EECS 498/598 Deep Learning - Homework 2

March 08th, 2019

1 [15 points] Transfer learning

1. See the code
2. See the code
3. See the code
4. See the code
5. validation accuracy for 'finetune' scenario: 0.947712; validation accuracy for 'freeze' scenario: 0.954248. (There could be some randomness in the running, so reasonable value around the number is correct.)

2 [15 points] Style Transfer

1. See the code
2. See the code
3. See the code

tubingen+composition_vii tubingen+the_scream tubingen+starry_night

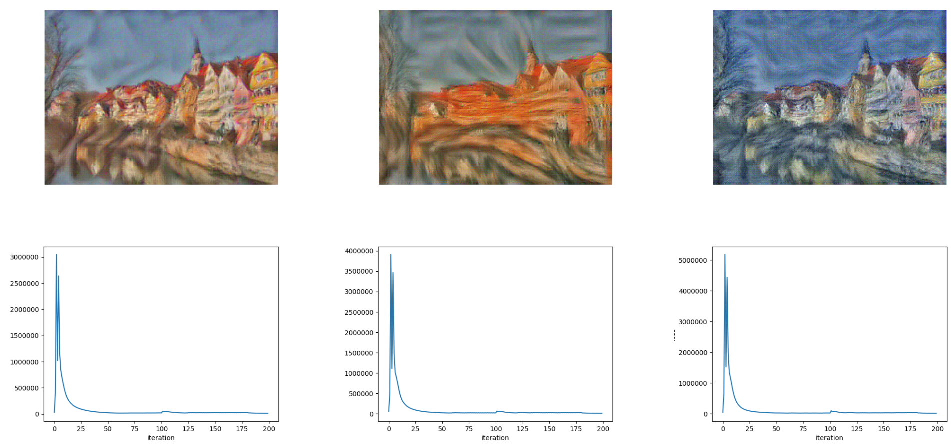

Figure 1: Question 2 style transfer results

4.

3 [15 points] Forward and Backward propagation module for RNN

1. See the code
2. Consider the forward pass, $y_t = W_x x_t + W_h h_{t-1} + b$ and $h_t = \tanh(y_t)$.

$$\frac{\partial L}{\partial y_t} = \frac{\partial L}{\partial h_t} * (1 - h_t * h_t) \text{ (Lemma .2)}$$

$$\frac{\partial L}{\partial W_x} = \frac{\partial L}{\partial y_t} x_t^T = \frac{\partial L}{\partial h_t} * (1 - h_t * h_t) x_t^T \text{ (Chain rule and Lemma .1)}$$

$$\frac{\partial L}{\partial x_t} = W_x^T \frac{\partial L}{\partial y_t} = W_x^T \frac{\partial L}{\partial h_t} * (1 - h_t * h_t) \text{ (Chain rule and Lemma .1)}$$

$$\frac{\partial L}{\partial W_h} = \frac{\partial L}{\partial y_t} h_{t-1}^T = \frac{\partial L}{\partial h_t} * (1 - h_t * h_t) h_{t-1}^T \text{ (Chain rule and Lemma .1)}$$

$$\frac{\partial L}{\partial h_{t-1}} = W_h^T \frac{\partial L}{\partial y_t} = W_h^T \frac{\partial L}{\partial h_t} * (1 - h_t * h_t) \text{ (Chain rule and Lemma .1)}$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial y_t} = \frac{\partial L}{\partial h_t} * (1 - h_t * h_t) \text{ (Chain rule and Lemma .1)}$$

3. See the code
4. **Possible answer 1**: Here $\frac{\partial L}{\partial h_t}$ is the gradient of total loss with respect to h_t. Then $\frac{\partial L}{\partial h_t} = \frac{\partial L_t}{\partial h_t}$ according to the notation in lecture slides.

$$\frac{\partial L}{\partial h_0} = W_h^T \frac{\partial L}{\partial h_1} * (1 - h_1 * h_1)$$

$$\frac{\partial L}{\partial x_t} = W_x^T \frac{\partial L}{\partial h_t} * (1 - h_t * h_t), \forall 1 \leq t \leq T$$

$$\frac{\partial L}{\partial W_x} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial W_x} = \sum_{t=1}^{T} \frac{\partial L}{\partial h_t} * (1 - h_t * h_t) x_t^T$$

$$\frac{\partial L}{\partial W_h} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial W_h} = \sum_{t=1}^{T} \frac{\partial L}{\partial h_t} * (1 - h_t * h_t) h_{t-1}^T$$

$$\frac{\partial L}{\partial b} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial b} = \sum_{t=1}^{T} \frac{\partial L}{\partial h_t} * (1 - h_t * h_t)$$

