

# HW4

---

## 1) Your model design and concept

Q1:

1. Chinese Preprocessing and Tokenizer 我使用的algorithm是character based tokenization，我將所有的中文都以label encoder的方式轉換成數字，每個中文字都會有對應的數字，然後把各個character合在一起存成一個list
2. Answer Prediction model會選擇機率最高的情緒作為輸出的答案

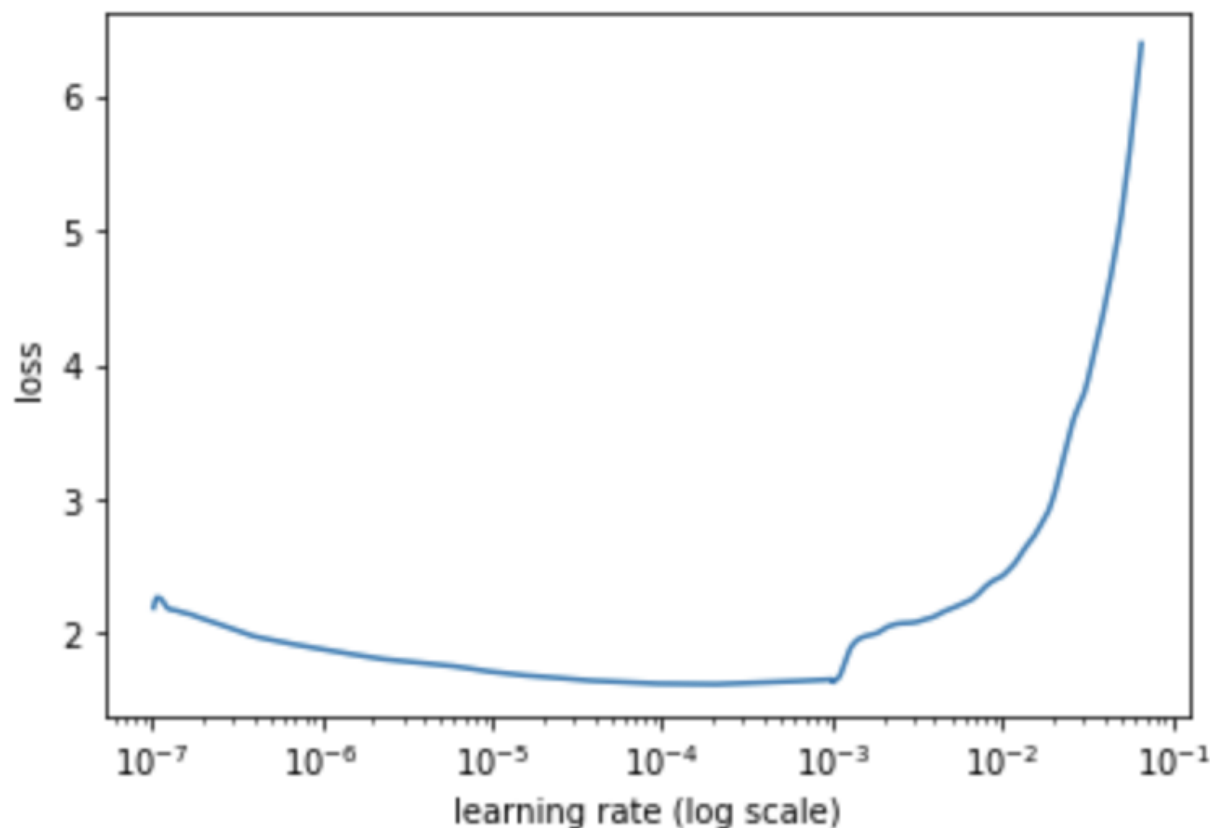
Q2:

1. Describe 我的model是以ktrain為工具來實現的BERT-like的pre-trained language model。batch-size大小為5，learning rate為 $2e-5$ ，兩次epoch，每次epoch都會得到一個反饋。training data分為10份，8份為training set，2份為validation set。performance如下：

evaluation(epoch1)	evaluation(epoch2)	test(Kaggle)
0.5046	0.6266	0.6040

2. plot

- learning curve of loss



## 2) Error Analysis and Experiment Discussion

error example:

