EECS 598 Deep Learning

HW2

Chuan Cen / chuancen / March 3, 2019

Q1. Transfer learning

1. Finetune the pre-trained model

   Best val Acc: 0.960784

2. Freeze the parameters in pre-trained model and train the final fc layer

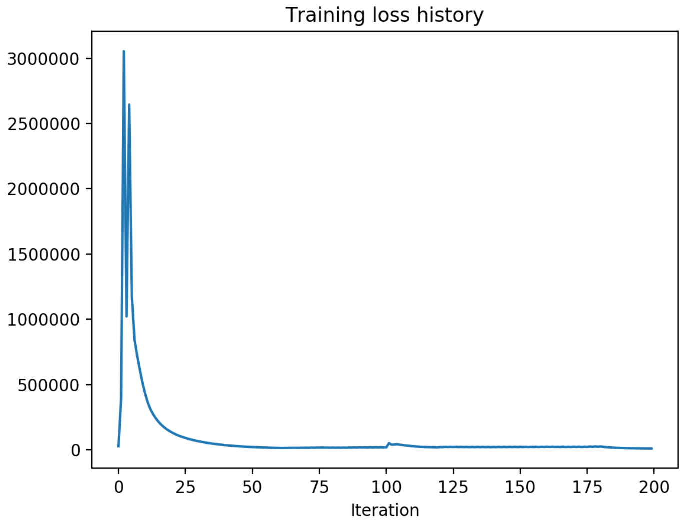
   Best val Acc: 0.954248

Q2. Style transfer

1. Composition VII + Tubingen

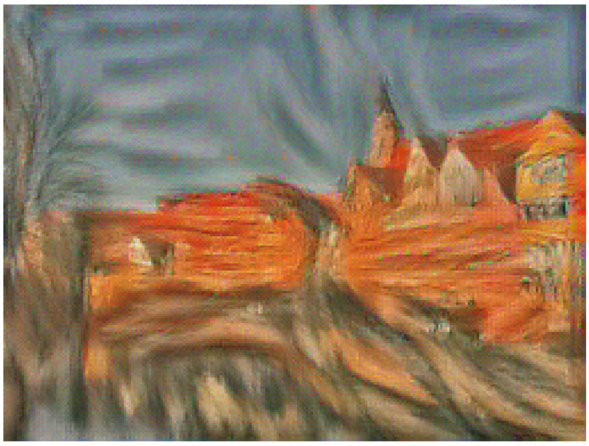
   output:

   learning curve:

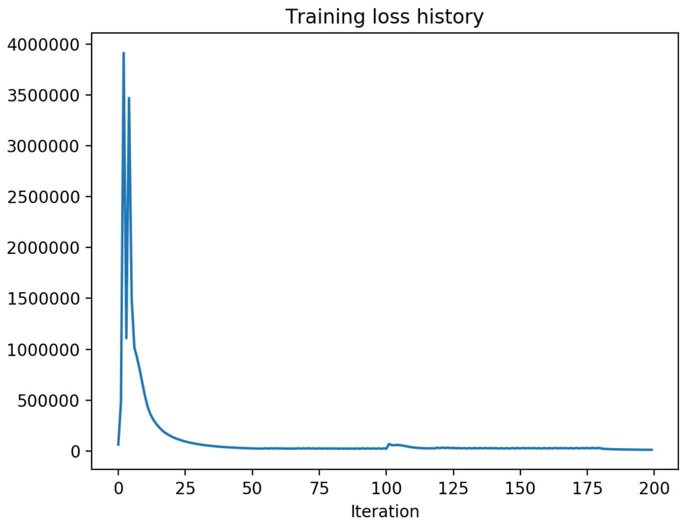


2. Scream + Tubingen

output:

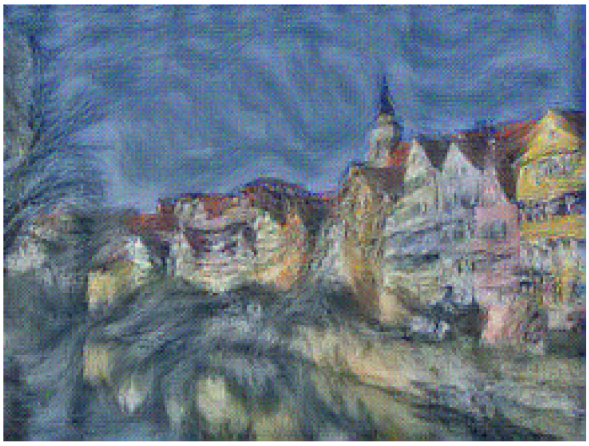


learning curve:

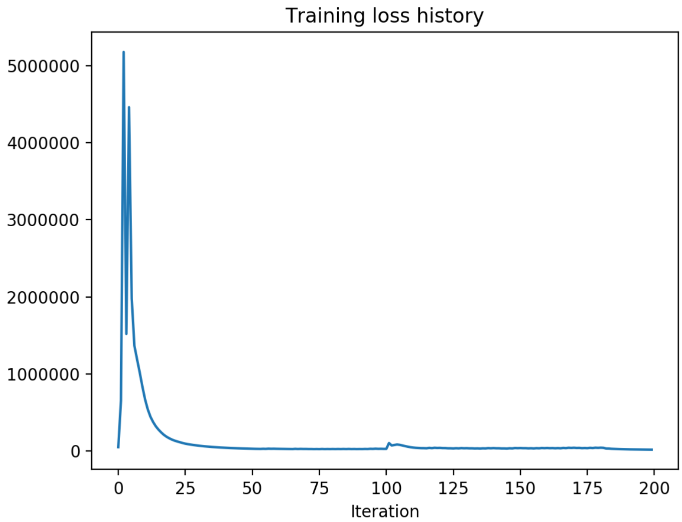


3. Starry Night + Tubingen

output:



learning curve:



Q3.