Possible answer 2: Here $\frac{\partial L}{\partial h_t}$ is the gradient of total loss with respect to the hidden feature from RNN. Then $\frac{\partial L}{\partial h_t} = \frac{\partial L_t}{\partial h_t}$ according to the notation in lecture slides. $\frac{\partial D(y_t, \hat{y_t})}{\partial h_t}$ is the upstream gradients of all hidden states as `dh` in function `rnn_backward`.

Recursively get the gradient of the loss accumulated from t up to T with respect to each hidden state at step t

At the last step,

$$\frac{\partial L}{\partial h_T} = \frac{\partial L_T}{\partial h_T} = \frac{\partial D(y_T, \hat{y_T})}{\partial h_T}$$

At the intermediate step,

$$\frac{\partial L}{\partial h_t} = \frac{\partial L_t}{\partial h_t} = \frac{\partial L_{t+1}}{\partial h_t} + \frac{\partial D(y_t, \hat{y_t})}{\partial h_t}\ , \forall 1 \leq t \leq T-1$$

$$\frac{\partial L_{t+1}}{\partial h_t} = W_h^T \frac{\partial L_{t+1}}{\partial h_{t+1}} * (1 - h_{t+1} * h_{t+1})\ , \forall 1 \leq t \leq T-1 \text{ (See answer in part 2)}$$

At the first step, we know

$$\frac{\partial L}{\partial h_0} = \frac{\partial L_1}{\partial h_0} = W_h^T \frac{\partial L}{\partial h_1} * (1 - h_1 * h_1)$$

Calculate the gradient for x_t, W_x, W_h, b, based on the gradient of loss accumulated from t up to T with respect to h_t

$$\frac{\partial L}{\partial x_t} = \frac{\partial L_t}{\partial x_t} = W_x^T \frac{\partial L}{\partial h_t} * (1 - h_t * h_t), \forall 1 \leq t \leq T$$

$$\frac{\partial L}{\partial W_x} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial W_x} = \sum_{t=1}^{T} \frac{\partial L}{\partial h_t} * (1 - h_t * h_t) x_t^T$$

$$\frac{\partial L}{\partial W_h} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial W_h} = \sum_{t=1}^{T} \frac{\partial L}{\partial h_t} * (1 - h_t * h_t) h_{t-1}^T$$

$$\frac{\partial L}{\partial b} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial b} = \sum_{t=1}^{T} \frac{\partial L}{\partial h_t} * (1 - h_t * h_t) h_{t-1}^T$$

4 [15 points] Forward and Backward propagation module for LSTM

1. See the code
2. As introduced in lecture, we can calculate back-propagation gradient based on the graph of computation in Figure 2. For this single step in LSTM, $L = L_t$ according to notation in lecture slides.

   According to the last two equation in forward pass, we have:

$$\frac{\partial L}{\partial o_t} = \frac{\partial L}{\partial h_t} * \tanh(c_t)$$

$$\frac{\partial L}{\partial c_t} = \frac{\partial L}{\partial h_t} * o_t * (1 - o_t^2) + \frac{\partial L_{t+1}}{\partial c_t}$$

