ICS 4102 Machine Learning

July 5th, 2023, ***Deadline 1:50 pm***

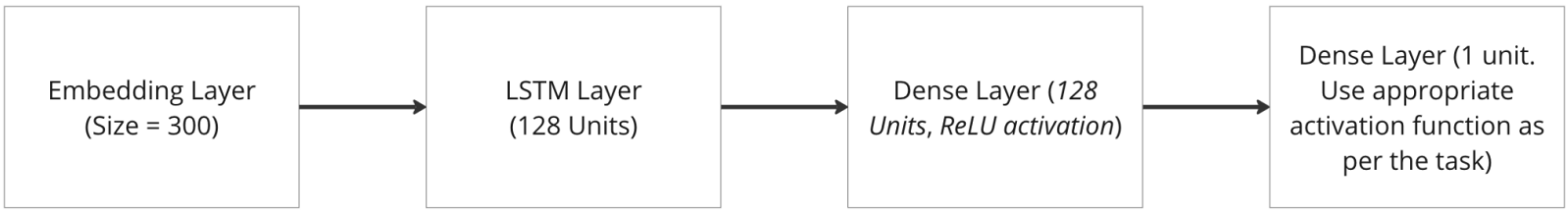
Background:

Medical query classification using Natural Language Processing (NLP) can be a beneficial approach to improving medical assistance in African countries. NLP techniques can help automate the process of understanding and categorizing medical queries, enabling faster responses.

Task

Primary Task

By using the knowledge you've acquired so far in the unit ICS 4104 Machine Learning, **implement a simple ANN pipeline** capable of classifying medical queries into two classes. Use the reference image below to design your ANN to match this exact image. The last Desne layer is the output layer as such use an appropriate activation function for this layer.



Sub Tasks

Since you are building an entire pipeline, you will need to also cover the following areas.

1. Data Cleaning: The data given to you is in CSV format This data is a snippet of real medical queries from real humans hence expect it to need some significant amount of cleaning. It is up to you to decide how you'd want to clean the data.
2. Encoding Labels (where necessary)
3. Obtaining Features using Tensflows (version 2.8 +) Embedding Layer with dimension 300.
4. Plotting a Loss Curve after training for a minimum of 30 iterations (epochs)
5. Saving your final Neural network weights as an **.h5** file extension with your student number as the file name.

Deliverables

After completing the above tasks, you will have all the required information and materials to upload as well as respond to some questions. Uploads and Question responses are to be done via this Google Form linked below.

Link: https://forms.gle/DHAFySeVWNpdunsY7

In case, you have problems uploading files via Google Forms, use the alternative links shared on E-Learning. Use the link that matches your group registration.

Files to upload include:

1. **Your final Colab / Jupyter Notebook** with all cell output included. Uploading a file with no outputs will lead to an automatic zero score since there won't be any outputs to validate your responses.
2. **Your final artificial neural network weights** stored in a **h5 file.**

Uploaded files should be named using the format:

Notebooks

- A - 111111.ipynb OR
- B - 111111.ipynb OR
- C - 111111.ipynb

Weight Files (.h5)

- A - 111111.h5 OR
- B - 111111.h5 OR
- C - 111111.h5

Provided

To aid you in this assessment, a dataset (csv) file has been shared along with this PDF document. Additionally, a Jupiter Notebook has also been provided. **It is mandatory that you use this notebook as it will read the given dataset and randomly select two classes for you to use.** This notebook also randomly deletes 50 rows of text and as such you should handle this during pre-processing.

Alternative Notebook Link:

https://colab.research.google.com/drive/1to78da52cK_8HOSJikh3fP_r6mPptIer?usp=sharing