1 Transfer Learning

1.5

The validation accuracy for finetune training is 0.934641.
The validation accuracy for training when freezing all weights except fc layer is 0.954248.
Below are results from running two models.

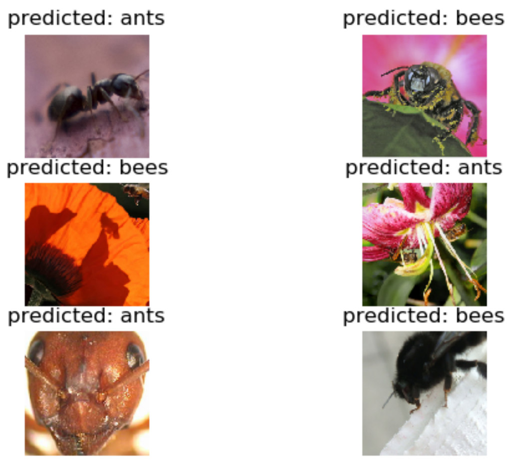


Figure 1: Predicted results from finetune training

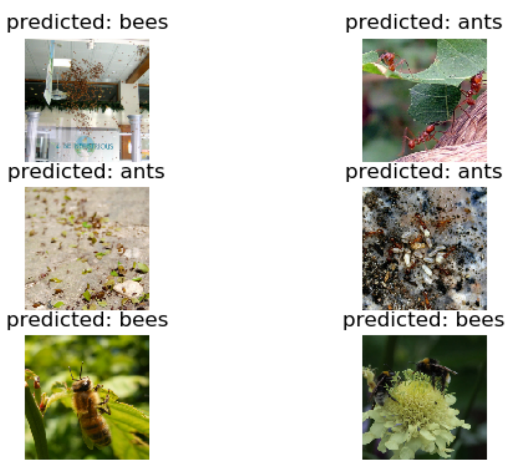


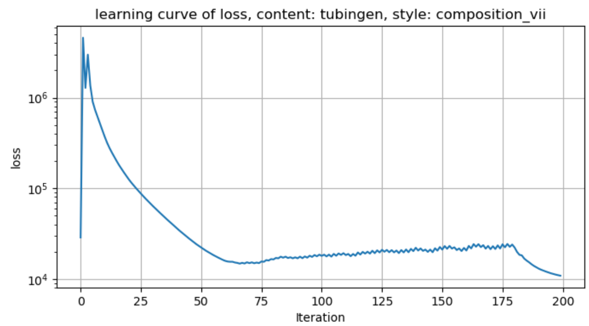
Figure 2: Predicted results from partially freezing training

2 Style Transfer

2.4

The generated images and learning curves of loss are shown below.

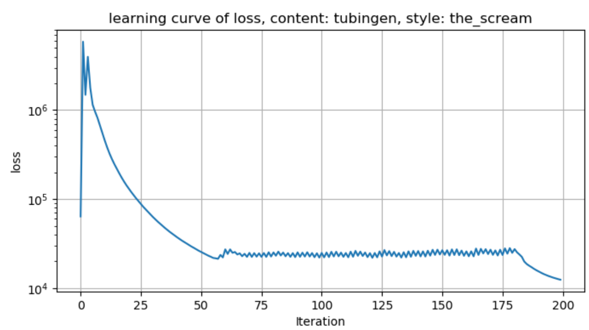
(a) Generated image



(b) Learnign curve of loss

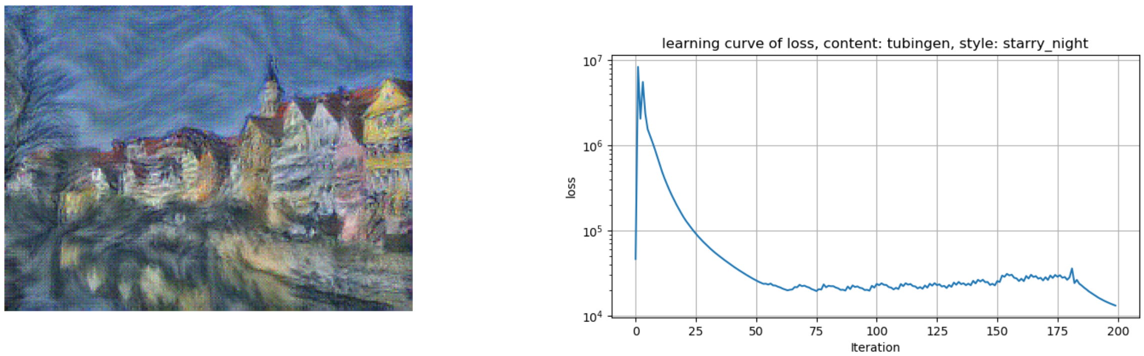
Figure 3: Style: Composition vii

(a) Generated image



(b) Learnign curve of loss

Figure 4: Style: the Scream



(a) Generated image

(b) Learnign curve of loss

Figure 5: Style: Starry night

3 Forward and Backward propagation module for RNN

3.2 RNN step backward

When deriving expressions using chain rule, remember that if there exists matrix multiplication, we have to sum up all contributions. But for the element-wise expression, we can just use the same subscript. So, we can derive $\frac{\partial L}{\partial W_x}$ as

$$
\begin{aligned}
\frac{\partial L}{\partial W_x} &= \frac{\partial L}{\partial h_t} \frac{\partial h_t}{\partial \tanh^{-1}(h_t)} \frac{\partial \tanh^{-1}(h_t)}{\partial W_x} \\
\left[\frac{\partial L}{\partial W_x}\right]_{i,j} &= \sum_k \frac{\partial L}{\partial (h_t)_k} \frac{\partial (h_t)_k}{\partial \tanh^{-1}(h_t)_k} \frac{\partial \tanh^{-1}(h_t)_k}{\partial (W_x)_{i,j}} \\
&= \sum_k \frac{\partial L}{\partial (h_t)_k} (1 - h_t \odot h_t)_k \frac{\partial \sum_m (W_x)_{k,m} (x_t)_m}{\partial (W_x)_{i,j}} \\
&= \sum_{k,m} \frac{\partial L}{\partial (h_t)_k} (1 - h_t \odot h_t)_k (x_t)_m \mathbb{1}_{k=i,m=j} \\
&= \left[\frac{\partial L}{\partial (h_t)} \odot (1 - h_t \odot h_t)\right]_i (x_t)_j \\
\Rightarrow \frac{\partial L}{\partial W_x} &= \left[\frac{\partial L}{\partial h_t} \odot (1 - h_t \odot h_t)\right] x_t^T
\end{aligned}
\tag{1}
$$

where $\odot$ represents the element-wise multiplication. Similarly,

$$
\frac{\partial L}{\partial W_h} = \frac{\partial L}{\partial h_t} \frac{\partial h_t}{\partial \tanh^{-1}(h_t)} \frac{\partial \tanh^{-1}(h_t)}{\partial W_h} = \left[\frac{\partial L}{\partial h_t} \odot (1 - h_t \odot h_t)\right] h_{t-1}^T \tag{2}
$$