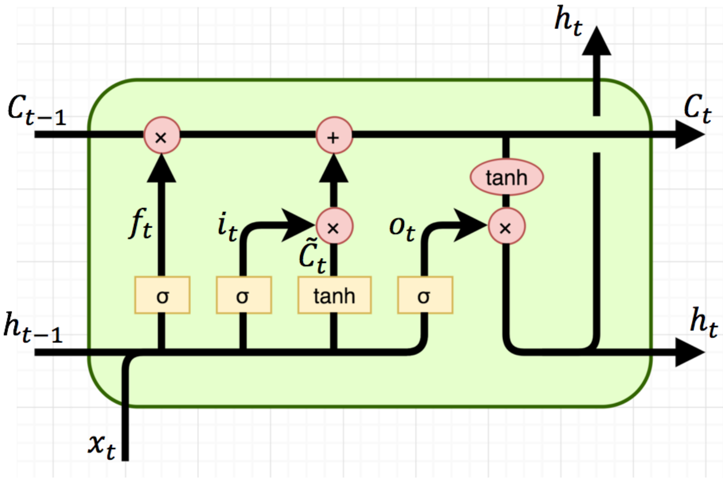

Figure 2: Computation graph for LSTM cell

$$\frac{\partial L}{\partial f_t} = \frac{\partial L}{\partial c_t} * c_{t-1}$$
$$\frac{\partial L}{\partial c_{t-1}} = \frac{\partial L}{\partial c_t} * f_t$$
$$\frac{\partial L}{\partial i_t} = \frac{\partial L}{\partial c_t} * \tilde{c_t}$$
$$\frac{\partial L}{\partial \tilde{c_t}} = \frac{\partial L}{\partial c_t} * i_t$$

According to the equation of forget gate in forward pass, we have:

$$\frac{\partial L}{\partial W_x^f} = \frac{\partial L}{\partial f_t} * f_t * (1 - f_t) x_t^T \text{ (Chain rule, Lemma .1, Lemma .3)}$$

$$\frac{\partial L}{\partial W_h^f} = \frac{\partial L}{\partial f_t} * f_t * (1 - f_t) h_{t-1}^T$$
$$\frac{\partial L}{\partial b^f} = \frac{\partial L}{\partial f_t} * f_t * (1 - f_t)$$

According to the equation of input gate in forward pass, we have:

$$\frac{\partial L}{\partial W_x^i} = \frac{\partial L}{\partial i_t} * i_t * (1 - i_t) x_t^T$$

$$\frac{\partial L}{\partial W_h^i} = \frac{\partial L}{\partial i_t} * i_t * (1 - i_t) h_{t-1}^T$$
$$\frac{\partial L}{\partial b^i} = \frac{\partial L}{\partial i_t} * i_t * (1 - i_t)$$

According to the equation of concurrent gate in forward pass, we have:

$$\frac{\partial L}{\partial W_x^c} = \frac{\partial L}{\partial \tilde{c_t}} * (1 - \tilde{c_t}^2)x_t^T$$

$$\frac{\partial L}{\partial W_h^c} = \frac{\partial L}{\partial \tilde{c_t}} * (1 - \tilde{c_t}^2)h_{t-1}^T$$

$$\frac{\partial L}{\partial b^c} = \frac{\partial L}{\partial \tilde{c_t}} * (1 - \tilde{c_t}^2)$$

According to the equation of output gate in forward pass, we have:

$$\frac{\partial L}{\partial W_x^o} = \frac{\partial L}{\partial o_t} * o_t * (1 - o_t)x_t^T$$

$$\frac{\partial L}{\partial W_h^o} = \frac{\partial L}{\partial o_t} * o_t * (1 - o_t)h_{t-1}^T$$

$$\frac{\partial L}{\partial b^o} = \frac{\partial L}{\partial o_t} * o_t * (1 - o_t)$$

According to the first 4 equations in forward pass and Lemmas:

$$\frac{\partial L}{\partial x_t} = (W_x^f)^T \frac{\partial L}{\partial f_t} * f_t * (1 - f_t) + (W_x^i)^T \frac{\partial L}{\partial i_t} * i_t * (1 - i_t) + (W_x^c)^T \frac{\partial L}{\partial \tilde{c_t}} * (1 - \tilde{c_t}^2) + (W_x^o)^T \frac{\partial L}{\partial o_t} * o_t * (1 - o_t)$$

$$\frac{\partial L}{\partial h_{t-1}} = (W_h^f)^T \frac{\partial L}{\partial f_t} * f_t * (1 - f_t) + (W_h^i)^T \frac{\partial L}{\partial i_t} * i_t * (1 - i_t) + (W_h^c)^T \frac{\partial L}{\partial \tilde{c_t}} * (1 - \tilde{c_t}^2) + (W_h^o)^T \frac{\partial L}{\partial o_t} * o_t * (1 - o_t)$$

3. See the code
4. As introduced in lecture, we can calculate back-propagation gradient based on the graph of computation in Figure 3.

