Machine Learning - Homework 5

資工四 B05902023 李澤諺

December 13, 2019

Part 1. Programming Problem

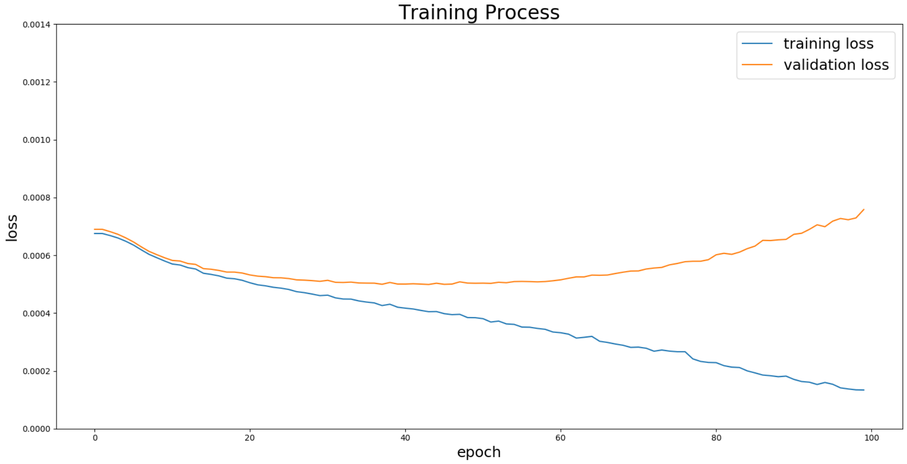
1. (1%) 請說明你實作之 RNN 模型架構及使用的 word embedding 方法，回報模型的正確率並繪出訓練曲線。

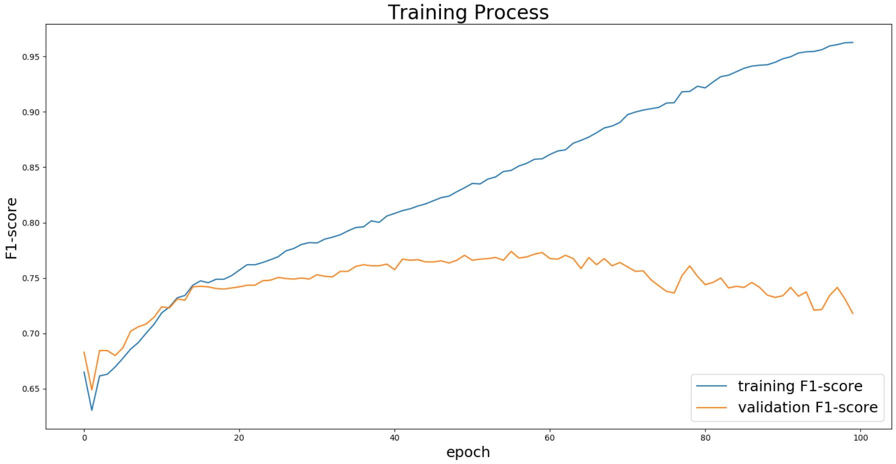
在 word embedding 中，首先，我使用 spaCy 的 en_core_web_lg 將 training data 和 testing data 中的所有句子進行斷詞，接著將標點符號、stop word、emoji，以及非數字或非字母等等的詞去掉，再將所有的詞轉爲 basic form，並將所有句子進行 trim 或 pad，以此進行 preprocessing，接著，將所有句子輸入 gensim 的 Word2Vec model 進行訓練，而在 Word2Vec model 的訓練中，word vector 爲 256 維，使用 skip gram 訓練，並將出現次數小於 5 的詞視爲 UNK，以此訓練了 200 個 iteration 得到 Word2Vec model。

接著，以下爲我於本次作業中所實作的 RNN 架構：

embedding	Embedding(embedding.size(0) , embedding.size(1) , padding_idx = padding_index)
recurrent	LSTM(embedding.size(1) , 128 , batch_first = True , bias = True , num_layers = 2 , dropout = 0.3 , bidirectional = True)
linear	Linear(768 , 100 , bias = True)
	BatchNorm1d(100)
	ReLU()
	Dropout()
	Linear(100 , 100 , bias = True)
	BatchNorm1d(100)
	ReLU()
	Dropout()
	Linear(100 , 100 , bias = True)
	BatchNorm1d(100)
	ReLU()
	Dropout()
	Linear(100 , 2 , bias = True)