Then for $\frac{\partial L}{\partial b}$,

$$\begin{aligned}
\frac{\partial L}{\partial b} &= \frac{\partial L}{\partial h_t} \frac{\partial h_t}{\partial \tanh^{-1}(h_t)} \frac{\partial \tanh^{-1}(h_t)}{\partial b} \\
\left[\frac{\partial L}{\partial b}\right]_i &= \sum_j \frac{\partial L}{\partial (h_t)_j} \frac{\partial (h_t)_j}{\partial \tanh^{-1}(h_t)_j} \frac{\partial \tanh^{-1}(h_t)_j}{\partial b_i} \\
&= \sum_j \frac{\partial L}{\partial (h_t)_j} (1 - h_t \odot h_t)_j \frac{\partial b_j}{\partial b_i} \\
&= \sum_j \frac{\partial L}{\partial (h_t)_j} (1 - h_t \odot h_t)_j \mathbb{1}_{j=i} \\
&= \left(\frac{\partial L}{\partial (h_t)} \odot (1 - h_t \odot h_t) \right)_i \\
\Rightarrow \frac{\partial L}{\partial b} &= \frac{\partial L}{\partial h_t} \odot (1 - h_t \odot h_t)
\end{aligned} \tag{3}$$

For $\frac{\partial L}{\partial x_t}$ and $\frac{\partial L}{\partial h_{t-1}}$, they are similar to $\frac{\partial L}{\partial W}$, but of different order in matrix multiplication. Hence we need to transpose the matrix dimension.

$$\begin{aligned}
\frac{\partial L}{\partial x_t} &= \frac{\partial L}{\partial h_t} \frac{\partial h_t}{\partial \tanh^{-1}(h_t)} \frac{\partial \tanh^{-1}(h_t)}{\partial x_t} \\
\left[\frac{\partial L}{\partial x_t}\right]_i &= \sum_k \frac{\partial L}{\partial (h_t)_k} \frac{\partial (h_t)_k}{\partial \tanh^{-1}(h_t)_k} \frac{\partial \tanh^{-1}(h_t)_k}{\partial (x_t)_i} \\
&= \sum_k \frac{\partial L}{\partial (h_t)_k} (1 - h_t \odot h_t)_k \frac{\partial \sum_m (W_x)_{k,m} (x_t)_m}{\partial (x_t)_i} \\
&= \sum_{k,m} \frac{\partial L}{\partial (h_t)_k} (1 - h_t \odot h_t)_k (W_x)_{k,m} \mathbb{1}_{m=i} \\
&= \sum_k \left[\frac{\partial L}{\partial (h_t)} \odot (1 - h_t \odot h_t) \right]_k (W_x)_{k,i} \\
\Rightarrow \frac{\partial L}{\partial x_t} &= W_x^T \left[\frac{\partial L}{\partial h_t} \odot (1 - h_t \odot h_t) \right]
\end{aligned} \tag{4}$$

Similarly,

$$\frac{\partial L}{\partial h_{t-1}} = \frac{\partial L}{\partial h_t} \frac{\partial h_t}{\partial \tanh^{-1}(h_t)} \frac{\partial \tanh^{-1}(h_t)}{\partial h_{t-1}} = W_h^T \left[\frac{\partial L}{\partial h_t} \odot (1 - h_t \odot h_t) \right] \tag{5}$$

For implementation, see codes in **rnn_layers.py**.

3.4 RNN backward

Note that for the whole RNN, h_t will contributes to all $h_{t'}$ ($t' \geq t$), and notice that only h_{t-1} contributes directly to h_t, hence when doing propagation for h_{t-1}, we can sum up all contributions from $h_{t'}$ to get $[\frac{\partial L}{\partial h_t}]_{new}$, and use it to calculate $[\frac{\partial L}{\partial h_{t-1}}]_{new}$, where the sub-new means the sum of all contributions from $\frac{\partial L}{\partial h_{t'}}$,

$$\begin{aligned}\left[\frac{\partial L}{\partial h_{t-1}}\right]_{new} &= \frac{\partial L}{\partial h_{t-1}} + \left[\frac{\partial L}{\partial h_t}\right]_{new}\frac{\partial h_t}{\partial h_{t-1}} \\ &= \frac{\partial L}{\partial h_{t-1}} + W_h^T\left[\left[\frac{\partial L}{\partial h_t}\right]_{new}\odot(1-h_t\odot h_t)\right]\end{aligned} \tag{6}$$

where the second term is derived in Eqn.5. Also, when $t-1=T$,

$$\left[\frac{\partial L}{\partial h_T}\right]_{new} = \frac{\partial L}{\partial h_T} \tag{7}$$

then, for $\frac{\partial L}{\partial W_x}$, $\frac{\partial L}{\partial W_h}$ and $\frac{\partial L}{\partial b}$ using the solutions in RNN Step backward derivation, we can write

$$\begin{aligned}\frac{\partial L}{\partial W_x} &= \sum_{t=1}^{T}\left[\frac{\partial L}{\partial h_t}\right]_{new}\frac{\partial h_t}{\partial W_x} \\ &= \sum_{t=1}^{T}\left[\left[\frac{\partial L}{\partial h_t}\right]_{new}\odot(1-h_t\odot h_t)\right]x_t^T\end{aligned} \tag{8}$$

$$\begin{aligned}\frac{\partial L}{\partial W_h} &= \sum_{t=1}^{T}\left[\frac{\partial L}{\partial h_t}\right]_{new}\frac{\partial h_t}{\partial W_h} \\ &= \sum_{t=1}^{T}\left[\left[\frac{\partial L}{\partial h_t}\right]_{new}\odot(1-h_t\odot h_t)\right]h_{t-1}^T\end{aligned} \tag{9}$$

$$\begin{aligned}\frac{\partial L}{\partial b} &= \sum_{t=1}^{T}\left[\frac{\partial L}{\partial h_t}\right]_{new}\frac{\partial h_t}{\partial b} \\ &= \sum_{t=1}^{T}\left[\frac{\partial L}{\partial h_t}\right]_{new}\odot(1-h_t\odot h_t)\end{aligned} \tag{10}$$

