ECS 189G-001

Deep Learning

Winter 2024

Course Project: Stage 4 Report

Team Information

NeuralNinjas		
Student 1: Srihita Ramini	Student 1: 919221527	Student 1: sramini@ucdavis.edu
Student 2: Trishna Sharma	Student 2: 918135782	Student 2: tksharma@ucdavis.edu
Student 3: Sai Sindura Vuppu	Student 3: 918091017	Student 3: svuppu@ucdavis.edu
Student 4: Olivia Shen	Student 4: 919293157	Student 4: oshen@ucdavis.edu

Section 1: Task Description

RNN Text Generation

In this task, we were given a dataset containing various short jokes. Using this dataset, we had to build an RNN model in order to generate the rest of a new joke based on the first 3 words that are inputted into the model.

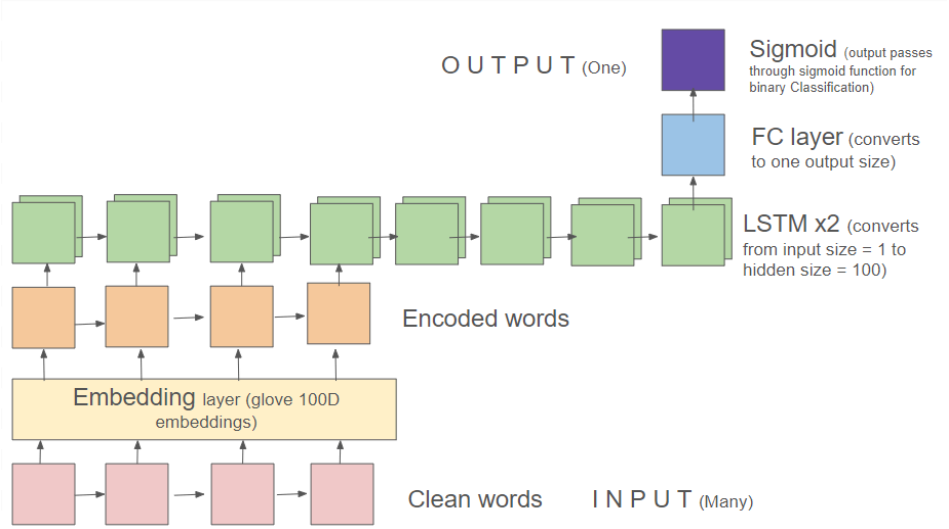
RNN Text Classification

In this task, we were given a dataset containing movie reviews. Using this dataset we are asked to perform sentiment analysis to train an RNN model to classify the movie reviews to positive or negative.