我隨機選出 2000 個句子作爲 validation data，剩下的句子則作爲 training data，並且，我使用了 Adam 訓練 RNN，其中 learning rate 爲 0.0001，batch size 爲 1024，以此訓練了 100 個 epoch，所得到的 loss 和 F1-score 如下圖所示：





以下爲我訓練出來的 RNN 所得到的 F1-score：

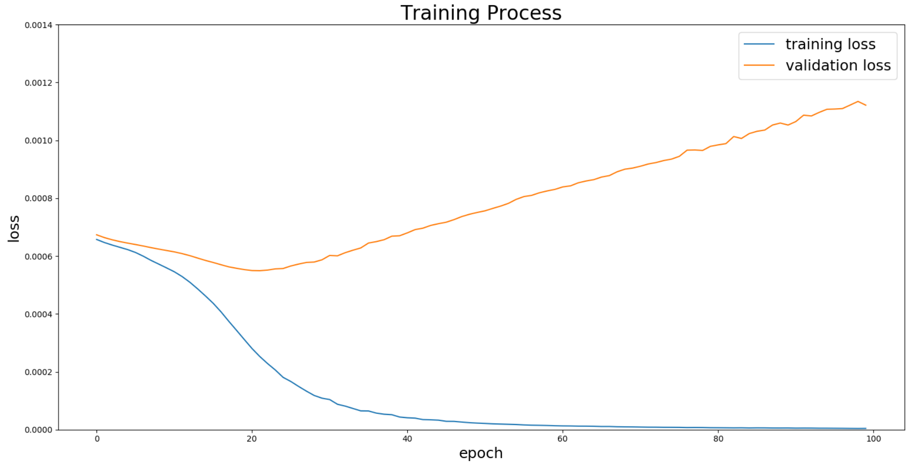
Train	Validation	Test	
		Public	Private
0.82260	0.75400	0.78372	0.81627

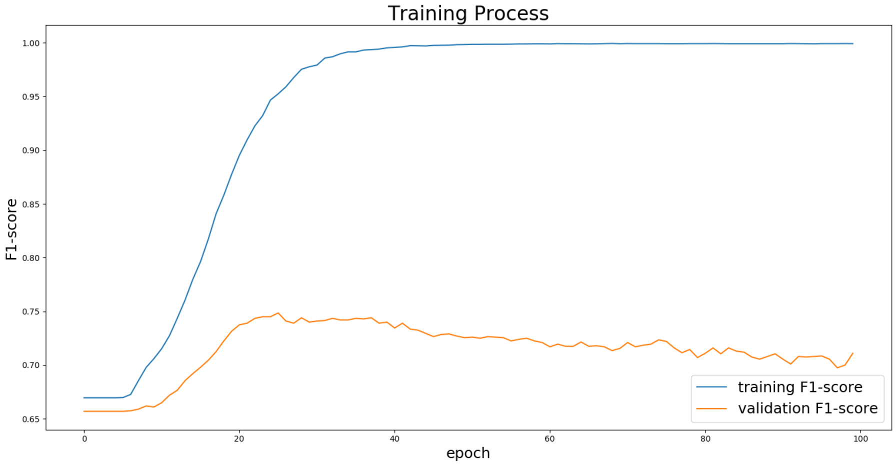
2. (1%) 請實作 BOW+DNN 模型，敘述你的模型架構，回報模型的正確率並繪出訓練曲線。

以下爲我於本次作業中所實作的 DNN 架構：

linear	Linear(dimension , 100 , bias = True)
	BatchNorm1d(100)
	ReLU()
	Dropout()
	Linear(100 , 100 , bias = True)
	BatchNorm1d(100)
	ReLU()
	Dropout()
	Linear(100 , 100 , bias = True)
	BatchNorm1d(100)
	ReLU()
	Dropout()
	Linear(100 , 100 , bias = True)
	BatchNorm1d(100)
	ReLU()
	Dropout()
	Linear(100 , 2 , bias = True)

我隨機選出 2000 個句子作爲 validation data，剩下的句子則作爲 training data，並且，我使用了 Adam 訓練 DNN，其中 learning rate 爲 0.0001，batch size 爲 1024，以此訓練了 100 個 epoch，所得到的 loss 和 F1-score 如下圖所示：