For $\frac{\partial L}{\partial x_t}$, since x_t only contributes to h_t directly, then

$$\begin{aligned}\frac{\partial L}{\partial x_t} &= \left[\frac{\partial L}{\partial h_t}\right]_{new}\frac{\partial h_t}{\partial x_t} \\ &= W_x^T\left[\left[\frac{\partial L}{\partial h_t}\right]_{new}\odot(1-h_t\odot h_t)\right]\end{aligned} \tag{11}$$

For $\frac{\partial L}{\partial h_0}$,

$$\begin{aligned}\frac{\partial L}{\partial h_0} &= \left[\frac{\partial L}{\partial h_1}\right]_{new}\frac{\partial h_1}{\partial h_0} \\ &= W_h^T\left[\left[\frac{\partial L}{\partial h_1}\right]_{new}\odot(1-h_1\odot h_1)\right]\end{aligned} \tag{12}$$

For implementation, see codes in **rnn_layers.py**.

4 Forward and Backward propagation module for LSTM

4.2 LSTM step backward

Before deriving partial derivatives with respect to x, W, b, h, we first calculate partial derivatives $\frac{\partial L}{\partial f_t}$, $\frac{\partial L}{\partial i_t}$, $\frac{\partial L}{\partial \tilde{c_t}}$ and $\frac{\partial L}{\partial o_t}$ with given $\frac{\partial L}{\partial c_t}$ and $\frac{\partial L}{\partial h_t}$. Note that c_t also do contributions to h_t, so

$$\frac{\partial h_t}{\partial c_t} = \frac{\partial h_t}{\partial \tanh(c_t)} \frac{\partial \tanh(c_t)}{\partial c_t} = o_t \odot (1 - \tanh(c_t) \odot \tanh(c_t))$$

where $\odot$ represents the element-wise multiplication. Then we can derive

$$\frac{\partial L}{\partial f_t} = \frac{\partial L}{\partial h_t}\frac{\partial h_t}{\partial c_t}\frac{\partial c_t}{\partial f_t} + \frac{\partial L}{\partial c_t}\frac{\partial c_t}{\partial f_t} = \left(\frac{\partial L}{\partial c_t} + \frac{\partial L}{\partial h_t} \odot o_t \odot (1 - \tanh(c_t) \odot \tanh(c_t))\right) \odot c_{t-1} \tag{13}$$

$$\frac{\partial L}{\partial i_t} = \frac{\partial L}{\partial h_t}\frac{\partial h_t}{\partial c_t}\frac{\partial c_t}{\partial i_t} + \frac{\partial L}{\partial c_t}\frac{\partial c_t}{\partial i_t} = \left(\frac{\partial L}{\partial c_t} + \frac{\partial L}{\partial h_t} \odot o_t \odot (1 - \tanh(c_t) \odot \tanh(c_t))\right) \odot \tilde{c}_t \tag{14}$$

$$\frac{\partial L}{\partial \tilde{c_t}} = \frac{\partial L}{\partial h_t}\frac{\partial h_t}{\partial c_t}\frac{\partial c_t}{\partial \tilde{c_t}} + \frac{\partial L}{\partial c_t}\frac{\partial c_t}{\partial \tilde{c_t}} = \left(\frac{\partial L}{\partial c_t} + \frac{\partial L}{\partial h_t} \odot o_t \odot (1 - \tanh(c_t) \odot \tanh(c_t))\right) \odot i_t \tag{15}$$

$$\frac{\partial L}{\partial o_t} = \frac{\partial L}{\partial h_t}\frac{\partial h_t}{\partial o_t} = \frac{\partial L}{\partial h_t} \odot \tanh(c_t) \tag{16}$$

Next, we will derive required expressions. Remember that if there exists matrix multiplication during the chain rule, we have to sum up all contributions. For example,

$$\begin{aligned}
\frac{\partial L}{\partial W_x^f} &= \frac{\partial L}{\partial f_t}\frac{\partial f_t}{\partial \sigma^{-1}(f_t)}\frac{\partial \sigma^{-1}(f_t)}{\partial W_x^f} \\
\left[\frac{\partial L}{\partial W_x^f}\right]_{i,j} &= \sum_k \frac{\partial L}{\partial (f_t)_k}\frac{\partial (f_t)_k}{\partial \tanh^{-1}(f_t)_k}\frac{\partial \tanh^{-1}(f_t)_k}{\partial (W_x^f)_{i,j}} \\
&= \sum_k \frac{\partial L}{\partial (f_t)_k}(f_t \odot (1-f_t))_k \frac{\partial \sum_m (W_x^f)_{k,m}(x_t)_m}{\partial (W_x^f)_{i,j}} \\
&= \sum_{k,m} \frac{\partial L}{\partial (f_t)_k}(f_t \odot (1-f_t))_k (x_t)_m \mathbb{1}_{k=i,m=j} \\
&= \left[\frac{\partial L}{\partial f_t} \odot (f_t \odot (1-f_t))\right]_i (x_t)_j \\
\Rightarrow \frac{\partial L}{\partial W_x^f} &= \frac{\partial L}{\partial f_t} \odot (f_t \odot (1-f_t))\, x_t^T \\
&= \left[\left(\frac{\partial L}{\partial c_t} + \frac{\partial L}{\partial h_t} \odot o_t \odot (1 - \tanh(c_t) \odot \tanh(c_t))\right) \odot c_{t-1} \odot (f_t \odot (1-f_t))\right] x_t^T
\end{aligned} \tag{17}$$

