# Machine Learning - 2021Fall - HW4

系級：電機所碩一 學號：R10921098 姓名：楊仁傑

**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。**

ANS:

本次作業中使兩層雙向LSTM相接一起，證明說雖然字詞的順序在理解語言關係很重要，但是使用的順序並不那麼重要。使用雙向LSTM的方法可以取得不錯的準確率。

後續的Dense層是想說如果使用Dense映射神經單元是否可以取得較好的成果，在準確率上而言是有，但提升效果幾乎微乎其微。我認為應該還是要在處理文字上多下工夫才能提高準確率。

```
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, None, 50)          500000

bidirectional (Bidirectional (None, None, 256)         183296

dropout (Dropout)            (None, None, 256)         0

bidirectional_1 (Bidirection (None, 256)               394240

dense (Dense)                (None, 64)                16448

dropout_1 (Dropout)          (None, 64)                0

dense_1 (Dense)              (None, 32)                2080

dropout_2 (Dropout)          (None, 32)                0

dense_2 (Dense)              (None, 1)                 33
=================================================================
Total params: 1,096,097
Trainable params: 1,096,097
Non-trainable params: 0
```

**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。**

ANS:

**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。**

ANS:

在部分單字中會有n ' t的設定，我認為如果拆成單字 not的話可能會有助於提高準確率。
在模型相同的情況下，將n ' t拆解為not後準確率由0.80550提高到0.81530左右。

**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。**

ANS:

```
RNN:

Today is hot, but I am happy. [0.08012492]

I am happy, but today is hot. [0.29768157]
```

RNN model給 I am happy, but today is hot.的分數較高(不過還是0.5以下所以不算惡意評論)。可見RNN有學到語意的部份，進而判斷說第一句是較為開心的。即RNN 則可以判斷字詞的不同順序會產生不同的語意，並分辨出這是兩個不同的句子，讓兩者輸出的值有些微的不同。

## 5. (4%) Refer to math problem

[(https://hackmd.io/@hAe95tLdTVqEePbZsJyqrw/BkWSTuqPF)](https://hackmd.io/@hAe95tLdTVqEePbZsJyqrw/BkWSTuqPF)

**1. Uni-gram language model and Maximum Likelihood Estimation (1%)**

ANS:

**2. LSTM Cell (1%)**

when $t = 1$,

$$
\begin{aligned}
z &= w \cdot x + b = [0,0,0,1] \cdot [0,1,0,3] + 0 = 3 + 0 = 3 \\
z_i &= w_i \cdot x + b_i = [100,100,0,0] \cdot [0,1,0,3] + (-10) = 100 + (-10) = 90 \\
z_f &= w_f \cdot x + b_f = [-100,-100,0,0] \cdot [0,1,0,3] + (110) = (-100) + 110 = 10 \\
z_o &= w_o \cdot x + b_o = [0,0,100,0] \cdot [0,1,0,3] + (-10) = 0 + (-10) = -10 \\
c' &= f(z_i)g(z) + cf(z_f) = \frac{1}{1+e^{-90}} \cdot 3 + 0 \cdot \frac{1}{1+e^{-10}} = 3 \\
y &= f(z_o)h(c') = \frac{1}{1+e^{-(-10)}} \cdot 3 = 0
\end{aligned}
$$

when $t = 2$,

$$
\begin{aligned}
z &= w \cdot x + b = [0,0,0,1] \cdot [1,0,1,-2] + 0 = -2 \\
z_i &= w_i \cdot x + b_i = [100,100,0,0] \cdot [1,0,1,-2] + (-10) = 90 \\
z_f &= w_f \cdot x + b_f = [-100,-100,0,0] \cdot [1,0,1,-2] + (110) = 10 \\
z_o &= w_o \cdot x + b_o = [0,0,100,0] \cdot [1,0,1,-2] + (-10) = 90 \\
c' &= f(z_i)g(z) + cf(z_f) = \frac{1}{1+e^{-90}} \cdot 3 + 3 \cdot \frac{1}{1+e^{-10}} = 1 \\
y &= f(z_o)h(c') = \frac{1}{1+e^{-90}} \cdot 1 = 1
\end{aligned}
$$

