

checkpoint: https://drive.google.com/file/d/1Ai9EMmXrz_xhos2CRTXTcEaLRXf1ZvZT/view?usp=sharing

學號：r10922111 系級：資工所碩一 姓名：黃冠璋

1. (1%) 請以block diagram或是文字的方式說明這次表現最好的 model 使用哪些 layer module(如 Conv/RNN/Linear 和各類 normalization layer) 及連接方式(如一般forward 或是使用 skip/residual connection) , 並概念性逐項說明選用該 layer module 的理由。

2層Lstm → Linear → Dropout → Linear → Sigmoid

因為要讀懂一整句話的意思，所以要使用Lstm，參考之前的結果來訓練。

接著接linear、dropout 來防止overfit、再接linear，最後用sigmoid來輸出介於0~1的數值作為機率。

2. (1%) 請比較 word2vec embedding layer 初始設為 non-trainable/trainable 的差別，列上兩者在 validation/public private testing 的結果，並嘗試在訓練過程中設置一策略改變 non-trainable/trainable 設定，描述自己判斷改變設定的機制以及該結果。

non-trainable: 0.8054 / 0.811 / 0.8125

trainable: 0.8246 / 0.821 / 0.8255

策略為在使用validation data測試時，計算沒看過的字，若是太多則改成使用 trainable，結果為每次都會使用trainable，因此跟從一開使就使用trainable沒有任何區別

3. (1%) 請敘述你如何對文字資料進行前處理，並概念性的描述你在資料中觀察到什麼因此你決定採用這些處理，並描述使用這些處理時作細節，以及比較其實際結果，該結果可以不用具備真正改進。如果你沒有作任何處理，請給出一段具體描述來說服我們為什麼不做處理可以得到好的結果，這個理由不能是因為表現比較好。

訓練資料中有許多類似 I' m , can' t 簡寫的用法，我將中間的頓號移除並將兩個字連結再一起。最後將字串做stemming，去除字尾，保留有用的部分。

4. (1%) 請「自行設計」兩句具有相同單字但擺放位置不同的語句，使得你表現最好的模型產生出不同的預測結果，例如 "Today is hot, but I am happy" 與 "I am happy, but today is hot"，並討論造成差異的原因。

The restaurant look beauty , but the food is not good

The restaurant look not beauty , but the food is good

因為中間有一個轉換語氣的詞語：but，通常整句話中若有這種詞語，則這句話的意思就會跟著 but 後面連結的那半句。

5. (4%) Math problem:

<https://hackmd.io/@hAe95tLdTVqEePbZsJyqrw/BkWSTuqPF>

1. To maximize $P(d | \theta_1, \dots, \theta_N)$, is equivalent to maximize $\log P(d | \theta_1, \dots, \theta_N)$

Let $L(\theta_1, \dots, \theta_N) = \log(c(w_1) \dots c(w_N)) + \sum_{i=1}^N c(w_i) \log \theta_i + \lambda \left(\sum_{i=1}^N \theta_i - 1 \right)$

$\sum_{i=1}^N \theta_i = 1$

$$\frac{\partial L}{\partial \theta_i} = \frac{c(w_i)}{\theta_i} + \lambda = 0 \Rightarrow \theta_i = \frac{c(w_i)}{-\lambda}$$

$$\Rightarrow \sum_{i=1}^N \theta_i = 1 = \sum_{i=1}^N \frac{c(w_i)}{-\lambda}, \text{ so } \lambda = - \sum_{i=1}^N c(w_i)$$

\therefore by MLE, $\theta_i = \frac{c(w_i)}{\sum_{i=1}^N c(w_i)} = \frac{c(w_i)}{1.1}$

t	1	2	3	4	5	6	7	8
z	3	-2	4	0	2	-4	1	2
z ₁	90	90	190	90	90	-10	190	90
z ₂	10	10	-90	10	10	110	-90	10
z ₃	-10	90	90	90	-10	90	90	90

$$t=1: \left(3 \times \frac{1}{1+e^{90}} + 0 \right) \times \frac{1}{1+e^{10}} \doteq \frac{3}{1+e^{10}}, \quad c \rightarrow 3$$

$$t=2: \left(-2 \times \frac{1}{1+e^{-90}} + 3 \times \frac{1}{1+e^{10}} \right) \times \frac{1}{1+e^{-90}} \doteq 1, \quad c \rightarrow 1$$