2.

$$\frac{\partial \mathrm{L}}{\partial \mathrm{x_t}} = \frac{\partial L}{\partial h_t}\frac{\partial h_t}{\partial x_t}$$

$$= \frac{\partial L}{\partial h_t}\frac{\partial h_t}{\partial Z}\frac{\partial Z}{\partial x_t}, \quad \text{where } Z = W_x x_t + W_h h_{t-1} + b$$

$$= \frac{\partial L}{\partial h_t} diag\left(\vec{1} - \tanh^2(W_x x_t + W_h h_{t-1} + b)\ \right)W_x$$

$$= \left(\frac{\partial L}{\partial h_t} \odot \left(\vec{1} - h_t^2\right)\right) W_x \qquad \text{note that } \odot \text{ gives a row vector}$$

$$\frac{\partial \mathrm{L}}{\partial \mathrm{W_x}} = \frac{\partial L}{\partial h_t}\frac{\partial h_t}{\partial W_x}$$

$$= \sum_i \frac{\partial L}{\partial h_{t,i}}\frac{\partial h_{t,i}}{\partial W_x}$$

$$= \left(\frac{\partial L}{\partial h_t} \odot \left(\vec{1} - h_t^2\right)\right)^T x_t^T$$

$$\frac{\partial \mathrm{L}}{\partial \mathrm{h_{t-1}}} = \frac{\partial L}{\partial h_t}\frac{\partial h_t}{\partial h_{t-1}}$$

$$= \frac{\partial L}{\partial h_t}\frac{\partial h_t}{\partial Z}\frac{\partial Z}{\partial h_{t-1}}, \quad \text{where } Z = W_x x_t + W_h h_{t-1} + b$$

$$= \left(\frac{\partial L}{\partial h_t} \odot \left(\vec{1} - h_t^2\right)\right) W_h$$

$$\frac{\partial \mathrm{L}}{\partial \mathrm{W_h}} = \frac{\partial L}{\partial h_t}\frac{\partial h_t}{\partial W_h}$$

$$= \sum_i \frac{\partial L}{\partial h_{t,i}}\frac{\partial h_{t,i}}{\partial W_h}$$

$$= \left(\frac{\partial L}{\partial h_t} \odot \left(\vec{1} - h_t^2\right)\right)^T h_{t-1}^T$$

$$\frac{\partial \mathrm{L}}{\partial \mathrm{b}} = \frac{\partial L}{\partial h_t}\frac{\partial h_t}{\partial b}$$

$$= \frac{\partial L}{\partial h_t}\frac{\partial h_t}{\partial Z}\frac{\partial Z}{\partial b}, \quad \text{where } Z = W_x x_t + W_h h_{t-1} + b$$

$$= \frac{\partial L}{\partial h_t} \odot \left(\vec{1} - h_t^2\right)$$

4.

In this question, $\frac{\partial \mathrm{L}}{\partial \mathrm{h_t}}$ is actually $\frac{\partial \mathrm{D}(\mathrm{y}_t, \hat{\mathrm{y}}_t)}{\partial h_t}$, and we denote $\frac{\partial \mathrm{L_t}}{\partial \mathrm{h_t}}$ as the partial derivative back-propagated from all time steps from t to T. We first derive $\frac{\partial \mathrm{L_t}}{\partial \mathrm{h_t}}$ using $\frac{\partial \mathrm{D}(\mathrm{y}_t, \hat{\mathrm{y}}_t)}{\partial h_t}$ (i.e. $\frac{\partial \mathrm{L}}{\partial \mathrm{h_t}}$ in the question), then all the $\frac{\partial \mathrm{L_t}}{\partial \mathrm{h_t}}$ below can be substituted to become an answer.

$$\frac{\partial \mathrm{L_t}}{\partial \mathrm{h_t}} = \frac{\partial \mathrm{D}(\mathrm{y}_t, \hat{\mathrm{y}}_t)}{\partial h_t} + \frac{\partial L_{t+1}}{\partial h_t}$$

$$= \frac{\partial \mathrm{D}(\mathrm{y}_t, \hat{\mathrm{y}}_t)}{\partial h_t} + \frac{\partial L_{t+1}}{\partial h_{t+1}}\frac{\partial h_{t+1}}{\partial h_t}$$

$$= \frac{\partial \mathrm{D}(\mathrm{y}_t, \hat{\mathrm{y}}_t)}{\partial h_t} + \frac{\partial L_{t+1}}{\partial h_{t+1}} \odot \left(\left(\vec{1} - h_t^2\right)^T \mathrm{W_h}\right)$$

$$= \frac{\partial \mathrm{D}(\mathrm{y}_t, \hat{\mathrm{y}}_t)}{\partial h_t} + \sum_{\tau=t+1}^{\mathrm{T}} \frac{\partial \mathrm{D}(\mathrm{y}_\tau, \hat{\mathrm{y}}_\tau)}{\partial h_\tau} \odot \prod_{\beta=t}^{\tau-1} \left(\left(\vec{1} - h_\beta^2\right)^T \mathrm{W_h}\right)$$

Then we express all the terms we want using $\frac{\partial \mathrm{L_t}}{\partial \mathrm{h_t}}$:

$\frac{\partial \mathrm{L_t}}{\partial \mathrm{x_t}}$:

$$\frac{\partial \mathrm{L_t}}{\partial \mathrm{x_t}} = \left(\frac{\partial L_t}{\partial h_t} \odot \left(\vec{1} - h_t^2\right)\right) W_x$$

$\frac{\partial \mathrm{L_1}}{\partial \mathrm{W_x}}$:

$$\frac{\partial \mathrm{L_1}}{\partial \mathrm{W_x}} = \sum_{t=1}^{\mathrm{T}} \frac{\partial L_t}{\partial h_t}\frac{\partial h_t}{\partial W_x}$$

$$= \sum_{t=1}^{\mathrm{T}} \left(\frac{\partial L_t}{\partial h_t} \odot \left(\vec{1} - h_t^2\right)\right)^T x_t^T$$