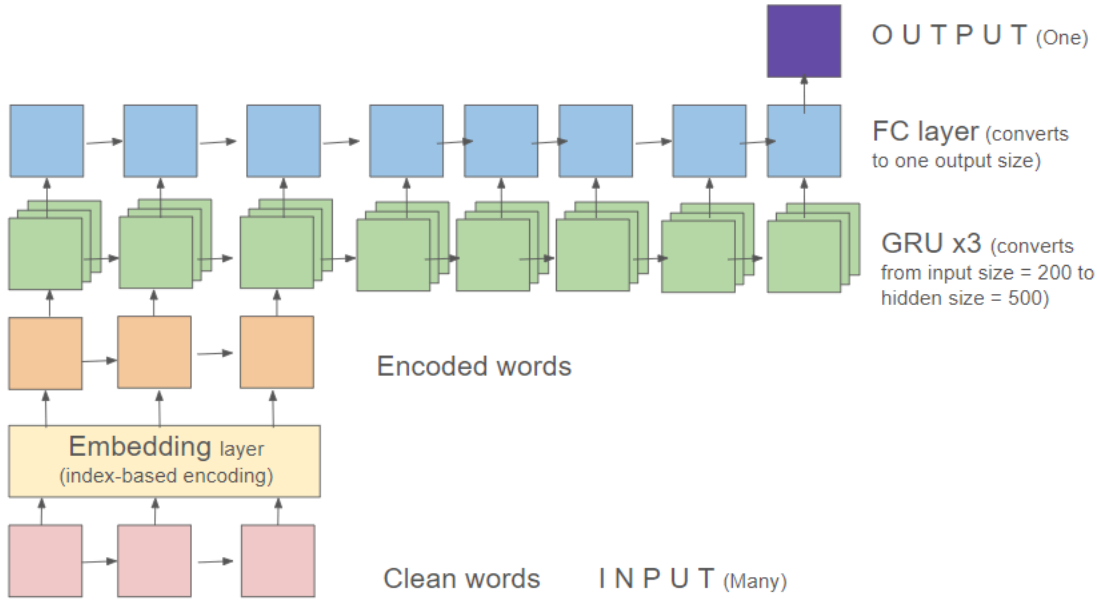
Section 2: Model Description

Text Classification

This is a many-to-one model architecture, as we have many input (sequence of many words of a review) which passes through the embedding, RNN, Dropout and Fully Connected layer to receive one output (positive or negative) through the Sigmoid function.



Text Generation
This is a many to one architecture that takes a sequence of 3 tokens. The embedding layer has size 200. The GRU layer has size 500, 3 layers, and 20% dropout. We only take the output of the current GRU and pass to a fully connected layer with output size 4429, which is the vocabulary size.



Section 3: Experiment Settings

3.1 Dataset Description

Text Classification
The dataset consists of 50k reviews, each in a text file, with a positive or negative encoding (>= 7 as positive and <=4 as negative) divided into 25k positive and 25k negative reviews. Which is equally divided into training and testing datasets

RNN Text Generation
The dataset consisted of 1622 pieces of short jokes which was all given in text format.

3.2 Detailed Experimental Setups

RNN/LSTM Text Classification
For this part of the assignment, we first cleaned the data. For each review, we got ecah of the words. Then we removed any punctuation, tags, and stop words such is “is”, “the” that are not important for classifying the information. We then tried out various encoding methods to create the embeddings. For this we created a vocab set with 69966 unique words and assigned a index value to each of the words in the set. We tried BERT encoding, glove 50D and glove 100D. And settled with glove 50D. We processed our data into mini batches of 64. We used the glove embeddings as an embedding layer for our model (with vocab size of 69966 and embedding size of 100), than ran a RNN model (with input size = 1, hidden size = 256 and 2 layers), followed by a dropout layer and a fully connected layer, outputing one value (choosing one between positive and negative) and ended with a Sigmoid function. We used BCELoss (Binary Cross Entropy Loss) since we’re dealing with classification of two labels, 0 (negative) or 1 (positive). And we used Adam optimizer to run the model for 10 epochs with a learning rate of 0.001.

RNN Text Generation
For this task, we chose to try out different types of encoding for the data preprocessing. One hot encoding and index based encoding were both methods that were tried and we decided to proceed with the index based method. We assigned a unique integer to every word that appeared in our dataset (vocabulary) as per our encoding method and fed it into an RNN with an embedding layer, GRU layer, and a fully connected layer. The hyperparameters consisted of an epoch number of 20, hidden size of 40 neurons, learning rate of 0.001, embedding size of 200., rnn_size of 200, 3 layers of stacked rnn This model was then trained used a batch size of 1000.

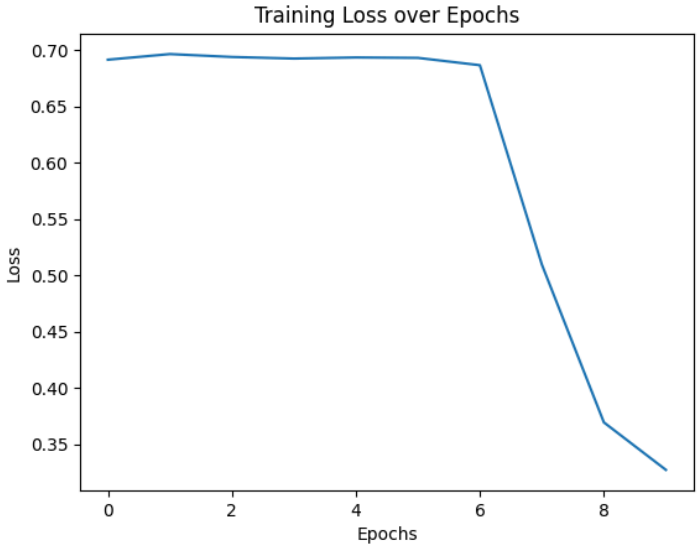
3.3 Evaluation Metrics
We use Accuracy, Precision, Recall and F1 as our metrics.

