CS224D Spring 2015: Homework 3

SUNet ID: dlchang
Name: Daryl Chang
Collaborators: April Yu, Leon Yao, Derrick Liu

By turning in this assignment, I agree by the Stanford honor code and declare that all of this is my own work.

Problem 1

Part a

Node 1

$$\delta_3 = \hat{y} - y$$
$$\delta_2 = U^T \delta_3 \circ 1\{h^{(1)} \neq 0\}$$
$$\frac{\partial J}{\partial U} = \delta_3 h^{(1)T}$$
$$\frac{\partial J}{\partial b^{(s)}} = \delta_3$$
$$\frac{\partial J}{\partial W^{(1)}} = \delta_2 \begin{bmatrix} h^{(1)}_{left} \\ h^{(1)}_{right} \end{bmatrix}^T$$
$$\frac{\partial J}{\partial b^{(1)}} = \delta_2$$
$$\frac{\partial J}{\partial L_i} = 0$$

Node 2

$$\delta_3 = \hat{y} - y$$
$$\delta_2 = U^T \delta_3 \circ 1\{h^{(1)} \neq 0\} + W^T_{[:,:d]} \delta_{above} \circ 1\{h^{(1)} \neq 0\}$$
$$\frac{\partial J}{\partial U} = \delta_3 h^{(1)T}$$
$$\frac{\partial J}{\partial b^{(s)}} = \delta_3$$
$$\frac{\partial J}{\partial W^{(1)}} = \delta_2 \begin{bmatrix} L_7 \\ L_{145} \end{bmatrix}^T$$
$$\frac{\partial J}{\partial b^{(1)}} = \delta_2$$
$$\frac{\partial J}{\partial L_i} = 0$$

Node 3

$$\delta_3 = \hat{y} - y$$
$$\delta_2 = U^T \delta_3 \circ 1\{h^{(1)} \neq 0\} + W^T_{[:,d:]} \delta_{above} \circ 1\{h^{(1)} \neq 0\}$$
$$\frac{\partial J}{\partial U} = \delta_3 h^{(1)T}$$
$$\frac{\partial J}{\partial b^{(s)}} = \delta_3$$
$$\frac{\partial J}{\partial W^{(1)}} = \delta_2 \begin{bmatrix} L_{29} \\ L_{430} \end{bmatrix}^T$$
$$\frac{\partial J}{\partial b^{(1)}} = \delta_2$$
$$\frac{\partial J}{\partial L_i} = 0$$

Leaves

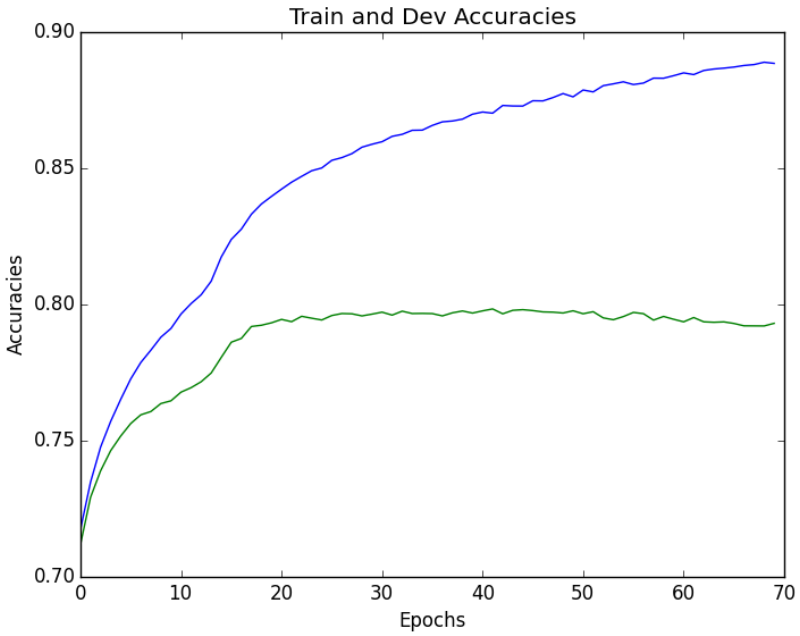
$$\delta_3 = \hat{y} - y$$
$$\delta_2 = U^T \delta_3 \circ 1\{h^{(1)} \neq 0\} + W^T_{[:,:d]} \delta_{above} \circ 1\{h^{(1)} \neq 0\} \text{ if left child}$$
$$\delta_2 = U^T \delta_3 \circ 1\{h^{(1)} \neq 0\} + W^T_{[:,d:]} \delta_{above} \circ 1\{h^{(1)} \neq 0\} \text{ if right child}$$
$$\frac{\partial J}{\partial U} = \delta_3 L_i^T$$
$$\frac{\partial J}{\partial b^{(s)}} = \delta_3$$
$$\frac{\partial J}{\partial W^{(1)}} = 0$$
$$\frac{\partial J}{\partial b^{(1)}} = 0$$
$$\frac{\partial J}{\partial L_i} = \delta_2$$

