NGEE ANN POLYTECHNIC

School of InfoComm Technology

Deep Learning Assignment

Diploma in CSF / FI / IT

April 2020 Semester

ASSIGNMENT 2

(40% of DL Module)

13th Jul 2020 – 21st Aug 2020

Submission Deadline:

Presentation: 11th Aug 2020 (Tuesday), 11:59PM

Report: 23rd Aug 2020 (Friday), 11:59PM

Tutorial Group	**:**	**P01 / P02 /P03 /P04**
Student Name	**:**	
Student Number	**:**	

Penalty for late submission:

10% of the marks will be deducted every calendar day after the deadline.

NO submission will be accepted after 28th Aug 2020 (Friday), 11:59PM.

A. Assignment Specifications

Problem 1 (50%)

1.1 Objective

Build a sentiment analysis model to predict the emoticon for each text input.

Dataset

Download the dataset from MEL:

- **dataset.csv**: the full dataset which includes text inputs and target labels
- **mapping.csv**: the emoji dictionary which maps each label to a emoticon

1.2 Suggested Tasks

You are recommended to tackle this problem by following the steps below. A Jupyter Notebook (**Assignment_2_p1.ipynb**) is provided for you to write the codes based on these steps.

Step 1 – Data Loading and Processing

- Data Loading
    - Load the emoji_dictionary from Mapping.csv
    - Load the dataset (texts and labels) from dataset.csv
    - Check the maximum length of texts
- Data Processing
    - Convert the texts and labels into numeric tensors (X & y)
    - It may be helpful to cleanse the texts before convert them into numbers
- Data Sampling
    - Split the dataset (X & y) into training set (X_train & y_train) and testing set (X_test & y_test). Please refer to the Appendix and use the random_state assigned to you.

Step 2 – Develop a Sentiment Analysis Model using Training Data

- Each student is required to develop **at least TWO different base models**, which must consist of:
    - Word Embedding Layer: You can use the pre-trained word embeddings (e.g. GloVe), or simply train the word embeddings using the provided dataset.
    - RNN Layers (e.g. LSTM, GRU and etc.)
- During training phase, please split the training data into training samples and validation samples. (**Hint**: use `validation_split` in model fit() function)
- You are suggested to follow the universal machine learning workflow to develop the model and improve the model performance, i.e.

- Start with a baseline model
- Scale up the model until it overfits
- Regularize the model accordingly

- Analyze the Model Performance during training phase
  - **Remember to record all the model performance curves** (i.e. training and validation accuracy, training and validation loss scores) during training phase

Step 3 – Evaluate the Model using Testing Data

- Evaluate the Model Performance using X_test & y_test
- Compare and discuss the model performance
- Recommend the best model

Step 4 – Use the Model to make prediction

- Use the input() function to record the user input
- Convert the user input into numeric tensor
- Feed the numeric tensor into your model and see whether the model can output a correct emoticon

1.3 Report Format & Content Guidelines

Write a report with the following sections and content guidelines. You are free to include other relevant information you deem necessary in the sections.

*(Note: For a page with 1 inch margins, 12 point Arial font, and minimal spacing elements, a good rule of thumb is **500 words** for a single spaced page)*

	Suggested Report Sections & Content Guidelines	**Word Count**
1.	Overview • Describe the problem, the objective and the approach.	Min: 500 words Max: 1000 words
2.	Data Loading and Processing • Describe how you load data into Jupyter Notebook • Describe how you convert the data into numeric tensors • Describe how you sample the data and why you need to sample the data	Min: 1000 words Max: 2000 words
3.	Develop the Sentiment Analysis Models using Training Data • Describe how you build and train the models • Analyze the model performance during training phase	Min: 1000 words Max: 2000 words
4.	Model Evaluation using Testing Data • Analyze the model performance during testing phase • Compare and Discuss the models performance • Recommend the best model and explain why	Min: 1000 words Max: 2000 words
5.	Use the Best Model to Make Prediction • How to apply the model on a real life text input • Explain and Analyze the model prediction	Min: 500 words Max: 1000 words
6.	Summary • Summarize your model performance and provide suggestions for further improvement.	Min: 500 words Max: 1000 words

Problem 2 (50%)

2.1 Objective

Implement a Recurrent Neural Network (RNN) to create an English language character generator capable of building semi-coherent English sentences from scratch, by building them up character-by-character. In particular, we will be using a complete version of Sir Arthur Conan Doyle's classic book The Adventures of Sherlock Holmes to train the models.

We can train a deep learning model to generate text automatically, character-by-character, by showing the model many training examples so that it can learn a pattern between text inputs and potential character outputs. With this type of text generation, each input is a string of valid characters like this one:

"dogs are grea"

The corresponding output is the next character in the sentence - which is 't' (since the complete sentence is 'dogs are great'). We need to show a model many such examples in order for it to make reasonable predictions.