when $t = 3$,

$$
\begin{aligned}
z &= w \cdot x + b = [0,0,0,1] \cdot [1,1,1,4] + 0 = 4 \\
z_i &= w_i \cdot x + b_i = [100,100,0,0] \cdot [1,1,1,4] + (-10) = 190 \\
z_f &= w_f \cdot x + b_f = [-100,-100,0,0] \cdot [1,1,1,4] + (110) = -90 \\
z_o &= w_o \cdot x + b_o = [0,0,100,0] \cdot [1,1,1,4] + (-10) = 90 \\
c' &= f(z_i)g(z) + cf(z_f) = \frac{1}{1+e^{-190}} \cdot 3 + (-2) \cdot \frac{1}{1+e^{-(-90)}} = 4 \\
y &= f(z_o)h(c') = \frac{1}{1+e^{-90}} \cdot 4 = 4
\end{aligned}
$$

when $t = 4$,

$$
\begin{aligned}
z &= w \cdot x + b = [0,0,0,1] \cdot [0,1,1,0] + 0 = 0 \\
z_i &= w_i \cdot x + b_i = [100,100,0,0] \cdot [0,1,1,0] + (-10) = 90 \\
z_f &= w_f \cdot x + b_f = [-100,-100,0,0] \cdot [0,1,1,0] + (110) = 10 \\
z_o &= w_o \cdot x + b_o = [0,0,100,0] \cdot [0,1,1,0] + (-10) = 90 \\
c' &= f(z_i)g(z) + cf(z_f) = \frac{1}{1+e^{-90}} \cdot 0 + 4 \cdot \frac{1}{1+e^{-10}} = 4 \\
y &= f(z_o)h(c') = \frac{1}{1+e^{-90}} \cdot 4 = 4
\end{aligned}
$$

when $t = 5$,

$$
\begin{aligned}
z &= w \cdot x + b = [0,0,0,1] \cdot [0,1,0,2] + 0 = 2 \\
z_i &= w_i \cdot x + b_i = [100,100,0,0] \cdot [0,1,0,2] + (-10) = 90 \\
z_f &= w_f \cdot x + b_f = [-100,-100,0,0] \cdot [0,1,0,2] + (110) = 10 \\
z_o &= w_o \cdot x + b_o = [0,0,100,0] \cdot [0,1,0,2] + (-10) = 90 \\
c' &= f(z_i)g(z) + cf(z_f) = \frac{1}{1+e^{-90}} \cdot 2 + 4 \cdot \frac{1}{1+e^{-10}} = 6 \\
y &= f(z_o)h(c') = \frac{1}{1+e^{-90}} \cdot 6 = 0
\end{aligned}
$$

when $t = 6$,

$$
\begin{aligned}
z &= w \cdot x + b = [0,0,0,1] \cdot [0,0,1,-4] + 0 = -4 \\
z_i &= w_i \cdot x + b_i = [100,100,0,0] \cdot [0,0,1,-4] + (-10) = (-10) \\
z_f &= w_f \cdot x + b_f = [-100,-100,0,0] \cdot [0,0,1,-4] + (110) = 110 \\
z_o &= w_o \cdot x + b_o = [0,0,100,0] \cdot [0,0,1,-4] + (-10) = 90 \\
c' &= f(z_i)g(z) + cf(z_f) = \frac{1}{1+e^{-(-10)}} \cdot -4 + 6 \cdot \frac{1}{1+e^{-110}} = 1 \\
y &= f(z_o)h(c') = \frac{1}{1+e^{-(-10)}} \cdot 6 = 6
\end{aligned}
$$

when $t = 7$,

$$
\begin{aligned}
z &= w \cdot x + b = [0,0,0,1] \cdot [1,1,1,1] + 0 = 1 \\
z_i &= w_i \cdot x + b_i = [100,100,0,0] \cdot [1,1,1,1] + (-10) = 190 \\
z_f &= w_f \cdot x + b_f = [-100,-100,0,0] \cdot [1,1,1,1] + (110) = -90 \\
z_o &= w_o \cdot x + b_o = [0,0,100,0] \cdot [1,1,1,1] + (-10) = 90 \\
c' &= f(z_i)g(z) + cf(z_f) = \frac{1}{1+e^{-190}} \cdot 1 + 6 \cdot \frac{1}{1+e^{-(-90)}} = 1 \\
y &= f(z_o)h(c') = \frac{1}{1+e^{-90}} \cdot 1 = 1
\end{aligned}
$$

