Word Vectors

$$P(O = o \mid C = c) = \frac{exp(u_0^T v_c)}{\sum_{w \in W} exp(u_w^T v_c)}$$

$$J_{naive-softmax}(v_c, o, U) = -logP(O = o \mid C = c)$$

(a)

$softmax(x) = softmax(x + c)$:

$$softmax(x+c)_i = \frac{exp((x+c)_i)}{\sum_j exp((x+c)_j)} = \frac{exp(x_i + c)}{\sum_j exp(x_j + c)} = \frac{exp(x_i) \cdot exp(c)}{\sum_j exp(x_j) \cdot exp(c)} =$$

$$\frac{exp(x_i)}{\sum_j exp(x_j)} = softmax(x)_i$$

(b)

$-\sum_{w \in W} y_w log(\hat{y}_w) = -log(\hat{y}_o)$:

$$-\sum_{w \in W} y_w log(\hat{y}_w) = -\sum_{w \in W, w \neq o} y_w log(\hat{y}_w) - y_o log(\hat{y}_o) \underset{(1)}{=} -\sum_{w \in W, w \neq o} 0 \cdot log(\hat{y}_w) - 1 \cdot log(\hat{y}_o) = -log(\hat{y}_o)$$

as the transition in (1) follows from the definition of $\hat{y}_w = \begin{cases} 0, w = o \\ 1, w = o \end{cases}$

(c)

$\frac{\partial}{\partial v_c} J_{naive-softmax}(v_c, o, U) =$

$$\frac{\partial}{\partial v_c}[-logP(O = o \mid C = c)] = \frac{\partial}{\partial v_c}[-log(\frac{exp(u_0^T v_c)}{\sum_{w \in W} exp(u_w^T v_c)})] =$$

$$\frac{\partial}{\partial v_c}[-log(exp(u_0^T v_c)) + log(\sum_{w \in W} exp(u_w^T v_c))] = \frac{\partial}{\partial v_c}[-u_0^T v_c + log(\sum_{w \in W} exp(u_w^T v_c))] =$$

$$\frac{\partial}{\partial v_c}[-u_0^T v_c] + \frac{\partial}{\partial v_c}[log(\sum_{w \in W} exp(u_w^T v_c))] \underset{(1)}{=} -u_0 + \frac{\partial}{\partial v_c}[log(\sum_{w \in W} exp(u_w^T v_c))] \underset{(2)}{=}$$

$$[-u_0 + \frac{\frac{\partial}{\partial v_c} \sum_{w \in W} exp(u_w^T v_c))}{\sum_{w \in W} exp(u_w^T v_c))} = -u_0 + \frac{\sum_{w \in W} \frac{\partial}{\partial v_c} exp(u_w^T v_c))}{\sum_{w \in W} exp(u_w^T v_c))} \underset{(3)}{=}$$

$$-u_0 + \frac{\sum_{w \in W} \frac{\partial}{\partial v_c}(u_w^T v_c) \cdot exp(u_w^T v_c))}{\sum_{w \in W} exp(u_w^T v_c))} = -u_0 + \frac{\sum_{w \in W} u_w \cdot exp(u_w^T v_c))}{\sum_{w \in W} exp(u_w^T v_c))} =$$

$$-u_0 + \sum_{w \in W} u_w \cdot \frac{exp(u_w^T v_c))}{\sum_{w' \in W} exp(u_{w'}^T v_c))} = -u_0 + \sum_{w \in W} u_w \cdot P(O = w \mid C = c) \underset{(4)}{=}$$

$$-U \cdot y + U \cdot \hat{y} = U[\hat{y} - y]$$

$$\implies \frac{\partial}{\partial v_c} J_{naive-softmax}(v_c, o, U) = -U \cdot y + U \cdot \hat{y} = U[\hat{y} - y]$$

Transitions follow from:
(1) $\frac{\partial}{\partial a} b^T a = b$
(2) $\frac{\partial}{\partial x} log(f(x)) = \frac{\frac{\partial}{\partial x} f(x)}{f(x)}$
(3) $\frac{\partial}{\partial x} exp(f(x)) = \frac{\partial}{\partial x} f(x) \cdot exp(f(x))$
(4) $U \cdot y = u_o$, $U \cdot \hat{y} = \sum_{w \in W} y_w \cdot u_w = \sum_{w \in W} P(O = w \mid C = c) \cdot u_w$

(d)