以下爲我訓練出來的 DNN 所得到的 F1-score：

Train	Validation	Test	
		Public	Private
0.99804	0.70200	0.76976	0.78837

3. (1%) 請敘述你如何 improve performance (preprocess embedding，架構等)，並解釋爲何這些做法可以使模型進步。

(1) 每個句子所作的 preprocessing 如第 1 題所述，標點符號、stop word、emoji，以及非數字或字母等等的詞，幾乎都不會影響到句子的語意，但是其在 RNN 的訓練過程中，可能會變成一種 noise，因此將其去掉，可能可以增進 RNN 的 performance。

(2) 將每個詞轉爲 basic form，可以避免同一個詞因爲文法變化而產生許多不同的 word vector，使得同樣語意的句子透過 RNN 所得到的分數不同的問題。

(3) 在將句子輸入 RNN 之前，會先將各個句子進行 trim 或 pad，首先會計算所有句子的平均長度，並將所有句子 trim 到平均長度以內，而若是長度不到平均長度的句子，則會將其 pad 到平均長度。雖然將句子中的詞去掉可能會遺失一些 information，但是非常長的句子其實占了少數，若爲了保留這些非常長的句子的 information，而將大部分的句子加入 padding，padding 其實也算是一種 noise，如此便會在 data 中加入大量的 noise，使得 RNN 的 performance 變差，因此比較好的做法可能是將所有句子 trim 或 pad 到平均長度，既能保留大部分句子的 information，也能盡可能減少 padding，在 trim 丟掉 information 和 padding 加入 noise 之間取得平衡。

(4) 在 LSTM 的 output 要傳給 linear 之前，我將 LSTM 的 output 對 sequence 的各個 dimension 取最大值、最小值以及平均值，並將這三個值 concatenate 之後才傳給 linear。RNN 輸出的 sequence 的各個 dimension 都代表了一種 feature，因此各個 dimension 的最大值和最小值就很有可能代表了這句話之中哪個詞最能表現出該 feature，以此加上 sequence 的平均值再傳給 linear，可能可以做出更正確的分類。

4. (%) 請比較不做斷詞 (e.g. 用空白分開) 與有做斷詞，兩種方法實作出來的效果差異，並解釋爲何有此差別。

我將 training data 和 testing data 中的所有句子分別以 spaCy 和空白進行斷詞後，爲作比較而不將任何詞去掉，也不將詞轉爲 basic form，以此直接給第 1 題所述的 Word2Vec model 以及 RNN 進行訓練，兩者所得到的 F1-score 如下：

	Train	Validation	Test	
			Public	Private
spaCy	0.79262	0.72050	0.79609	0.80232
空白	0.77224	0.70050	0.77209	0.74418

使用 spaCy 進行斷詞的 performance 比使用空白進行斷詞來得好，我認爲可能的原因之一爲有時英文的詞會縮寫在一起，例如”He’s my bro.” 中，”he” 和”is” 縮寫在一起而未以空白隔開，此時用 spaCy 可以正確的將”he” 和”is” 斷詞，從而得到較爲正確的 information，因此使用 spaCy 進行斷詞所得到的 performance 較好。

5. (1%) 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。

以下爲 RNN 和 BOW+DNN 分別對這兩句話進行預測的結果 ((p, q) 的意思爲 model 預測該句子不爲 malicious 的機率爲 p，預測該句子爲 malicious 的機率爲 q：

	”Today is hot, but I am happy.”	”I am happy, but today is hot.”
RNN	$(0.9014, 0.0986)$	$(0.7443, 0.2557)$
BOW+DNN	$(0.99998, 0.00002)$	$(0.99998, 0.00002)$

由此可以看出 RNN 可以分辨出這兩句話在語意上的不同，但是 BOW+DNN 卻無法做到，原因在於 RNN 會考慮到字詞的順序是如何影響語意，但是 BOW+DNN 則不考慮字詞順序，因此對 BOW+DNN 來說這兩句話是一樣的，故 BOW+DNN 無法分辨出這兩句話語意的不同。

Part 2. Math Problem

LSTM Cell (1%)