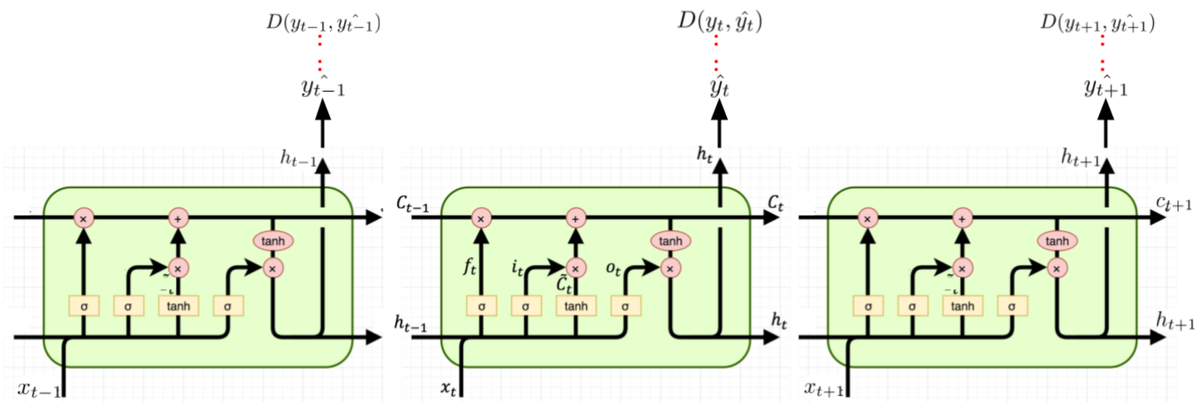

Figure 3: Computation graph for LSTM with multiple steps

Possible answer 1: Here $\frac{\partial L}{\partial h_t}$ is the gradient of total loss with respect to h_t. Then $\frac{\partial L}{\partial h_t} = \frac{\partial L_t}{\partial h_t}$ according to the notation in lecture slides.

Based on the anwser in part 2:

$$\frac{\partial L}{\partial o_t} = \frac{\partial L}{\partial h_t} * \tanh(c_t)$$

$$\frac{\partial L}{\partial c_t} = \frac{\partial L}{\partial h_t} * o_t * (1 - \tanh(c_t)^2) + \frac{\partial L_{t+1}}{\partial c_t}, \forall 0 \leq t \leq T-1, \frac{\partial L}{\partial c_T} = \frac{\partial L}{\partial h_T} * o_T * (1 - \tanh(c_T)^2)$$

$$\frac{\partial L}{\partial f_t} = \frac{\partial L}{\partial c_t} * c_{t-1}$$

$$\frac{\partial L_t}{\partial c_{t-1}} = \frac{\partial L}{\partial c_t} * f_t$$

$$\frac{\partial L}{\partial i_t} = \frac{\partial L}{\partial c_t} * \tilde{c_t}$$

$$\frac{\partial L}{\partial \tilde{c_t}} = \frac{\partial L}{\partial c_t} * i_t$$

$$\frac{\partial L_t}{\partial c_{t-1}} = \frac{\partial L}{\partial c_t} * f_t$$

At the first step, we know

$$\frac{\partial L}{\partial h_0} = (W_h^f)^T \frac{\partial L}{\partial f_1} * f_1 * (1 - f_1) + (W_h^i)^T \frac{\partial L}{\partial i_1} * i_1 * (1 - i_1) + (W_h^c)^T \frac{\partial L}{\partial \tilde{c_1}} * (1 - \tilde{c_1}^2) + (W_h^o)^T \frac{\partial L}{\partial o_t} * o_1 * (1 - o_1)$$

Calculate the gradient for x_t, W_x, W_h, b, based on the gradient of loss accumulated from t up to T with respect to h_t

$$\frac{\partial L}{\partial x_t} = (W_x^f)^T \frac{\partial L}{\partial f_t} * f_t * (1 - f_t) + (W_x^i)^T \frac{\partial L}{\partial i_t} * i_t * (1 - i_t) + (W_x^c)^T \frac{\partial L}{\partial \tilde{c_t}} * (1 - \tilde{c_t}^2) + (W_x^o)^T \frac{\partial L}{\partial o_t} * o_t * (1 - o_t), \forall 1 \leq t \leq T$$

For the forget gate,

$$\frac{\partial L}{\partial W_x^f} = \sum_{t=1}^{T} \frac{\partial L}{\partial f_t} * f_t * (1 - f_t) x_t^T$$

$$\frac{\partial L}{\partial W_h^f} = \sum_{t=1}^{T} \frac{\partial L}{\partial f_t} * f_t * (1 - f_t) h_{t-1}^T$$

$$\frac{\partial L}{\partial b^f} = \sum_{t=1}^{T} \frac{\partial L}{\partial f_t} * f_t * (1 - f_t)$$

