Midterm Review Week-3 (Recurrent Neural Network)

Q1. Which of the following task is name classification?

a) Sequence generation `model generates a sequence of outputs, (eg sentence/paragraph, series of actions)`

b) Sequence-to-sequence `transforms an input sequence into an output sequence, (eg machine translation)`

c) Sequence labeling `each element in the input sequence is assigned a label`

d) Text generation

Q2. What is a potential downside of gradient clipping? `limit max value of gradient to prevent exploding gradients not solution to vanishing gradients`

a) Underfitting `can limit the step size during training and hinder learning, this is because it artificially restricts the updates to the model parameters`

b) Slower convergence `can distort the gradients/learning process resulting in suboptimal convergence or longer training time`

c) Exploding gradients

d) Difficult optimization

Q3. In RNNs, what issue is associated with initializing all weights to zero?

a) It prevents the network from learning

b) It leads to exploding gradients during training

c) It results in slow convergence

d) It causes the vanishing gradient problem

Q4 How does Monte Carlo dropout provide regularization for RNNs?

a) By randomly dropping connections during training

b) By training multiple models in an ensemble `same model runs multiple times with different dropout masks, not multiple different models`

c) By adding Gaussian noise to activations `noise from randomly dropping units (setting their outputs to zero), not by adding Gaussian noise to the activations (stochastic noise injection)`

d) By using dropout at test time to sample predictions

Q5. For a simple RNN, if hidden vector is of size 1x5, input vector is of size 1x5, output vector is of size 1x5, then calulate the number of trainable parameters. RNN uses these equations for forward pass:

$$h_t = \sigma\left(Ux_t + Wh_{t-1}\right)$$

$$y_t = \phi\left(Vh_t\right)$$

a) 75

b) 50

c) 25

d) 15

Q6. Suppose a droupout layer has 10 neurons and the dropout probability is 0.3. What are the different number of configurations obtained by this setup.

a) 120

b) 45

```
10 choose 7
or
10 choose 3
```

c) 21

d) 210

Q7.Which of the following activation function has derivative of the form y' = (1 - y)(1 + y)

a) Relu

b) Sigmoid

c) Tanh

d) Exponential Linear Unit

Q8. Where does the input gate place new information in an LSTM?

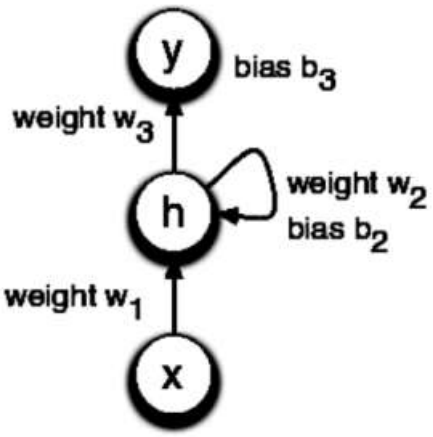
a) Hidden state

b) Output layers

c) Cell state

d) Forget gate

Q9. In the following RNN:



$h_t = f(w_1x_t + w_2h_{t-1} + b_2)$

$y_t = g(w_3h_t + b_3)$

f is Relu, g(x) = abs(x).

Hidden State is initialized as 0.

$w_1 = -1$, $w_2=1$, $w_3=2$, $b_2 = 2$, $b_3 = -1$

Find out output to the input sequence 1 1 5.

a) 1 3 1

b) 1 3 0

c) 1 2 1

d) 1 1 0

Q10. For an RNN language model with a vocabulary of 10,000 words, if you're using one-hot encoding for the input data, what will be the size of the input vector at each time step?

a) 1,000

b) 10,000

c) Isn't related to the vocabulary size

d) 200