In this exercise, we will simulate the forward pass of a simple LSTM cell. Figure.1 shows a single LSTM cell, where z is the cell input, z_i, z_f, z_o are the control inputs of the gates, c is the cell memory, and f, g, h are activation functions. Given an input x, the cell input and the control inputs can be calculated by the following equations:

$$\begin{aligned} z &= w \cdot x + b \\ z_i &= w_i \cdot x + b_i \\ z_f &= w_f \cdot x + b_f \\ z_o &= w_o \cdot x + b_o \end{aligned}$$

where w, w_i, w_f, w_o are weights and b, b_i, b_f, b_o are biases. The final output can be calculated by

$$y = f(z_o)h(c')$$

where the value stored in cell memory is updated by

$$c' = f(z_i)g(z) + cf(z_f)$$

Given an input sequence x^t ($t = 1,\ 2,\ \cdots,\ 8$), please derive the output sequence y_t. The input sequence, the weights, and the activation functions are provided below. The initial value in cell memory is 0. Please note that your calculation process is required to receive full credit.

$w = [0,\ 0,\ 0,\ 1]$, $b = 0$
$w_i = [100,\ 100,\ 0,\ 0]$, $b_i = -10$
$w_f = [-100, -100,\ 0,\ 0]$, $b_f = 110$
$w_o = [0,\ 0,\ 100,\ 0]$, $b_o = -10$

t	1	2	3	4	5	6	7	8
x^t	0	1	1	0	0	0	1	1
	1	0	1	1	1	0	1	0
	0	1	1	1	0	1	1	1
	3	-2	4	0	2	-4	1	2

$$f(z) = \frac{1}{1+e^{-z}} \qquad g(z) = z \qquad h(z) = z$$

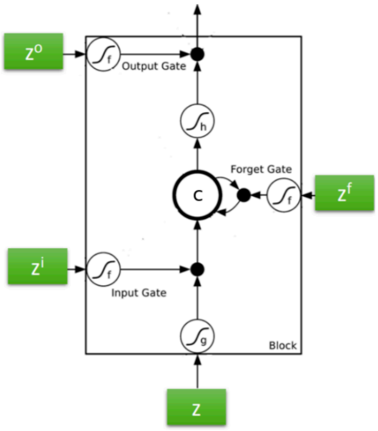


Figure. 1

solution

當 $t = 1$ 時

$$\begin{aligned}
z &= w \cdot x^1 + b = (0,0,0,1) \cdot (0,1,0,3) + 3 = 3 \\
z_i &= w_i \cdot x^1 + b_i = (100,100,0,0) \cdot (0,1,0,3) - 10 = 90 \\
z_f &= w_f \cdot x^1 + b_f = (-100,-100,0,0) \cdot (0,1,0,3) + 110 = 10 \\
z_o &= w_o \cdot x^1 + b_o = (0,0,100,0) \cdot (0,1,0,3) - 10 = -10 \\
c' &= f(z_i)g(z) + cf(z_f) = \frac{1}{1+e^{-90}} \cdot 3 + 0 \cdot \frac{1}{1+e^{-10}} \approx 3 \\
y_1 &= f(z_o)h(c') = \frac{1}{1+e^{-(-10)}} \cdot 3 \approx 0
\end{aligned}$$

當 $t = 2$ 時

$$\begin{aligned}
z &= w \cdot x^2 + b = (0,0,0,1) \cdot (1,0,1,-2) + 3 = -2 \\
z_i &= w_i \cdot x^2 + b_i = (100,100,0,0) \cdot (1,0,1,-2) - 10 = 90 \\
z_f &= w_f \cdot x^2 + b_f = (-100,-100,0,0) \cdot (1,0,1,-2) + 110 = 10 \\
z_o &= w_o \cdot x^2 + b_o = (0,0,100,0) \cdot (1,0,1,-2) - 10 = 90 \\
c' &= f(z_i)g(z) + cf(z_f) = \frac{1}{1+e^{-90}} \cdot (-2) + 3 \cdot \frac{1}{1+e^{-10}} \approx 1 \\
y_2 &= f(z_o)h(c') = \frac{1}{1+e^{-90}} \cdot 1 \approx 1
\end{aligned}$$