$\frac{\partial \mathrm{L_1}}{\partial \mathrm{W_h}}$:

$$\frac{\partial \mathrm{L}_1}{\partial \mathrm{W}_\mathrm{h}} = \sum_{\mathrm{t}=1}^{\mathrm{T}} \frac{\partial L_t}{\partial h_t} \frac{\partial h_t}{\partial W_h}$$

$$= \sum_{\mathrm{t}=1}^{\mathrm{T}} \left(\frac{\partial L_t}{\partial h_t} \odot \left(\vec{1} - h_t^2 \right) \right)^T h_{t-1}^T$$

$\frac{\partial \mathrm{L}_1}{\partial \mathrm{b}}$:

$$\frac{\partial \mathrm{L}_1}{\partial \mathrm{b}} = \sum_{\mathrm{t}=1}^{\mathrm{T}} \frac{\partial L_t}{\partial h_t} \frac{\partial h_t}{\partial b}$$

$$= \sum_{\mathrm{t}=1}^{\mathrm{T}} \frac{\partial L_t}{\partial h_t} \frac{\partial h_t}{\partial Z} \frac{\partial Z}{\partial b}, \quad \text{where } Z = W_x x_t + W_h h_{t-1} + b$$

$$= \sum_{\mathrm{t}=1}^{\mathrm{T}} \frac{\partial L_t}{\partial h_t} \odot \left(\vec{1} - h_t^2 \right)$$

$\frac{\partial \mathrm{L}_1}{\partial \mathrm{h}_0}$:

$$\frac{\partial \mathrm{L}_1}{\partial \mathrm{h}_0} = \frac{\partial L_1}{\partial h_1} \frac{\partial h_1}{\partial h_0}$$

$$= \frac{\partial L_1}{\partial h_1} \frac{\partial h_1}{\partial Z} \frac{\partial Z}{\partial h_0}, \quad \text{where } Z = W_x x_t + W_h h_{t-1} + b$$

$$= \frac{\partial L_1}{\partial h_1} \odot \left(\vec{1} - h_1^2 \right) W_h$$

Q4.

2. Derive for single time step.

In this question, "L" is assumed to include "all the losses that we need to consider". This is different with what we assume in question 4 below.

- $\frac{\partial \mathrm{L}}{\partial \mathrm{x}_\mathrm{t}}$ **:**

$$\frac{\partial \mathrm{L}}{\partial \mathrm{x}_\mathrm{t}} = \frac{\partial \mathrm{L}}{\partial \mathrm{f}_\mathrm{t}} \frac{\partial f_t}{\partial x_t} + \frac{\partial \mathrm{L}}{\partial \mathrm{i}_\mathrm{t}} \frac{\partial i_t}{\partial x_t} + \frac{\partial \mathrm{L}}{\partial \tilde{\mathrm{c}}_\mathrm{t}} \frac{\partial \tilde{c}_t}{\partial x_t} + \frac{\partial \mathrm{L}}{\partial \mathrm{o}_\mathrm{t}} \frac{\partial o_t}{\partial x_t}$$

$$= \frac{\partial \mathrm{L}}{\partial \mathrm{c}_\mathrm{t}} \frac{\partial c_t}{\partial f_t} \frac{\partial f_t}{\partial x_t} + \frac{\partial \mathrm{L}}{\partial c_t} \frac{\partial c_t}{\partial i_t} \frac{\partial i_t}{\partial x_t} + \frac{\partial \mathrm{L}}{\partial c_t} \frac{\partial c_t}{\partial \tilde{\mathrm{c}}_\mathrm{t}} \frac{\partial \tilde{c}_t}{\partial x_t} + \frac{\partial \mathrm{L}}{\partial \mathrm{h}_\mathrm{t}} \frac{\partial h_t}{\partial o_t} \frac{\partial o_t}{\partial x_t}$$

$$= \left(\frac{\partial \mathrm{L}}{\partial \mathrm{c}_\mathrm{t}} \odot c_{t-1} \odot f_t \odot (1 - f_t) \right) W_x^f + \left(\frac{\partial \mathrm{L}}{\partial \mathrm{c}_\mathrm{t}} \odot \tilde{c}_t \odot i_t \odot (1 - i_t) \right) W_x^i +$$

$$\left(\frac{\partial \mathrm{L}}{\partial \mathrm{c_t}} \odot i_t \odot (1-\tilde{c}_t^2)\right) W_x^c + \left(\frac{\partial \mathrm{L}}{\partial \mathrm{h_t}} \odot \tanh(c_t) \odot o_t \odot (1-o_t)\right) W_x^o$$

where $\frac{\partial \mathrm{L}}{\partial \mathrm{c_t}} = \frac{\partial L}{\partial h_t} \odot o_t \odot (1-\tanh^2(c_t)) + \frac{\partial L}{\partial c_{t+1}} \odot f_{t+1}$.

** Note that the notation "$\frac{\partial \mathrm{L}}{\partial \mathrm{c_t}}$" in the problem description (instead of the "$\frac{\partial \mathrm{L}}{\partial \mathrm{c_t}}$" used above) as well as the "dnext_c" variable given in the code may in fact, in my understanding, refer to "$\frac{\partial \mathrm{L}}{\partial \mathrm{c_{t+1}}}\frac{\partial c_{t+1}}{\partial c_t}$", which is only one of the two ways for the gradients to back-propagate from L to $\mathrm{c_t}$. Thus, in my answer, for strictness and consistency, we express $\frac{\partial \mathrm{L}}{\partial \mathrm{c_t}}$ strictly as the sum of two paths, i.e. $\mathrm{L} \to \mathrm{h_t} \to c_t\ and\ L \to \mathrm{c_{t+1}} \to \mathrm{c_t}$.

