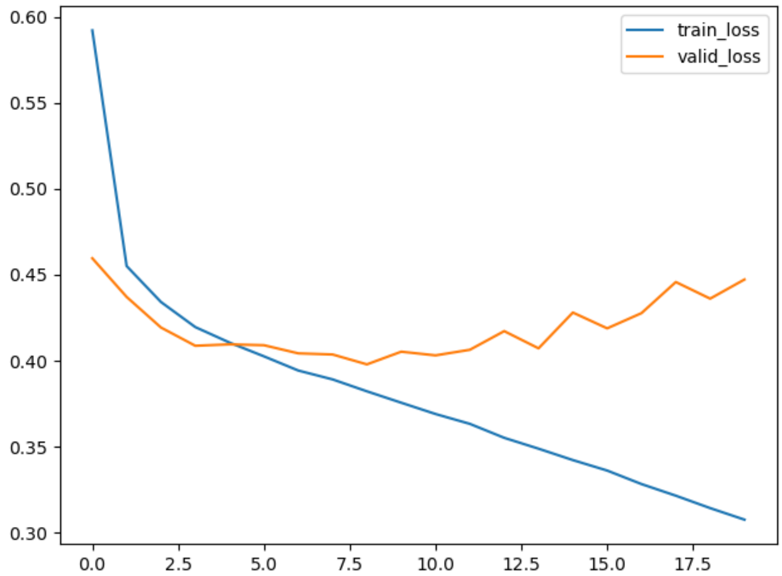
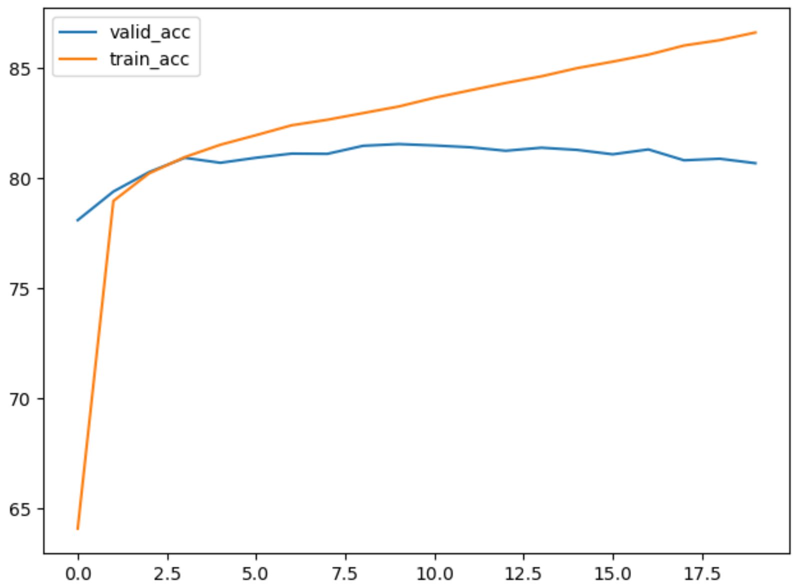
學號：B07902143 系級：資訊三 姓名：陳正康