- utterance: 正鵬，你還裝朦朧是嗎？你還是不是男子漢？
- real\_emotion: angry
- predict\_emotion: neutral

這一題是一個類似激問句，或是反諷結構的句子，所以句子當中並沒有太多直接表達情緒的字眼，比如說：“我一點也不稀罕！”、“快跟我滾開！”等，而是一種表面上平淡，實際上挖苦的方式，剛好我的model沒有辦法分辨這種反諷句構。

### 3) 請描述嘗試過的方法，並且討論曾經遇到的問題以及解決的方法

剛開始的data都是原本的句子而已，結果每次都距離baseline一點點，所以我採用教授的意見，增加了上下文的内容來預測，結果效果還是不太明顯，因此我想到可能是因為我沒有對主要句子和上下文關係做切割，導致訓練過程把上下文當作内容一起計算，所以我在句子和句子中間加入'///'的符號代表不同句子，結果效果蠻好的，很輕易就突破了baseline

### 4) 請根據實驗結果分析哪一類的情緒較容易預測?哪一類的情緒較難預測?較難預測的情緒容易被誤認為何種情緒?並探討可能的原因且提出解決辦法

我把training set中的資料做了以下統計

- real\_emotion: angry

angry	disgust	fear	happiness	neutral	sadness
264	11	30	219	53	442

- real\_emotion: disgust

angry	disgust	fear	happiness	neutral	sadness
1	766	33	315	52	303

- real\_emotion: fear

angry	disgust	fear	happiness	neutral	sadness
11	17	2779	707	29	205

- real\_emotion: happiness

angry	disgust	fear	happiness	neutral	sadness
22	148	454	5367	138	592

- real\_emotion: neutral

angry	disgust	fear	happiness	neutral	sadness
7	70	25	178	641	66

- real\_emotion: sadness

angry	disgust	fear	happiness	neutral	sadness
46	135	155	436	76	5207

predict angry and is angry:  $264/1019 = 0.26$  predict disgust and is disgust:  $766/1470 = 0.52$  predict fear and is fear:  $2779/3748 = 0.74$  predict happiness and is happiness:  $5367/6721 = 0.79$  predict neutral and is neutral:  $641/987 = 0.65$  predict sadness and is sadness:  $5207/6055 = 0.86$

sadness最容易預測，angry最難預測，最容易被預測為sadness，可能是因為在表達憤怒的同時也會包含悲傷的情緒，比如："媽，你們這些人的心腸好毒辣呀！她病成這樣了，你們還瞞着我，是誰叫你們這樣做的，我饒不了他！我若早點知道把她送到大醫院去治療，也不至於她今天會癱在牀上，哎喲，我的可憐妻子，我的心肝寶貝，我對不起你呀！我的心好疼的....." 這句話就同時包含對母親的憤怒和對妻子的不捨

我認為比較好的方式是在多分出一類情緒，label為sadness and angry at the same time












### 5) 請分析給定不同上下文的資訊量(句數)，對於預測該句話情緒的影響

s1 = 這麼簡單的事情，你為什麼非得那天回答我？你真是婆婆媽媽的。 s2 = 這麼簡單的事情，你為什麼非得那天回答我？你真是婆婆媽媽的。正鵬，我父母表示沒意見，我自己也吃下了定心丸。我願意嫁給你。

s1的預測結果是surprise s2的預測結果是neutral

我的理解是：單看第一句話，是一個驚訝的句子，但是如果加上第二句話，重心就變成在第二句，而第二句相對起來比較中立，所以會預測為中立。

- 最終結果

24	0716026		0.60397	16	2h
 <b>Your Best Entry!</b> Your most recent submission scored 0.60397, which is the same as your previous score. Keep trying!					
25	Uei-Dar Chen		0.60235	10	1d
26	0716018		0.60154	6	4d
27	0716307		0.60032	10	17h
28	0713347		0.59991	6	1d
29	0816162		0.59991	11	11h
30	0816166		0.59829	1	8d
31	Jia072		0.59627	9	2d
32	Joe Huang		0.59586	12	21h
	submission.csv		0.59505		