Similarly,

$$\begin{aligned}
\frac{\partial L}{\partial W_h^f} &= \frac{\partial L}{\partial f_t}\frac{\partial f_t}{\partial \sigma^{-1}(f_t)}\frac{\partial \sigma^{-1}(f_t)}{\partial W_h^f} \\
&= \left[\left(\frac{\partial L}{\partial c_t} + \frac{\partial L}{\partial h_t} \odot o_t \odot (1 - \tanh(c_t) \odot \tanh(c_t))\right) \odot c_{t-1} \odot (f_t \odot (1-f_t))\right] h_{t-1}^T
\end{aligned} \tag{18}$$

$$\begin{aligned}
\frac{\partial L}{\partial W_x^i} &= \frac{\partial L}{\partial i_t}\frac{\partial i_t}{\partial \sigma^{-1}(i_t)}\frac{\partial \sigma^{-1}(i_t)}{\partial W_x^i} \\
&= \left[\left(\frac{\partial L}{\partial c_t} + \frac{\partial L}{\partial h_t} \odot o_t \odot (1 - \tanh(c_t) \odot \tanh(c_t))\right) \odot \tilde{c}_t \odot (i_t \odot (1-i_t))\right] x_t^T
\end{aligned} \tag{19}$$

$$\begin{aligned}\frac{\partial L}{\partial W_h^i} &= \frac{\partial L}{\partial i_t}\frac{\partial i_t}{\partial \sigma^{-1}(i_t)}\frac{\partial \sigma^{-1}(i_t)}{\partial W_h^i} \\ &= \left[\left(\frac{\partial L}{\partial c_t} + \frac{\partial L}{\partial h_t}\odot o_t \odot (1-\tanh(c_t)\odot\tanh(c_t))\right)\odot \tilde{c}_t \odot (i_t\odot(1-i_t))\right] h_{t-1}^T\end{aligned} \tag{20}$$

$$\begin{aligned}\frac{\partial L}{\partial W_x^c} &= \frac{\partial L}{\partial \tilde{c}_t}\frac{\partial \tilde{c}_t}{\partial \tanh^{-1}(\tilde{c}_t)}\frac{\partial \tanh^{-1}(\tilde{c}_t)}{\partial W_x^c} \\ &= \left[\left(\frac{\partial L}{\partial c_t} + \frac{\partial L}{\partial h_t}\odot o_t \odot (1-\tanh(c_t)\odot\tanh(c_t))\right)\odot i_t \odot (1-\tilde{c}_t\odot\tilde{c}_t)\right] x_t^T\end{aligned} \tag{21}$$

$$\begin{aligned}\frac{\partial L}{\partial W_h^c} &= \frac{\partial L}{\partial \tilde{c}_t}\frac{\partial \tilde{c}_t}{\partial \tanh^{-1}(\tilde{c}_t)}\frac{\partial \tanh^{-1}(\tilde{c}_t)}{\partial W_h^c} \\ &= \left[\left(\frac{\partial L}{\partial c_t} + \frac{\partial L}{\partial h_t}\odot o_t \odot (1-\tanh(c_t)\odot\tanh(c_t))\right)\odot i_t \odot (1-\tilde{c}_t\odot\tilde{c}_t)\right] h_{t-1}^T\end{aligned} \tag{22}$$

$$\frac{\partial L}{\partial W_x^o} = \frac{\partial L}{\partial o_t}\frac{\partial o_t}{\partial \sigma^{-1}(o_t)}\frac{\partial \sigma^{-1}(\sigma o_t)}{\partial W_x^o} = \left[\frac{\partial L}{\partial h_t}\odot \tanh(c_t)\odot(o_t\odot(1-o_t))\right] x_t^T \tag{23}$$

$$\frac{\partial L}{\partial W_h^o} = \frac{\partial L}{\partial o_t}\frac{\partial o_t}{\partial \sigma^{-1}(o_t)}\frac{\partial \sigma^{-1}(\sigma o_t)}{\partial W_h^o} = \left[\frac{\partial L}{\partial h_t}\odot \tanh(c_t)\odot(o_t\odot(1-o_t))\right] h_{t-1}^T \tag{24}$$

Then for $\frac{\partial L}{\partial b^f}$,

$$\begin{aligned}\frac{\partial L}{\partial b^f} &= \frac{\partial L}{\partial f_t}\frac{\partial f_t}{\partial \sigma^{-1}(f_t)}\frac{\partial \sigma^{-1}(f_t)}{\partial b^f} \\ \left[\frac{\partial L}{\partial b^f}\right]_i &= \sum_j \frac{\partial L}{\partial (f_t)_j}\frac{\partial (f_t)_j}{\partial \sigma^{-1}(f_t)_j}\frac{\partial \sigma^{-1}(f_t)_j}{\partial b_i^f} \\ &= \sum_j \frac{\partial L}{\partial (f_t)_j}(f_t\odot(1-f_t))_j\frac{\partial b_j^f}{\partial b_i^f} \\ &= \sum_j \frac{\partial L}{\partial (f_t)_j}(f_t\odot(1-f_t))_j \mathbb{1}_{j=i} \\ &= \left[\frac{\partial L}{\partial (f_t)}\odot(f_t\odot(1-f_t))\right]_i \\ \Rightarrow \frac{\partial L}{\partial b^f} &= \left(\frac{\partial L}{\partial c_t} + \frac{\partial L}{\partial h_t}\odot o_t \odot (1-\tanh(c_t)\odot\tanh(c_t))\right)\odot c_{t-1}\odot(f_t\odot(1-f_t))\end{aligned} \tag{25}$$

Similarly,

$$\begin{aligned}\frac{\partial L}{\partial b^i} &= \frac{dL}{di_t}\frac{\partial i_t}{\partial \sigma^{-1}(i_t)}\frac{\partial \sigma^{-1}(i_t)}{\partial b^i} \\ &= \left(\frac{\partial L}{\partial c_t} + \frac{\partial L}{\partial h_t}\odot o_t \odot (1-\tanh(c_t)\odot\tanh(c_t))\right)\odot \tilde{c}_t \odot (i_t\odot(1-i_t))\end{aligned} \tag{26}$$

$$\begin{aligned}\frac{\partial L}{\partial b^c} &= \frac{dL}{d\tilde{c}_t}\frac{\partial \tilde{c}_t}{\partial \tanh^{-1}(\tilde{c}_t)}\frac{\partial \tanh^{-1}(\tilde{(c_t)})}{\partial b^c} \\ &= \left(\frac{\partial L}{\partial c_t} + \frac{\partial L}{\partial h_t}\odot o_t \odot (1-\tanh(c_t)\odot\tanh(c_t))\right)\odot i_t \odot (1-\tilde{c}_t\odot\tilde{c}_t)\end{aligned} \tag{27}$$

$$\frac{\partial L}{\partial b^o} = \frac{\partial L}{\partial o_t}\frac{\partial o_t}{\partial \sigma^{-1} o_t}\frac{\partial \sigma^{-1} o_t}{\partial b^o} = \frac{\partial L}{\partial h_t} \odot \tanh(c_t) \odot (o_t \odot (1 - o_t)) \tag{28}$$

