Question 3

a. $\hat{y}_O = p(o|c) = \frac{e^{u_0^T v_c}}{{}^{e}{}_{w=1}^{|W|}\; e^{u_w^T v_c}}$

Assuming cross-entropy cost is applied to this prediction and word O is the expected word.

$$J_{softmax-CE}(o, v_c, U) = CE(y, \hat{y}) = -\log \frac{e^{u_0^T v_c}}{{}^{e}{}_{w=1}^{|W|}\; e^{u_w^T v_c}}$$

let us denote:

$$z_j = u_j^T v_c$$

$$1_{i,j} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

Now, let us first calculate the partial derivative of J w.r.t any z_k:

$$\frac{\partial J}{\partial z_k} = -1_{k,o} + \frac{e^{z_k}}{{}^{e}{}_{w=1}^{|W|}\; e^{z_w}}$$

And now to the partial derivative w.r.t v_c:

$$\frac{\partial J}{\partial v_c} \underbrace{=}_{chain\,rule} \sum_{w=1}^{|W|} \frac{\partial J}{\partial z_w}\frac{\partial z_w}{\partial v_c} = \sum_{w=1}^{|W|}\left[\left(-1_{w,o} + \frac{e^{z_w}}{{}^{e}{}_{k=1}^{|W|}\; e^{k}}\right)\frac{\partial z_w}{\partial v_c}\right]$$

$$= \sum_{w=1}^{|W|}\left(-1_{w,o} + \frac{e^{z_w}}{{}^{e}{}_{k=1}^{|W|}\; e^{k}}\right)u_w = -u_o + \sum_{w=1}^{|W|} p(w|c)u_w$$

b. Derived by using the chain rule:

$$\frac{\partial J}{\partial u_i} = \frac{\partial J}{\partial z_i}\frac{\partial z_i}{\partial u_i} = \left(P(i|c) - 1_{i,o}\right)\cdot v_c$$

c. In this case, the new cost function is defined as:

$$J_{neg-sample}(o, v_c, U) = -\log\left(\sigma(u_o^T v_c)\right) - \sum_{k=1}^{K}\log\left(\sigma(-u_k^T v_c)\right)$$

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

$$= -\frac{1}{\sigma(u_o^T v_c)}\cdot\left[\left(1 - \sigma(u_o^T v_c)\right)\sigma(u_o^T v_c)\right]u_o - \sum_{k=1}^{K}\frac{\partial}{\partial v_c}\log\left(\sigma(-u_k^T v_c)\right)$$

$$= (\sigma(u_o^T v_c) - 1)u_o + \sum_{k=1}^{K}\frac{\left(1 - \sigma(-u_k^T v_c)\right)\sigma(-u_k^T v_c)u_k}{\sigma(-u_k^T v_c)}$$

$$= (\sigma(u_o^T v_c) - 1)u_o + \sum_{k=1}^{K}\left(1 - \sigma(-u_k^T v_c)\right)u_k$$

$$\underbrace{=}_{\sigma(-x)=1-\sigma(x)} (\sigma(u_o^T v_c) - 1)u_o + \sum_{k=1}^{K}\sigma(u_k^T v_c)u_k$$

c. (continued):

- $\frac{\partial J}{\partial u_o} = (\sigma(u_o^T v_c) - 1)v_c$ Since we're assuming $o \notin K$.
- For $k \in K$: $\frac{\partial J}{\partial u_k} = -\frac{\partial}{\partial u_k}\log\Big(\sigma(-u_k^T v_c)\Big) = -\frac{\big(1-\sigma(-u_k^T v_c)\big)\sigma(-u_k^T v_c)(-v_c)}{\sigma(-u_k^T v_c)}$
  $= v_c\Big(1 - \sigma(-u_k^T v_c)\Big) \underbrace{=}_{\sigma(-x)=1-\sigma(x)} v_c\sigma(u_k^T v_c)$
- For $i \notin K \cup \{o\}$: $\frac{\partial J}{\partial u_i} = 0$

It is much more efficient to compute the negative sample cost function because we only have to sum over $|K| \ll |W|$, instead of summing over the entire vocabulary.

d. Gradients: In the case of skip-gram, we can use the fact that the derivative of a sum is the derivative of every item in the sum. Therefore:

$$\frac{\partial J}{\partial U} = \sum_{-m \le j \le m, j \ne 0} \frac{\partial}{\partial U} F\left(w_{c+j}, v_c\right)$$

and for v_c:

$$\frac{\partial J}{\partial v_c} = \sum_{-m \le j \le m, j \ne 0} \frac{\partial}{\partial v_c} F\left(w_{c+j}, v_c\right)$$

In both cases, we already know how to calculate the derivatives. All that’s left is to sum them up.

For v_i, in cases where $i \ne c$ the derivative is simply 0 because the cost function is not a function of v_i for $i \ne c$. It is a function of o, v_c and U.