1. (0.5%) 請說明你實作之 RNN 模型架構及使用的 word embedding 方法，回報模型的正確率並繪出訓練曲線*

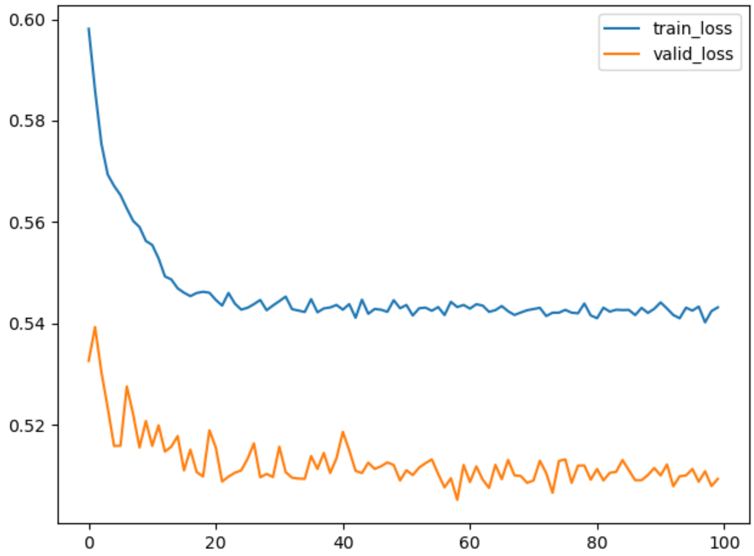


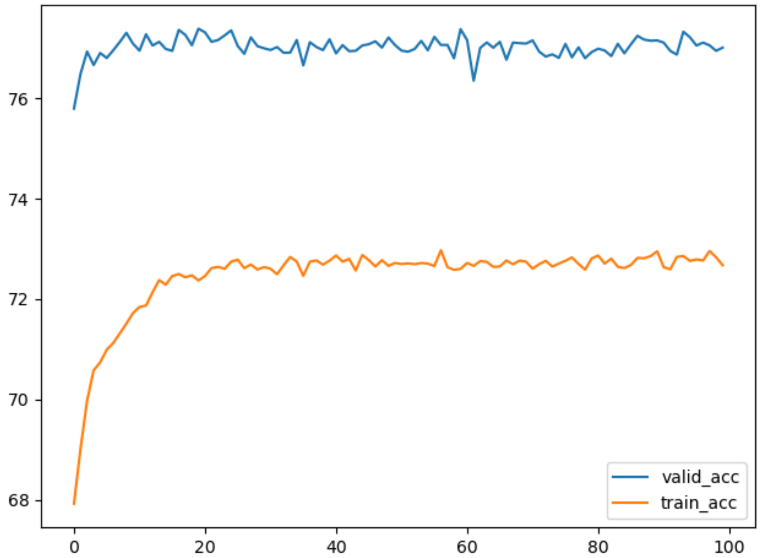


模型：1層Word2vec Embedding (dim 30) → 4層LSTM → Dropout (0.75) → 1層線性 (dim 64) → Sigmoid

正確率：Train 0.83266, Validation 0.81559, Kaggle 0.82560

2. (0.5%) 請實作 BOW+DNN 模型，敘述你的模型架構，回報模型的正確率並繪出訓練曲線*。





模型：4層線性 (dim 128)，每層間間隔dropout 0.5，再接sigmoid

正確率：Train 0.72369, Validation 0.77386, Kaggle 0.77180

這裡發現一個現象，validation的結果比較training還好。

3. (0.5%) **請敘述你如何 improve performance（preprocess, embedding, 架構等），並解釋為何這些做法可以使模型進步。**

i. **預處理**

英文有大量的縮寫字，在預處理時先展開這些字，如[can't] → [can, not]，而不是原本的[can, ', t]，因為撇號通常不帶有語意，把它當symbol不太合理。

ii. **改變word2vec的參數**

window size調高到10，dimension調高到300，這樣word可以看見更多的資訊，embedding有機會更準確。

另外，使用unlabeled data一起作word embedding，讓字之間距離的資訊更多。

iii. **改變rnn的參數與架講**

lstm用了4層，每層dimension是64，並加上dropout 0.75和L2 regularization (係數1e-7)，以減少overfitting。效果最明顯的是L2 regularization。
句子長度取45，可以涵蓋data中大部分的句子。

iv. **使用gru取代lstm**

因為gru的參數較少，比較不會overfitting。但在這個dataset似乎沒有明顯進步。

v. **使用Bidirectional RNN**

這樣可以讓training過程中看見之前和之後的word。

4. **(0.5%) 請比較 RNN 與 BOW 兩種不同 model 對於 "Today is hot, but I am happy" 與 "I am happy, but today is hot" 這兩句話的分數（model output），並討論造成差異的原因。**

	RNN	BOW
Today is hot, but I am happy:	0.8716	0.1350
I am happy, but today is hot:	0.1251	0.1350

這兩句話出現的字完全一樣，只有順序不一樣。bow只考慮出現的字，而rnn會參考往前數個字(以lstm為例)，因此rnn可以分辨因語序不同造成的語意差異。

5. **(3%) Math problem:**
https://drive.google.com/file/d/1fEu87banB4s6Yjku1dA5sMcnwCugEPBF/view?usp=sharing

