Natural Language Processing
Assignment №2

Pooya Kabiri

Department of Computer Science

Iran University of Science and Technology

April 2021

1 Written: Understanding word2Vec

1.1 a

y is a one-hot vector, $y_m = 1$ if $w = 1$ and $y_m = 0$ if $w \neq 0$, So all y_m elements are 0 except one of them which has a value of 1.

$$\sum_{w \in Vocab} y_w \log(\hat{y}_w) = -[y_1 \log(\hat{y}_1) + \cdots + y_o \log(\hat{y}_o) + \cdots + y_w \log(\hat{y}_w)]$$

$$= -y_o \log(\hat{y}_o) = -\log(\hat{y}_o)$$

1.2 b

Because the softmax loss is the same as cross-entropy loss (from a), the derivatives are also equal (if y was not a one-hot vector this term couldn't be applied). So we calculate the derivative for softmax loss with respect to input vector θ which equals to $v_c U$ (inner product of outside vectors with the current center word vector), then we use chain rule to calculate it with respect to v_c.

From softmax cross-entropy we know that:

$$\frac{\partial\ Softmax\ CrossEntropy}{\partial\ Input\ Vector} = \frac{\partial J}{\partial \theta}$$

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

now if we apply chain derivation:

$$\frac{\partial J}{\partial v_c} = \frac{\partial \theta}{\partial v_c} \frac{\partial J}{\partial \theta} = \frac{\partial v_c U}{\partial v_c} (\hat{y} - y) = U(\hat{y} - y)$$

1.3 c

With using the exact approach in (b) and applying the chain derivation rule:

$$\frac{\partial J}{\partial v_w} = \frac{\partial \theta}{\partial v_w}\frac{\partial J}{\partial \theta}$$

$$\xrightarrow{U=\sum v_w} \frac{\partial \theta}{\partial v_w}\frac{\partial J}{\partial \theta} = \frac{\partial \theta}{\partial U}\frac{\partial J}{\partial \theta} = \frac{\partial v_c U}{\partial U}(\hat{y}-y) = v_c^\intercal(\hat{y}-y)$$

1.4 d

The answer is actually a vector containing all the first order partial derivatives of J with respect to U, it is somehow similar to a Jacobian..

$$\frac{\partial J}{\partial U} = [\frac{\partial J}{\partial u_1}\ \frac{\partial J}{\partial u_2}\ \frac{\partial J}{\partial u_3}\ ...\ \frac{\partial J}{\partial u_{|vocab|}}]$$

1.5 e

For simpler quotient derivation, I multiply the sigmoid function by $\frac{e^x}{e^x}$ which does not reflect any change in the function values. So the sigmoid after is: $\frac{e^x}{e^x+1}$

$$\frac{\partial \sigma(x)}{\partial x} = \frac{\partial \frac{e^x}{e^x+1}}{\partial x} \xrightarrow{Quotient\ Rule} = \frac{e^x(e^x+1)-e^x e^x}{(e^x+1)^2} = \frac{e^x}{(e^x+1)^2} = \sigma(x)(1-\sigma(x))$$

1.6 f

- Respect to u_o

$$\frac{\partial J}{\partial u_o} = \frac{\partial(-\log(\sigma(u_o^\intercal v_c)) - \sum_{k=1}^{K}\log(\sigma(-u_k^\intercal v_c))}{\partial u_o} = \frac{\partial(-\log(\sigma(u_o^\intercal v_c))}{\partial u_o}$$

$$= -(1-\sigma(u_o^\intercal v_c))v_c$$

- Respect to u_k:

  Similar to u_o, but in the second phase of derivation the other term is eliminated:

$$\frac{\partial J}{\partial u_k} = \frac{\partial(-\log(\sigma(u_o^\intercal v_c)) - \sum_{k=1}^{K}\log(\sigma(-u_k^\intercal v_c))}{\partial u_k} = \frac{\partial(-\log(\sigma(-u_k^\intercal v_c))}{\partial u_k}$$

$$= (1-\sigma(-u_k^\intercal v_c))v_c$$

- Respect to v_c:

$$\frac{\partial J}{\partial v_c} = \frac{\partial(-\log(\sigma(u_o^\intercal v_c)) - \sum_{k=1}^{K}\log(\sigma(-u_k^\intercal v_c))}{\partial v_c}$$
$$= -(1-\sigma(u_o^\intercal v_c))u_o + \sum_{k=1}^{K}(1-\sigma(-u_k^\intercal v_c))u_k$$

