NLP - HW2

Omri Efroni, 204037840
Meitar Shechter, 307938217

December 12th, 2020

1

1.1

Assuming j is the entry where $y_j = 1$:
$\frac{\partial CE(y,\hat{y})}{\partial \theta_l} = \frac{\partial(-log(softmax(\theta)_j))}{\partial \theta_l} = \frac{\partial(-\theta_j + log(\sum_k e^{\theta_k}))}{\partial \theta_l}$
For $l = j$:
$= -1 + softmax(\theta_l)$
For $l \neq j$:
$= softmax(\theta_l)$
Or in vector notation:
$\hat{y} - y$

1.2

Note $\theta = hW_2 + b_2$:

$$\frac{\partial CE(y,\hat{y})}{\partial x} = \frac{\partial CE(y,\hat{y})}{\partial \theta} \cdot \frac{\partial \theta}{\partial x} \tag{1}$$

Now:

$$\frac{\partial \theta}{\partial x} = \frac{\partial(hW_2 + b_2)}{\partial x} = \frac{\partial \theta}{\partial h} \cdot \frac{\partial h}{\partial x} = [\frac{\partial \theta}{\partial h} = W_2^T] = W_2^T \cdot \frac{\partial h}{\partial x} \tag{2}$$

Note $\theta^* = xW_1 + b_1$:

$$\frac{\partial h}{\partial x} = \frac{\partial h}{\partial \theta^*} \cdot \frac{\partial \theta^*}{\partial x} = diag(\sigma(xW_1 + b_1) \circ (1 - \sigma(xW_1 + b_1))) \cdot W_1^T \tag{3}$$

Plugging this into equation (2):

$$\frac{\partial \theta}{\partial x} = W_2^T \cdot diag(\sigma(xW_1 + b_1) \circ (1 - \sigma(xW_1 + b_1))) \cdot W_1^T \tag{4}$$

And together with subsection (a):
With the same notation as the previous section, for $l = j$:

$$\frac{\partial CE(y,\hat{y})}{\partial x} = (\hat{y} - y) \cdot W_2^T \cdot diag(\sigma(xW_1 + b_1) \circ (1 - \sigma(xW_1 + b_1))) \cdot W_1^T \tag{5}$$

1.3

CODE

1.4

Our perplexity is: ”dev perplexity : 115.34486812512428”.

2

2.1

Let’s note $\theta_1 = h^{(t)}U + b_2$, $\theta_2 = h^{(t-1)}H + e^{(t)}I + b_1$ and $\delta = \frac{\partial CE(y,\hat{y})}{\partial \theta_1} \cdot \frac{\partial \theta_1}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial \theta_2}$. Now:

$$\frac{\partial J^{(t)}}{\partial b_2} = \frac{\partial CE(y,\hat{y})}{\partial \theta_1} \cdot \frac{\partial \theta_1}{\partial b_2} = (\hat{y} - y) \tag{6}$$

$$\frac{\partial J^{(t)}}{\partial U} = \frac{\partial CE(y,\hat{y})}{\partial \theta_1} \cdot \frac{\partial \theta_1}{\partial U} = (h^{(t)})^T \cdot (\hat{y} - y) \tag{7}$$

$$\frac{\partial J^{(t)}}{\partial b_1}|_{(t)} = \delta \cdot \frac{\partial \theta_2}{\partial b_1} = (\hat{y} - y) \cdot U^T \cdot diag(h^{(t)} \circ (1 - h^{(t)})) \tag{8}$$