For the input gate

$$\frac{\partial L}{\partial W_x^i} = \sum_{t=1}^{T} \frac{\partial L}{\partial i_t} * i_t * (1 - i_t) x_t^T$$

$$\frac{\partial L}{\partial W_h^i} = \sum_{t=1}^{T} \frac{\partial L}{\partial i_t} * i_t * (1 - i_t) h_{t-1}^T$$

$$\frac{\partial L}{\partial b^i} = \sum_{t=1}^{T} \frac{\partial L}{\partial i_t} * i_t * (1 - i_t)$$

For the concurrent gate

$$\frac{\partial L}{\partial W_x^c} = \sum_{t=1}^{T} \frac{\partial L}{\partial \tilde{c_t}} * (1 - \tilde{c_t}^2) x_t^T$$

$$\frac{\partial L}{\partial W_h^c} = \sum_{t=1}^{T} \frac{\partial L}{\partial \tilde{c_t}} * (1 - \tilde{c_t}^2) h_{t-1}^T$$

$$\frac{\partial L}{\partial b^c} = \sum_{t=1}^{T} \frac{\partial L}{\partial \tilde{c_t}} * (1 - \tilde{c_t}^2)$$

For the output gate

$$\frac{\partial L}{\partial W_x^o} = \sum_{t=1}^{T} \frac{\partial L}{\partial o_t} * o_t * (1 - o_t) x_t^T$$

$$\frac{\partial L}{\partial W_h^o} = \sum_{t=1}^{T} \frac{\partial L}{\partial o_t} * o_t * (1 - o_t) h_{t-1}^T$$

$$\frac{\partial L}{\partial b^o} = \sum_{t=1}^{T} \frac{\partial L}{\partial o_t} * o_t * (1 - o_t)$$

Possible answer 2: Here $\frac{\partial L}{\partial h_t}$ is the gradient of total loss with respect to the hidden feature from LSTM. Then $\frac{\partial L}{\partial h_t} = \frac{\partial L_t}{\partial h_t}$ according to the notation in lecture slides. $\frac{\partial D(y_t, \hat{y_t})}{\partial h_t}$ is the upstream gradients of all hidden states as `dh` in function `lstm_backward`.

Recursively get the gradient of the loss accumulated from t up to T with respect to each hidden state at step t.

At the last step,

$$\frac{\partial L}{\partial h_T} = \frac{\partial L_T}{\partial h_T} = \frac{\partial D(y_T, \hat{y_T})}{\partial h_T}$$

At the intermediate step,

$$\frac{\partial L}{\partial h_t} = \frac{\partial L_t}{\partial h_t} = \frac{\partial L_{t+1}}{\partial h_t} + \frac{\partial D(y_t, \hat{y_t})}{\partial h_t}, \ , \forall 1 \le t \le T-1$$

where $\frac{\partial L_{t+1}}{\partial h_t}$ can be calculated based on the answer in part 2.

$$\frac{\partial L_t}{\partial o_t} = \frac{\partial L_t}{\partial h_t} * \tanh(c_t)$$

$$\frac{\partial L_t}{\partial c_t} = \frac{\partial L_t}{\partial h_t} * o_t * (1 - \tanh(c_t)^2) + \frac{\partial L_{t+1}}{\partial c_t}, \forall 0 \le t \le T-1, \frac{\partial L_T}{\partial c_T} = \frac{\partial L_T}{\partial h_T} * o_T * (1 - \tanh(c_T)^2)$$

$$\frac{\partial L_t}{\partial f_t} = \frac{\partial L_t}{\partial c_t} * c_{t-1}$$

$$\frac{\partial L_t}{\partial c_{t-1}} = \frac{\partial L_t}{\partial c_t} * f_t$$
$$\frac{\partial L_t}{\partial i_t} = \frac{\partial L_t}{\partial c_t} * \tilde{c_t}$$
$$\frac{\partial L_t}{\partial \tilde{c_t}} = \frac{\partial L_t}{\partial c_t} * i_t$$
$$\frac{\partial L_t}{\partial c_{t-1}} = \frac{\partial L_t}{\partial c_t} * f_t$$

$$\begin{aligned}\frac{\partial L_t}{\partial h_{t-1}} &= (W_h^f)^T \frac{\partial L_t}{\partial f_t} * f_t * (1 - f_t) + (W_h^i)^T \frac{\partial L_t}{\partial i_t} * i_t * (1 - i_t) \\ &+ (W_h^c)^T \frac{\partial L_t}{\partial \tilde{c_t}} * (1 - \tilde{c_t}^2) + (W_h^o)^T \frac{\partial L_t}{\partial o_t} * o_t * (1 - o_t) \ , \forall 1 \le t \le T\end{aligned}$$

