Quiz-07

- Due Mar 2 at 11:59pm
- Points 10
- Questions 10
- Available Feb 28 at 6pm - Mar 2 at 11:59pm
- Time Limit None
- Allowed Attempts 3

Take the Quiz Again

Attempt History

	Attempt	Time	Score
KEPT	Attempt 1	139 minutes	7 out of 10
LATEST	Attempt 2	159 minutes	6 out of 10
	Attempt 1	139 minutes	7 out of 10

Correct answers are hidden.

Score for this attempt: 6 out of 10
Submitted Mar 2 at 4:14pm
This attempt took 159 minutes.

Incorrect Question 1
0 / 1 pts

Which of the following are true for the Elman and Jordan networks:

Hint: Lecture 13- slides 43 - 46

- [ ] The training processes for neither type of network captures true recurrence
- [ ] Jordan networks have infinite memory since they factor in all the past outputs during inference
- [ ]

Both networks are not actually recurrent during inference because the history is not fully represented in the networks.

Elman networks maintain a running average of outputs from previous time instants within memory unit, while Jordan networks store hidden unit values for one time instant in a “context” unit.

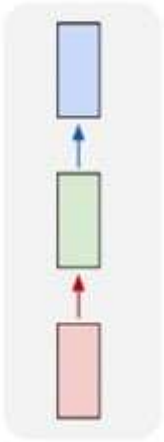
Question 2

1 / 1 pts

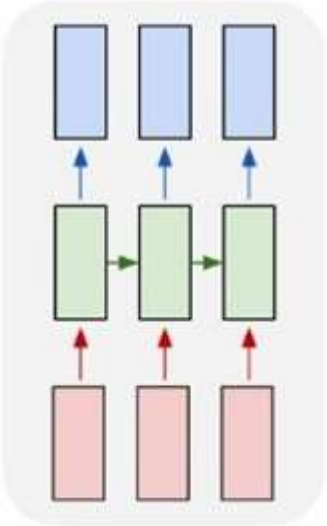
In a problem where an RNN is used to analyze a sequence of words to predict the next word, what kind of model would best fit the problem. Please refer to the definition of different types of recurrent networks mentioned in the lecture.

Hint: Lecture 13- slides 68, 70

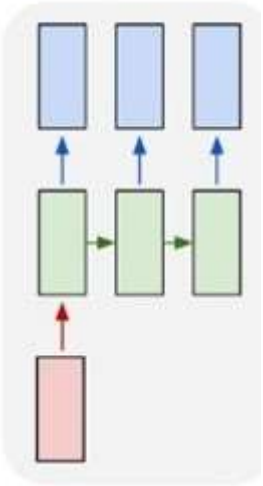
one to one



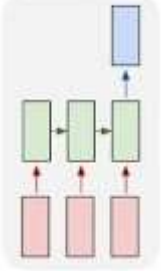
many to many



one to many



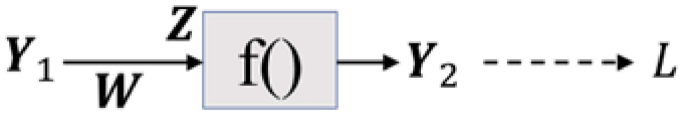
many to one



IncorrectQuestion 3

0 / 1 pts

A network layer operates on an input Y1 (from the previous layer) to compute an affine value Z = WY1 + b. It then applies an activation function f() on Z to compute its output Y2. Subsequent layers of the net operate on Y1 to compute the network output, for which the loss L is computed.



