DEEP LEARNING WEEK 11

1. We construct an RNN for the sentiment classification of text where a text can have positive sentiment or negative sentiment. Suppose the dimension of one-hot encoded-words is $R^{100\times 1}$, dimension of state vector s_i is $R^{50\times 1}$. What is the total number of parameters in the network? (Don't include biases also in the network) (NAT)

   Answer: Range (7599.5,7601.5)

   Solution: No. of weight parameters in the network is given by 100x50(input to si)+50x50(s_i to s_{i+1})+50x2(s_i to output classes positive and negative).. So the total number of parameters in the network is 2500+5000+100=7600

2. Arrange the following sequence in the order they are performed by LSTM at time step t. [Selectively read, Selectively write, Selectively forget]

   a)Selectively read, Selectively write, Selectively forget
   b)Selectively write, Selectively read, Selectively forget
   c)Selectively read, Selectively forget, Selectively write
   d)Selectively forget, Selectively write, Selectively read

   Answer: c)
   Solution: At time step t we first selectively read from the state s_{t-1}, then selectively forget to create the state s_t. Then we selectively write to create the state h_t from s_t which will be used in the t+1 time step.

3. What are the problems in the RNN architecture? (MSQ)

   a)Morphing of information stored at each time step.
   b)Exploding and Vanishing gradient problem.
   c)Errors caused at time step t_n can't be related to previous time steps faraway
   d)All of the above

   Answer: d)
   Solution: Information stored in the network gets morphed at every time step due to new input. Exploding and vanishing gradient problems are caused by the long dependency chains in RNN.

4. We are given an RNN where max eigenvalue λ of Weight matrix is 0.9. The activation function used in the RNN is logistic. What can we say about $\nabla = ||\frac{\partial s_{20}}{\partial s_1}||$?

   a)Value of ∇ is close to 0.
   b)Value of ∇ is very high.
   c)Value of ∇ is 3.5.
   d)Insufficient information to say anything.

   Answer: a)
   Solution:Derivative of logistic is always less than $\frac{1}{4}$. Hence the gradient
   $\nabla = ||\frac{\partial s_n}{\partial s_{n-1}}|| = \gamma * \lambda < 1$. Due to backpropagation through other states s_i we get

$||\frac{\partial s_{20}}{\partial s_1}|| < (\gamma * \lambda)^{19}$ which is very close to 0.

5. What is the objective(loss) function in the RNN?

   a)Cross Entropy
   b)Sum of cross-entropy
   c)Squared error
   d)Accuracy

   Answer: b)
   Solution: RNN is used for sequential tasks. At each state s we have some predicted and actual output where loss between two is measured by cross-entropy. The loss across in RNN is measured by the sum of cross entropy across all such states in the network.

6. Which of the following is a limitation of traditional feedforward neural networks in handling sequential data? (MSQ)
   a) They can only process fixed-length input sequences
   b) They can handle variable-length input sequences
   c) They can't model temporal dependencies between sequential data
   d) All of These

   Answer: a),c) They can only process fixed-length input sequences

   Solution: Traditional feedforward neural networks are limited in their ability to handle sequential data because they can only process fixed-length input sequences. In contrast, recurrent neural networks (RNNs) can handle variable-length input sequences and model the temporal dependencies between sequential data.

7. Which of the following techniques can be used to address the exploding gradient problem in RNNs?
   a) Gradient clipping
   b) Dropout
   c) L1 regularization
   d) L2 regularization

   Answer: a) Gradient clipping

   Solution: Gradient clipping is a technique used to address the exploding gradient problem in RNNs. It involves capping the magnitude of the gradients during backpropagation, which helps prevent them from becoming too large and destabilizing the network.

8. Which of the following is a formula for computing the output of an LSTM cell?
   a) $o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$
   b) $f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$
   c) $c_t = f_t * c_{t-1} + i_t * g_t$
   d) $h_t = o_t * tanh(c_t)$

   Answer: d)

   Solution: The formula for computing the output of an LSTM cell is $h_t = o_t * tanh(c_t)$ where o_t is the output gate, c_t is the cell state, and h_t is the output at time t.

9. What is the purpose of the reset gate in a GRU network?

A) To decide how much of the previous hidden state to forget
B) To decide how much of the current input to add to the cell state
C) To decide how much of the previous hidden state to keep for the current time step
D) None of These

Answer: A, C)

Solution: To decide how much of the previous hidden state to keep for the current time step

10. Which of the following is true about LSTM and GRU networks?
A) LSTM networks have more gates than GRU networks
B) GRU networks have more gates than LSTM networks
C) LSTM and GRU networks have the same number of gates
D) Both LSTM and GRU networks have no gates

Answer: A) LSTM networks have more gates than GRU networks

Explanation: LSTM networks have three gates (input, output, and forget gates), while GRU networks have two gates (reset and update gates). Therefore, LSTM networks have more gates than GRU networks.