Solutions to assignment1 of CS224n

陈承勃

May 22, 2018

1 Softmax

(a) Omitted. (b) See **q1_softmax.py**

2 Neural Network Basics

(a)

$$\sigma'(x) = \sigma(x)\sigma(1-x) \tag{1}$$

(b) Assume k is the correct class, then

$$CE(\boldsymbol{y}, \widehat{\boldsymbol{y}}) = -y_k \log \widehat{y_k} = -\log \widehat{y_k} = -\log \frac{\exp(\boldsymbol{\theta}_k)}{\sum_i \exp(\boldsymbol{\theta}_i)} = -\boldsymbol{\theta}_k + \log \textstyle\sum_i \exp(\boldsymbol{\theta}_i) \tag{2}$$

$$\therefore \frac{\partial CE(\boldsymbol{y}, \widehat{\boldsymbol{y}})}{\partial \boldsymbol{\theta}_k} = -1 + \frac{\exp(\boldsymbol{\theta}_k)}{\sum_i \exp(\boldsymbol{\theta}_i)} = \widehat{y_k} - 1, \tag{3}$$

$$\frac{\partial CE(\boldsymbol{y}, \widehat{\boldsymbol{y}})}{\partial \boldsymbol{\theta}_j} = \frac{\exp(\boldsymbol{\theta}_j)}{\sum_i \exp(\boldsymbol{\theta}_i)} = \widehat{y_j},\ j \neq k \tag{4}$$

$$\therefore \frac{\partial CE(\boldsymbol{y}, \widehat{\boldsymbol{y}})}{\partial \boldsymbol{\theta}} = \widehat{\boldsymbol{y}} - \boldsymbol{y} \tag{5}$$

(c) The forward propagation steps:

$$\boldsymbol{Z_1} = \boldsymbol{x}\boldsymbol{W_1} + \boldsymbol{b_1} \tag{6}$$

$$\boldsymbol{h} = sigmoid(\boldsymbol{Z_1}) \tag{7}$$

$$\boldsymbol{Z_2} = \boldsymbol{h}\boldsymbol{W_2} + \boldsymbol{b_2} \tag{8}$$

$$\widehat{\boldsymbol{y}} = sigmoid(\boldsymbol{Z_2}) \tag{9}$$

$$\boldsymbol{J} = CE(\boldsymbol{y}, \widehat{\boldsymbol{y}}) \tag{10}$$

The backward propagation:

$$\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{Z_2}} = \widehat{\boldsymbol{y}} - \boldsymbol{y} \triangleq \boldsymbol{\delta_1} \tag{11}$$

$$\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{h}} = \boldsymbol{\delta_1}\boldsymbol{W_2^{\mathrm{T}}} \triangleq \boldsymbol{\delta_2} \tag{12}$$

$$\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{Z_1}} = \boldsymbol{\delta_2} * \sigma'(\boldsymbol{Z_1}) \triangleq \boldsymbol{\delta_3},\ * \, denotes\ element-wise\ product. \tag{13}$$

$$\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{x}} = \boldsymbol{\delta_3}\boldsymbol{W_1^{\mathrm{T}}} \tag{14}$$

(d) $(1 + D_x) \times H + (1 + H) \times D_y$

(e) See **q2_sigmoid.py**

(f) See **q2_gradcheck.py**

(g) See **q2_neural.py**

3 word2vec

(a)

$$\begin{aligned} J_{softmax\_CE}(o, \boldsymbol{v}_c, \boldsymbol{U}) &= CE(\boldsymbol{y}, \widehat{\boldsymbol{y}}) = -\textstyle\sum_i yi \log(\widehat{y}_i) = -\log \widehat{y}_o \\ &= -\log \frac{\exp(\boldsymbol{u}_o^{\mathrm{T}} \boldsymbol{v}_c)}{\sum_{w=1}^{V} \exp(\boldsymbol{u}_w^{\mathrm{T}} \boldsymbol{V}_c)} = -\boldsymbol{u}_o^{\mathrm{T}} \boldsymbol{v}_c + \log \textstyle\sum_{w=1}^{V} \exp(\boldsymbol{u}_w^{\mathrm{T}} \boldsymbol{v}_c) \end{aligned} \tag{15}$$

