10-707: Deep Learning, Fall 2017

Homework 3

Lecturer: Russ Salakhutdinov *Name: Yuan Liu(yuanl4)*

1 Problem 1, 4-gram language model

- **Embedding layer:**
  - **Input**: three word index:$w_{i-1}, w_{i-2}, w_{i-3}$.
  - **Parameters:** A 2-dimensional matrix $\boldsymbol{C}$ of size $V \times D$.
  - **Output:** Vector representations for these three words: $\boldsymbol{C}_{w_{i-1},:}, \boldsymbol{C}_{w_{i-2},:}, \boldsymbol{C}_{w_{i-3},:}$, where $\boldsymbol{C}_{j,:}$ means the j^{th} row of the matrix $\boldsymbol{C}$.
- **Embedding to Hidden:**
  - **Input**: $\boldsymbol{C}_{w_{i-1},:}, \boldsymbol{C}_{w_{i-2},:}, \boldsymbol{C}_{w_{i-3},:}$ from the output of the embedding layer. The size of each element of input $1 \times D$.
  - **Parameters**: Three embed_to_hidden_weights matrix $\boldsymbol{W}^{(1)}, \boldsymbol{W}^{(2)}, \boldsymbol{W}^{(3)}$, each matrix's size is $D \times H$. A hidden_bias vector $\boldsymbol{b}^{hidden}$ of size $1 \times H$.
  - **Output**: $\boldsymbol{C}_{w_{i-1},:}\boldsymbol{W}^{(1)} + \boldsymbol{C}_{w_{i-2},:}\boldsymbol{W}^{(2)} + \boldsymbol{C}_{w_{i-3},:}\boldsymbol{W}^{(3)} + \boldsymbol{b}^{hidden}$
- **Tanh Layer**:
  - **Input**: $\boldsymbol{A} = \boldsymbol{C}_{w_{i-1},:}\boldsymbol{W}^{(1)} + \boldsymbol{C}_{w_{i-2},:}\boldsymbol{W}^{(2)} + \boldsymbol{C}_{w_{i-3},:}\boldsymbol{W}^{(3)} + \boldsymbol{b}^{hidden}$ of size $1 \times H$
  - **Parameters:** None
  - **Output**: $tanh(\boldsymbol{A})$
- **Hidden to Output:**
  - **Input**: $\boldsymbol{H} = tanh(\boldsymbol{A})$ of size $1 \times H$.
  - **Parameters**: The hidden_to_output_weight $\boldsymbol{W}^{out}$ of size $H \times V$and the output_bias $\boldsymbol{b}^{out}$ of size $1 \times V$.
  - **Output**: $\boldsymbol{H}\boldsymbol{W}^{out} + \boldsymbol{b}^{out}$
- **Softmax layer:**
  - **Input**: $\boldsymbol{O} = \boldsymbol{H}\boldsymbol{W}^{out} + \boldsymbol{b}^{out}$ of size $1 \times V$.
  - **Parameters**: None
  - **Output**: $\boldsymbol{S}_i = \frac{e^{\boldsymbol{O}_i}}{\sum_{j=1}^{V} e^{\boldsymbol{O}_j}}$, $\boldsymbol{S}$ is a matrix of size $1 \times V$.
- **Loss:** $loss = -\log \boldsymbol{S}_{w_i}$

Now we can calculate the derivation:

$$\frac{\partial loss}{\partial \boldsymbol{S}_{w_i}} = -\frac{1}{\boldsymbol{S}_{w_i}}$$

$$\frac{\partial \boldsymbol{S}_{w_i}}{\partial \boldsymbol{O}_j} = \frac{1(w_i = j)e^{\boldsymbol{O}_{w_i}}(\sum_{k=1}^{V} e^{\boldsymbol{O}_k}) - e^{\boldsymbol{O}_{w_i}} e^{\boldsymbol{O}_j}}{(\sum_{k=1}^{V} e^{\boldsymbol{O}_k})^2}$$