when $t = 8$,

$$
\begin{aligned}
z &= w \cdot x + b = [0,0,0,1] \cdot [1,0,1,2] + 0 = 2 \\
z_i &= w_i \cdot x + b_i = [100,100,0,0] \cdot [1,0,1,2] + (-10) = 90 \\
z_f &= w_f \cdot x + b_f = [-100,-100,0,0] \cdot [1,0,1,2] + (110) = 10 \\
z_o &= w_o \cdot x + b_o = [0,0,100,0] \cdot [1,0,1,2] + (-10) = 90 \\
c' &= f(z_i)g(z) + cf(z_f) = \frac{1}{1+e^{-90}} \cdot 2 + 1 \cdot \frac{1}{1+e^{-10}} = 1 \\
y &= f(z_o)h(c') = \frac{1}{1+e^{-90}} \cdot 3 = 3
\end{aligned}
$$

$\therefore$ Output $= 3$

**3. Backpropagation through time via Simple RNN (1%)**

ANS:

**4. Multiclass AdaBoost (1%)**

ANS:

$$
\begin{aligned}
L(g_t^1 \cdots g_t^k) &= \sum_{i=1}^{n} exp(\frac{1}{k} \sum g_t^k(x_i) - g_t^{\hat{y}_i}(x_i)) \\
&= \sum_{i=1}^{n} exp(\frac{1}{k-1} \sum_{k \neq \hat{y}_i} g_{t-1}^k(x_i) + \frac{a_t^k f_t(x)}{k-1} - g_t^{\hat{y}_i}(x_i)) + \sum_{i=1}^{n} exp(\frac{1}{k-1} \sum_{k \neq \hat{y}_i} g_{t-1}^k(x_i) - a_t^k f_t(x) - g_t^{\hat{y}_i}(x_i))
\end{aligned}
$$

$$
\begin{aligned}
\because \Delta L &= L(g_T^k) - L(g_{T-1}^k) \\
&= \sum_{i=1}^{n} exp(\frac{1}{k} \sum g_t^k(x_i) - g_t^{\hat{y}_i}(x_i))(exp(\frac{a_t^k f_t(x)}{k-1}) - 1) + \sum_{i=1}^{n} exp(\frac{1}{k} \sum g_t^k(x_i) - g_t^{\hat{y}_i}(x_i))(exp(a_t^k f_t(x)) - 1) \\
&= \sum_{i=1}^{n} exp(\frac{1}{k} \sum g_t^k(x_i) - g_t^{\hat{y}_i}(x_i))(exp(\frac{a_t^k f_t(x)}{k-1}) - exp(a_t^k f_t(x)))
\end{aligned}
$$

$$
\therefore \frac{\partial \Delta L}{\partial a_t} = \frac{f_t(x)}{k-1} \sum exp(\frac{1}{k} \sum g_t^k(x_i) - g_t^{\hat{y}_i}(x_i)) - f_t(x_i) \sum exp(\frac{1}{k-1} \sum g_{t-1}^k(x_i) - g_t^{\hat{y}_i}(x_i))
$$

Let $\frac{\partial \Delta L}{\partial a_t} = 0$

$$\frac{\partial \Delta L}{\partial a_t} = \sum_{k=\hat{y}_i} u_t^n e^{a_t} \frac{1}{k-1} - \sum_{k=\hat{y}_i} u_t^n e^{a_t} = 0$$

$$\Rightarrow \ z_t \varepsilon_t e^{a_t} \frac{1}{k-1} - z_t (1 - g_t) e^{-a_t} = 0$$

$$\Rightarrow \ e^{at} = (k-1)(\frac{1-\varepsilon_t}{\varepsilon_t})$$

$$\Rightarrow \ a_t = ln(\sqrt{(k-1)(\frac{1-\varepsilon_t}{\varepsilon_t})})$$