1.7 g

For derivation with respect to v_c and u_o there will be no difference. With respect to a negative sample u_k, we have a sum over all the negative samples that in derivation with respect to u_k all other terms are eliminated and only the one with u_k remians. When sampling words are not distinct, so there could be n samples which are equal to u_k, so in derivation simply this n comes out of the derivation as a constant and is mirrored in the final derivation term as a simple multiplication:

$$\frac{\partial J}{\partial u_k} = \frac{\partial(-\log(\sigma(u_o^\intercal v_c)) - \sum_{k=1}^{K}\log(\sigma(-u_k^\intercal v_c))}{\partial u_k} = \frac{\partial(-\log(\sigma(-u_k^\intercal v_c))}{\partial u_k}$$
$$= -\frac{\partial[\log(\sigma(-u_1^\intercal v_c) + \log(\sigma(-u_2^\intercal v_c) + ... + \log(\sigma(-u_k^\intercal v_c) + \log(\sigma(-u_k^\intercal v_c) + \log(\sigma(-u_k^\intercal v_c)]}{\partial u_k}$$
$$= -\frac{\partial[\log(\sigma(-u_1^\intercal v_c) + \log(\sigma(-u_2^\intercal v_c) + ... + n*(\log(\sigma(-u_k^\intercal v_c))}{\partial u_k}$$
$$= n(1-\sigma(-u_k^\intercal v_c))v_c$$

For example if there are 3 samples of the same ($n = 3$), the derivative is multiplied by 3.

1.8 h

Same as derivation of J but it is now summed over the window elements.

- Respect to U:

$$\frac{\partial J_{skip-gram}}{\partial U} = \sum_{\substack{-m\leq j\leq m \\ j\neq 0}} \frac{\partial J(v_c, w_{t+j}, U)}{\partial U}$$

- Respect to v_c:

$$\frac{\partial J_{skip-gram}}{\partial v_c} = \sum_{\substack{-m\leq j\leq m \\ j\neq 0}} \frac{\partial J(v_c, w_{t+j}, U)}{\partial v_c}$$

- Respect to v_w, when $w \neq c$:

$$\frac{\partial J_{skip-gram}}{\partial v_w} = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \frac{\partial J(v_c, w_{t+j}, U)}{\partial v_w} = 0$$

  Which is always equal to 0.

2 Coding: Implementing word2vec

2.1 c

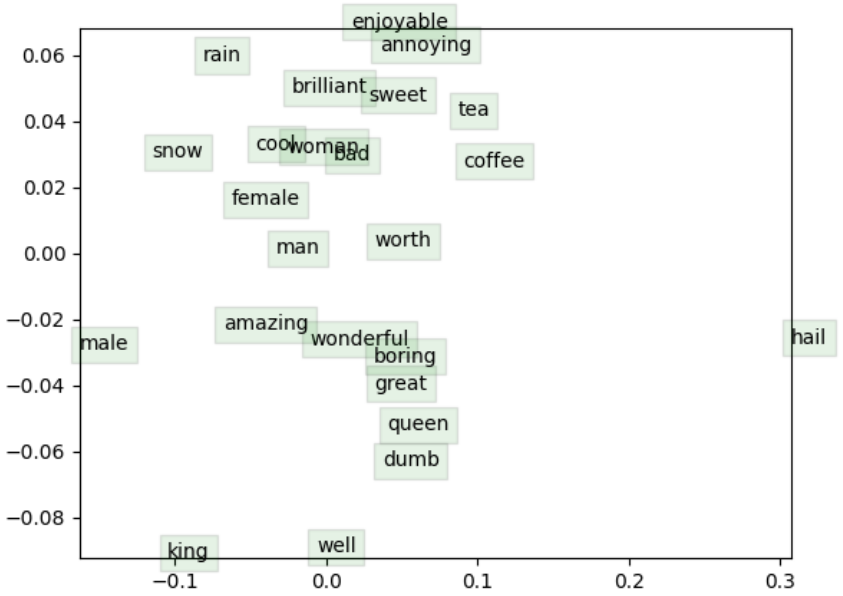

Figure 1: word2vec result

Explanation: In the plot we can see that word with related meaning are clustered together, like female and woman are close to each other, or. adjectives are clustered together like amazing, boring, great, wonderful. We can see some errors also, like male is really apart from man and female, but they share the same semantic group. Also, king and queen are far from each other. If we look more carefully we can see a lot of displacements in the plot.