Let Y1 be an Mx1 column vector, W be an NxM matrix, and Y2 be an Nx1 column vector. We use the notation $\nabla_X Y$ to represent the derivative of Y with respect to X. Which of the following is true (you may want to check the sizes of the left and right hand sides of the equations to verify):

- [x] $\nabla_W L = \nabla_Z L Y_1$
- [x] $\nabla_{Y1} L = W \nabla_Z L$
- [ ] $\nabla_Z L = \nabla_Z Y_2 \nabla_{Y_2} L$
- [ ] $\nabla_{Y1} L = \nabla_Z L W$
- [ ] $\nabla_W L = \nabla_Z L Y_1^T$
- [ ] $\nabla_W L = Y_1 \nabla_Z L$
- [x] $\nabla_Z L = \nabla_{Y_2} L \nabla_Z Y_2$

Question 4

1 / 1 pts

The following statements relate to backpropagation in a bi-directional RNN. Select all that are true. All statements must be assumed to be complete, and not partial facts.

Hint: Lecture 13- slides 115 - 123

- [x] Backpropagation propagates gradients forwards through time in some parts of the network and backwards through time in others.
- [x] Backpropagation is independently performed in the forward and backward components of a bidirectional block.
- [ ] The derivatives computed from the forward network are subsequently propagated onwards into the backward network.
- [ ] Backpropagation in a bi-directional RNN only propagates gradients backwards through time.

IncorrectQuestion 5

0 / 1 pts

Select all statements that are true. You will have to refer to:

http://deeplearning.cs.cmu.edu/F20/document/readings/Bidirectional%20Recurrent%20Neural%20 (https://ieeexplore.ieee.org/document/650093)

- [x] Bi-directional RNNs capture distant long-distance (i.e. over time) dependence between the input and the output, whereas uni-directional RNNs only capture short-distance dependencies.
- [x] Bi-directional RNNs are unsuited to problems where output predictions must be made online, as newer samples from the input sequence arrive.
- [x] A uni-directional RNN can incrementally compute its outputs in a streaming manner as inputs are processed sequentially over time, whereas a bi-directional RNN must process the entire input sequence before computing its outputs.
- [x] You can get the same output as a bi-directional RNN, as it is defined in the paper, by training two uni-directional RNNs, one processing the input left-to-right, and the other right-to-left, and averaging their outputs.

IncorrectQuestion 6

0 / 1 pts

In RNN, given that the following operation holds for the hidden unit $h(t) = f(wh(t-1) + cx(t))$, where f is a non-linear activation function and w and c are positive, which of the following is true about

$h(t)$ when $x(0)$ is negative, and all subsequent $x(t)$ values are 0? (Select all that apply) Assume $h(-1)$ is 0. In these statements the word “activation” refers to the function $f()$

Hint: Draws concepts from Lecture 14 Slides 37-43

- [x] Using sigmoid, the output of the activation saturates at 0
- [x] Using tanh activation, the output eventually saturates.
- [x] Using Relu activation, the output is zero and nothing passes on to the network
- [x] Using Relu activation, it is sensitive and can blow up

Question 7
1 / 1 pts

A recurrent neural network with a single hidden layer has 3 neurons in its recurrent hidden layer (i.e. the hidden representation is a 3-dimensional vector). The recurrent weights matrix of the hidden layer is given by

$$W_r = \begin{bmatrix} 0.75 & 0 & 0.25 \\ 0.75 & -0.2 & 0.0 \\ 0.5 & 0. & -0.35 \end{bmatrix}$$

The hidden unit activation is the identity function $f(x) = x$. All bias values are 0.

Let us define the “memory duration” of the network as the minimum number of time steps N such that for an isolated input at time t (with no previous or subsequent inputs) the length of the hidden activation vector at time t+N certainly falls to less than 1/100th of its value at time t. This is effectively the amount of time in which the influence of the input at time t all but vanishes from the network.

What is the memory duration of the above network (choose the closest answer). Assume h(t-1) = 0 (the zero vector).

Hint: Lecture 14 Slides 28-30

Hint: Remember that for any eigenvector E, $W_r E = \lambda_e E$, where λ_e is the Eigenvalue corresponding to E. If the hidden response of the system is exactly E (with length 1) at t = 0, what is the response of the system after t time steps, and what is its length?