$$\therefore \frac{\partial J}{\partial \boldsymbol{v}_c} = -\boldsymbol{u_o} + \frac{\sum_{w=1}^{V} \exp(\boldsymbol{u}_w^{\mathrm{T}} \boldsymbol{v}_c) \boldsymbol{u_w}}{\sum_{w=1}^{V} \exp(\boldsymbol{u}_w^{\mathrm{T}} \boldsymbol{v}_c)} = -\boldsymbol{u}_o + \sum_{w=1}^{V} \frac{\exp(\boldsymbol{u}_w^{\mathrm{T}} \boldsymbol{v}_c)}{\sum_{w=1}^{V} \exp(\boldsymbol{u}_w^{\mathrm{T}} \boldsymbol{v}_c)} \boldsymbol{u}_w = -\boldsymbol{u}_o + \textstyle\sum_{w=1}^{V} \widehat{y_w} \boldsymbol{u}_w \tag{16}$$

$$\therefore \frac{\partial J}{\partial \boldsymbol{v}_c} = \textstyle\sum_{w=1}^{V} \widehat{y_w} \boldsymbol{u}_w - \sum_{w=1}^{V} y_w \boldsymbol{u}_o = \boldsymbol{U}(\widehat{\boldsymbol{y}} - \boldsymbol{y}) \tag{17}$$

(b)

$$\frac{\partial J}{\partial \boldsymbol{u}_o} = -\boldsymbol{v}_c + \frac{\exp(\boldsymbol{u}_o^{\mathrm{T}} \boldsymbol{v}_c) \boldsymbol{v}_c}{\sum_{w=1}^{V} \exp(\boldsymbol{u}_w^{\mathrm{T}} \boldsymbol{V}_c)} = (\widehat{\boldsymbol{y}}_o - 1) \boldsymbol{v}_c \tag{18}$$

$$\frac{\partial J}{\partial \boldsymbol{u}_k} = \frac{\exp(\boldsymbol{u}_k^{\mathrm{T}} \boldsymbol{v}_c) \boldsymbol{v}_c}{\sum_{w=1}^{V} \exp(\boldsymbol{u}_w^{\mathrm{T}} \boldsymbol{V}_c)} = \widehat{\boldsymbol{y}}_k \boldsymbol{v}_c, \;\; for\ k \neq o. \tag{19}$$

$$\therefore \frac{\partial J}{\partial \boldsymbol{U}} = \boldsymbol{v}_c (\widehat{\boldsymbol{y}} - \boldsymbol{y})^{\mathrm{T}}, \;\; or \;\; \frac{\partial J}{\partial \boldsymbol{u}_k} = \begin{cases} (\widehat{\boldsymbol{y}}_o - 1) \boldsymbol{v}_c, & k = o \\ \widehat{\boldsymbol{y}}_k \boldsymbol{v}_c, & k \neq o \end{cases} \tag{20}$$

(c)

$$\begin{aligned} \frac{\partial J}{\partial \boldsymbol{v}_c} &= -\frac{1}{\sigma(\boldsymbol{u}_o^{\mathrm{T}} \boldsymbol{v}_c)} \sigma(\boldsymbol{u}_o^{\mathrm{T}} \boldsymbol{v}_c)(1 - \sigma(\boldsymbol{u}_o^{\mathrm{T}} \boldsymbol{v}_c)) \boldsymbol{u}_o + \sum_{k=1}^{K} \frac{1}{\sigma(-\boldsymbol{u}_k^{\mathrm{T}} \boldsymbol{v}_c)} \sigma(-\boldsymbol{u}_k^{\mathrm{T}} \boldsymbol{v}_c)(1 - \sigma(-\boldsymbol{u}_k^{\mathrm{T}} \boldsymbol{v}_c)) \boldsymbol{u}_k \\ &= (\sigma(\boldsymbol{u}_o^{\mathrm{T}} \boldsymbol{v}_c) - 1) \boldsymbol{u}_o - \sum_{k=1}^{K} (\sigma(-\boldsymbol{u}_k^{\mathrm{T}} \boldsymbol{v}_c) - 1) \boldsymbol{u}_k \end{aligned} \tag{21}$$