Text Classification:
Accuracy will be a good enough measure for text classification since the dataset is balanced with no more than 30 reviews per movie and equal number of positive and negative reviews for both training and testing.

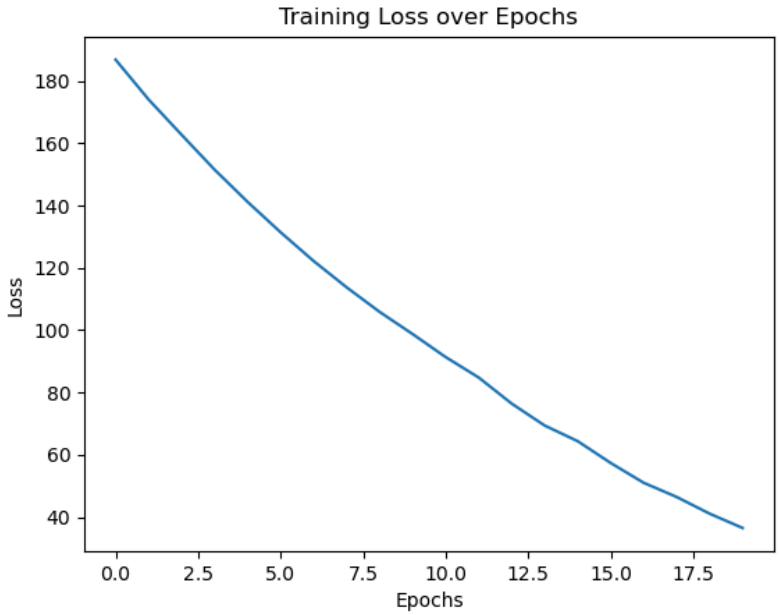
Text Generation:
No formal evaluation metrics used. The model is judged on if the text output sounds like an English sentence.

3.4 Source Code
Source code: https://github.com/srihita123/189G-

3.5 Training Convergence Plot
Text Classification:



Text Generation:



3.6 Model Performance

Text Classification:

The LSTM model performs well, with a batch size of 64, hidden layer of size 100, 2 LSTM layers stacked, running for 10 epochs with a learning rate of 0.001, gives an accuracy of 88%

```
Accuracy: 0.8787, Loss: 0.2910
              precision    recall  f1-score   support

           0       0.85      0.90      0.88     11869
           1       0.90      0.86      0.88     13131

    accuracy                           0.88     25000
   macro avg       0.88      0.88      0.88     25000
weighted avg       0.88      0.88      0.88     25000

saving results...
evaluating performance...
Accuracy: 0.87868
Precision: 0.8796474291201659
Recall: 0.8786799999999999
F1: 0.8786026629037831
```