Part b

See code.

Part c

Subpart a



Best dev set accuracy occurred at 41 epochs.

Subpart b

The dev accuracy starts to decrease after a certain point because the neural network overfits the training data, resulting in a loss of generalization.

Subpart c

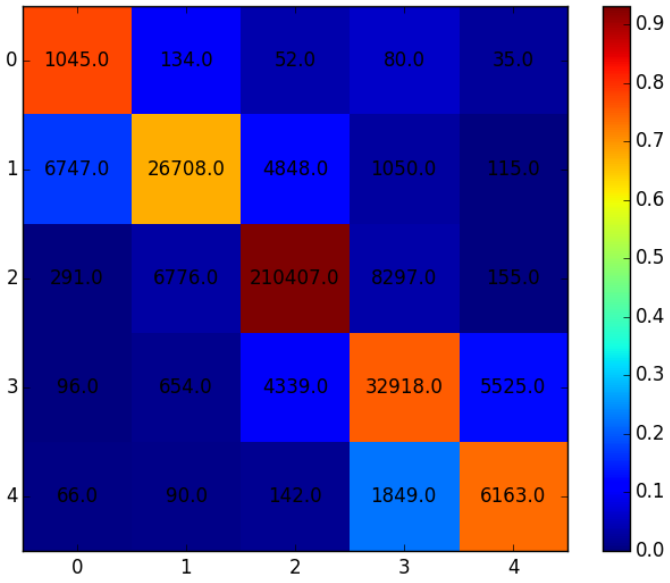


Figure 1: Confusion matrix on training data

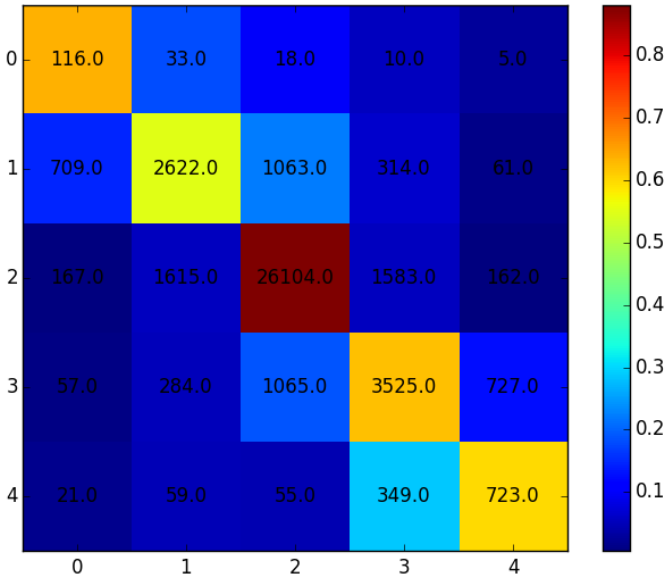
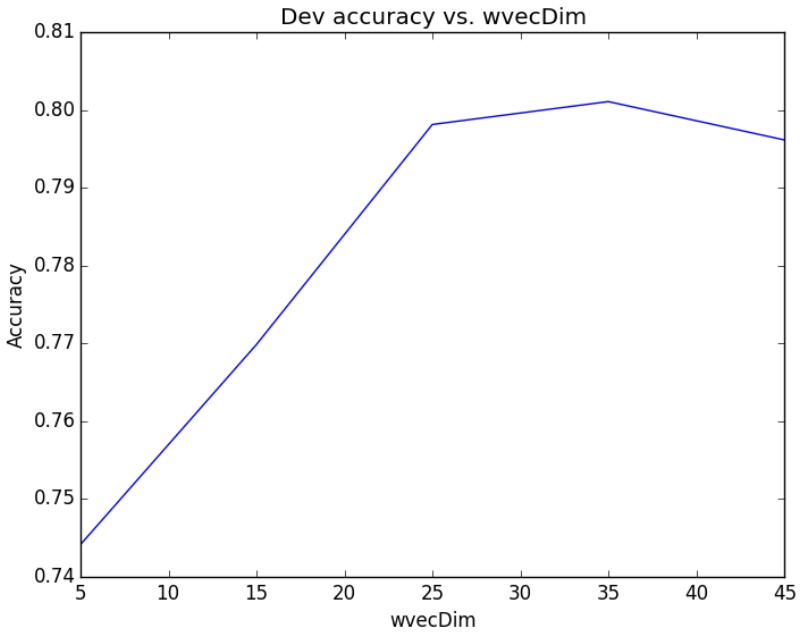