$$\frac{\partial J}{\partial \boldsymbol{u}_o} = (\sigma(\boldsymbol{u}_o^{\mathrm{T}} \boldsymbol{v}_c) - 1) \boldsymbol{v}_c \tag{22}$$

$$\frac{\partial J}{\partial \boldsymbol{u}_k} = -(\sigma(-\boldsymbol{u}_k^{\mathrm{T}} \boldsymbol{v}_c) - 1) \boldsymbol{v}_c \tag{23}$$

(d) Let U be the collection of all output vectors for all words in the vocabulary.

For Skip-Gram model, the gradients for the cost of one context window are:

$$\frac{\partial J_{skip\_gram}}{\partial \boldsymbol{v}_c} = \sum_{-m \le j \le m, j \neq 0} \frac{\partial \boldsymbol{F}(w_{t+j}, \boldsymbol{v}_c)}{\partial \boldsymbol{v}_c}, \quad \frac{\partial J_{skip\_gram}}{\partial \boldsymbol{U}} = \sum_{-m \le j \le m, j \neq 0} \frac{\partial \boldsymbol{F}(w_{t+j}, \boldsymbol{v}_c)}{\partial \boldsymbol{U}} \tag{24}$$

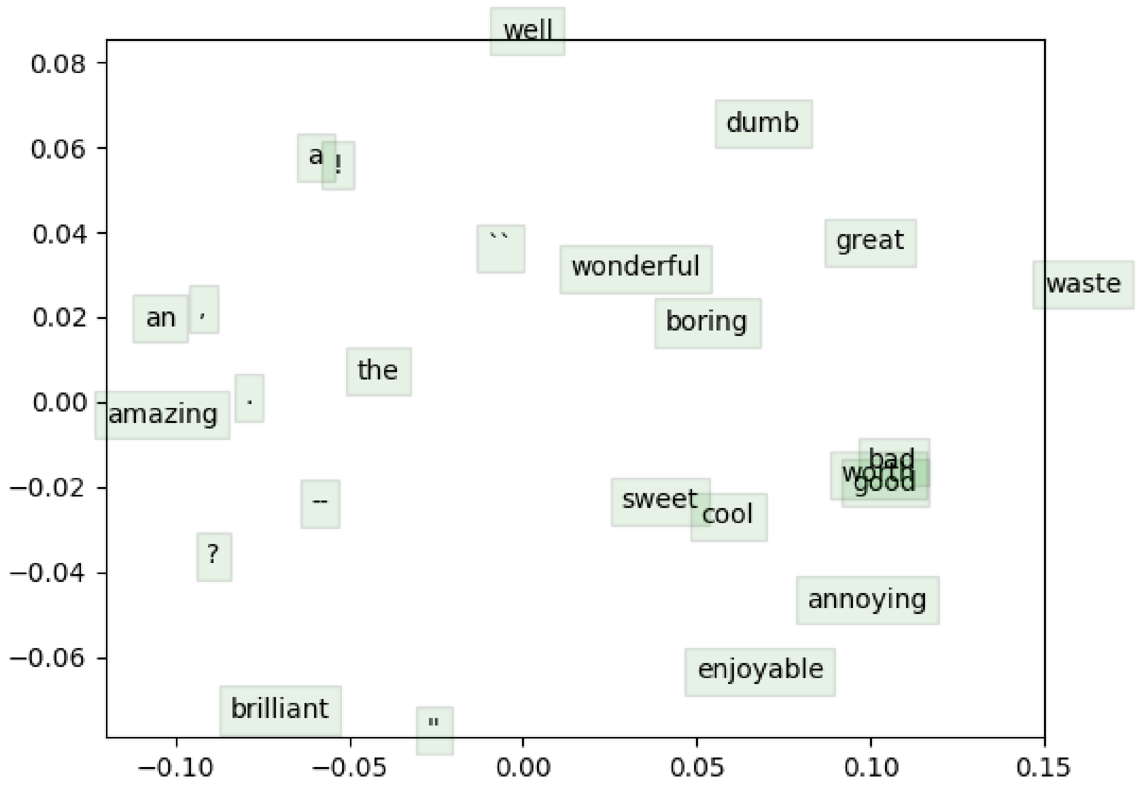