Text Generation:
The Text Generation performs differently with RNN, LSTM and GRU models. The GRU model works the best jokes generated. Here are jokes generated with 3 words input:
---------- LSTM ----------
why do cheetahs sleeping taste ? a dock catholic .
why do cheetahs you keep a elephant ? to get to a ride
why do cheetahs blind wear fall ? acids ! ! .
why do cheetahs fish abroad .
why do cheetahs get pikmin ? a buccaneer !
why did will smith say when the rooster was play to say to grain
how do dogs buy pikmin their drop in a salad car and its
how do dogs know if it 14 such no dinner ? lefts a
how do dogs find about a f1 ledge fe zombie
---------- RNN ----------
how do dogs drive ? hundred
what do you call a fish who wants off foot tuna and as
what do you call a fish who is hard https brand fired
what does chicken call go , what 's red and they have been
how do i hate count sick talk trying typo : way : mad
how do they always say to her ? nice lane
how do cheetahs unlock you know not a time ! it bird .
how do cheetahs dhaka when it would affect the chicken . today who
how do cheetahs also duck it outside for ... but i 'm good
why did the pope ask the door was a)
why did the name bar his work dressing
why did the life use when itself ? investigator
---------- GRU ----------
why did the chicken use skydiving greek a baseball from the rash top
why did the grocery delivery thing ?
how do cheetahs chickens when we his he ran to the center ?
how do cheetahs plant like all these ? thesaurus could moving breakfast .

how do cheetahs stay ? there 's both scary 's ? flowers their
why do girls hate ? faithbook
why do girls like ? hot .
why do girls wear ? because impatient all over by body with help
why do girls hate sofas at spacious icecream ? she needed new over
what is red gump ? blue : (man who could only solo
what is red and stay ? bunny beef .
a bunny walks into a bar ... two can once on the bottom
a bunny walks out .
a bunny walks into a bar i typist her off there .
a bunny walks into a bar the bartender says , `` keep frozen
how do dogs like you can see your bucket .
how do dogs above when fun = : & buns ; butterfly out
how do dogs people have going sleeping ... but yep i 'm 45.
how do dogs drive ? bread `` quack color . "

3.7 Ablation Studies

Text Classification:

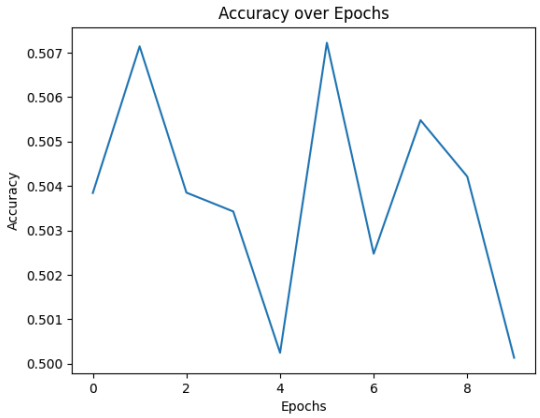
1. LSTM
    a. LSTM performs really well with 88% accuracy in 10 epochs.
    b.

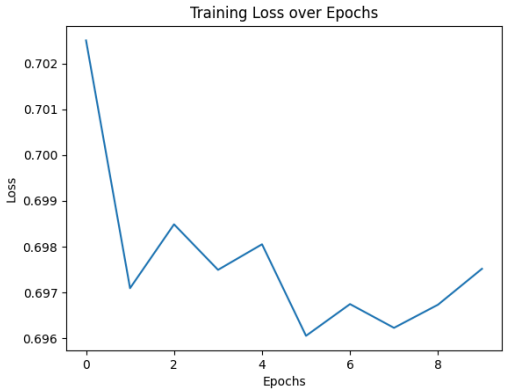
```
Accuracy: 0.8787, Loss: 0.2910
              precision    recall  f1-score   support

           0       0.85      0.90      0.88     11869
           1       0.90      0.86      0.88     13131

    accuracy                           0.88     25000
   macro avg       0.88      0.88      0.88     25000
weighted avg       0.88      0.88      0.88     25000

saving results...
evaluating performance...
Accuracy: 0.87868
Precision: 0.8796474291201659
Recall: 0.8786799999999999
F1: 0.8786026629037831
```

2. RNN: performed worse than LSTM (with all the parameters as the same)
    a.

```
evaluating performance...
Accuracy: 0.50888
Precision: 0.5305824889627522
Recall: 0.50888
F1: 0.4029593531926746
```

b.





Text Generation

1. LSTM
    a. Took the longest
    b. Results were better than RNN
2. GRU
    a. Best output among all three
3. RNN
    a. Not very desirable outputs