ML HW4

1. Let $X^{raw} = \begin{pmatrix} 1 & 4 & 3 & 1 & 5 & 7 & 9 & 3 & 11 & 10 \\ 2 & 8 & 12 & 8 & 14 & 4 & 8 & 8 & 5 & 11 \\ 3 & 5 & 9 & 5 & 2 & 1 & 9 & 1 & 6 & 7 \end{pmatrix} = \begin{pmatrix} x_1^{raw} & \cdots & x_{10}^{raw} \end{pmatrix}$

Mean $\mu := \begin{pmatrix} 5.4 \\ 8 \\ 4.8 \end{pmatrix} = \frac{1}{10} \sum_{i=1}^{10} x_i^{raw}$

$X := X^{raw} - \mu = \begin{pmatrix} x_1^{raw} - \mu & \cdots & x_{10}^{raw} - \mu \end{pmatrix} = \begin{pmatrix} x_1 & \cdots & x_{10} \end{pmatrix} = \begin{pmatrix} -4.4 & \cdots \\ -6 & \\ -1.8 & \end{pmatrix}$

Do singular-value decomposition:

$$X = UDV^T$$

$$\Sigma = \frac{1}{10} \sum_{i=1}^{10} x_i x_i^T = \frac{1}{10} XX^T = \frac{1}{10} UDV^T VDU^T = U\left(\frac{1}{10} DD^T\right)U^T$$

Where $U = \begin{pmatrix} u_1 & u_2 & u_3 \end{pmatrix} = \begin{pmatrix} -0.62 & 0.68 & -0.40 \\ -0.59 & -0.73 & -0.34 \\ -0.52 & 0.03 & 0.85 \end{pmatrix}$

$D = \begin{pmatrix} 12.37 & 0 & 0 & 0 & & 0 \\ 0 & 10.78 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 7.40 & 0 & & 0 \end{pmatrix}_{3\times 10}$

(a) Principle Axes are : $u_1 = \begin{pmatrix} -0.62 \\ -0.59 \\ -0.52 \end{pmatrix}$, $u_2 = \begin{pmatrix} 0.68 \\ -0.73 \\ 0.03 \end{pmatrix}$, $u_3 = \begin{pmatrix} -0.40 \\ -0.34 \\ 0.85 \end{pmatrix}$
(Approximately)

(b) $U^T X^{raw} = \begin{pmatrix} \boxed{\begin{matrix} -3.36 \\ -0.71 \\ 1.48 \end{matrix}} & \boxed{\begin{matrix} -9.79 \\ -3.03 \\ -0.04 \end{matrix}} & \boxed{\begin{matrix} -13.62 \\ -6.53 \\ 2.42 \end{matrix}} & \begin{matrix} -7.94 \\ -5.06 \\ 1.16 \end{matrix} & \begin{matrix} -12.37 \\ -6.84 \\ -5.02 \end{matrix} & \begin{matrix} -7.19 \\ 1.84 \\ -3.30 \end{matrix} & \begin{matrix} -14.96 \\ 0.47 \\ 1.37 \end{matrix} & \begin{matrix} -7.08 \\ -3.81 \\ -3.05 \end{matrix} & \begin{matrix} -12.86 \\ 3.95 \\ -0.97 \end{matrix} & \boxed{\begin{matrix} -16.3 \\ -1.11 \\ -1.75 \end{matrix}} \end{pmatrix}$

pc of x_1^{raw} ... pc of x_{10}^{raw}

(c) Use u_1 and u_2 since they have largest eigenvalue

$\Rightarrow$ Principle Component (pc) $= \begin{pmatrix} u_1^T \\ u_2^T \end{pmatrix} X^{raw}$

Reconstruction Error $= \left\| \begin{pmatrix} u_1 & u_2 \end{pmatrix} \begin{pmatrix} u_1^T \\ u_2^T \end{pmatrix} X^{raw} - X^{raw} \right\|^2 = 60.64$

ML HW4

2. (a) $A \in R^{m \times n}$

Symmetric: $(AA^T)^T = AA^T$

$$(A^TA)^T = A^TA$$

Positive Semi-definite:

$\forall x \quad x^T AA^T x = \|A^T x\|^2 \geq 0$

$\forall y \quad y^T A^T A y = \|Ay\|^2 \geq 0$

Sharing same eigenvalues:

For matrix A^TA, since it's positive semidefinite, its eigenvalues are non-negative. Let positive eigenvalues be $\lambda_1, \lambda_2 \cdots \lambda_k$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k > 0$, $k \leq n$, and corresponding eigenvectors be $v_1 \cdots v_k$

So $A^TAv_i = \lambda_i v_i$, $i \leq k$

Define $\sigma_i = \sqrt{\lambda_i}$, $u_i = \frac{1}{\sigma_i} A v_i$, for $i \leq k$

$\Rightarrow A^T u_i = A^T \frac{1}{\sigma_i} A v_i = \frac{1}{\sigma_i} A^TAv_i = \frac{1}{\sigma_i} \sigma_i^2 v_i = \sigma_i v_i$

$\Rightarrow AA^T u_i = \sigma_i A v_i = \sigma_i^2 u_i$

$\Rightarrow \sigma_i^2 = \lambda_i$ is also an eigenvalue of AA^T

$\Rightarrow$ Let $B = A^T$, we can show the inverse direction (positive eigenvalues of AA^T are eigenvalues of A^TA)

$\Rightarrow A^TA$ and AA^T share same non-zero eigenvalues

2. (b) $\because \Sigma$ is symmetric and positive definite

$\Rightarrow \Sigma$ can be diagonalize as:

$\Sigma = U \Lambda U^T$ where $U \in R^{m\times m}$ is orthonormal

$\Lambda \in R^{m\times m}$ is diagonal

$\Lambda = diag\left(\frac{\sigma_1^2}{n}, \frac{\sigma_2^2}{n} \cdots, \frac{\sigma_m^2}{n}\right)$ for some $\begin{cases} \sigma_1 \cdots \sigma_m, \sigma_i \geq \\ n, \; n > m \end{cases}$

$\Rightarrow \Sigma = U\left(\frac{1}{n}DD^T\right)U^T$ where $D \in R^{m\times n}$, $D = \begin{pmatrix} \sigma_1 & & 0 & 0 & 0 \\ & \sigma_2 & & \vdots & \cdots \vdots \\ 0 & & \sigma_m & 0 & 0 \end{pmatrix}$

$= \frac{1}{n} UD\underbrace{V^TV}_{=I_n}D^TU^T$ for some orthonormal $V \in R^{n\times n}$

$= \frac{1}{n}(UDV^T)(UDV^T)^T$

$= \frac{1}{n} XX^T$] Let $X = UDV^T = (\hat{x}_1 \cdots \hat{x}_n)$

$= \frac{1}{n}\sum_{i=1}^{n} \hat{x}_i \hat{x}_i^T$

$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)(x_i-\mu)^T$] Let $x_i = \hat{x}_i + \mu$

There always exists V such that $\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i = 0$ since:

$\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i = 0 \Leftrightarrow X\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix} = 0 \Leftrightarrow UDV^T\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix} = 0$

$\Leftrightarrow V^T\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix} \in Null(UD)$

Since $dim(Null(UD)) = Nullity(UD) = n - \underbrace{Rank(UD)}_{\leq m} \geq n - m > 0$ ($UD \in R^{m\times n}$)

$\Rightarrow \exists y = V^T\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}$

So $\{x_1, x_2 \cdots x_n\}$ satisfies $\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i + \mu) = \mu$

$\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T = \Sigma$

2.(c) $1 \le k \le m$

$$\begin{cases} \text{min: } Tr(\phi^T \Sigma \phi) \\ \text{subject to } \phi^T\phi = I_k \\ \qquad\qquad \phi \in R^{m \times k} \end{cases}$$

$\text{Trace}(\phi^T \Sigma \phi)$

$= \frac{1}{n}\text{Trace}(\phi^T X X^T \phi)$ ← By 2.(b), Σ is symmetric and positive semi-definite, so $\Sigma = \frac{1}{n} XX^T$ $\exists X \in R^{m\times n}$ $\exists n$
$X = (x_1\ x_2 \cdots x_n)$