2.2 Dataset

Download the dataset from MeL:

- **holmes.txt**: the full text of The Adventures of Sherlock Holmes

2.3 Suggested Tasks

You are recommended to tackle this problem by following the steps below. A Jupyter Notebook (**Assignment_2_p2.ipynb**) is provided for you to write the codes based on these steps.

Step 1 – Data Loading and Processing

- Data Loading
  - Open and read the holmes.txt file
  - Check the total number of characters in the original text
- Data Processing
  - Perform basic data cleansing by removing unnecessary characters
  - Identify a list of unique characters and punctuations in the clean text
  - Prepare data into training text and labels (X & y) using the "sliding window" method
  - Perform one-hot encoding on the unique characters and punctuations

Step 2 – Develop a Sequence Generator Model using Training Data

- Each student is required to develop **at least TWO different base models**, which must consist of RNN Layers (e.g. LSTM, GRU and etc.)
- You are recommended to experiment with different "sliding window" sizes
- Follow the universal machine learning workflow to develop the model and improve the model performance
- **Remember to record all the model performance curves** after every round of fine-tuning

Step 3 – Use the Model to make prediction

- Use the input() function to record new input from user (or assign random starting indices from training data to extract input data to be used for prediction)
- Encode the input using one-hot encoding
- Feed the encoded input into your model and see whether the model can output a correct character prediction
- Compare and discuss the model performance
- Recommend the best model

2.4 Report Format & Content Guidelines

Write a report with the following sections and content guidelines. You are free to include other relevant information you deem necessary in the sections.

*(Note: For a page with 1 inch margins, 12 point Arial font, and minimal spacing elements, a good rule of thumb is **500 words** for a single spaced page)*

	Suggested Report Sections & Content Guidelines	**Word Count**
1.	Overview • Describe the problem, the objective and the approach.	Min: 500 words Max: 1000 words
2.	Data Loading and Processing • Describe how you load and cleanse the data in Jupyter Notebook • Describe how you perform one-hot encoding • Describe how you sample the data and why you need to sample the data	Min: 1000 words Max: 2000 words
3.	Develop the Character Generation Models using Training Data • Describe how you build and train the models • Analyze the model performance during training phase	Min: 1000 words Max: 2000 words
4.	Use the Models to Make Predictions • Compare and discuss the models' performance • Explain and analyze the model prediction • Recommend the best model and explain why	Min: 1000 words Max: 2000 words
5.	Summary • Summarize your model performance and provide suggestions for further improvement.	Min: 500 words Max: 1000 words

B. Presentation and Demonstration

You are required to submit a **video recorded presentation** to showcase and demo your work for both problems. The video recorded presentation **should not exceed 10 minutes**. Video recorded presentations which exceed the allotted time will be penalized.

You must record your video presentation using **Microsoft Teams**.

After completion of your video recorded presentation, you are required to **submit the link to your video** (from Microsoft Stream). Instructions to submit your video recorded presentation link are provided in the following section.

You are also required to **submit the presentation slides used in your video recorded presentation** in MeL.

C. Deliverables

For this assignment, you must submit all the following:

1. A set of **Final Presentation Slides** in MeL
   - This is the set of final presentation slides which you will use to conduct your video recorded presentation
   - Deadline for the slides submission is **Tuesday 11th Aug 2020, 2359 hours**

2. The **link to your video recorded presentation for both problems 1 and 2**
   - Submit the link to your video recorded presentation using the link below:
     - **Assignment 2 Video Presentation Submission Link**

       (Login using only your NP student account.

       Remember to grant view access to your tutor.)
   - Deadline for the link submission is **Tuesday 11th Aug 2020, 2359 hours**

3. A softcopy **Final Report for both problems 1 and 2** via **SafeAssign** in MeL
   - Deadline for report submission is **Sunday 23rd Aug 2020, 2359 hours**
4. The zipped file containing the **completed "assignment_2_p1.ipynb" and "assignment_2_p2.ipynb"** Jupyter Notebook Files in MeL
   - Deadline for Jupyter Notebook submission is **Sunday 23rd Aug 2020, 2359 hours**

D. Grading Criteria

	Grading Criteria	Component Weightage
Video Recorded Presentation	a) Flow of presentation based on content guidelines b) Quality of presentation slides c) Presentation and articulation skills d) Presentation kept within 5 min limit	**30%**
Final Report & Jupyter Notebooks	a) Completeness of report based on suggested report sections and content guidelines b) Quality of model building and evaluation c) Clarity of report and use of proper grammar d) Quality of recommendations for further improvements	**70%**

E. Appendix

Assignment of random_state number

Each student is assigned a **random_state** number to be used for random splitting of data into training and testing sets.