For the remaining three terms, $\frac{\partial L}{\partial x_t}$, $\frac{\partial L}{\partial h_{t-1}}$ and $\frac{\partial L}{\partial c_{t-1}}$, we can easily derive

$$\begin{aligned}\frac{\partial L}{\partial c_{t-1}} &= \frac{\partial L}{\partial h_t}\frac{\partial h_t}{\partial c_t}\frac{\partial c_t}{\partial c_{t-1}} + \frac{\partial L}{\partial c_t}\frac{\partial c_t}{\partial c_{t-1}} \\ &= \left(\frac{\partial L}{\partial c_t} + \frac{\partial L}{\partial h_t} \odot o_t \odot (1 - \tanh(c_t) \odot \tanh(c_t))\right) \odot f_t\end{aligned} \tag{29}$$

Finally, for the last two terms, they are similar to $\frac{\partial L}{\partial W}$, however, considering the order of matrix multiplication (W times x and W times h_{t-1}), we need to transpose the matrix so that the derived expression matches exact dimensions of partial derivatives.

$$\begin{aligned}\frac{\partial L}{\partial x_t} =& \frac{\partial L}{\partial f_t}\frac{\partial f_t}{\partial x_t} + \frac{\partial L}{\partial i_t}\frac{\partial i_t}{\partial x_t} + \frac{\partial L}{\partial \tilde{d}_t}\frac{\partial \tilde{c}_t}{\partial x_t} + \frac{\partial L}{\partial o_t}\frac{\partial o_t}{\partial x_t} \\ =& (W_x^f)^T \left[\left(\frac{\partial L}{\partial c_t} + \frac{\partial L}{\partial h_t} \odot o_t \odot (1 - \tanh(c_t) \odot \tanh(c_t))\right) \odot c_{t-1} \odot (f_t \odot (1 - f_t))\right] \\ &+ (W_x^i)^T \left[\left(\frac{\partial L}{\partial c_t} + \frac{\partial L}{\partial h_t} \odot o_t \odot (1 - \tanh(c_t) \odot \tanh(c_t))\right) \odot \tilde{c}_t \odot (i_t \odot (1 - i_t))\right] \\ &+ (W_x^c)^T \left[\left(\frac{\partial L}{\partial c_t} + \frac{\partial L}{\partial h_t} \odot o_t \odot (1 - \tanh(c_t) \odot \tanh(c_t))\right) \odot i_t \odot (1 - \tilde{c}_t \odot \tilde{c}_t)\right] \\ &+ (W_x^o)^T \left[\frac{\partial L}{\partial h_t} \odot \tanh(c_t) \odot (o_t \odot (1 - o_t))\right]\end{aligned} \tag{30}$$

$$\begin{aligned}\frac{\partial L}{\partial h_{t-1}} =& \frac{\partial L}{\partial f_t}\frac{\partial f_t}{\partial h_{t-1}} + \frac{\partial L}{\partial i_t}\frac{\partial i_t}{\partial h_{t-1}} + \frac{\partial L}{\partial \tilde{d}_t}\frac{\partial \tilde{c}_t}{\partial h_{t-1}} + \frac{\partial L}{\partial o_t}\frac{\partial o_t}{\partial h_{t-1}} \\ =& (W_h^f)^T \left[\left(\frac{\partial L}{\partial c_t} + \frac{\partial L}{\partial h_t} \odot o_t \odot (1 - \tanh(c_t) \odot \tanh(c_t))\right) \odot c_{t-1} \odot (f_t \odot (1 - f_t))\right] \\ &+ (W_h^i)^T \left[\left(\frac{\partial L}{\partial c_t} + \frac{\partial L}{\partial h_t} \odot o_t \odot (1 - \tanh(c_t) \odot \tanh(c_t))\right) \odot \tilde{c}_t \odot (i_t \odot (1 - i_t))\right] \\ &+ (W_h^c)^T \left[\left(\frac{\partial L}{\partial c_t} + \frac{\partial L}{\partial h_t} \odot o_t \odot (1 - \tanh(c_t) \odot \tanh(c_t))\right) \odot i_t \odot (1 - \tilde{c}_t \odot \tilde{c}_t)\right] \\ &+ (W_h^o)^T \left[\frac{\partial L}{\partial h_t} \odot \tanh(c_t) \odot (o_t \odot (1 - o_t))\right]\end{aligned} \tag{31}$$

For implementation, see codes in **rnn_layers.py**.

4.4 LSTM backward

Similar to RNN backward, for the whole LSTM, c_t, h_t will contributes to all $h_{t'}$, $c_{t'}$ ($t' \geq t$), and notice that only h_{t-1} and c_{t-1} contributes directly to h_t, c_t, hence when doing propagation for h_{t-1}, c_{t-1}, we can sum up all contributions from $h_{t'}$ and $c_{t'}$ to get $[\frac{\partial L}{\partial h_t}]_{new}$, $[\frac{\partial L}{\partial c_t}]_{new}$, and use it to calculate $[\frac{\partial L}{\partial h_{t-1}}]_{new}$, $[\frac{\partial L}{\partial c_{t-1}}]_{new}$, where the sub-new means the sum of all contributions from $\frac{\partial L}{\partial h_{t'}}$ and $\frac{\partial L}{\partial c_{t'}}$.

$$
\begin{aligned}
\left[\frac{\partial L}{\partial h_{t-1}}\right]_{new} =& \frac{\partial L}{\partial h_{t-1}} + \left[\frac{\partial L}{\partial h_t}\right]_{new} \frac{\partial h_t}{\partial h_{t-1}} + \left[\frac{\partial L}{\partial c_t}\right]_{new} \frac{\partial c_t}{\partial h_{t-1}} \\
=& \frac{\partial L}{\partial h_{t-1}} + (W_h^f)^T \left[\left(\left[\frac{\partial L}{\partial c_t}\right]_{new} + \left[\frac{\partial L}{\partial h_t}\right]_{new} \odot o_t \odot (1 - \tanh(c_t) \odot \tanh(c_t)) \right) \odot c_{t-1} \odot (f_t \odot (1 - f_t)) \right] \\
& + (W_h^i)^T \left[\left(\left[\frac{\partial L}{\partial c_t}\right]_{new} + \left[\frac{\partial L}{\partial h_t}\right]_{new} \odot o_t \odot (1 - \tanh(c_t) \odot \tanh(c_t)) \right) \odot \tilde{c}_t \odot (i_t \odot (1 - i_t)) \right] \\
& + (W_h^c)^T \left[\left(\left[\frac{\partial L}{\partial c_t}\right]_{new} + \left[\frac{\partial L}{\partial h_t}\right]_{new} \odot o_t \odot (1 - \tanh(c_t) \odot \tanh(c_t)) \right) \odot i_t \odot (1 - \tilde{c}_t \odot \tilde{c}_t) \right] \\
& + (W_h^o)^T \left[\left[\frac{\partial L}{\partial h_t}\right]_{new} \odot \tanh(c_t) \odot (o_t \odot (1 - o_t)) \right]
\end{aligned} \tag{32}
$$