$= \frac{1}{n} \| \phi^T X \|_F^2$

$= \frac{1}{n} \sum_{i=1}^{n} \| \phi^T x_i \|^2$

$= \frac{1}{n} \sum_{i=1}^{n} \| \hat{x}_i^{(S)} \|^2$ ← By $\phi^T\phi = I_k$, ϕ^T is a projection matrix onto a k-dimensional space of R^m (let's say, S)
$\hat{x}_i^{(S)}$ denotes the projection of x_i onto S

Let $\hat{X}^{(S)} = (\hat{x}_1^{(S)} \cdots \hat{x}_n^{(S)})$

$\hat{X}^{(P)} = (\hat{x}_1^{(P)} \cdots \hat{x}_n^{(P)}) = \sum_{i=1}^{k} \sigma_i u_i v_i^T$ where $\sigma_1 \le \sigma_2 \le \cdots \le \sigma_m$ are singular values of X
$u_1 \cdots u_m$ are left singular vectors
$v_1 \cdots v_m$ are right singular vectors

By Eskart-Young-Mirsky thm.

$$\| X - \hat{X}^{(P)} \|_F \ge \| X - \hat{X}^{(S)} \|_F$$

$$\Rightarrow \sum_{i=1}^{n} \| x_i - \hat{x}_i^{(P)} \|^2 \ge \sum_{i=1}^{n} \| x_i - \hat{x}_i^{(S)} \|^2$$

$$\Rightarrow \sum_{i=1}^{n} \| \hat{x}_i^{(P)} \|^2 \le \sum_{i=1}^{n} \| \hat{x}_i^{(S)} \|^2 = \text{Trace}(\phi^T \Sigma \phi)$$

So

$$\text{Trace}(\phi^T \Sigma \phi) \ge \sum_{i=1}^{n} \| \hat{x}_i^{(P)} \|^2 = \sum_{i=1}^{n} \| \sum_{i=1}^{k} \sigma_i u_i v_i^T \|^2$$

$$= \| \hat{X}^{(P)} \|_F^2 = \| \sum_{i=1}^{k} \sigma_i u_i v_i^T \|_F^2 = \sum_{i=1}^{k} \sigma_i \| u_i v_i^T \|_F^2$$

3. Define : True label $\hat{y}_i = \begin{pmatrix} -\frac{1}{K-1} \\ \vdots \\ 1 \\ \vdots \\ -\frac{1}{K-1} \end{pmatrix}$ where $\hat{y}_i^{(p)} = \begin{cases} 1 & \text{if this data is class } p \\ -\frac{1}{K-1} & \text{otherwise} \end{cases}$

Function $g_t(x) = \begin{pmatrix} g_t^1(x) \\ \vdots \\ g_t^K(x) \end{pmatrix}$, $\alpha_t = \begin{pmatrix} \alpha_t^1 \\ \vdots \\ \alpha_t^K \end{pmatrix}$

The loss function becomes : $L(g_t) = \sum_{i=1}^{n} \exp\left\{ -\hat{y}_i^T g_t(x_i) \right\}$

$$= \sum_{i=1}^{n} \exp\left\{ -\hat{y}_i^T \left(g_{t-1}(x_i) + f_t(x_i)\alpha_t \right) \right\} = \sum_{i=1}^{n} w_i \exp\left\{ -f_t(x_i)\, \hat{y}_i^T \alpha_t \right\}$$

where $w_i = \exp\left\{ -\hat{y}_i^T g_{t-1}(x_i) \right\}$

$$\Rightarrow \nabla_{\alpha_t} L = \sum_{i=1}^{n} w_i \exp\left\{ -f_t(x_i)\, \hat{y}_i^T \alpha_t \right\} \left(-f_t(x_i)\, \hat{y}_i \right) = 0$$

$$\Rightarrow \alpha_t = \frac{(K-1)^2}{K} \left[\log\left(\frac{1-\varepsilon_t}{\varepsilon_t} \right) + \log(K-1) \right]$$

where $\varepsilon_t = \sum_{\hat{y}_i \neq f_t(x_i)} w_i \Big/ \sum_{i=1}^{n} w_i$