Figure 2: Confusion matrix on dev data

Subpart d



Problem 2

Part a

Node 1

$$\delta_3 = \hat{y} - y$$

$$\delta_2 = U^T \delta_3 \circ 1\{h^{(2)} \neq 0\}$$

$$\delta_1 = W^{(2)T} \delta_2 \circ 1\{h^{(1)} \neq 0\}$$

$$\frac{\partial J}{\partial U} = \delta_3 h^{(2)T}$$

$$\frac{\partial J}{\partial b^{(s)}} = \delta_3$$

$$\frac{\partial J}{\partial W^{(1)}} = \delta_1 \begin{bmatrix} h^{(1)}_{Left} \\ h^{(1)}_{Right} \end{bmatrix}^T$$

$$\frac{\partial J}{\partial b^{(1)}} = \delta_1$$

$$\frac{\partial J}{\partial W^{(2)}} = \delta_2 h^{(1)T}$$

$$\frac{\partial J}{\partial b^{(2)}} = \delta_2$$
$$\frac{\partial J}{\partial L_i} = 0$$

Node 2

$$\delta_3 = \hat{y} - y$$
$$\delta_2 = U^T \delta_3 \circ 1\{h^{(2)} \neq 0\}$$
$$\delta_1 = W^{(2)T} \delta_2 \circ 1\{h^{(1)} \neq 0\} + W^{(1)T}_{[:,:d]} \delta_{above} \circ 1\{h^{(1)} \neq 0\}$$
$$\frac{\partial J}{\partial U} = \delta_3 h^{(2)T}$$
$$\frac{\partial J}{\partial b^{(s)}} = \delta_3$$
$$\frac{\partial J}{\partial W^{(1)}} = \delta_1 \begin{bmatrix} L_7 \\ L_{145} \end{bmatrix}^T$$
$$\frac{\partial J}{\partial b^{(1)}} = \delta_1$$
$$\frac{\partial J}{\partial W^{(2)}} = \delta_2 h^{(1)T}$$
$$\frac{\partial J}{\partial b^{(2)}} = \delta_2$$
$$\frac{\partial J}{\partial L_i} = 0$$

Node 3

$$\delta_3 = \hat{y} - y$$
$$\delta_2 = U^T \delta_3 \circ 1\{h^{(2)} \neq 0\}$$
$$\delta_1 = W^{(2)T} \delta_2 \circ 1\{h^{(1)} \neq 0\} + W^{(1)T}_{[:,d:]} \delta_{above} \circ 1\{h^{(1)} \neq 0\}$$
$$\frac{\partial J}{\partial U} = \delta_3 h^{(2)T}$$
$$\frac{\partial J}{\partial b^{(s)}} = \delta_3$$
$$\frac{\partial J}{\partial W^{(1)}} = \delta_1 \begin{bmatrix} L_{29} \\ L_{430} \end{bmatrix}^T$$
$$\frac{\partial J}{\partial b^{(1)}} = \delta_1$$

$$\frac{\partial J}{\partial W^{(2)}} = \delta_2 h^{(1)T}$$
$$\frac{\partial J}{\partial b^{(2)}} = \delta_2$$
$$\frac{\partial J}{\partial L_i} = 0$$