$$
\begin{aligned}
\left[\frac{\partial L}{\partial c_{t-1}}\right]_{new} &= \left[\frac{\partial L}{\partial h_t}\right]_{new} \frac{\partial h_t}{\partial c_{t-1}} + \left[\frac{\partial L}{\partial c_t}\right]_{new} \frac{\partial c_t}{\partial c_{t-1}} \\
&= \left(\left[\frac{\partial L}{\partial c_t}\right]_{new} + \left[\frac{\partial L}{\partial h_t}\right]_{new} \odot o_t \odot (1 - \tanh(c_t) \odot \tanh(c_t)) \right) \odot f_t
\end{aligned} \tag{33}
$$

where the second term is derived in LSTM step backward. Also, when $t-1 = T$,

$$
\left[\frac{\partial L}{\partial h_T}\right]_{new} = \frac{\partial L}{\partial h_T} \tag{34}
$$

$$
\left[\frac{\partial L}{\partial c_T}\right]_{new} = 0 \tag{35}
$$

Note that all $\frac{\partial L}{\partial c_t}$ can be derived from all $\frac{\partial L}{\partial h_t}$ for any $t < T$. Hence the only input we need to determine the derivatives are $\frac{\partial L}{\partial h_t}$. Then for $\frac{\partial L}{\partial W_x^f}$, $\frac{\partial L}{\partial W_x^i}$, $\frac{\partial L}{\partial W_x^c}$, $\frac{\partial L}{\partial W_x^o}$, $\frac{\partial L}{\partial W_h^f}$, $\frac{\partial L}{\partial W_h^i}$, $\frac{\partial L}{\partial W_h^c}$, $\frac{\partial L}{\partial W_h^o}$, $\frac{\partial L}{\partial b_f}$, $\frac{\partial L}{\partial b_i}$, $\frac{\partial L}{\partial b_c}$, $\frac{\partial L}{\partial b_o}$, we can derive expressions based on results of the LSTM step backward.

$$
\begin{aligned}
\frac{\partial L}{\partial W_x^f} &= \sum_{t=1}^{T} \left(\left[\frac{\partial L}{\partial h_t}\right]_{new} \frac{\partial h_t}{\partial W_x^f} + \left[\frac{\partial L}{\partial c_t}\right]_{new} \frac{\partial c_t}{\partial W_x^f} \right) \\
&= \left[\left(\left[\frac{\partial L}{\partial c_t}\right]_{new} + \left[\frac{\partial L}{\partial h_t}\right]_{new} \odot o_t \odot (1 - \tanh(c_t) \odot \tanh(c_t)) \right) \odot c_{t-1} \odot (f_t \odot (1 - f_t)) \right] x_t^T
\end{aligned} \tag{36}
$$

$$
\begin{aligned}
\frac{\partial L}{\partial W_h^f} &= \sum_{t=1}^{T} \left(\left[\frac{\partial L}{\partial h_t}\right]_{new} \frac{\partial h_t}{\partial W_h^f} + \left[\frac{\partial L}{\partial c_t}\right]_{new} \frac{\partial c_t}{\partial W_h^f} \right) \\
&= \left[\left(\left[\frac{\partial L}{\partial c_t}\right]_{new} + \left[\frac{\partial L}{\partial h_t}\right]_{new} \odot o_t \odot (1 - \tanh(c_t) \odot \tanh(c_t)) \right) \odot c_{t-1} \odot (f_t \odot (1 - f_t)) \right] h_{t-1}^T
\end{aligned} \tag{37}
$$

$$
\begin{aligned}
\frac{\partial L}{\partial W_x^i} &= \sum_{t=1}^{T} \left(\left[\frac{\partial L}{\partial h_t}\right]_{new} \frac{\partial h_t}{\partial W_x^i} + \left[\frac{\partial L}{\partial c_t}\right]_{new} \frac{\partial c_t}{\partial W_x^i} \right) \\
&= \left[\left(\left[\frac{\partial L}{\partial c_t}\right]_{new} + \left[\frac{\partial L}{\partial h_t}\right]_{new} \odot o_t \odot (1 - \tanh(c_t) \odot \tanh(c_t)) \right) \odot \tilde{c}_t \odot (i_t \odot (1 - i_t)) \right] x_t^T
\end{aligned} \tag{38}
$$

$$\begin{aligned}\frac{\partial L}{\partial W_h^i} &= \sum_{t=1}^{T}\left(\left[\frac{\partial L}{\partial h_t}\right]_{new}\frac{\partial h_t}{\partial W_h^i} + \left[\frac{\partial L}{\partial c_t}\right]_{new}\frac{\partial c_t}{\partial W_h^i}\right) \\ &= \left[\left(\left[\frac{\partial L}{\partial c_t}\right]_{new} + \left[\frac{\partial L}{\partial h_t}\right]_{new} \odot o_t \odot (1 - \tanh(c_t) \odot \tanh(c_t))\right) \odot \tilde{c}_t \odot (i_t \odot (1 - i_t))\right] h_{t-1}^T\end{aligned} \tag{39}$$

$$\begin{aligned}\frac{\partial L}{\partial W_x^c} &= \sum_{t=1}^{T}\left(\left[\frac{\partial L}{\partial h_t}\right]_{new}\frac{\partial h_t}{\partial W_x^c} + \left[\frac{\partial L}{\partial c_t}\right]_{new}\frac{\partial c_t}{\partial W_x^c}\right) \\ &= \left[\left(\left[\frac{\partial L}{\partial c_t}\right]_{new} + \left[\frac{\partial L}{\partial h_t}\right]_{new} \odot o_t \odot (1 - \tanh(c_t) \odot \tanh(c_t))\right) \odot i_t \odot (1 - \tilde{c}_t \odot \tilde{c}_t)\right] x_t^T\end{aligned} \tag{40}$$