- (•) 30
- ( ) 40
- ( ) 10
- ( ) 20

Question 8
1 / 1 pts

A recurrent neural network with a single hidden layer has 3 neurons in its recurrent hidden layer (i.e. the hidden representation is a 3-dimensional vector). The recurrent weights matrix of the hidden layer is given by:

$$W_r = \begin{bmatrix} 0.75 & 0.25 & 0 \\ 0.2 & -0.1 & 0.7 \\ -0.2 & 0.65 & 0.15 \end{bmatrix}$$

The hidden units have ReLU() as their activation function. All bias values are 0.

Let us define the "memory duration" of the network as the minimum number of time steps N such that for an isolated input at time t (with no previous or subsequent inputs) the length of the hidden activation vector at time t+N almost certainly falls to less than 1/100th of its value at time t. This is effectively the amount of time in which the influence of the input at time t all but vanishes from the network. The weight matrix for the inputs to the RNN (W_c in the slides) is an Identity matrix (3×3). Consider the input vector at time step t, $x(t)$ be $\left[\frac{1}{\sqrt{(3)}}, \frac{1}{\sqrt{(3)}}, \frac{1}{\sqrt{(3)}}\right]^\top$.

What is the memory duration of the above network (choose the closest answer). Assume h(t-1) = 0 (the zero vector). You may want to simulate the network to determine this value (an analytical solution is difficult to get).

Hint: Draws concepts from Lecture 13. For practical implementation, slides discussing the RNN architecture, equations governing the state updates, and examples of how inputs are processed through time would be most relevant. (Slide 73 and beyond)

- 13
- 7
- 31
- 23

Question 9
1 / 1 pts

An RNN has a single recurrent hidden layer. The hidden state activations are all $sigmoid$. You will recall from class that the recurrence of the hidden layer in such a network has the form:

$$\mathbf{z}_t = W_{hh}\mathbf{h}_{t-1} + W_{xh}\mathbf{x}_t + \mathbf{b}$$

$$\mathbf{h}_t = sigmoid(\mathbf{z}_t)$$

where $\mathbf{h}_t$ is the recurrent hidden state (vector) at time t, $\mathbf{x}_t$ is the input (vector) at time t, W_{hh} is the recurrent weight matrix, W_{xh} is the input weight matrix and $\mathbf{b}$ is the bias.

During backpropagation, the length of the derivative vector $\nabla_{\mathbf{h}_t} Div$ for the hidden activation at time t is found to be exactly 1. This derivative is further backpropagated through the activation of the hidden layer to obtain $\nabla_{\mathbf{z}_t} Div$. What is the *maximum* length of $\nabla_{\mathbf{z}_t} Div$?

Please provide the answer in X.XX numeric format (i.e. in explicit decimal format, truncating the answer after the second digit after the decimal)

Hint: Lecture 13- slides 80 - 100

0.25

Question 10
1 / 1 pts

A deep neural network has 2 neurons in its $(l-1)$-th layer and 3 neurons in its l-th layer. The weight matrix for the l-th layer (i.e. the connections between the $(l-1)$-th layer and the l-th layer) is

$$\mathrm{W}_l = \begin{bmatrix} \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ 0.25 & -0.25 \\ -0.25 & 0.25 \end{bmatrix}$$

During backpropagation, the derivative vector $\nabla_{\mathbf{z}_l} Div$ for the affine input to the l-th layer, $\mathbf{z}_l$, is found to have length exactly equal to 1.0. What is the maximum length for the derivative $\nabla_{\mathbf{y}_{l-1}} Div$ for the output of the $(l-1)$-th layer, $\mathbf{y}_{l-1}$?

Hint: Recall that for any relation $\mathbf{u} = W\mathbf{v}$, the maximum value for the ratio of the lengths of $\mathbf{u}$ and $\mathbf{v}$ is the largest singular value of W, i.e. $\frac{|\mathbf{u}|}{|\mathbf{v}|} \leq \mathbf{max}(singular\ value(W))$

0.5

Quiz Score: 6 out of 10