當 $t = 3$ 時

$$\begin{aligned}
z &= w \cdot x^3 + b = (0,0,0,1) \cdot (1,1,1,4) + 3 = 4 \\
z_i &= w_i \cdot x^3 + b_i = (100,100,0,0) \cdot (1,1,1,4) - 10 = 190 \\
z_f &= w_f \cdot x^3 + b_f = (-100,-100,0,0) \cdot (1,1,1,4) + 110 = -90 \\
z_o &= w_o \cdot x^3 + b_o = (0,0,100,0) \cdot (1,1,1,4) - 10 = 90 \\
c' &= f(z_i)g(z) + cf(z_f) = \frac{1}{1+e^{-190}} \cdot 4 + 1 \cdot \frac{1}{1+e^{-(-90)}} \approx 4 \\
y_3 &= f(z_o)h(c') = \frac{1}{1+e^{-90}} \cdot 4 \approx 4
\end{aligned}$$

當 $t = 4$ 時

$$\begin{aligned}
z &= w \cdot x^4 + b = (0,0,0,1) \cdot (0,1,1,0) + 3 = 0 \\
z_i &= w_i \cdot x^4 + b_i = (100,100,0,0) \cdot (0,1,1,0) - 10 = 90 \\
z_f &= w_f \cdot x^4 + b_f = (-100,-100,0,0) \cdot (0,1,1,0) + 110 = 10 \\
z_o &= w_o \cdot x^4 + b_o = (0,0,100,0) \cdot (0,1,1,0) - 10 = 90 \\
c' &= f(z_i)g(z) + cf(z_f) = \frac{1}{1+e^{-90}} \cdot 0 + 4 \cdot \frac{1}{1+e^{-10}} \approx 4 \\
y_4 &= f(z_o)h(c') = \frac{1}{1+e^{-90}} \cdot 4 \approx 4
\end{aligned}$$

當 $t = 5$ 時

$$
\begin{aligned}
z &= w \cdot x^5 + b = (0,0,0,1) \cdot (0,1,0,2) + 3 = 2 \\
z_i &= w_i \cdot x^5 + b_i = (100,100,0,0) \cdot (0,1,0,2) - 10 = 90 \\
z_f &= w_f \cdot x^5 + b_f = (-100,-100,0,0) \cdot (0,1,0,2) + 110 = 10 \\
z_o &= w_o \cdot x^5 + b_o = (0,0,100,0) \cdot (0,1,0,2) - 10 = -10 \\
c' &= f(z_i)g(z) + cf(z_f) = \frac{1}{1+e^{-90}} \cdot 2 + 4 \cdot \frac{1}{1+e^{-10}} \approx 6 \\
y_5 &= f(z_o)h(c') = \frac{1}{1+e^{-(-10)}} \cdot 6 \approx 0
\end{aligned}
$$

當 $t = 6$ 時

$$
\begin{aligned}
z &= w \cdot x^6 + b = (0,0,0,1) \cdot (0,0,1,-4) + 3 = -4 \\
z_i &= w_i \cdot x^6 + b_i = (100,100,0,0) \cdot (0,0,1,-4) - 10 = -10 \\
z_f &= w_f \cdot x^6 + b_f = (-100,-100,0,0) \cdot (0,0,1,-4) + 110 = 110 \\
z_o &= w_o \cdot x^6 + b_o = (0,0,100,0) \cdot (0,0,1,-4) - 10 = 90 \\
c' &= f(z_i)g(z) + cf(z_f) = \frac{1}{1+e^{-(-10)}} \cdot (-4) + 6 \cdot \frac{1}{1+e^{-110}} \approx 6 \\
y_6 &= f(z_o)h(c') = \frac{1}{1+e^{-90}} \cdot 6 \approx 6
\end{aligned}
$$

當 $t = 7$ 時

$$
\begin{aligned}
z &= w \cdot x^7 + b = (0,0,0,1) \cdot (1,1,1,1) + 3 = 1 \\
z_i &= w_i \cdot x^7 + b_i = (100,100,0,0) \cdot (1,1,1,1) - 10 = 190 \\
z_f &= w_f \cdot x^7 + b_f = (-100,-100,0,0) \cdot (1,1,1,1) + 110 = -90 \\
z_o &= w_o \cdot x^7 + b_o = (0,0,100,0) \cdot (1,1,1,1) - 10 = 90 \\
c' &= f(z_i)g(z) + cf(z_f) = \frac{1}{1+e^{-190}} \cdot 1 + 6 \cdot \frac{1}{1+e^{-(-90)}} \approx 1 \\
y_7 &= f(z_o)h(c') = \frac{1}{1+e^{-90}} \cdot 1 \approx 1
\end{aligned}
$$