$$\begin{aligned}\frac{\partial L}{\partial W_h^c} &= \sum_{t=1}^{T}\left(\left[\frac{\partial L}{\partial h_t}\right]_{new}\frac{\partial h_t}{\partial W_h^c} + \left[\frac{\partial L}{\partial c_t}\right]_{new}\frac{\partial c_t}{\partial W_h^c}\right) \\ &= \left[\left(\left[\frac{\partial L}{\partial c_t}\right]_{new} + \left[\frac{\partial L}{\partial h_t}\right]_{new} \odot o_t \odot (1 - \tanh(c_t) \odot \tanh(c_t))\right) \odot i_t \odot (1 - \tilde{c}_t \odot \tilde{c}_t)\right] h_{t-1}^T\end{aligned} \tag{41}$$

$$\frac{\partial L}{\partial W_x^o} = \left(\sum_{t=1}^{T}\left[\frac{\partial L}{\partial h_t}\right]_{new}\frac{\partial h_t}{\partial W_x^o}\right) = \left[\left[\frac{\partial L}{\partial h_t}\right]_{new} \odot \tanh(c_t) \odot (o_t \odot (1 - o_t))\right] x_t^T \tag{42}$$

$$\frac{\partial L}{\partial W_h^o} = \left(\sum_{t=1}^{T}\left[\frac{\partial L}{\partial h_t}\right]_{new}\frac{\partial h_t}{\partial W_h^o}\right) = \left[\left[\frac{\partial L}{\partial h_t}\right]_{new} \odot \tanh(c_t) \odot (o_t \odot (1 - o_t))\right] h_{t-1}^T \tag{43}$$

$$\begin{aligned}\frac{\partial L}{\partial b^f} &= \sum_{t=1}^{T}\left(\left[\frac{\partial L}{\partial h_t}\right]_{new}\frac{\partial h_t}{\partial b^f} + \left[\frac{\partial L}{\partial c_t}\right]_{new}\frac{\partial c_t}{\partial b^f}\right) \\ &= \left(\left[\frac{\partial L}{\partial c_t}\right]_{new} + \left[\frac{\partial L}{\partial h_t}\right]_{new} \odot o_t \odot (1 - \tanh(c_t) \odot \tanh(c_t))\right) \odot c_{t-1} \odot (f_t \odot (1 - f_t))\end{aligned} \tag{44}$$

$$\begin{aligned}\frac{\partial L}{\partial b^i} &= \sum_{t=1}^{T}\left(\left[\frac{\partial L}{\partial h_t}\right]_{new}\frac{\partial h_t}{\partial b^i} + \left[\frac{\partial L}{\partial c_t}\right]_{new}\frac{\partial c_t}{\partial b^i}\right) \\ &= \left(\left[\frac{\partial L}{\partial c_t}\right]_{new} + \left[\frac{\partial L}{\partial h_t}\right]_{new} \odot o_t \odot (1 - \tanh(c_t) \odot \tanh(c_t))\right) \odot \tilde{c}_t \odot (i_t \odot (1 - i_t))\end{aligned} \tag{45}$$

$$\begin{aligned}\frac{\partial L}{\partial b^c} &= \sum_{t=1}^{T}\left(\left[\frac{\partial L}{\partial h_t}\right]_{new}\frac{\partial h_t}{\partial b^c} + \left[\frac{\partial L}{\partial c_t}\right]_{new}\frac{\partial c_t}{\partial b^c}\right) \\ &= \left(\left[\frac{\partial L}{\partial c_t}\right]_{new} + \left[\frac{\partial L}{\partial h_t}\right]_{new} \odot o_t \odot (1 - \tanh(c_t) \odot \tanh(c_t))\right) \odot i_t \odot (1 - \tilde{c}_t \odot \tilde{c}_t)\end{aligned} \tag{46}$$

$$\frac{\partial L}{\partial b^o} = \sum_{t=1}^{T}\left(\left[\frac{\partial L}{\partial h_t}\right]_{new}\frac{\partial h_t}{\partial b^o}\right) = \left[\frac{\partial L}{\partial h_t}\right]_{new} \odot \tanh(c_t) \odot (o_t \odot (1 - o_t)) \tag{47}$$

For $\frac{\partial L}{\partial x_t}$,

$$\begin{aligned}\frac{\partial L}{\partial x_t} &= \left[\frac{\partial L}{\partial h_t}\right]_{new} \frac{\partial h_t}{\partial x_t} + \left[\frac{\partial L}{\partial c_t}\right]_{new} \frac{\partial c_t}{\partial x_t} \\ &= (W_x^f)^T \left[\left(\left[\frac{\partial L}{\partial c_t}\right]_{new} + \left[\frac{\partial L}{\partial h_t}\right]_{new} \odot o_t \odot (1 - \tanh(c_t) \odot \tanh(c_t))\right) \odot c_{t-1} \odot (f_t \odot (1 - f_t))\right] \\ &\quad + (W_x^i)^T \left[\left(\left[\frac{\partial L}{\partial c_t}\right]_{new} + \left[\frac{\partial L}{\partial h_t}\right]_{new} \odot o_t \odot (1 - \tanh(c_t) \odot \tanh(c_t))\right) \odot \tilde{c_t} \odot (i_t \odot (1 - i_t))\right] \\ &\quad + (W_x^c)^T \left[\left(\left[\frac{\partial L}{\partial c_t}\right]_{new} + \left[\frac{\partial L}{\partial h_t}\right]_{new} \odot o_t \odot (1 - \tanh(c_t) \odot \tanh(c_t))\right) \odot i_t \odot (1 - \tilde{c_t} \odot \tilde{c_t})\right] \\ &\quad + (W_x^o)^T \left[\left[\frac{\partial L}{\partial h_t}\right]_{new} \odot \tanh(c_t) \odot (o_t \odot (1 - o_t))\right]\end{aligned} \tag{48}$$

