CSE514, Fall 2021, HW 4 Name: Anh Le Student ID:488493

Note: This homework is worth a total of 15 points

Answers to the following questions should be based on your reading of a research paper. Pick one following:

OptCoNet: an optimized convolutional neural network for an automatic diagnosis of COVID-19
Downloadable from Canvas with file name: **HW4 Goel2021_CNN.pdf**

A Neural Language Model for Query Auto-Completion
Downloadable from Canvas with file name: **HW4 Park20217_RNN.pdf**

Please indicate which paper you chose (or just circle it above):

A Neural Language Model for Query Auto-Completion

Q1 (5pts): What is the problem that the network is meant to solve?
What features of the problem motivated the authors to choose their proposed network structure (CNN or RNN)?

The network is trying to suggest a list of queries that complete the user's input text. Goal is to build an auto completion system for search query without past queries in the storage so that even a completely new input would suggest the list of queries for users.

The author proposed using an RNN-model network structure because it "*encodes variable-length prefix in a fixed-length state vector*", and it will predict the most probable query for each state. Moreover, since the number of possible prefixes increases exponentially as the number of words increases. Therefore, an RNN will be a better solution since the information from all previous texts can be represented as a low-dimensional state vector.

Q2 (5pts): Give a written or graphical description of the network's architecture.
For each unique substructure (e.g. MPL or LSTM), give a brief definition of its purpose.

The architecture of the neural network is as follow:
1. Input layer: The user's search query
2. Embedding layer:
- Place each character's one-hot encoded vector and embedded vector to combine the original input character and its distributed representation
3. Layers of Long Short-Term Memory and dropout:
- Each character in the search query is map to its concatenated vector, which is passed to 2 layers of LSTMs to prevent over-fitting
4. SoftMax layer:
- After the last LSTM later, the output character's probability is estimated by applying a softmax function in this layer.
5. Output layer:
- The predicted character that comes after the input character.

Q3 (5pts): How was the network evaluated? Define at least two performance metrics used.
For one of the performance metrics, explain how it's calculated, and give an example case

e.g. For the performance metric of variance explained on evaluation data, the formula is 1 - MSE/Variance.
An example case would be predicting [1.2, 1.8, 4] when the actual data is [1, 2, 3]
The MSE = [(1 - 1.2)^2 + (2 - 1.8)^2 + (3 - 4)^2] / 3 = 0.27
The Variance = [(1 - 2)^2 + (2 - 2)^2 + (3 - 2)^2] / 3 = 0.67
The variance explained = 1 – (0.27/0.67) = 0.595

$$MRR = \frac{1}{|P|} \sum_{p \in P} \frac{1}{r_p}$$

P = ["New York", "restaurant"]

First prefix p_1 = "New York"
- original query : "New York Giants"
- query candidates : ["New York Grill", "New York Times", "New York Yankees", "New York post", "New York Giants"]

Second prefix p_2 = "restaurant"
- original query : "restaurant near me"
- query candidates : ["restaurant near me", "restaurants st louis", "restaurant open near me"]

$$\Rightarrow MRR = \frac{1}{2}\left(\frac{1}{5} + \frac{1}{1}\right) = 0.6$$