S/No	Class	Student Name	random_state
1	P01	AIDIL FARHAN B AMRAN	1
2	P01	ANG SI HAO	2
3	P01	CHIA KAI ZER	3
4	P01	DARIEN TAN WEI HAO	4
5	P01	DO LI FANG, SARAH	5
6	P01	ELIJAH NG DING JIE	6
7	P01	EWEN KECK JUN YUAN	7
8	P01	GERRON LEE YAN FONG	8
9	P01	GLADYS CHUA LING HUI	9
10	P01	GOH JUN JIE, NICHOLAS	10
11	P01	HANNAH LEONG JIA WEN	11
12	P01	JEWEL JACE LIM	12
13	P01	LIEW JING DE BENJAMIN	13
14	P01	LIM RAY'EN	14
15	P01	LIM YI JIE	15
16	P01	MUHAMMAD ANAS B ISMAIL	16
17	P01	NG TIANYU JERRIC	17
18	P01	NG WAI KEET	18
19	P01	PAE XIANG SHENG	19
20	P01	SOH LIU JING MABEL	20
21	P01	SOH QI HUI SELINA	21
22	P01	TAN WEN HAO	22
23	P01	TEO ZHI HAO	23
24	P01	TSEN FAN LOONG	24
25	P01	XIE ZHUOHAN	25
26	P02	ALYSSA CHWEE BEI ER	26
27	P02	BRADLEY GOH	27
28	P02	BRYAN LEE YIXIAN	28
29	P02	CHUA ZHE YU	29
30	P02	DAINEL KOH CHYE LEK	30
31	P02	DEBBIE HII WENXIN	31
32	P02	EZRA HO JINCHENG	32
33	P02	JASON CHUA YUAN ZHUANG	33
34	P02	KERVIN ONG GUAN CHENG	34
35	P02	LI ZIBIN	35
36	P02	LIM JUN HAO	36

37	P02	LIM KAI XIAN	37
38	P02	MATTHIAS WEN-ZHONG BRUNO-JEAN MOREL GAN	38
39	P02	NEO SAY PING	39
40	P02	NG CHIN TIONG RYAN	40
41	P02	NG RAY SON	41
42	P02	ONG CHEE KUAN	42
43	P02	RON JOSHUA ABES POLOYAPOY	43
44	P02	SEAH LE	44
45	P02	TAN JIA SHUN	45
46	P02	TAN SHAW HERNG LUCAS	46
47	P02	TAY QUAN YI	47
48	P02	TEO SHI JIE	48
49	P02	THADDEUS TEO E KAI	49
50	P02	ZURIEL SHANLEY TANYORY	50
51	P03	BENAIAH MARK MO DI	51
52	P03	CHUA ZONG HAN, LIONEL	52
53	P03	DARREN YEO YU XIONG	53
54	P03	GERALD TAN LIANG CHEE	54
55	P03	HENG GUAN XIANG	55
56	P03	HENG WEI YAO	56
57	P03	HO ZHEN XIAN	57
58	P03	KENNY TAN SENG TENG	58
59	P03	KIRTANARAM S/O HARIDASZ SHUNMUGAM	59
60	P03	LAU YI LIN	60
61	P03	LEE ZHI HONG, TIMOTHY	61
62	P03	LEONG JING FENG	62
63	P03	LEW JIAJUN	63
64	P03	LIANG SHI YIN, MARCUS	64
65	P03	LIM WEI XUAN, MARCUS	65
66	P03	ONG SI HUI	66
67	P03	P DHARSHANA NAIDU	67
68	P03	SOO QIN LOONG MARCUS	68
69	P03	SOPHIA CHONG JIA ROU	69
70	P03	TAN SHI HAO	70
71	P03	TAY JEUNG HONG	71
72	P03	VINCENT SEAH CHONG KENG	72
73	P03	ZHOU JIN CHENG	73
74	P04	AARON TEO YUAN CAI	74
75	P04	CHAN ZHI XIU	75
76	P04	CHUA WEI KANG	76
77	P04	DANIEL LEE JIA XIONG	77
78	P04	GINNA TAI YUN MIN	78
79	P04	KOH YAO HAO	79
80	P04	LEE KAI XIN	80
81	P04	LEE LI HAO	81

82	P04	LEE WEI KIAT	82
83	P04	LIM XUE ER	83
84	P04	LOW JIN YIK	84
85	P04	MANICKAVASAGAM SUSHMITHA	85
86	P04	NATHANIEL SEE WEI	86
87	P04	NG JUN HAO	87
88	P04	NG KAR WAI, ANDRE	88
89	P04	NICHOLAS CHENG DE FEI	89
90	P04	PHOEBE CHEONG QIAN MING	90
91	P04	SEAH PEI EN	91
92	P04	SIAH LI LING	92
93	P04	TAN BUN HARN JASON	93
94	P04	TAN KEE XIANG	94
95	P04	TAN ZHE KAI	95
96	P04	TOH SHAN FENG	96
97	P04	WONG SHEEN KERR	97
98	P04	ZHU JIAYUAN	98