Leaves

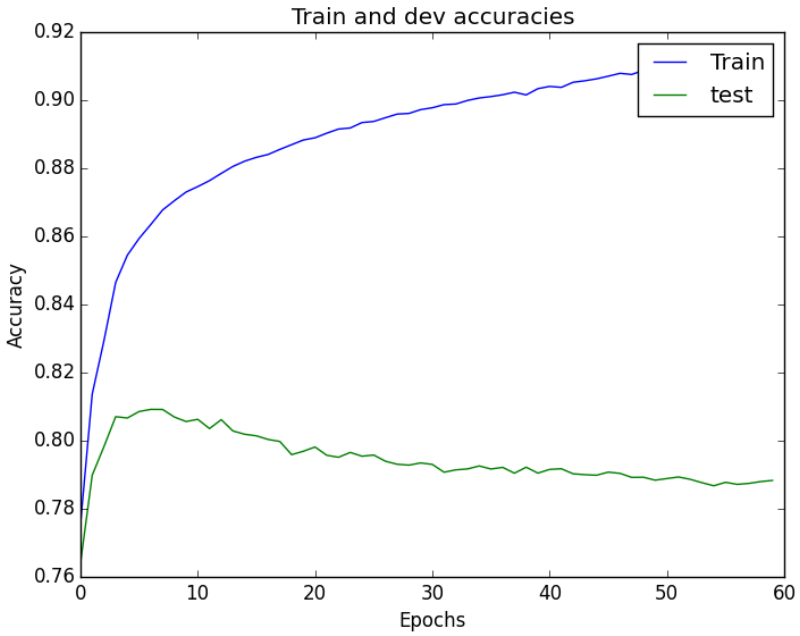
$$\delta_3 = \hat{y} - y$$
$$\delta_2 = U^T \delta_3 \circ 1\{h^{(2)} \neq 0\}$$
$$\delta_1 = W^{(2)T}\delta_2 \circ 1\{L_i \neq 0\} + W^{(1)T}_{[:,:d]}\delta_{above} \circ 1\{L_i \neq 0\} \text{ if left child}$$
$$\delta_1 = W^{(2)T}\delta_2 \circ 1\{L_i \neq 0\} + W^{(1)T}_{[:,d:]}\delta_{above} \circ 1\{L_i \neq 0\} \text{ if right child}$$
$$\frac{\partial J}{\partial U} = \delta_3 h^{(2)T}$$
$$\frac{\partial J}{\partial b^{(s)}} = \delta_3$$
$$\frac{\partial J}{\partial W^{(1)}} = 0$$
$$\frac{\partial J}{\partial b^{(1)}} = 0$$
$$\frac{\partial J}{\partial W^{(2)}} = \delta_2 L_i^T$$
$$\frac{\partial J}{\partial b^{(2)}} = \delta_2$$
$$\frac{\partial J}{\partial L_i} = W^{(1)T}_{[:,:d]}\delta_{above} \circ 1\{L_i \neq 0\} + W^{(2)T}\delta^2 \circ 1\{L_i \neq 0\} \text{ if left child}$$
$$\frac{\partial J}{\partial L_i} = W^{(1)T}_{[:,d:]}\delta_{above} \circ 1\{L_i \neq 0\} + W^{(2)T}\delta^2 \circ 1\{L_i \neq 0\} \text{ if right child}$$

Part b

See code.

Part c

Subpart a



The best dev accuracy occurred at 6 epochs.

