ASSIGNMENT 6

1. **What is the *vanishing gradient problem* in deep learning and how can it be mediated?**

Generally, a neural network contains a thousand or millions of hidden layers which makes the model complex. We keep on adding more hidden layers to the neural network so that the learning rate of model increases. Along with this we also do back propagation to find the gradient of loss(Error) with respect to the weights which keeps on decreasing as we move backward in the network. In a complex neural network with millions of hidden layers the gradient seems to disappear which is known as vanishing gradient problem.

There can be multiple ways to avoid the Vanishing Gradient Problem. One of them depends on the choice of Activation function. The Rectified Linear Unit[RELU] is one of the solution to this problem as it uses max(0,x) as activation function which means the gradient is 1 when output > 0, and zero otherwise. Hence multiplying the RELU derivatives together in the backprop equations can be 1 or zero which means the update is either nothing or takes contributions entirely from the other weights and biases.

Another recently discovered technique is Residual Networks [1], which introduces identity function id(x) = x, given x is the input. A residual network calculates y=f(x)+id(x) = f(x)+x, due to which the ResNets propagate the gradient throughout the model.

2. **What is the result when convolving the array A with the filter B without padding?**

```
A= array([[0, 0, 0, 2],            B= array([[2, 0, 1],
          [2, 1, 0, 0],                      [0, 0, 0],
          [2, 1, 0, 2],                      [2, 0, 1]])
          [2, 2, 2, 1]])

                                   array([[4,6],
                                          [10,7]])
```

3. **What is a pooling operation in convolutional neural networks and why is this operation important?**

Pooling in Convolution Neural Network is a form on non-linear down sampling. The objective is to reduce the dimensionality of the input representation (images, input matrix, etc.) by partitioning the image into non-overlapping rectangles (sub regions) and extracting one element from this sub region to form a pooling layer. Max pooling is one of the most common pooling operations in which the element with max value in this sub region is used to form the pooling layer. The pooling operation is important in CNN because it reduces the dimensionality which results in lesser number of parameters in the neural network preventing overfitting. The operational cost also decreases because of the lesser dimensions.

4. What is a gated recurrent network? Name an example of such a neural network.

A gated recurrent network solved the gradient vanishing problem of the standard RNN by using two gates update gate and reset gate which are two vectors. The update gate decides how much information from past should be passed to next level whereas the reset gate decides how much information from the past should be forgotten [6].

Update Gate z_t at time t is given by,

$$z_t = \sigma (W^{(z)}x_t + U^{(z)}h_{t-1})$$

X_t is passed in the network and multiplied by its own weight W(z) and h_(t-1) holds information from previous state multiplied by its own weight U(z).

Similarly, the reset gate r_t at time t is given by,

$$r_t =\sigma(W^{(r)}x_t + U^{(r)}h_{t-1})$$

The reset gate is used to store the relevant information from the past by using following equation,

$$h'_t = \tanh(Wx_t + r_t \odot Uh_{t-1})$$

$\odot$ is a Hadamard (element-wise) product symbol

Finally, the network calculated ht-1 which holds information from current layer and passes it to next layer, the update gates is used for this purpose,

$$h_t=z_t \odot h_{t-1} + (1-z) \odot h_t$$

LSTM (Long Short-Term Memory) is a simpler version of gated recurrent network that is made up of a cell to remember arbitrary values over time and following three gates:

1. Forget gate, to remove the information from cell state which is no longer required.
2. Input gate, to add new information to the cell state which is important and not redundant.
3. Output gate, to select only useful information from current cell state and transferring as output.

5. In reinforcement learning, what is a policy?

In a state of reinforcement learning, a policy is a function that defines which action *a* should take at a given time. The control policy function is given by,

$$a_t =\pi(s_t)$$

where $\mathbf{a}_t$ is the action at time t for being in state $\mathbf{s}_t$.

6. **Explain the difference between the SARSA and Q-Learning algorithm.**
Q-Learning is off policy, the Q function is updated by assuming the post state function Q(s t+1,a) will be maximized by taking action a.

SARSA (State Action Reward State Action) is on policy, we use the same function used in previous action at to generate the next action at+1.

7. **Briefly explain `dropout' and why it is used in deep networks.**
In neural networks, dropout is a regularization technique to prevent overfitting by randomly dropping out units from the hidden and visible layers. This stops from co-adapting too much. For example, if dropout is set to 0.5, half of the neurons will be ignored in the input layer.

References:

[1] https://arxiv.org/abs/1512.03385

[2] https://cs224d.stanford.edu/notebooks/vanishing_grad_example.html

[3] https://www.utc.fr/~bordesan/dokuwiki/_media/en/glorot10nipsworkshop.pdf

[4] https://stats.stackexchange.com/questions/176794/how-does-rectilinear-activation-function-solve-the-vanishing-gradient-problem-in

[5] https://colah.github.io/posts/2015-08-Understanding-LSTMs/

[6] https://towardsdatascience.com/understanding-gru-networks-2ef37df6c9be

[7] http://jmlr.org/papers/volume15/srivastava14a.old/srivastava14a.pdf

[8] https://www.quora.com/What-is-the-difference-between-Q-learning-and-SARSA-learning

[9] https://www.analyticsvidhya.com/blog/2017/12/fundamentals-of-deep-learning-introduction-to-lstm/

[10] https://en.wikipedia.org/wiki/Long_short-term_memory

[11] https://math.stackexchange.com/questions/20412/element-wise-or-pointwise-operations-notation