- $\frac{\partial \mathrm{L}}{\partial \mathrm{h_{t-1}}}$: (similar with $\frac{\partial \mathrm{L}}{\partial \mathrm{x_t}}$)

$$\frac{\partial \mathrm{L}}{\partial \mathrm{h_{t-1}}} = \left(\frac{\partial \mathrm{L}}{\partial \mathrm{c_t}} \odot c_{t-1} \odot f_t \odot (1-f_t)\right) W_h^f + \left(\frac{\partial \mathrm{L}}{\partial \mathrm{c_t}} \odot \tilde{c}_t \odot i_t \odot (1-i_t)\right) W_h^i +$$

$$\left(\frac{\partial \mathrm{L}}{\partial \mathrm{c_t}} \odot i_t \odot (1-\tilde{c}_t^2)\right) W_h^c + \left(\frac{\partial \mathrm{L}}{\partial \mathrm{h_t}} \odot \tanh(c_t) \odot o_t \odot (1-o_t)\right) W_h^o$$

where $\frac{\partial \mathrm{L}}{\partial \mathrm{c_t}} = \frac{\partial L}{\partial h_t} \odot o_t \odot (1-\tanh^2(c_t)) + \frac{\partial L}{\partial c_{t+1}} \odot f_{t+1}$.

- $\frac{\partial \mathrm{L}}{\partial \mathrm{c_{t-1}}}$:

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

where $\frac{\partial \mathrm{L}}{\partial \mathrm{c_t}} = \frac{\partial L}{\partial h_t} \odot o_t \odot (1-\tanh^2(c_t)) + \frac{\partial L}{\partial c_{t+1}} \odot f_{t+1}$.

- $\frac{\partial \mathrm{L}}{\partial \mathrm{W_x^f}}$:

$$\frac{\partial \mathrm{L}}{\partial \mathrm{W_x^f}} = \frac{\partial L}{\partial c_t}\frac{\partial c_t}{\partial f_t}\frac{\partial f_t}{\partial W_x^f} = \left(\frac{\partial L}{\partial c_t} \odot \mathrm{c_{t-1}} \odot f_t \odot (1-\mathrm{f_t})\right)^T x_t^T \in \mathbb{R}^{\mathrm{mxd}}$$

where $\frac{\partial \mathrm{L}}{\partial \mathrm{c_t}} = \frac{\partial L}{\partial h_t} \odot o_t \odot (1-\tanh^2(c_t)) + \frac{\partial L}{\partial c_{t+1}} \odot f_{t+1}$.

- $\frac{\partial \mathrm{L}}{\partial \mathrm{W}_{\mathrm{h}}^{f}}$:

$$\frac{\partial \mathrm{L}}{\partial \mathrm{W}_{\mathrm{h}}^{f}} = \left(\frac{\partial L}{\partial c_t} \odot \mathrm{c}_{\mathrm{t}-1} \odot f_t \odot (1 - \mathrm{f}_{\mathrm{t}})\right)^T h_{t-1}^T \in \mathbb{R}^{\mathrm{mxm}}$$

where $\frac{\partial \mathrm{L}}{\partial \mathrm{c}_{\mathrm{t}}} = \frac{\partial L}{\partial h_t} \odot o_t \odot (1 - \tanh^2(c_t)) + \frac{\partial L}{\partial c_{t+1}} \odot f_{t+1}$.

- $\frac{\partial \mathrm{L}}{\partial b^{f}}$:

$$\frac{\partial \mathrm{L}}{\partial f} = \frac{\partial L}{\partial c_t} \odot \mathrm{c}_{\mathrm{t}-1} \odot f_t \odot (1 - \mathrm{f}_{\mathrm{t}})$$

where $\frac{\partial \mathrm{L}}{\partial \mathrm{c}_{\mathrm{t}}} = \frac{\partial L}{\partial h_t} \odot o_t \odot (1 - \tanh^2(c_t)) + \frac{\partial L}{\partial c_{t+1}} \odot f_{t+1}$.

- $\frac{\partial \mathrm{L}}{\partial \mathrm{W}_{\mathrm{x}}^{i}}$:

$$\frac{\partial \mathrm{L}}{\partial \mathrm{W}_{\mathrm{x}}^{i}} = \left(\frac{\partial L}{\partial c_t} \odot \tilde{\mathrm{c}}_{\mathrm{t}} \odot i_t \odot (1 - \mathrm{i}_{\mathrm{t}})\right)^T x_t^T \in \mathbb{R}^{\mathrm{mxd}}$$

where $\frac{\partial \mathrm{L}}{\partial \mathrm{c}_{\mathrm{t}}} = \frac{\partial L}{\partial h_t} \odot o_t \odot (1 - \tanh^2(c_t)) + \frac{\partial L}{\partial c_{t+1}} \odot f_{t+1}$.

- $\frac{\partial \mathrm{L}}{\partial \mathrm{W}_{\mathrm{h}}^{i}}$:

$$\frac{\partial \mathrm{L}}{\partial \mathrm{W}_{\mathrm{h}}^{i}} = \left(\frac{\partial L}{\partial c_t} \odot \tilde{\mathrm{c}}_{\mathrm{t}} \odot i_t \odot (1 - \mathrm{i}_{\mathrm{t}})\right)^T h_{t-1}^T \in \mathbb{R}^{\mathrm{mxm}}$$

where $\frac{\partial \mathrm{L}}{\partial \mathrm{c}_{\mathrm{t}}} = \frac{\partial L}{\partial h_t} \odot o_t \odot (1 - \tanh^2(c_t)) + \frac{\partial L}{\partial c_{t+1}} \odot f_{t+1}$.

- $\frac{\partial \mathrm{L}}{\partial b^{i}}$:

$$\frac{\partial \mathrm{L}}{\partial b^{i}} = \frac{\partial L}{\partial c_t} \odot \tilde{\mathrm{c}}_{\mathrm{t}} \odot i_t \odot (1 - \mathrm{i}_{\mathrm{t}})$$

where $\frac{\partial \mathrm{L}}{\partial \mathrm{c}_{\mathrm{t}}} = \frac{\partial L}{\partial h_t} \odot o_t \odot (1 - \tanh^2(c_t)) + \frac{\partial L}{\partial c_{t+1}} \odot f_{t+1}$.