<u>For CBOW</u>: In this case, we want to predict the center words given its context:

$$\hat{v} = \sum_{-m \le j \le m, j \ne 0} v_{c+j}$$

$$J_{CBOW}\left(word_{c-m,\dots,c+m}\right) = F(w_c, \hat{v})$$

$$\frac{\partial J}{\partial v_i} = \frac{\partial}{\partial v_i} F(w_c, \hat{v}) = \frac{\partial F}{\partial \hat{v}} \frac{\partial \hat{v}}{\partial v_i} = \begin{cases} \frac{\partial F}{\partial \hat{v}} & i \in \{c-m, \dots, c-1, c+1, \dots, c+m\} \\ 0 & else \end{cases}$$

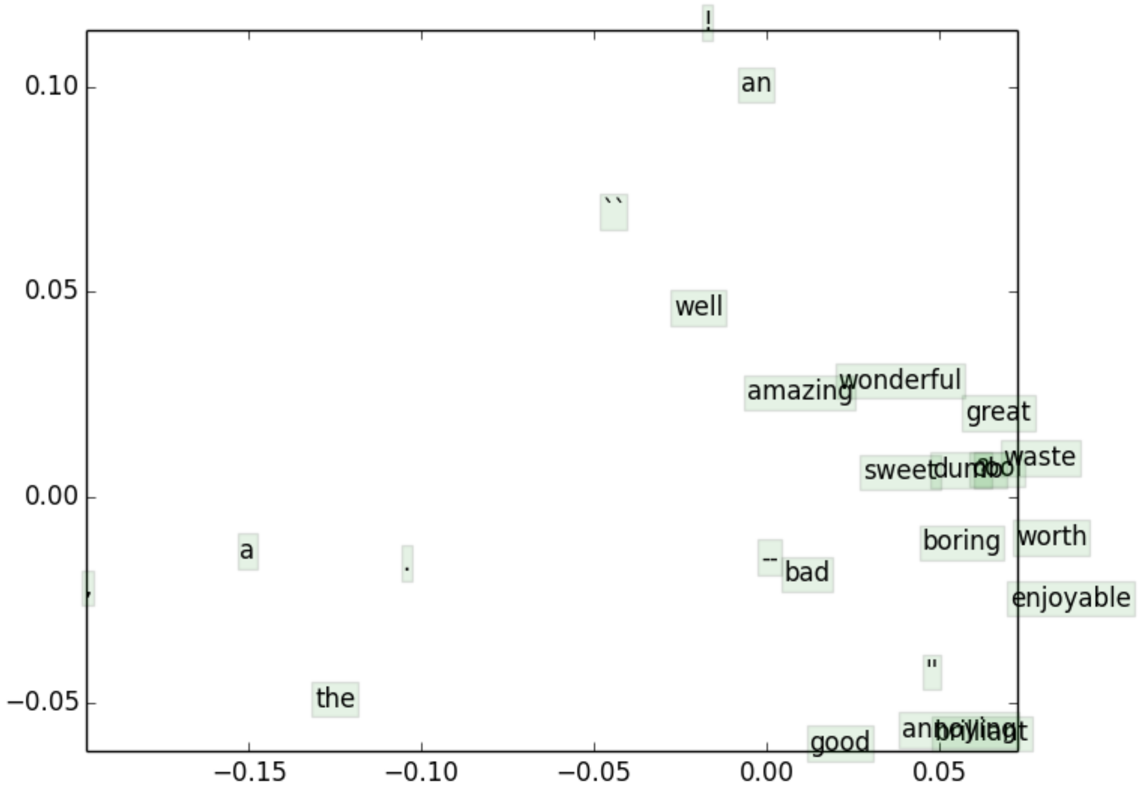
The reasoning is as follows:

Since $\hat{v} = \sum_{-m \le j \le m, j \ne 0} v_{c+j}$, the derivative of the sum is the derivative of every part of the sum and so:

$$\frac{\partial}{\partial v_i} \hat{v} = \frac{\partial}{\partial v_i} \sum_{-m \le j \le m, j \ne 0} v_{c+j} = \sum_{-m \le j \le m, j \ne 0} \frac{\partial}{\partial v_i} v_{c+j}$$
$$= \begin{cases} 1 & i \in \{c-m, \dots, c-1, c+1, \dots, c+m\} \\ 0 & else \end{cases}$$

$$\frac{\partial J}{\partial U} = \frac{\partial F(w_c, \hat{v})}{\partial U}$$

g. In the following plot we can see a 2D representation of two features in the word vectors, and how the words are spaced in the subspace spanned by those features. We are taking the two features with the highest variance (i.e. most “info”), from the SVD decomposition.


!
an
` `
well
amazing
wonderful
great
waste
sweet
dumb
worth
boring
a
bad
enjoyable
"
the
good
annoying
brilliant