Subpart b

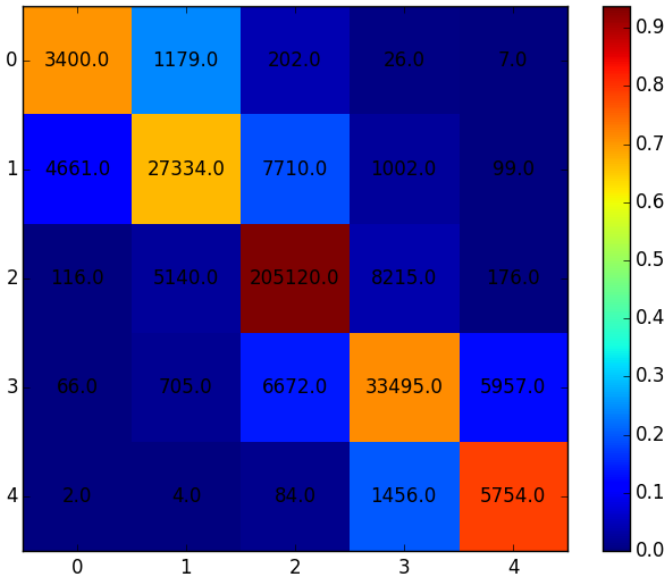


Figure 3: Confusion matrix on training data

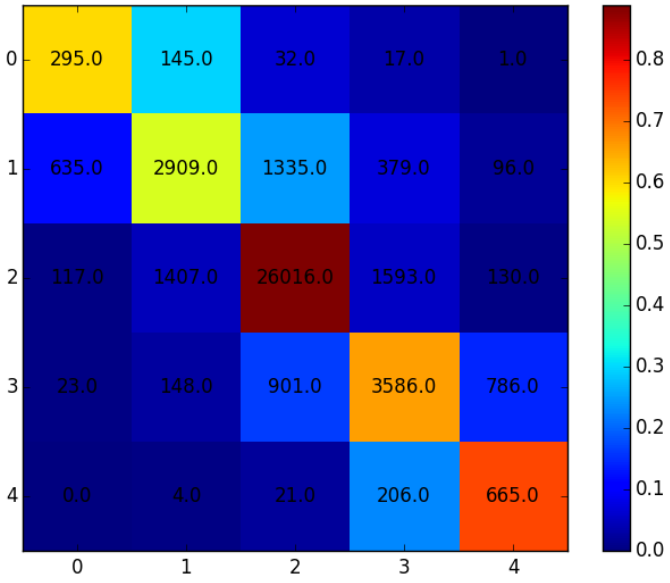
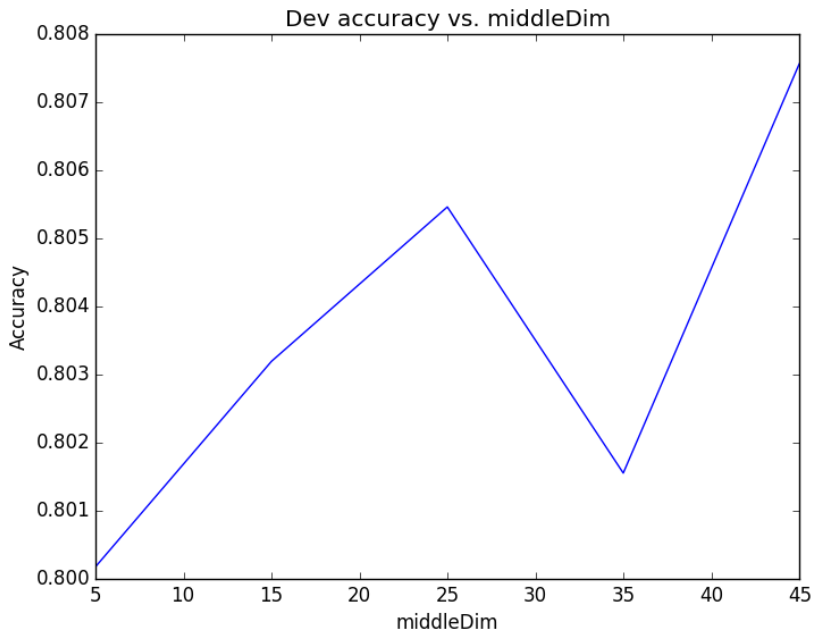


Figure 4: Confusion matrix on dev data

Subpart c

The RNN does slightly better than the original one. This is most likely because the added layer allows the network to learn higher-order features, which in turn increases accuracy.

Subpart d



Part d

One possible way to improve the performance of the neural network is to make it into a recursive autoencoder (i.e. to base the structure of the network on the syntactic parse of the sentence). The recursive autoencoder would essentially encode groups of words into a hidden representation until the entire sentence is encoded; it minimizes the reconstruction error.This would account for the syntactic structure of the sentence, which should in turn increase the performance of the network.