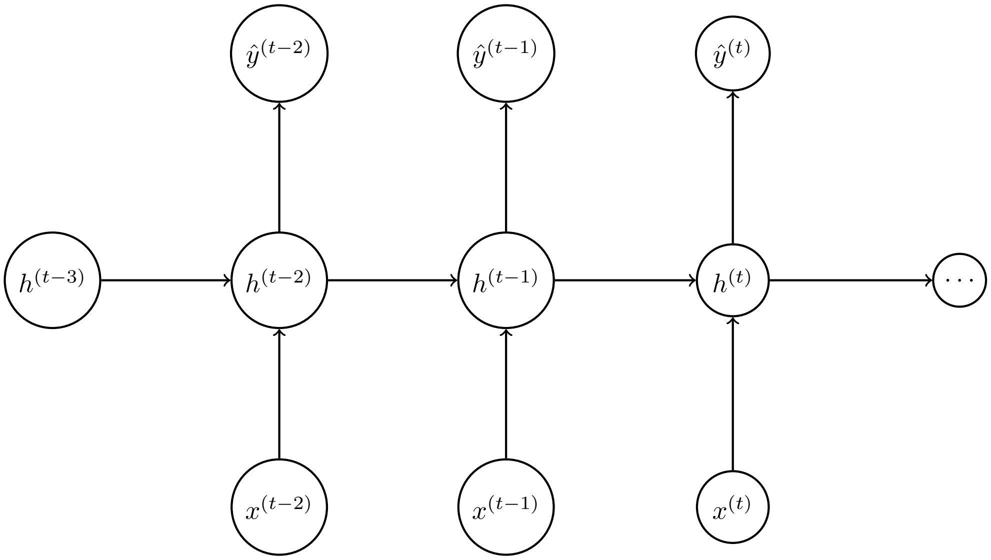
$$\frac{\partial J^{(t)}}{\partial H}|_{(t)} = \delta \cdot \frac{\partial \theta_2}{\partial H} = (h^{(t-1)})^T \cdot \frac{\partial J^{(t)}}{\partial b_1} \tag{9}$$

$$\frac{\partial J^{(t)}}{\partial I}|_{(t)} = \delta \cdot \frac{\partial \theta_2}{\partial I} = (e^{(t)})^T \cdot \frac{\partial J^{(t)}}{\partial b_1} \tag{10}$$

$$\frac{\partial J^{(t)}}{\partial L_{x^{(t)}}} = \delta \cdot \frac{\partial \theta_2}{\partial e^{(t)}} \cdot \frac{\partial e^{(t)}}{\partial L_{x^{(t)}}} = \frac{\partial J^{(t)}}{\partial b_1} \cdot I^T \tag{11}$$

$$\frac{\partial J^{(t)}}{\partial h^{(t-1)}} = \delta \cdot \frac{\partial \theta_2}{\partial h^{(t-1)}} = \frac{\partial J^{(t)}}{\partial b_1} \cdot H \tag{12}$$

2.2



In this subsection our notation is adjusted to the current time step, meaning: $\theta_2 = h^{(t-2)}H + e^{(t-1)}I + b_1$

$$\frac{\partial J^{(t)}}{\partial b_1}|_{(t-1)} = \frac{\partial J^{(t)}}{\partial h^{(t-1)}} \cdot \frac{\partial h^{(t-1)}}{\partial \theta_2} \cdot \frac{\partial \theta_2}{\partial b_1} = \frac{\partial J^{(t)}}{\partial h^{(t-1)}} \cdot diag(h^{(t-1)} \circ (1 - h^{(t-1)})) \tag{13}$$

$$\frac{\partial J^{(t)}}{\partial I}|_{(t-1)} = \frac{\partial J^{(t)}}{\partial h^{(t-1)}} \cdot \frac{\partial h^{(t-1)}}{\partial \theta_2} \cdot \frac{\partial \theta_2}{\partial I} = (e^{(t-1)})^T \cdot \frac{\partial J^{(t)}}{\partial b_1}|_{(t-1)} \tag{14}$$

$$\frac{\partial J^{(t)}}{\partial H}|_{(t-1)} = \frac{\partial J^{(t)}}{\partial h^{(t-1)}} \cdot \frac{\partial h^{(t-1)}}{\partial \theta_2} \cdot \frac{\partial \theta_2}{\partial H} = (h^{(t-2)})^T \cdot \frac{\partial J^{(t)}}{\partial b_1}|_{(t-1)} \tag{15}$$

$$\frac{\partial J^{(t)}}{\partial L_{x^{(t-1)}}} = \frac{\partial J^{(t)}}{\partial h^{(t-1)}} \cdot \frac{\partial h^{(t-1)}}{\partial \theta_2} \cdot \frac{\partial \theta_2}{\partial e^{(t-1)}} \cdot \frac{\partial e^{(t-1)}}{\partial L_{x^{(t-1)}}} = \frac{\partial J^{(t)}}{\partial b_1}|_{(t-1)} \cdot I^T \tag{16}$$