At the first step, we know

$$\frac{\partial L}{\partial h_0} = \frac{\partial L_1}{\partial h_0} = (W_h^f)^T \frac{\partial L_1}{\partial f_1} * f_1 * (1 - f_1) + (W_h^i)^T \frac{\partial L_1}{\partial i_1} * i_1 * (1 - i_1) + (W_h^c)^T \frac{\partial L_1}{\partial \tilde{c_1}} * (1 - \tilde{c_1}^2) + (W_h^o)^T \frac{\partial L_1}{\partial o_t} * o_1 * (1 - o_1)$$

Calculate the gradient for x_t, W_x, W_h, b, based on the gradient of loss accumulated from t up to T with respect to h_t

$$\frac{\partial L}{\partial x_t} = \frac{\partial L_t}{\partial x_t} = (W_x^f)^T \frac{\partial L_t}{\partial f_t} * f_t * (1 - f_t) + (W_x^i)^T \frac{\partial L_t}{\partial i_t} * i_t * (1 - i_t) + (W_x^c)^T \frac{\partial L_t}{\partial \tilde{c_t}} * (1 - \tilde{c_t}^2) + (W_x^o)^T \frac{\partial L_t}{\partial o_t} * o_t * (1 - o_t)$$

For the forget gate,

$$\frac{\partial L}{\partial W_x^f} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial W_x^f} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial f_t} * f_t * (1 - f_t) x_t^T$$
$$\frac{\partial L}{\partial W_h^f} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial W_h^f} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial f_t} * f_t * (1 - f_t) h_{t-1}^T$$
$$\frac{\partial L}{\partial b^f} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial b^f} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial f_t} * f_t * (1 - f_t)$$

For the input gate

$$\frac{\partial L}{\partial W_x^i} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial W_x^i} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial i_t} * i_t * (1 - i_t) x_t^T$$
$$\frac{\partial L}{\partial W_h^i} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial W_h^i} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial i_t} * i_t * (1 - i_t) h_{t-1}^T$$
$$\frac{\partial L}{\partial b^i} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial b^i} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial i_t} * i_t * (1 - i_t)$$

For the concurrent gate

$$\frac{\partial L}{\partial W_x^c} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial W_x^c} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial \tilde{c_t}} * (1 - \tilde{c_t}^2) x_t^T$$

$$\frac{\partial L}{\partial W_h^c} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial W_h^c} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial \tilde{c_t}} * (1 - \tilde{c_t}^2) h_{t-1}^T$$

$$\frac{\partial L}{\partial b^c} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial b^c} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial \tilde{c_t}} * (1 - \tilde{c_t}^2)$$

For the output gate

$$\frac{\partial L}{\partial W_x^o} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial W_x^o} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial o_t} * o_t * (1 - o_t) x_t^T$$

$$\frac{\partial L}{\partial W_h^o} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial W_h^o} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial o_t} * o_t * (1 - o_t) h_{t-1}^T$$

$$\frac{\partial L}{\partial b^o} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial b^o} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial o_t} * o_t * (1 - o_t)$$

5 [20 points] Application to Image Captioning

1. See the code
2. See the code
3. See the code
4. As shown in Figure 4, the training overfits on training dataset. So caption for image from training data is good, but caption for image from validation data is not reasonable

6 [20 points] Application to text classification

The test accuracy for all 5 parts shall be around 90%. There won't be significant difference in performances. A more specific breakdown could be:

1. Bag of Words: test accuracy 92%
2. Word Embeddings: test accuracy 90%
3. GloVe: test accuracy 93%
4. RNN: test accuracy 95%
5. LSTM: test accuracy 95%

Note that a more generous threshold will be set when grading this problem.