Figure 1: Visualization of word2vec

For CBOW model,

$$\frac{\partial J_{CBOW}}{\partial \boldsymbol{v}_{w_{t+j}}} = \frac{\partial \boldsymbol{F}(w_t, \widehat{\boldsymbol{v}})}{\partial \widehat{\boldsymbol{v}}}, \; for \; all \; j \in \{-m, ..., -1, 1, ...m\} \tag{25}$$

$$\frac{\partial J_{CBOW}}{\partial \boldsymbol{v}_{w_{t+j}}} = 0, \; for \; all \; j \notin \{-m, ..., -1, 1, ...m\} \tag{26}$$

$$\frac{\partial J_{CBOW}}{\partial \boldsymbol{U}} = \frac{\partial \boldsymbol{F}(w_t, \widehat{\boldsymbol{v}})}{\partial \boldsymbol{U}} \tag{27}$$

(e) See **q3_word2vec.py**

(f) See **q3_sgd.py**

(h) See **q3_word2vec.py**

4 Sentiment Analysis

(a) See **q4_sentiment.py**

(c) See **q4_sentiment.py**

(d) See **q4_sentiment.py**

(e) See **q4_sentiment.py**

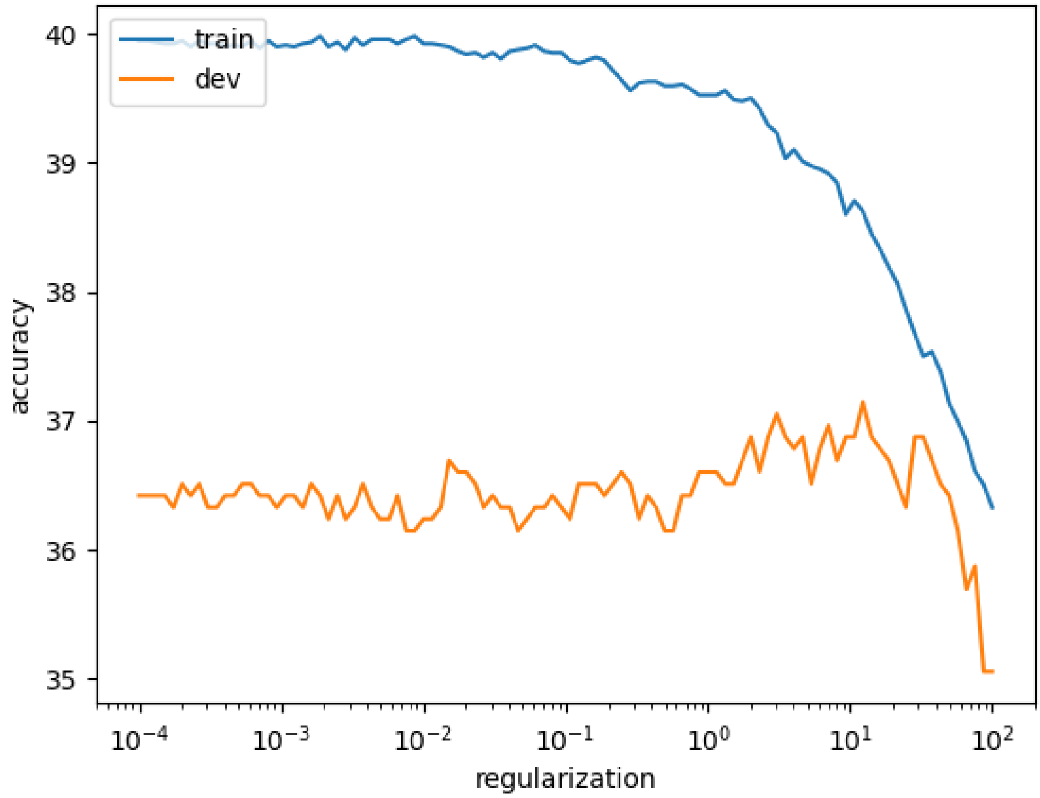

Figure 2: Visualization of word2vec

Table 1: Examples of errors.