- $\frac{\partial \mathrm{L}}{\partial \mathrm{W}_{\mathrm{x}}^{c}}$:

$$\frac{\partial \mathrm{L}}{\partial \mathrm{W}_{\mathrm{x}}^{c}} = \left(\frac{\partial L}{\partial c_t} \odot \mathrm{i}_{\mathrm{t}} \odot (1 - \tilde{\mathrm{c}}_t^2)\right)^T x_t^T \in \mathbb{R}^{\mathrm{mxd}}$$

where $\frac{\partial \mathrm{L}}{\partial \mathrm{c_t}} = \frac{\partial L}{\partial h_t} \odot o_t \odot (1 - \tanh^2(c_t)) + \frac{\partial L}{\partial c_{t+1}} \odot f_{t+1}$.

- $\frac{\partial \mathrm{L}}{\partial \mathrm{W_h^c}}$:

$$\frac{\partial \mathrm{L}}{\partial \mathrm{W_h^c}} = \left(\frac{\partial L}{\partial c_t} \odot \mathrm{i_t} \odot (1 - \tilde{c}_t^2)\right)^T h_{t-1}^T \in \mathbb{R}^{\mathrm{mxm}}$$

where $\frac{\partial \mathrm{L}}{\partial \mathrm{c_t}} = \frac{\partial L}{\partial h_t} \odot o_t \odot (1 - \tanh^2(c_t)) + \frac{\partial L}{\partial c_{t+1}} \odot f_{t+1}$.

- $\frac{\partial \mathrm{L}}{\partial b^c}$:

$$\frac{\partial \mathrm{L}}{\partial b^c} = \frac{\partial L}{\partial c_t} \odot \mathrm{i_t} \odot (1 - \tilde{c}_t^2)$$

where $\frac{\partial \mathrm{L}}{\partial \mathrm{c_t}} = \frac{\partial L}{\partial h_t} \odot o_t \odot (1 - \tanh^2(c_t)) + \frac{\partial L}{\partial c_{t+1}} \odot f_{t+1}$.

- $\frac{\partial \mathrm{L}}{\partial \mathrm{W_x^o}}$:

$$\frac{\partial \mathrm{L}}{\partial \mathrm{W_x^o}} = \left(\frac{\partial L}{\partial h_t} \odot \tanh(\mathrm{c_t}) \odot o_\mathrm{t} \odot (1 - o_\mathrm{t})\right)^T x_t^T \in \mathbb{R}^{\mathrm{mxd}}$$

- $\frac{\partial \mathrm{L}}{\partial \mathrm{W_h^o}}$:

$$\frac{\partial \mathrm{L}}{\partial \mathrm{W_h^o}} = \left(\frac{\partial L}{\partial h_t} \odot \tanh(\mathrm{c_t}) \odot o_\mathrm{t} \odot (1 - o_\mathrm{t})\right)^T h_{t-1}^T \in \mathbb{R}^{\mathrm{mxm}}$$

- $\frac{\partial \mathrm{L}}{\partial b^o}$:

$$\frac{\partial \mathrm{L}}{\partial b^o} = \frac{\partial L}{\partial h_t} \odot \tanh(\mathrm{c_t}) \odot o_\mathrm{t} \odot (1 - o_\mathrm{t})$$

4. Derive for all time steps ($\forall 1 \le \mathrm{t} \le \mathrm{T}$).

Similar with the case in Problem 2, question 4, here the $\frac{\partial \mathrm{L}}{\partial \mathrm{h_t}}$ is actually $\frac{\partial \mathrm{D}(\mathrm{y}_t, \hat{\mathrm{y}}_t)}{\partial h_t}$, and we denote $\frac{\partial \mathrm{L_t}}{\partial \mathrm{h_t}}$ as the partial derivative back-propagated from all time steps from t to T. We once again first derive $\frac{\partial \mathrm{L_t}}{\partial \mathrm{h_t}}$ using $\frac{\partial \mathrm{D}(\mathrm{y}_t, \hat{\mathrm{y}}_t)}{\partial h_t}$, then all the $\frac{\partial \mathrm{L_t}}{\partial \mathrm{h_t}}$ below can be substituted to become an answer.

$$\frac{\partial \mathrm{L_t}}{\partial \mathrm{h_t}} = \frac{\partial \mathrm{D}(\mathrm{y}_t, \hat{\mathrm{y}}_t)}{\partial h_t} + \frac{\partial L_{t+1}}{\partial h_t}$$

$$= \frac{\partial \mathrm{D}(\mathrm{y}_t, \hat{\mathrm{y}}_t)}{\partial h_t} + \frac{\partial L_{t+1}}{\partial h_{t+1}} \frac{\partial h_{t+1}}{\partial h_t}$$

$$= \frac{\partial \mathrm{D}(\mathrm{y}_t, \hat{\mathrm{y}}_t)}{\partial h_t} + \frac{\partial L_{t+1}}{\partial h_{t+1}} \odot \left(\frac{\partial h_{t+1}}{\partial f_{t+1}} \frac{\partial f_{t+1}}{\partial h_t} + \frac{\partial h_{t+1}}{\partial i_{t+1}} \frac{\partial i_{t+1}}{\partial h_t} + \frac{\partial h_{t+1}}{\partial \tilde{c}_{t+1}} \frac{\partial \tilde{c}_{t+1}}{\partial h_t} + \frac{\partial h_{t+1}}{\partial o_{t+1}} \frac{\partial o_{t+1}}{\partial h_t}\right)$$

$$= \frac{\partial \mathrm{D}(\mathrm{y}_t, \hat{\mathrm{y}}_t)}{\partial h_t} + \frac{\partial L_{t+1}}{\partial h_{t+1}} \odot \left(\frac{\partial h_{t+1}}{\partial c_{t+1}} \frac{\partial c_{t+1}}{\partial f_{t+1}} \frac{\partial f_{t+1}}{\partial h_t} + \frac{\partial h_{t+1}}{\partial c_{t+1}} \frac{\partial c_{t+1}}{\partial i_{t+1}} \frac{\partial i_{t+1}}{\partial h_t} + \frac{\partial h_{t+1}}{\partial c_{t+1}} \frac{\partial c_{t+1}}{\partial \tilde{c}_{t+1}} \frac{\partial \tilde{c}_{t+1}}{\partial h_t} + \frac{\partial h_{t+1}}{\partial o_{t+1}} \frac{\partial o_{t+1}}{\partial h_t}\right)$$