$$t=3: \left(4 \times \frac{1}{1+e^{-190}} + 1 \times \frac{1}{1+e^{90}} \right) \times \frac{1}{1+e^{-90}} \doteq 4, \quad c \rightarrow 4$$

$$t=4: \left(0 + 4 \times \frac{1}{1+e^{10}} \right) \times \frac{1}{1+e^{-90}} \doteq 4, \quad c \rightarrow 4$$

$$t=5: \left(2 \times \frac{1}{1+e^{90}} + 4 \times \frac{1}{1+e^{10}} \right) \times \frac{1}{1+e^{10}} \doteq 0, \quad c \rightarrow 6$$

$$t=6: \left(-4 \times \frac{1}{1+e^{10}} + 6 \times \frac{1}{1+e^{-110}} \right) \times \frac{1}{1+e^{-90}} \doteq 6, \quad c \rightarrow 6$$

$$t=7: \left(1 \times \frac{1}{1+e^{-190}} + 6 \times \frac{1}{1+e^{90}} \right) \times \frac{1}{1+e^{-90}} \doteq 1, \quad c \rightarrow 1$$

$$t=8: \quad \left(2 - \frac{1}{1+e^{-90}} + 1 \cdot \frac{1}{1+e^{-10}}\right) \cdot \frac{1}{1+e^{-90}} = 3$$

$$y_t = (0, 1, 4, 4, 0, 6, 1, 3)$$

$$3. \quad h_1 = \tanh(w_1 x_1)$$

$$h_2 = \tanh(w_1 x_2 + w_h \tanh(w_1 x_1))$$

$$\hat{y} = \frac{1}{1+e^{-w_0 h_2}}$$

$$L(y, \hat{y}) = -y \ln \hat{y} - (1-y) \ln(1-\hat{y})$$

$$\begin{aligned} \frac{\partial L(y, \hat{y})}{\partial w_0} &= \frac{\partial L(y, \hat{y})}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_0} = \left[\frac{-y}{\hat{y}} - \frac{(1-y)}{-(1-\hat{y})} \right] \cdot -(1+e^{-w_0 h_2})^{-2} (-h_2) \\ &= \frac{-y + \hat{y}}{\hat{y}(1-\hat{y})} \cdot \frac{h_2}{(1+e^{-w_0 h_2})^2} \end{aligned}$$

$$\begin{aligned} \left(\text{let } \frac{1}{1+e^{-w_0 h_2}} = p \right) &= \frac{-y + \frac{1}{1+e^{-w_0 h_2}}}{\frac{1}{1+e^{-w_0 h_2}} \cdot \frac{e^{-w_0 h_2}}{1+e^{-w_0 h_2}}} \cdot \frac{h_2}{(1+e^{-w_0 h_2})^2} = \frac{-yp+1}{\frac{p}{p-1}} \cdot \frac{h_2}{p^2} \\ &= \frac{h_2(1-y(1+e^{-w_0 h_2}))}{(1+e^{-w_0 h_2}) \cdot e^{-w_0 h_2}} \end{aligned}$$

$$\frac{\partial L(y, \hat{y})}{\partial w_h} = \frac{\partial L(y, \hat{y})}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial h_2} \cdot \frac{\partial h_2}{\partial w_h} = \frac{(-yp+1)p}{p-1} \cdot w_0 \cdot \frac{1}{p^2} \cdot \text{sech}^2(w_1 x_2 + w_h \tanh(w_1 x_1)) \cdot \tanh(w_1 x_1)$$

$$= \frac{1-y(1+e^{-w_0 h_2})}{(1+e^{-w_0 h_2})(e^{-w_0 h_2})} \cdot w_0 \cdot \text{sech}^2(w_1 x_2 + w_h \tanh(w_1 x_1)) \cdot \tanh(w_1 x_1)$$

$$\frac{\partial L(y, \hat{y})}{\partial w_1} = \frac{\partial L(y, \hat{y})}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial h} \cdot \frac{\partial h}{\partial w_1} = \frac{(1-yP)}{P-1} \cdot w_0 \cdot \frac{1}{P^2} \cdot \text{sech}^2(w_1 x_2 + w_h \tanh(w_1 x_1)) \cdot (x_2 + w_h \text{sech}^2(w_1 x_1) \cdot x_1)$$

↑
 $P = 1 + e^{-w_0 h}$

4.