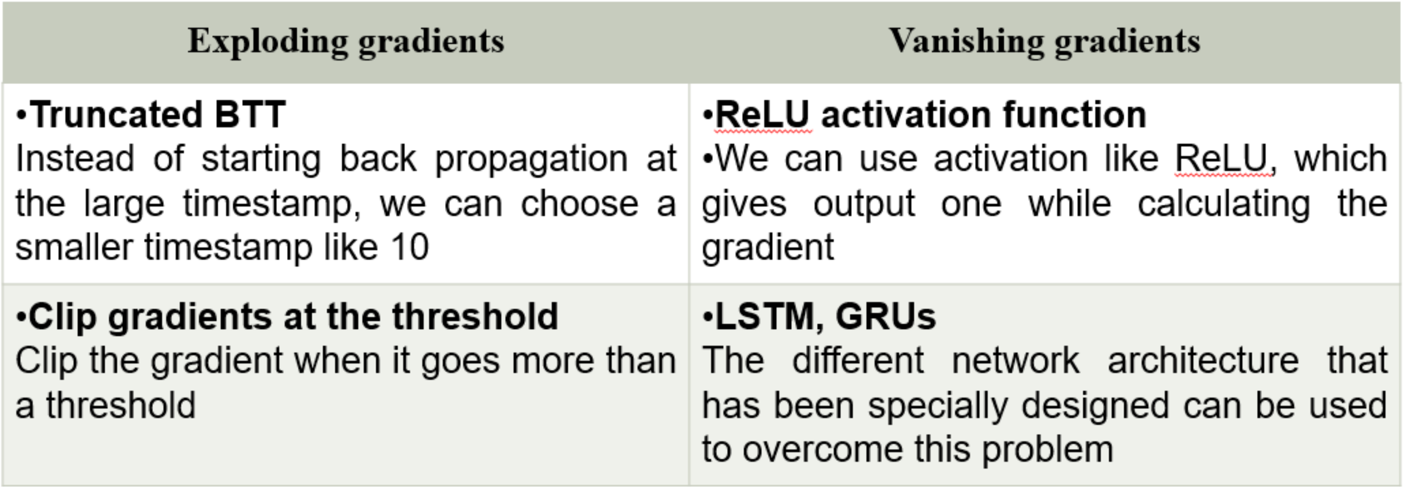
2. Vanishing Gradient Problem

RNNs suffer from the problem of vanishing gradients. The gradients carry information used in the RNN, and when the gradient becomes too small, the parameter updates become insignificant. This makes the learning of long data sequences difficult.

Long training time, poor performance, and bad accuracy are the major issues in gradient problems.

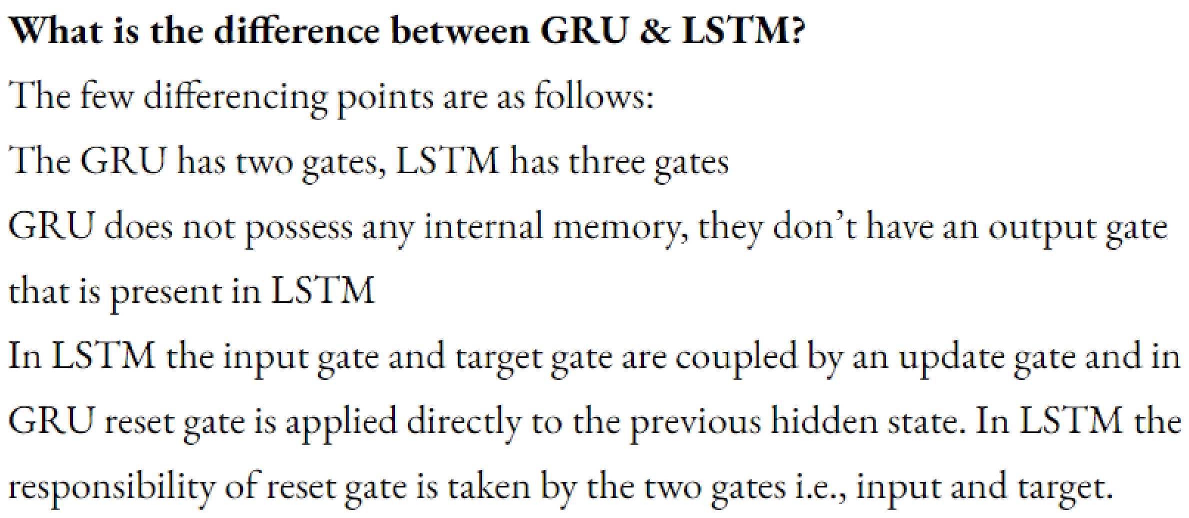
Gradient Problem Solutions





Exploding gradients	Vanishing gradients
•Truncated BTT Instead of starting back propagation at the large timestamp, we can choose a smaller timestamp like 10	**•ReLU activation function** •We can use activation like ReLU, which gives output one while calculating the gradient
•Clip gradients at the threshold Clip the gradient when it goes more than a threshold	**•LSTM, GRUs** The different network architecture that has been specially designed can be used to overcome this problem

D.) What is the difference between the workflow of LSTM and GRU? Give the examples where LSTM should be used over GRU.



What is the difference between GRU & LSTM?

The few differencing points are as follows:

The GRU has two gates, LSTM has three gates

GRU does not possess any internal memory, they don't have an output gate that is present in LSTM

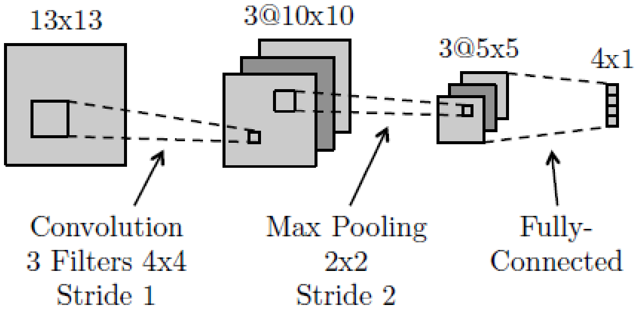
In LSTM the input gate and target gate are coupled by an update gate and in GRU reset gate is applied directly to the previous hidden state. In LSTM the responsibility of reset gate is taken by the two gates i.e., input and target.

One can choose LSTM if you are dealing with large sequences and accuracy is concerned, GRU is used when you have less memory consumption and want faster results

E.)

Q2. Convolutional Neural Nets

Below is a diagram of a small convolutional neural network that converts a 13x13 image into 4 output values. The network has the following layers/operations from input to output: convolution with 3 filters, max pooling, ReLu, and finally a fully-connected layer. For this network we will not be using any bias/offset parameters (b). Please answer the following questions about this network.



(a) How many weights in the convolutional layer do we need to learn?

48 weights. Three filters with 4x4=16 weights each.

(b) How many ReLu operations are performed on the forward pass?

75 ReLu operations. ReLu is performed after the pooling step. ReLu is performed on each pixel of the three 5x5 feature images.

(c) How many weights do we need to learn for the entire network?

348 weights. 48 for the convolutional layer. Fully-connected has 3x5x5=75 pixels each connected to four outputs, which is 300 weights. Pooling layer does not have any weights.

(d) True or false: A fully-connected neural network with the same size layers as the above network (13x13 → 3x10x10 → 3x5x5 → 4x1) can represent any classifier that the above convolutional network can represent.

● True ○ False

(e) What is the disadvantage of a fully-connected neural network compared to a convolutional neural network with the same size layers?

There are too many weights to effectively learn.