$$= \frac{\partial \mathrm{D}(\mathrm{y}_t, \hat{\mathrm{y}}_t)}{\partial h_t} + \frac{\partial L_{t+1}}{\partial h_{t+1}} \odot \big(o_{t+1} \odot (1 - \tanh^2(c_{t+1})) \odot c_t \odot f_{t+1} \odot (1 - f_{t+1}) W_h^f + o_{t+1} \odot (1 - \tanh^2(c_{t+1})) \odot \tilde{c}_{t+1} \odot i_{t+1} \odot (1 - i_{t+1}) W_h^i + o_{t+1} \odot (1 - \tanh^2(c_{t+1})) \odot i_{t+1} \odot (1 - \tilde{\mathrm{c}}_{t+1}^2) \mathrm{W_h^c} + \tanh(c_{t+1}) \odot o_{t+1} \odot (1 - o_{t+1}) W_h^o\big)$$

$$= \frac{\partial \mathrm{D}(\mathrm{y}_t, \hat{\mathrm{y}}_t)}{\partial h_t} + \sum_{\tau=t+1}^{T} \frac{\partial \mathrm{D}(\mathrm{y}_\tau, \hat{\mathrm{y}}_\tau)}{\partial h_\tau} \odot \prod_{\beta=t}^{\tau-1} (\Phi)$$

where $\Phi = \big(o_{t+1} \odot (1 - \tanh^2(c_{t+1})) \odot c_t \odot f_{t+1} \odot (1 - f_{t+1}) W_h^f + o_{t+1} \odot (1 - \tanh^2(c_{t+1})) \odot \tilde{c}_{t+1} \odot i_{t+1} \odot (1 - i_{t+1}) W_h^i + o_{t+1} \odot (1 - \tanh^2(c_{t+1})) \odot i_{t+1} \odot (1 - \tilde{\mathrm{c}}_{t+1}^2) W_{\mathrm{h}}^{\mathrm{c}} + \tanh(c_{t+1}) \odot o_{t+1} \odot (1 - o_{t+1}) W_h^o\big)$

Since the question requires to use $\frac{\partial \mathrm{L}}{\partial \mathrm{h_t}}$ only, we can further derive $\frac{\partial \mathrm{L_t}}{\partial \mathrm{c_t}}$ using $\frac{\partial \mathrm{L_t}}{\partial \mathrm{h_t}}$, and since we have just derived $\frac{\partial \mathrm{L_t}}{\partial \mathrm{h_t}}$ using $\frac{\partial \mathrm{L}}{\partial \mathrm{h_t}}$, after two times of substitutions, all expressions below can be "$\frac{\partial \mathrm{L_t}}{\partial c_t}$-free" (and also "$\frac{\partial \mathrm{L_t}}{\partial h_t}$-free" as well) and depend only on $\frac{\partial \mathrm{L}}{\partial \mathrm{h_t}}$.

$$\frac{\partial \mathrm{L_t}}{\partial \mathrm{c_t}} = \frac{\partial \mathrm{L_t}}{\partial \mathrm{h_t}} \odot o_t \odot (1 - \tanh^2(c_t)) + \frac{\partial L_t}{\partial c_{t+1}} \odot f_t$$

$$= \frac{\partial \mathrm{L_t}}{\partial \mathrm{h_t}} \odot o_t \odot (1 - \tanh^2(c_t)) + \frac{\partial \mathrm{L_t}}{\partial \mathrm{h_{t+1}}} \odot o_{t+1} \odot (1 - \tanh^2(c_{t+1})) \odot f_t$$

$$= \frac{\partial \mathrm{L_t}}{\partial \mathrm{h_t}} \odot o_t \odot (1 - \tanh^2(c_t)) + \sum_{\tau=t+1}^{T} \frac{\partial \mathrm{L_t}}{\partial \mathrm{h_\tau}} \odot \prod_{\beta=t}^{\tau-1} \big(o_{\beta+1} \odot \big(1 - \tanh^2\big(c_{\beta+1}\big)\big) \odot f_\beta\big)$$

- $\frac{\partial \mathrm{L_t}}{\partial \mathrm{x_t}}$:

$$\frac{\partial \mathrm{L_t}}{\partial \mathrm{x_t}} = \frac{\partial \mathrm{L_t}}{\partial \mathrm{f_t}}\frac{\partial f_t}{\partial x_t} + \frac{\partial \mathrm{L_t}}{\partial \mathrm{i_t}}\frac{\partial i_t}{\partial x_t} + \frac{\partial \mathrm{L_t}}{\partial \tilde{\mathrm{c}}_\mathrm{t}}\frac{\partial \tilde{c}_t}{\partial x_t} + \frac{\partial \mathrm{L_t}}{\partial \mathrm{o_t}}\frac{\partial o_t}{\partial x_t}$$

$$= \frac{\partial \mathrm{L_t}}{\partial \mathrm{h_t}}\frac{\partial h_t}{\partial c_t}\frac{\partial c_t}{\partial f_t}\frac{\partial f_t}{\partial x_t} + \frac{\partial \mathrm{L_t}}{\partial \mathrm{h_t}}\frac{\partial h_t}{\partial c_t}\frac{\partial c_t}{\partial i_t}\frac{\partial i_t}{\partial x_t} + \frac{\partial \mathrm{L_t}}{\partial \mathrm{h_t}}\frac{\partial h_t}{\partial c_t}\frac{\partial c_t}{\partial \tilde{\mathrm{c}}_\mathrm{t}}\frac{\partial \tilde{c}_t}{\partial x_t} + \frac{\partial \mathrm{L_t}}{\partial \mathrm{h_t}}\frac{\partial h_t}{\partial o_t}\frac{\partial o_t}{\partial x_t}$$

$$= \left(\frac{\partial \mathrm{L_t}}{\partial \mathrm{c_t}} \odot c_{t-1} \odot f_t \odot (1-f_t)\right) W_x^f + \left(\frac{\partial \mathrm{L_t}}{\partial \mathrm{c_t}} \odot \tilde{c}_t \odot i_t \odot (1-i_t)\right) W_x^i +$$

$$\left(\frac{\partial \mathrm{L_t}}{\partial \mathrm{c_t}} \odot i_t \odot (1-\tilde{c}_t^2)\right) W_x^c + \left(\frac{\partial \mathrm{L_t}}{\partial \mathrm{h_t}} \odot \tanh(c_t) \odot o_t \odot (1-o_t)\right) W_x^o$$

