Deep Learning CS583 Fall 2021
Quiz 3

December 1, 2021

Instructor: Jia Xu

Student name: Archana Kalburgi

Student ID: 10469491

Student email address: akalburg@stevens.edu

- Read these instructions carefully
- Fill-in your personal info, as indicated above.
- You have 24 hours.
- There are five items in one question. Each item is worth the same (1 point). • Both computer-typed and hand-writing in the very clear form are accepted. • This is an open-book test.
- You should work on the exam only by yourself.
- Submit your PDF/Doc/Pages by 12:30 Dec 3rd on Canvas under Final exam.

good luck!

1 Question

- What is the advantage and disadvantage of attentional models compared to RNNs.

Advantages of attention models compared to RNN:

- Attentional models in NLP are helpful in identifying the information in an input by adding attention weights to a model as trainable parameters to augment important parts of our input.
- Attentional models have the ability to selectively focus on segments of the sequence.
- The context vector encapsulates the most relevant information from the encoder, which is makes attention is a useful mechanism.

Disadvantages of attentional models

- The main disadvantage of the attention mechanism is that it adds more weight parameters to the model, which can increase training time especially if the input data for the model are long sequences.
- In a sequence-to-sequence model the attention mechanism combines the encoder's output at each time step, along with the decoder's output at time step t, to create the context vector for time step *t*.
- In humans, attention serves to reduce our workload by allowing us to ignore unimportant features. However, in a neural network, attention entails overhead as we are now generating attention distributions and training our attention weights.
- We are not actually ignoring the unimportant features, just diminishing their importance

Choose one correct answer from four candidates:

- In practice, what is the most accurate description of activation functions (such as Sigmoid, Sum, Tanh, ReLU) used in neural networks?

1. They must be differentiable.
2. They can be non-differentiable, but only for a small number of points.
3. They can be any continuous function.
4. They must be non-linear to be learnable.

Ans- **2**. They can be non-differentiable, but only for a small number of points.

- Given a neural network with *N* input nodes, no hidden layers, one output node, with entropy loss and sigmoid activation functions, which of the following algorithms (with the proper hyper-parameters and initialization) can be used to find the global opti mum?

    1. Stochastic Gradient Descent
    2. Batch Gradient Descent
    3. Mini-Batch Gradient Descent
    4. All of the above

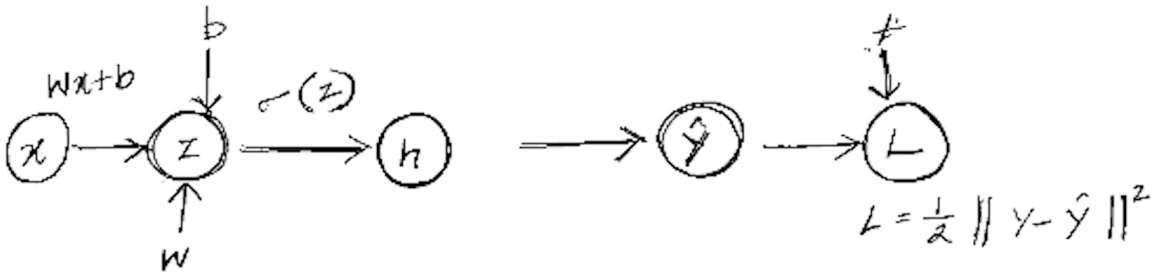
Ans- 4. All of the above

- You want to train a neural network to predict the next 30 daily prices using the previous 30 daily prices as inputs. Which model selection and explanation make the most sense?

    1. A fully connected deep feed-forward network because it considers all input prices in the hidden layers to make the best decision.
    2. A single one-directional RNN because it considers the order of the prices, and the output length is the same as the input length.
    3. A bidirectional RNN because the prediction benefits from future labels.
    4. A one-directional encoder-decoder architecture can generate a sequence of future prices based on all historical input prices.

Ans 4- A one-directional encoder-decoder architecture can generate a sequence of future prices based on all historical input prices.

- Draw the computational graph of a one-hidden layer feed-forward neural network and write the derivatives of each variable in the backpropagation.



Forward:

$z = wx + b$

$h = \sigma(z)$

$y = wh + b$

$L = \frac{1}{2} \| \hat{y} - y \|^2$

Backward:

$\bar{L} = 1$

$\bar{y} = \bar{L}(y - \hat{y})$

$\bar{w} = \bar{y} w$

$\bar{b} = \bar{y}$

$\bar{h} = w + \bar{y}$

where, $\bar{z} = \bar{h} \cdot \sigma(z)$

$w = z +$

$b = z$