當 $t = 8$ 時

$$
\begin{aligned}
z &= w \cdot x^8 + b = (0,0,0,1) \cdot (1,0,1,2) + 3 = 2 \\
z_i &= w_i \cdot x^8 + b_i = (100,100,0,0) \cdot (1,0,1,2) - 10 = 90 \\
z_f &= w_f \cdot x^8 + b_f = (-100,-100,0,0) \cdot (1,0,1,2) + 110 = 10 \\
z_o &= w_o \cdot x^8 + b_o = (0,0,100,0) \cdot (1,0,1,2) - 10 = 90 \\
c' &= f(z_i)g(z) + cf(z_f) = \frac{1}{1+e^{-90}} \cdot 2 + 1 \cdot \frac{1}{1+e^{-10}} \approx 3 \\
y_1 &= f(z_o)h(c') = \frac{1}{1+e^{-90}} \cdot 3 \approx 3 \; \square
\end{aligned}
$$

Word Embedding (1%)

Consider the Skip-Gram model below, let

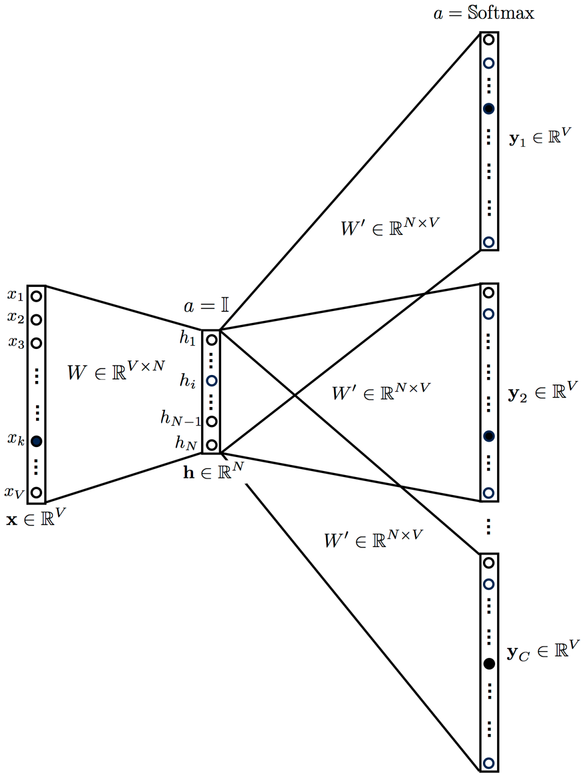
$$h = W^T x$$

$$u = W'^T h$$

$$y = Softmax(u) = Softmax(W'^T W^T x)$$

$$Loss = L = -log \prod_{c \in C} P(w_{output,c}, w_{input}) = -log \prod_{c \in C} \frac{exp(u_c)}{\Sigma_{i \in V} exp(u_i)}$$

where C = the context words of the input word. Calculate $\frac{\partial L}{\partial W^T_{ij}}$ and $\frac{\partial L}{\partial W'^T_{ij}}$. (Note: assume that softmax was performed on all output neuron, i.e., no negative sampling)



$a = \mathbb{S}$oftmax
$\mathbf{y}_1 \in \mathbb{R}^V$
$W' \in \mathbb{R}^{N \times V}$
x_1
x_2
x_3
$W \in \mathbb{R}^{V \times N}$
$a = \mathbb{I}$
h_1
h_i
h_{N-1}
h_N
x_k
x_V
$\mathbf{x} \in \mathbb{R}^V$
$\mathbf{h} \in \mathbb{R}^N$
$W' \in \mathbb{R}^{N \times V}$
$\mathbf{y}_2 \in \mathbb{R}^V$
$W' \in \mathbb{R}^{N \times V}$
$\mathbf{y}_C \in \mathbb{R}^V$