- $\frac{\partial \mathrm{L_1}}{\partial \mathrm{h_0}}$: (similar with $\frac{\partial \mathrm{L}}{\partial \mathrm{x_t}}$)

$$\frac{\partial \mathrm{L_1}}{\partial \mathrm{h_0}} = \left(\frac{\partial \mathrm{L_1}}{\partial \mathrm{c_1}} \odot c_0 \odot f_1 \odot (1-f_1)\right) W_h^f + \left(\frac{\partial \mathrm{L_1}}{\partial \mathrm{c_1}} \odot \tilde{c}_1 \odot i_1 \odot (1-i_1)\right) W_h^i +$$

$$\left(\frac{\partial \mathrm{L_1}}{\partial \mathrm{c_1}} \odot i_1 \odot (1-\tilde{c}_1^2)\right) W_h^c + \left(\frac{\partial \mathrm{L_1}}{\partial \mathrm{h_1}} \odot \tanh(c_1) \odot o_1 \odot (1-o_1)\right) W_h^o$$

- $\frac{\partial \mathrm{L_1}}{\partial \mathrm{W_x^{\mathit{f}}}}$:

$$\frac{\partial \mathrm{L_1}}{\partial \mathrm{W_x^{\mathit{f}}}} = \sum\nolimits_\mathrm{t} \left(\frac{\partial L_t}{\partial c_t} \odot \mathrm{c_{t-1}} \odot f_t \odot (1-\mathrm{f_t})\right)^T x_t^T \ \in \mathbb{R}^{\mathrm{mxd}}$$

- $\frac{\partial \mathrm{L_1}}{\partial \mathrm{W_h^{\mathit{f}}}}$:

$$\frac{\partial \mathrm{L_1}}{\partial \mathrm{W_h^{\mathit{f}}}} = \sum\nolimits_\mathrm{t} \left(\frac{\partial L_t}{\partial c_t} \odot \mathrm{c_{t-1}} \odot f_t \odot (1-\mathrm{f_t})\right)^T h_{t-1}^T \ \in \mathbb{R}^{\mathrm{mxm}}$$

- $\frac{\partial \mathrm{L_1}}{\partial b^f}$:

$$\frac{\partial \mathrm{L_1}}{\partial \mathrm{b}^f} = \sum\nolimits_\mathrm{t} \frac{\partial L_t}{\partial c_t} \odot \mathrm{c_{t-1}} \odot f_t \odot (1-\mathrm{f_t})$$

- $\frac{\partial \mathrm{L_1}}{\partial \mathrm{W_x^{\mathit{i}}}}$:

$$\frac{\partial \mathrm{L_1}}{\partial \mathrm{W_x^{\mathit{i}}}} = \sum\nolimits_\mathrm{t} \left(\frac{\partial L_t}{\partial c_t} \odot \tilde{\mathrm{c}}_\mathrm{t} \odot i_t \odot (1-\mathrm{i_t})\right)^T x_t^T \ \in \mathbb{R}^{\mathrm{mxd}}$$

- $\frac{\partial \mathrm{L_1}}{\partial \mathrm{W_h^{\mathit{i}}}}$:

$$\frac{\partial \mathrm{L}_1}{\partial \mathrm{W}_{\mathrm{h}}^{i}} = \sum_{\mathrm{t}} \left(\frac{\partial L_t}{\partial c_t} \odot \tilde{\mathrm{c}}_{\mathrm{t}} \odot i_t \odot (1 - \mathrm{i}_{\mathrm{t}}) \right)^T h_{t-1}^T \in \mathbb{R}^{\mathrm{mxm}}$$

- $\frac{\partial \mathrm{L}_1}{\partial b^i}$:

$$\frac{\partial \mathrm{L}_1}{\partial b^i} = \sum_t \frac{\partial L_t}{\partial c_t} \odot \tilde{\mathrm{c}}_{\mathrm{t}} \odot i_t \odot (1 - \mathrm{i}_{\mathrm{t}})$$

- $\frac{\partial \mathrm{L}_1}{\partial \mathrm{W}_{\mathrm{x}}^{c}}$:

$$\frac{\partial \mathrm{L}_1}{\partial \mathrm{W}_{\mathrm{x}}^{c}} = \sum_t \left(\frac{\partial L_t}{\partial c_t} \odot \mathrm{i}_{\mathrm{t}} \odot (1 - \tilde{\mathrm{c}}_t^2) \right)^T x_t^T \in \mathbb{R}^{\mathrm{mxd}}$$

- $\frac{\partial \mathrm{L}_1}{\partial \mathrm{W}_{\mathrm{h}}^{c}}$:

$$\frac{\partial \mathrm{L}}{\partial \mathrm{W}_{\mathrm{h}}^{c}} = \sum_t \left(\frac{\partial L_t}{\partial c_t} \odot \mathrm{i}_{\mathrm{t}} \odot (1 - \tilde{\mathrm{c}}_t^2) \right)^T h_{t-1}^T \in \mathbb{R}^{\mathrm{mxm}}$$