$$\frac{\partial loss}{\partial \boldsymbol{O}_j} = \frac{\partial loss}{\partial \boldsymbol{S}_{w_i}} \frac{\partial \boldsymbol{S}_{w_i}}{\partial \boldsymbol{O}_j} = -(1(w_i = j) - \boldsymbol{S}_j)$$

$$\frac{\partial \boldsymbol{O}_j}{\partial \boldsymbol{H}_i} = \boldsymbol{W}_{i,j}^{\boldsymbol{out}}, \ \frac{\partial \boldsymbol{O}_j}{\partial \boldsymbol{W}_{i,j}^{out}} = \boldsymbol{H}_i, \ \frac{\partial \boldsymbol{O}_j}{\partial \boldsymbol{b}_j^{out}} = 1$$

$$\frac{\partial loss}{\partial \boldsymbol{H}_i} = \sum_{j=1}^{V} \frac{\partial loss}{\partial \boldsymbol{O}_j} \frac{\partial \boldsymbol{O}_j}{\partial \boldsymbol{H}_i}$$

$$\frac{\partial loss}{\partial \boldsymbol{W}_{i,j}^{out}} = \frac{\partial loss}{\partial \boldsymbol{O}_j} \frac{\partial \boldsymbol{O}_j}{\partial \boldsymbol{W}_{i,j}^{out}} = -(1(w_i = j) - \boldsymbol{S}_j)\boldsymbol{H}_i \tag{1}$$

$$\frac{\partial loss}{\partial \boldsymbol{b}_i^{out}} = \frac{\partial loss}{\partial \boldsymbol{O}_i} \frac{\partial \boldsymbol{O}_i}{\partial \boldsymbol{b}_i^{out}} = -(1(w_i = i) - \boldsymbol{S}_i) \tag{2}$$

$$\frac{\partial \boldsymbol{H_j}}{\partial \boldsymbol{A}_i} = (1 - \boldsymbol{H}_j^2) * 1(i = j)$$

$$\frac{\partial \boldsymbol{A}_i}{\partial \boldsymbol{C}_{w_{i-1},j}} = \boldsymbol{W}_{j,i}^{(1)}, \ \frac{\partial \boldsymbol{A}_i}{\partial \boldsymbol{C}_{w_{i-2},j}} = \boldsymbol{W}_{j,i}^{(2)}, \ \frac{\partial \boldsymbol{A}_i}{\partial \boldsymbol{C}_{w_{i-3},j}} = \boldsymbol{W}_{j,i}^{(3)}$$

$$\frac{\partial loss}{\partial \boldsymbol{C}_{w_{i-1},j}} = \sum_{i,k} \frac{\partial loss}{\partial \boldsymbol{H}_i} \frac{\partial \boldsymbol{H_i}}{\partial \boldsymbol{A}_k} \frac{\partial \boldsymbol{A}_k}{\partial \boldsymbol{C}_{w_{i-1},j}} = \sum_{i} \frac{\partial loss}{\partial \boldsymbol{H}_i} \frac{\partial \boldsymbol{H_i}}{\partial \boldsymbol{A}_i} \frac{\partial \boldsymbol{A}_i}{\partial \boldsymbol{C}_{w_{i-1},j}} = \sum_{i} \frac{\partial loss}{\partial \boldsymbol{H}_i}(1 - \boldsymbol{H}_i^2)\boldsymbol{W}_{j,i}^{(1)}$$

$$\frac{\partial loss}{\partial \boldsymbol{C}_{w_{i-1},j}} = \sum_{i} \frac{\partial loss}{\partial \boldsymbol{H}_i}(1 - \boldsymbol{H}_i^2)\boldsymbol{W}_{j,i}^{(1)} \tag{3}$$

$$\frac{\partial loss}{\partial \boldsymbol{C}_{w_{i-2},j}} = \sum_{i} \frac{\partial loss}{\partial \boldsymbol{H}_i}(1 - \boldsymbol{H}_i^2)\boldsymbol{W}_{j,i}^{(2)} \tag{4}$$

$$\frac{\partial loss}{\partial \boldsymbol{C}_{w_{i-3},j}} = \sum_{i} \frac{\partial loss}{\partial \boldsymbol{H}_i}(1 - \boldsymbol{H}_i^2)\boldsymbol{W}_{j,i}^{(3)} \tag{5}$$

