David Freilich SkipGram Derivation

$$J_{softmax\text{-}CE}(o, v_c, U) = -\sum_{j=1}^{V} \log(\hat{y}_j) y_j \qquad \hat{y}_o = p(o|c) = \frac{\exp(u_o^T v_c)}{\sum \exp(u_w^T v_c)}$$

$$\frac{\partial soft}{\partial v_c} = u_w^T (y - \hat{y})$$

$$\frac{\partial soft_j}{\partial u_w} = \frac{\partial J}{\partial u_w} = -y \cdot \frac{\partial \log \hat{y}}{\partial u_w} \Rightarrow \text{we only need to calculate for } y = 1, \text{ at } y_o$$

$$= -y_o \cdot \frac{\partial \log \hat{y}}{\partial u_w} \Rightarrow \text{we'll split tederivations now, for } u_o \text{ \& } u_k$$

$\boxed{y_o = 1}$

$$\frac{\partial J}{\partial u_o} = -\frac{\partial \log \hat{y}}{\partial u_o} = -\left(\frac{\partial u_o^T v_c}{\partial u_o} - \frac{\partial \sum \exp(u_w^T v_c)}{\partial u_o}\right) = -\left(\frac{\partial u_o^T v_c}{\partial u_o} - \frac{1}{\sum \exp(u_w^T v_c)} \cdot \frac{\partial \sum \exp(u_w^T v_c)}{\partial u_o}\right)$$

Expanding out for each k, we'd get:

$$= -\left(\frac{\exp(u_k^T v_c)}{\sum \exp(u_w^T v_c)} \cdot \frac{\partial u_k^T v_c}{\partial u_o}\right)$$

which would be 0 for all $k \neq 0$, so we simplify as:

$$= -\left(\frac{\partial u_o^T v_c}{\partial u_o} - \frac{\exp(u_o^T v_c)}{\sum \exp(u_w^T v_c)} \cdot \frac{\partial u_o^T v_c}{\partial u_o}\right) = -\left(v_c^T - p(o|c) \cdot v_c^T\right) = \boxed{(\hat{y} - 1) v_c^T}$$

$$\frac{\partial J}{\partial u_k} = -y_o \cdot \frac{\partial \log \hat{y}}{\partial u_k} = -\left(\frac{\partial u_o^T v_c}{\partial u_k} - \frac{\partial \log \sum \exp(u_k^T v_c)}{\partial u_k}\right)$$

$$= -\left(0 - \frac{1}{\sum \exp(u_w^T v_c)} \cdot \frac{\sum \exp(u_w^T v_c) \cdot \partial u_w^T v_c}{\partial u_k}\right)$$

The only non zero one will be the k we are looking at

$= p(k|c) \cdot v_c$ for each k where $k \neq 0$

$$= \boxed{\hat{y}_k v_c^T}$$

David Freilich Negative Sampling

$$J_{neg\text{-}sample} = -\log(\sigma(v_o^T v_c)) - \sum_{k}^{K} \log(\sigma(-v_k^T v_c))$$

$$\frac{\partial J_{neg}}{\partial v_c} = -\left(\frac{1}{\sigma(v_o^T v_c)} \cdot \frac{\sigma(v_o^T v_c)\cdot(1-\sigma(v_o^T v_c))}{1} \cdot \frac{v_o^T}{1}\right) - \sum_{k}^{K}\left(\frac{1}{\sigma(-v_k^T v_c)} \cdot \frac{\sigma(-v_k^T v_c)(1-\sigma(-v_k^T v_c))}{1} \cdot \frac{-v_k^T}{1}\right)$$

$$\boxed{\frac{\partial J}{\partial v_c} = \left(\sigma(v_o^T v_c) - 1\right)\cdot v_o^T \qquad - \sum_{k}^{K}\left(\sigma(-v_k^T v_c) - 1\right)\cdot v_k^T}$$

Assumptions:

$$\frac{\partial \log x}{\partial x} = \frac{1}{x}$$

$$\frac{\partial \sigma x}{\partial x} = \frac{(\sigma x)(1-\sigma x)}{1}$$

$$\frac{\partial J_{neg}}{\partial v_k} = -\partial \log \sigma(v_o^T v_c) - \sum^{K} \log \sigma(-v_k^T v_c)$$

$$= -\left(\frac{1}{\sigma(v_o^T v_c)} \cdot \frac{\sigma(v_o^T v_c)(1-\sigma(v_o^T v_c)}{1} \cdot 0\right) - \sum_{k}^{K} \frac{1}{\sigma(-v_k^T v_c)} \cdot \frac{\sigma(-v_k^T v_c)(1-\sigma(-v_k^T v_c)}{1} \cdot \frac{-v_c}{1}$$

$$\frac{\partial J}{\partial v_k} = 0 \qquad - \sum_{k}^{K}\left(\sigma(-v_k^T v_c) - 1\right)\cdot v_c$$

Since we're ~~don't~~ derivate gradient of each k alone, the summation disappears, so we have:

$$\boxed{\frac{\partial J}{\partial v_k} = -\left(\sigma(-v_k^T v_c) - 1\right)\cdot v_c}$$