For $\frac{\partial L}{\partial h0}$,

$$\begin{aligned}\frac{\partial L}{\partial h_0} &= \left[\frac{\partial L}{\partial h_t}\right]_{new} \frac{\partial h_1}{\partial x_1} + \left[\frac{\partial L}{\partial c_1}\right]_{new} \frac{\partial c_1}{\partial h_1} \\ &= (W_h^f)^T \left[\left(\left[\frac{\partial L}{\partial c_1}\right]_{new} + \left[\frac{\partial L}{\partial h_1}\right]_{new} \odot o_1 \odot (1 - \tanh(c_1) \odot \tanh(c_t))\right) \odot c_0 \odot (f_1 \odot (1 - f_1))\right] \\ &\quad + (W_h^i)^T \left[\left(\left[\frac{\partial L}{\partial c_1}\right]_{new} + \left[\frac{\partial L}{\partial h_1}\right]_{new} \odot o_1 \odot (1 - \tanh(c_1) \odot \tanh(c_t))\right) \odot \tilde{c_1} \odot (i_1 \odot (1 - i_1))\right] \\ &\quad + (W_h^c)^T \left[\left(\left[\frac{\partial L}{\partial c_1}\right]_{new} + \left[\frac{\partial L}{\partial h_1}\right]_{new} \odot o_1 \odot (1 - \tanh(c_1) \odot \tanh(c_1))\right) \odot i_1 \odot (1 - \tilde{c_1} \odot \tilde{c_1})\right] \\ &\quad + (W_h^o)^T \left[\left[\frac{\partial L}{\partial h_1}\right]_{new} \odot \tanh(c_1) \odot (o_1 \odot (1 - o_1))\right]\end{aligned} \tag{49}$$

5 Image Captioning

5.4

The learnign curve of losses and learned captions are shown below.

5.4.1 RNN Results

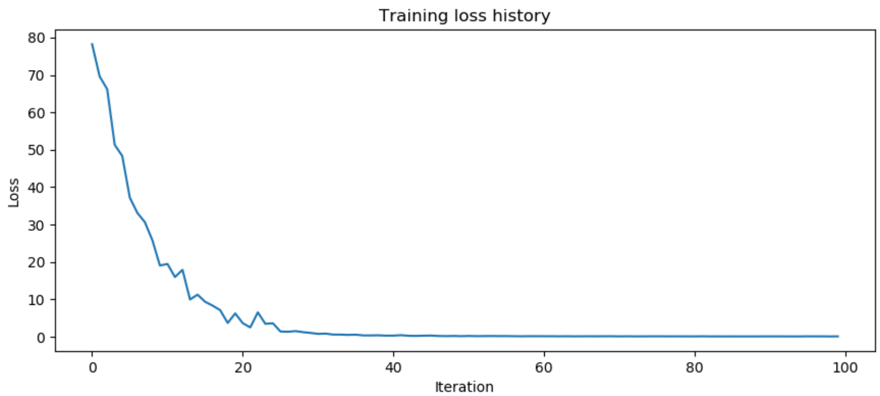


Figure 6: RNN traning loss

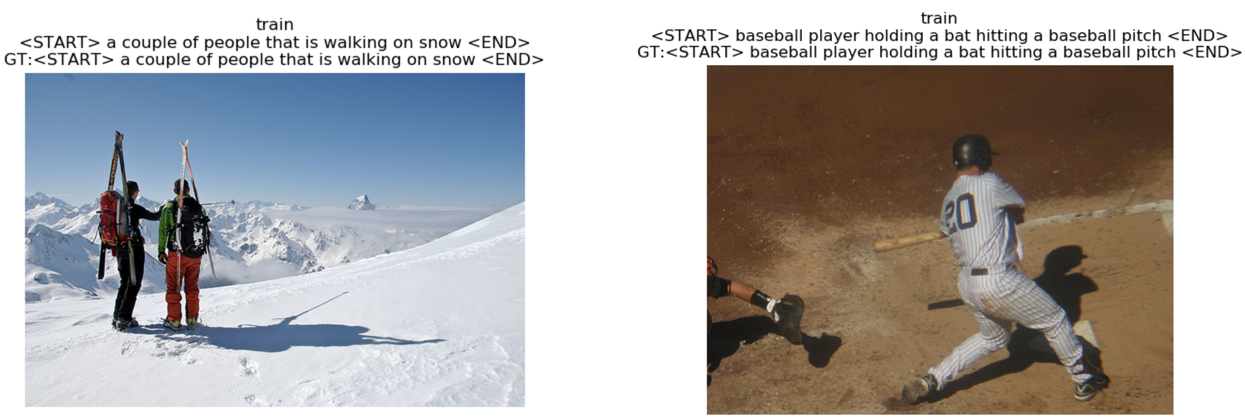


Figure 7: RNN learned captions for training dataset

Figure 8: RNN learned captions for validating dataset

5.4.2 LSTM Results

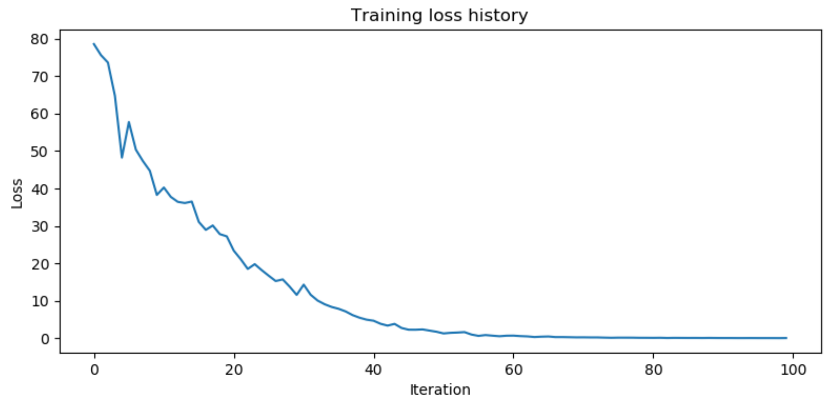


Figure 9: LSTM traning loss

Figure 10: LSTM learned captions for training dataset

Figure 11: LSTM learned captions for validating dataset

5.1 Text Classification

5.1.1 Bag of words

Using the bag-of-words method, and build a neural network structure as shown below,

Bag of words → Linear → sigmoid

the test accuracy is 0.953.

5.1.2 Word Embedding and Average Pool

Using an updated word embedding layer, and build a neural network structure as shown below,

Word embeddings → Average pooling → Linear → sigmoid

the test accuracy is 0.950.

5.1.3 Word Embedding with GloVe

Using the fixed embedding vector for every word from GloVe, and build a neural network structure as shown below,

Word embeddings (GloVe) → Average pooling → Linear → sigmoid

the test accuracy is 0.932.

5.1.4 RNN Model

Using the RNN model, and build a neural network structure as shown below,

Word embedding (GloVe) → RNN → Linear → sigmoid

the test accuracy is 0.922.

5.1.5 LSTM Model

Using the LSTM model, and build a neural network structure as shown below,

Word embedding (GloVe) → LSTM → Linear → sigmoid

the test accuracy is 0.937, which is considered as the baseline.