$$\frac{\partial \boldsymbol{A}_j}{\partial \boldsymbol{W}_{i,j}^{(1)}} = \boldsymbol{C}_{w_{i-1},i}, \ \frac{\partial \boldsymbol{A}_j}{\partial \boldsymbol{W}_{i,j}^{(2)}} = \boldsymbol{C}_{w_{i-2},i}, \ \frac{\partial \boldsymbol{A}_j}{\partial \boldsymbol{W}_{i,j}^{(3)}} = \boldsymbol{C}_{w_{i-3},i}$$

$$\frac{\partial loss}{\partial \boldsymbol{W}_{i,j}^{(1)}} = \frac{\partial loss}{\partial \boldsymbol{H}_j}(1 - \boldsymbol{H}_j)^2 \boldsymbol{C}_{w_{i-1},i} \tag{6}$$

$$\frac{\partial loss}{\partial \boldsymbol{W}_{i,j}^{(2)}} = \frac{\partial loss}{\partial \boldsymbol{H}_j}(1 - \boldsymbol{H}_j)^2 \boldsymbol{C}_{w_{i-2},i} \tag{7}$$

$$\frac{\partial loss}{\partial \boldsymbol{W}_{i,j}^{(3)}} = \frac{\partial loss}{\partial \boldsymbol{H}_j}(1 - \boldsymbol{H}_j)^2 \boldsymbol{C}_{w_{i-3},i} \tag{8}$$

$$\frac{\partial \boldsymbol{A}_i}{\partial \boldsymbol{b}_j^{hidden}} = 1(i = j)$$

$$\frac{\partial loss}{\partial \boldsymbol{b}_j^{hidden}} = \frac{\partial loss}{\partial \boldsymbol{H}_j}(1 - \boldsymbol{H}_j)^2 \tag{9}$$

The derivation of weights is listed in (1) (9)

2 Problem 2, LSTM & GRU

(a) LSTM contains 3 gates: *input* gates i_t, *forget* gates f_t and *output* gates o_t.
GRU contains 2 gates: *update* gate z_t and *reset* gate r_t.

(b) LSTM:

- *forget* gate: to decide what information should be throw away form the cell state. 1 means 'completely remember' and 0 means 'completely forget'.
- *input* gate: to decide which value in the cell state should be updated. 1 means 'add' and 0 means 'ignore'.
- *output* gate: to decide what to output. 1 means 'output' and 0 means 'don't output'.

GRU:

- *update* gate: to decide which value from $\tilde{h}_t$ should be added to h_t and what value from h_{t-1} should be forget. 1 means the corresponding value from $\tilde{h}_t$ should be remembered and the corresponding value from h_{t-1} should be forgot.
- *reset* gate: to decide what part of h_{t-1} should be computed to get $\tilde{h}_t$.

(c) The output of LSTM is the memory unit $\boldsymbol{C}$ and hidden content h. The output of GRU only contains the hidden content h. The LSTM controls the flow of information according to both $\boldsymbol{C}$ and h, while the GRU only expose the full hidden content without any control.

(d) LSTM: $W_f : n \times m$, $U_f : n \times n$, $b_f : n$, $W_i : n \times m$, $U_i : n \times n$, $b_i : n$,$W_o : n \times m$, $U_o : n \times n$, $b_o : n$, $W_c : n \times m$, $U_c : n \times n$, $b_c : n$. Totally, the LSTM contains $4(n^2 + mn + n)$ parameters.
GRU:$W_z : n \times m$, $U_z : n \times n$, $b_z : n$, $W_r : n \times m$, $U_r : n \times n$, $b_r : n$,$W : n \times m$, $U : n \times n$, $b : n$. Totally, the GRU contains $3(n^2 + mn + n)$ parameters.

(e) The GRU might take less time to train. Because GRU contains fewer parameters than the LSTM. Moreover, GRU only contains 2 gates and its structure is simpler, so it will be more computationally efficient.