Lemma .1. *Assume that* $y = Wx + b$ *where* $y \in \mathbb{R}^m, W \in \mathbb{R}^{m \times d}, x \in \mathbb{R}^d, b \in \mathbb{R}^m$, *then we have* $\frac{\partial L}{\partial x} = W^T \frac{\partial L}{\partial y}$, $\frac{\partial L}{\partial W} = \frac{\partial L}{\partial y} x^T$, $\frac{\partial L}{\partial b} = \frac{\partial L}{\partial y}$

Proof.

$$
\begin{aligned}
& y = Wx + b \\
& \Rightarrow y_i = \sum_j W_{ij} x_j + b_j
\end{aligned}
$$

Computing $\frac{\partial L}{\partial W}$:

$$
\begin{aligned}
\frac{\partial L}{\partial W_{mn}} &= \sum_i \frac{\partial L}{\partial y_i} \frac{\partial y_i}{\partial W_{mn}} \text{ (Chain rule)} \\
&= \sum_i \frac{\partial L}{\partial y_i} \frac{\partial y_i}{\partial W_{mn}} \delta[i = m] \\
&= \frac{\partial L}{\partial y_m} \frac{\partial y_m}{\partial W_{mn}} \\
&= \frac{\partial L}{\partial y_m} x_n \\
\Rightarrow \frac{\partial L}{\partial W} &= \frac{\partial L}{\partial y} x^T \text{ (Outer product)}
\end{aligned}
$$

Computing $\frac{\partial L}{\partial b}$:

$$
\begin{aligned}
\frac{\partial L}{\partial b_p} &= \frac{\partial L}{\partial y_p} \frac{\partial y_p}{\partial b_p} \\
&= \frac{\partial L}{\partial y_p} \cdot 1 \\
\Rightarrow \frac{\partial L}{\partial b} &= \frac{\partial L}{\partial y}
\end{aligned}
$$

Computing $\frac{\partial L}{\partial X}$:

$$
\begin{aligned}
\frac{\partial L}{\partial x_p} &= \sum_i \frac{\partial L}{\partial y_i} \frac{\partial y_i}{\partial x_p} \\
&= \sum_i \frac{\partial L}{\partial y_i} W_{ip} \\
\Rightarrow \frac{\partial L}{\partial x} &= W^T \frac{\partial L}{\partial y}
\end{aligned}
$$

□

Lemma .2. *Assume* $y = \tanh(x)$ *where* $x \in \mathbb{R}^n, y \in \mathbb{R}^n$, *then we have* $\frac{\partial L}{\partial x} = \frac{\partial L}{\partial y} * (1 - \tanh(x) * \tanh(x)) = \frac{\partial L}{\partial y} * (1 - y * y)$ *where* $*$ *means elementwise multiplication*

Proof.

$$\begin{aligned}
\frac{\partial L}{\partial x_p} &= \sum_i \frac{\partial L}{\partial y_i}\frac{\partial y_i}{\partial x_p} \\
&= \sum_i \frac{\partial L}{\partial y_i}\frac{\partial y_i}{\partial x_p}\delta[i=p] \\
&= \frac{\partial L}{\partial y_p}\frac{\partial y_p}{\partial x_p} \\
&= \frac{\partial L}{\partial y_p}(1-\tanh(x_p)^2) \\
\Rightarrow \frac{\partial L}{\partial x} &= \frac{\partial L}{\partial y} * (1-\tanh(x)*\tanh(x)))
\end{aligned}$$

□

Lemma .3. *Assume $y = sigmoid(x)$ where $x \in \mathbb{R}^n, y \in \mathbb{R}^n$, then we have $\frac{\partial L}{\partial x} = \frac{\partial L}{\partial y} * sigmoid(x) * (1 - sigmoid(x)) = \frac{\partial L}{\partial y} * y * (1-y)$ where $*$ means elementwise multiplication*

Proof.

$$\begin{aligned}
\frac{\partial L}{\partial x_p} &= \sum_i \frac{\partial L}{\partial y_i}\frac{\partial y_i}{\partial x_p} \\
&= \sum_i \frac{\partial L}{\partial y_i}\frac{\partial y_i}{\partial x_p}\delta[i=p] \\
&= \frac{\partial L}{\partial y_p}\frac{\partial y_p}{\partial x_p} \\
&= \frac{\partial L}{\partial y_p} sigmoid(x_p)(1-sigmoid(x_p)) \\
\Rightarrow \frac{\partial L}{\partial x} &= \frac{\partial L}{\partial y} * sigmoid(x) * (1-sigmoid(x))
\end{aligned}$$

□

RNN LSTM

Figure 4: Question 5 image captioning results