- $\frac{\partial \mathrm{L}_1}{\partial b^c}$:

$$\frac{\partial \mathrm{L}_1}{\partial b^c} = \sum_t \frac{\partial L_t}{\partial c_t} \odot \mathrm{i}_{\mathrm{t}} \odot (1 - \tilde{\mathrm{c}}_t^2)$$

- $\frac{\partial \mathrm{L}_1}{\partial \mathrm{W}_{\mathrm{x}}^{o}}$:

$$\frac{\partial \mathrm{L}_1}{\partial \mathrm{W}_{\mathrm{x}}^{o}} = \sum_t \left(\frac{\partial L_t}{\partial h_t} \odot \tanh(\mathrm{c}_{\mathrm{t}}) \odot o_{\mathrm{t}} \odot (1 - o_{\mathrm{t}}) \right)^T x_t^T \in \mathbb{R}^{\mathrm{mxd}}$$

- $\frac{\partial \mathrm{L}_1}{\partial \mathrm{W}_{\mathrm{h}}^{o}}$:

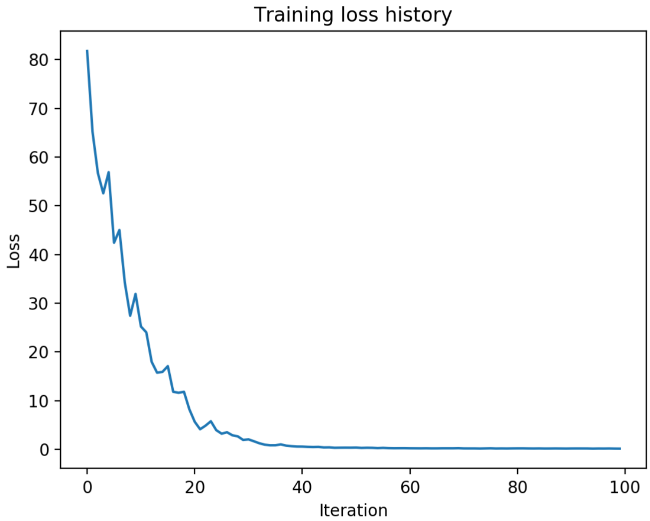
$$\frac{\partial \mathrm{L}_1}{\partial \mathrm{W}_{\mathrm{h}}^{o}} = \sum_t \left(\frac{\partial L_t}{\partial h_t} \odot \tanh(\mathrm{c}_{\mathrm{t}}) \odot o_{\mathrm{t}} \odot (1 - o_{\mathrm{t}}) \right)^T h_{t-1}^T \in \mathbb{R}^{\mathrm{mxm}}$$

- $\frac{\partial \mathrm{L}_1}{\partial b^o}$:

$$\frac{\partial \mathrm{L}_1}{\partial b^o} = \sum_t \frac{\partial L_t}{\partial h_t} \odot \tanh(\mathrm{c}_{\mathrm{t}}) \odot o_{\mathrm{t}} \odot (1 - o_{\mathrm{t}})$$

Q5.

RNN learning curve:



RNN samples:

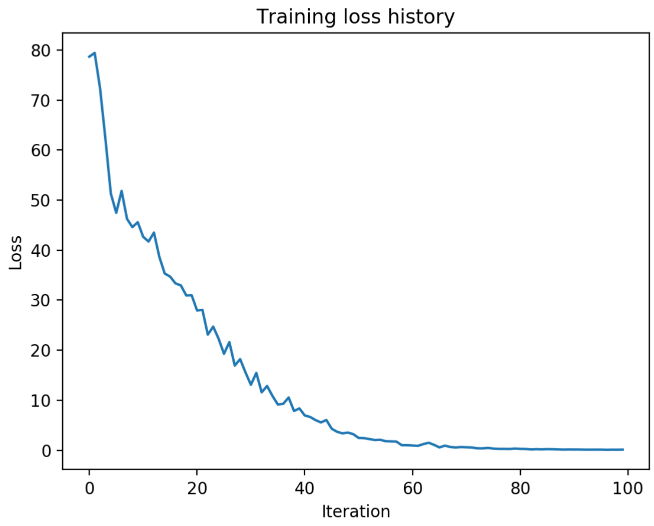
train
a large group of people in a building <END>
GT:<START> a large group of people in a building <END>

train
a man and a child standing out in the rain holding up an umbrella <END>
:START> a man and a child standing out in the rain holding up an umbrella <

val
: middle smiling holding a a wii remote and white with a broccoli platform so
Γ:<START> a female tennis player <UNK> to hit the ball with her racket <EN

val
a large up of bird busy a frisbee on a end on smiling grass <END>
GT:<START> a woman standing next to a pile of luggage <END>

LSTM learning curve:



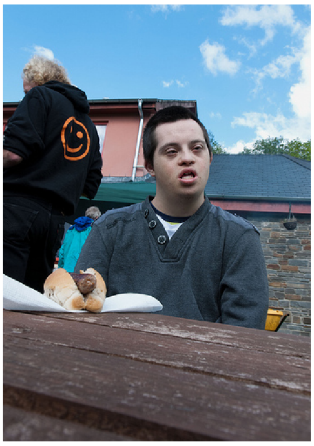
LSTM samples:

train
a man doing trick trick frisbee frisbee <END>
GT:<START> a man doing a trick with a frisbee on a beach <END>

train
a cat sitting in in in in <END>
GT:<START> a cat sitting in a <UNK> in a full refrigerator <END>

val
several are are are around around in on <END>
;T:<START> people are on the beach with buildings in the background <ENI

val
<UNK> <UNK> up up of of sign <END>
GT:<START> a man sitting at a table while reading a book <END>



Q6.

1. Bag of words → Linear → sigmoid

test acc: 0.9536

2. Word embeddings → Average pooling → Linear → sigmoid

test acc: 0.9489

3. Word embeddings → Average pooling → Linear → sigmoid with glove

test acc: 0.9503

4. Word embeddings → RNN → Linear → sigmoid

test acc: 0.9412

5. Word embeddings → LSTM → Linear → sigmoid

test acc: 0.9564