True	Pred	Text
3	1	and that 's a big part of why we go to the movies.
4	3	a quiet treasure – a film to be savored.
1	3	an absurdist comedy about alienation , separation and loss.

(f) See **q4_sentiment.py**

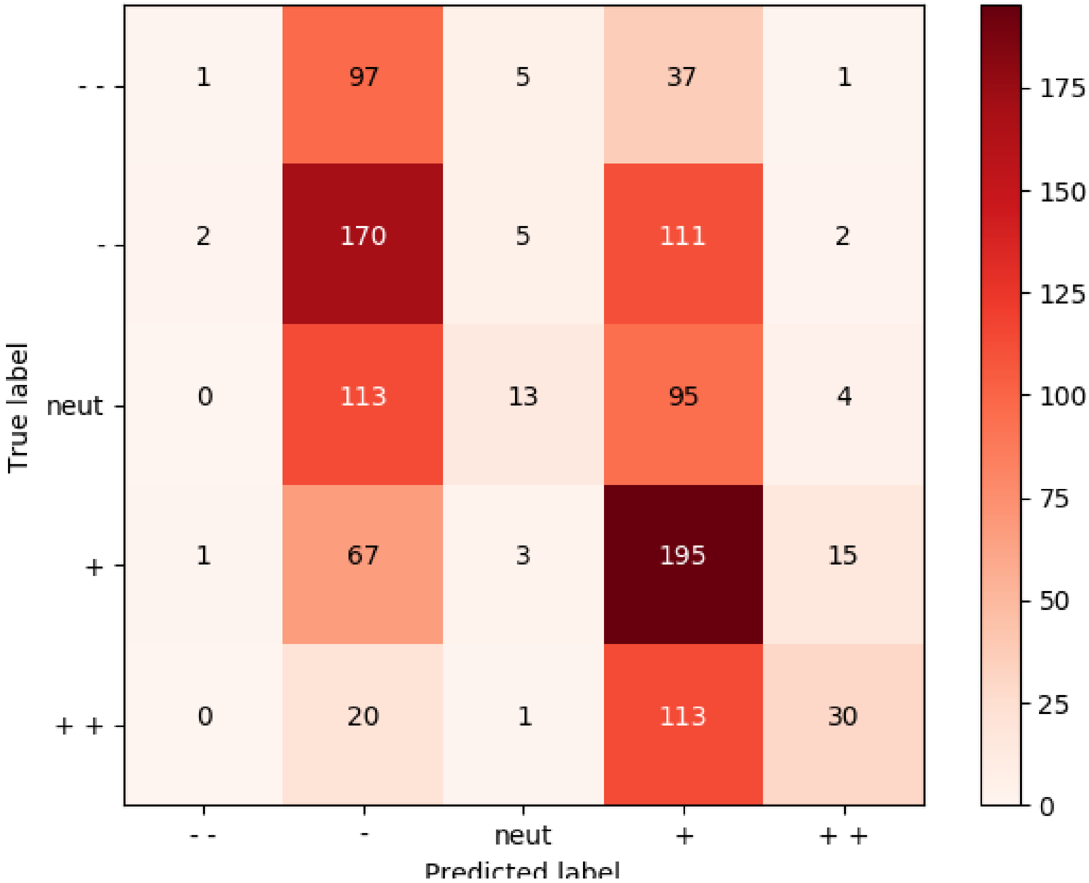

Figure 3: Visualization of word2vec