Figure. 2

solution

因爲

$$\begin{aligned} h = W^T x &= \begin{pmatrix} W_{11} & W_{21} & \cdots & W_{V1} \\ W_{12} & W_{22} & \cdots & W_{V2} \\ \vdots & \vdots & & \vdots \\ W_{1N} & W_{2N} & \cdots & W_{VN} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_V \end{pmatrix} \\ &= \begin{pmatrix} \sum_{p=1}^{V} W_{p1} x_p \\ \sum_{p=1}^{V} W_{p2} x_p \\ \vdots \\ \sum_{p=1}^{V} W_{pN} x_p \end{pmatrix} \end{aligned}$$

$$\begin{aligned} u = W'^T h &= \begin{pmatrix} W'_{11} & W'_{21} & \cdots & W'_{N1} \\ W'_{12} & W'_{22} & \cdots & W'_{N2} \\ \vdots & \vdots & & \vdots \\ W'_{1V} & W'_{2V} & \cdots & W'_{NV} \end{pmatrix} \begin{pmatrix} \sum_{p=1}^{V} W_{p1} x_p \\ \sum_{p=1}^{V} W_{p2} x_p \\ \vdots \\ \sum_{p=1}^{V} W_{pN} x_p \end{pmatrix} \\ &= \begin{pmatrix} \sum_{q=1}^{N} (W'_{q1} \sum_{p=1}^{V} W_{pq} x_p) \\ \sum_{q=1}^{N} (W'_{q2} \sum_{p=1}^{V} W_{pq} x_p) \\ \vdots \\ \sum_{q=1}^{N} (W'_{qV} \sum_{p=1}^{V} W_{pq} x_p) \end{pmatrix} \\ &= \begin{pmatrix} \sum_{p=1}^{V} \sum_{q=1}^{N} W_{pq} W'_{q1} x_p \\ \sum_{p=1}^{V} \sum_{q=1}^{N} W_{pq} W'_{q2} x_p \\ \vdots \\ \sum_{p=1}^{V} \sum_{q=1}^{N} W_{pq} W'_{qV} x_p \end{pmatrix} \end{aligned}$$

所以

$$\begin{aligned} L &= -log(\prod_{c \in C} \frac{exp(u_c)}{\sum_{r \in V} exp(u_r)}) \\ &= -\sum_{c \in C} log \frac{exp(u_c)}{\sum_{r \in V} exp(u_r)} \\ &= -\sum_{c \in C} (log(exp(u_c)) - log(\sum_{r \in V} exp(u_r))) \\ &= -\sum_{c \in C} (u_c - log(\sum_{r \in V} exp(u_r))) \\ &= -\sum_{c \in C} (\sum_{p=1}^{V} \sum_{q=1}^{N} W_{pq} W'_{qc} x_p - log(\sum_{r \in V} exp(\sum_{p=1}^{V} \sum_{q=1}^{N} W_{pq} W'_{qr} x_p))) \end{aligned}$$

因此可得

$$\begin{aligned}
\frac{\partial L}{\partial W^T_{ij}} &= \frac{\partial L}{\partial W_{ji}} \\
&= -\sum_{c\in C}(W'_{ic}x_j - \frac{\sum_{r\in V} W'_{ir}x_j exp(\sum_{p=1}^{V}\sum_{q=1}^{N} W_{pq}W'_{qr}x_p)}{\sum_{r\in V} exp(\sum_{p=1}^{V}\sum_{q=1}^{N} W_{pq}W'_{qr}x_p)}) \\
&= -\sum_{c\in C}(W'_{ic}x_j - \frac{\sum_{r\in V} W'_{ir}x_j exp(u_r)}{\sum_{r\in V} exp(u_r)})
\end{aligned}$$

$$\begin{aligned}
\frac{\partial L}{\partial W'^T_{ij}} &= \frac{\partial L}{\partial W'_{ji}} \\
&= [\![i\in C]\!](-\sum_{p=1}^{V} W_{pj}x_p) + \sum_{c\in C}\frac{(\sum_{p=1}^{V} W_{pj}x_p)exp(\sum_{p=1}^{V}\sum_{q=1}^{N} W_{pq}W'_{qi}x_p)}{\sum_{r\in V} exp(\sum_{p=1}^{V}\sum_{q=1}^{N} W_{pq}W'_{qr}x_p)} \\
&= [\![i\in C]\!](-\sum_{p=1}^{V} W_{pj}x_p) + \sum_{c\in C}\frac{(\sum_{p=1}^{V} W_{pj}x_p)exp(u_i)}{\sum_{r\in V} exp(u_r)} \ \square
\end{aligned}$$