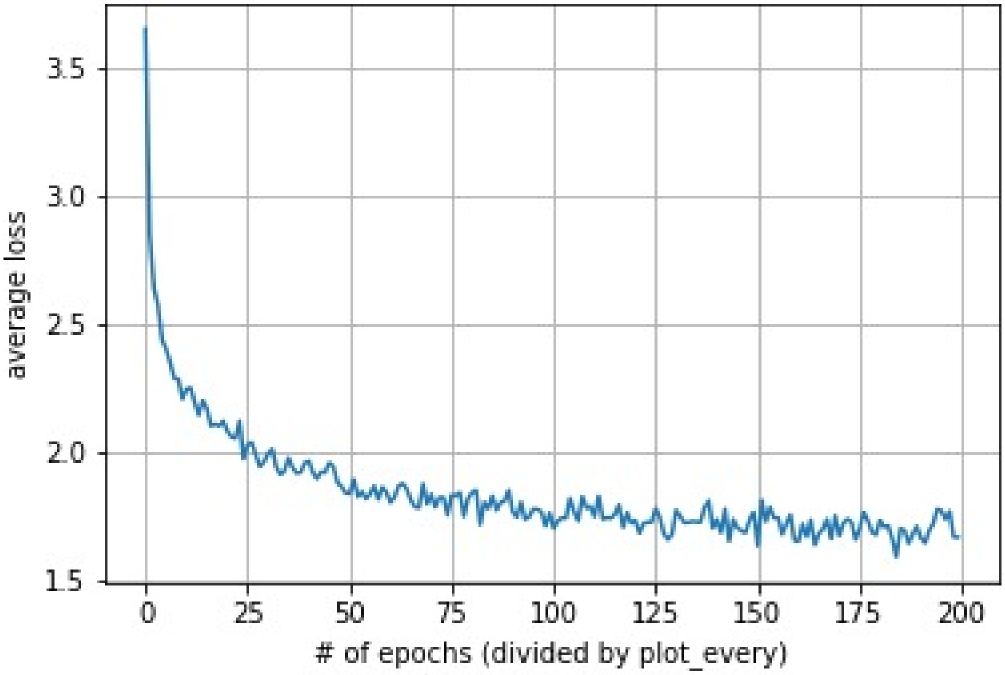
3

3.1

One obvious advantage of a character model is that the size of the vocabulary in much smaller compare to a word model (which reduces the complexity).
Another advantage is that a character-based model can learn temporal and grammar structures (for example) of the language, so the model can generate (theoretically) any word in the English language (for our case), while a word-based model is limited to the chosen vocabulary (for example, the model can learn to generalize how to change a verb to its past-tense, while the actual past-tense verb might not be in the seen data).
On the other hand, a big advantage of a word-based model is that every word

it outputs is a valid word, while a character-based model can output ”Gibrish”. Also, a word-based model has much more context comparing to a character-based model, so it can produce much more coherent sentences with much shorter dependencies.

3.2



4

From logarithms’ rules we know: $log_2(x) = log_2(e) \cdot ln(x)$, therefore:
$2^{-\frac{1}{M}\sum_{i=1}^{M} log_2 p(s_i|s_1,...,s_{i-1})} = 2^{-\frac{1}{M}\sum_{i=1}^{M} ln(p(s_i|s_1,...,s_{i-1}))\cdot log_2(e)} = 2^{-log_2(e)\frac{1}{M}\sum_{i=1}^{M} ln(p(s_i|s_1,...,s_{i-1}))} = e^{-\frac{1}{M}\sum_{i=1}^{M} ln(p(s_i|s_1,...,s_{i-1}))}$