$\underline{\frac{\partial}{\partial u_w} J_{naive-softmax}(v_c, o, U) =}$

$$\frac{\partial}{\partial u_w}[-logP(O = o \mid C = c)] = \frac{\partial}{\partial u_w}[-log(\frac{exp(u_0^T v_c)}{\sum_{w \in W} exp(u_w^T v_c)})] =$$

$$\frac{\partial}{\partial u_w}[-log(exp(u_0^T v_c)) + log(\sum_{w \in W} exp(u_w^T v_c))] = \frac{\partial}{\partial u_w}[-u_0^T v_c + log(\sum_{w \in W} exp(u_w^T v_c))] =$$

$$\frac{\partial}{\partial u_w}[-u_0^T v_c] + \frac{\partial}{\partial u_w}[log(\sum_{w \in W} exp(u_w^T v_c))] \underset{(2)}{=} \frac{\partial}{\partial u_w}[-u_0^T v_c] + \frac{\frac{\partial}{\partial u_w} \sum_{w \in W} exp(u_w^T v_c))}{\sum_{w \in W} exp(u_w^T v_c))} =$$

$$\frac{\partial}{\partial u_w}[-u_0^T v_c] + \frac{\sum_{w \in W} \frac{\partial}{\partial u_w} exp(u_w^T v_c))}{\sum_{w \in W} exp(u_w^T v_c))} \underset{(3)}{=} \frac{\partial}{\partial u_w}[-u_0^T v_c] + \frac{\sum_{w \in W} \frac{\partial}{\partial u_w}(u_w^T v_c) \cdot exp(u_w^T v_c))}{\sum_{w \in W} exp(u_w^T v_c))} =$$

$$\frac{\partial}{\partial u_w}[-u_0^T v_c] + \frac{\sum_{w' \in W, w' \neq w} \frac{\partial}{\partial u_w}(u_{w'}^T v_c) \cdot exp(u_{w'}^T v_c)) + \frac{\partial}{\partial u_w}(u_w^T v_c) \cdot exp(u_w^T v_c))}{\sum_{w' \in W} exp(u_{w'}^T v_c))} = (\cdot)$$

Transitions, again, follow from:
(2) $\frac{\partial}{\partial x} log(f(x)) = \frac{\frac{\partial}{\partial x} f(x)}{f(x)}$
(3) $\frac{\partial}{\partial x} exp(f(x)) = \frac{\partial}{\partial x} f(x) \cdot exp(f(x))$

if $w = o$ then

$$(\cdot) = \frac{\partial}{\partial u_o}[-u_0^T v_c] + \frac{\sum_{w' \in W, w' \neq o} \frac{\partial}{\partial u_o}(u_{w'}^T v_c) \cdot exp(u_{w'}^T v_c)) + \frac{\partial}{\partial u_o}(u_o^T v_c) \cdot exp(u_o^T v_c))}{\sum_{w' \in W} exp(u_{w'}^T v_c))} \underset{(1)}{=}$$

$$-v_c + \frac{v_c \cdot exp(u_o^T v_c)}{\sum_{w' \in W} exp(u_{w'}^T v_c))} = -v_c + v_c \cdot \frac{exp(u_o^T v_c)}{\sum_{w' \in W} exp(u_{w'}^T v_c))} = -v_c + v_c \cdot P(O = o \mid C = c) =$$
$$-v_c + v_c \cdot \hat{y}_o$$

and if $w \neq o$ then:

$$(\cdot) = \frac{\partial}{\partial u_w}[-u_0^T v_c] + \frac{\sum_{w' \in W, w' \neq w} \frac{\partial}{\partial u_w}(u_{w'}^T v_c) \cdot exp(u_{w'}^T v_c)) + \frac{\partial}{\partial u_w}(u_w^T v_c) \cdot exp(u_w^T v_c))}{\sum_{w' \in W} exp(u_{w'}^T v_c))} \underset{(1)}{=}$$

$$\frac{v_c \cdot exp(u_w^T v_c)}{\sum_{w' \in W} exp(u_{w'}^T v_c))} = v_c \cdot \frac{exp(u_w^T v_c)}{\sum_{w' \in W} exp(u_{w'}^T v_c))} = v_c \cdot P(O = w \mid C = c) = v_c \cdot \hat{y}_w$$

As for each $w \neq w', \frac{\partial}{\partial u_w} f(u_{w'}) = 0$ and again, (1) $\frac{\partial}{\partial a} b^T a = b$

$\implies \frac{\partial}{\partial u_w} J_{naive-softmax}(v_c, o, U) = \begin{cases} v_c \cdot \hat{y}_w, w \neq o \\ -v_c + v_c \cdot \hat{y}_o, w = o \end{cases}$

(e)

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

$\underline{\frac{\partial}{\partial x} \sigma(x) =}$

$$\frac{\partial}{\partial x}[\frac{exp(x)}{exp(x)+1}] \underset{(1)}{=} \frac{\frac{\partial}{\partial x}[exp(x)] \cdot (exp(x)+1) - \frac{\partial}{\partial x}[exp(x)+1] \cdot exp(x)}{(exp(x)+1)^2} =$$

$$\frac{exp(x) \cdot \frac{\partial}{\partial x}[exp(x)] + \frac{\partial}{\partial x}[exp(x)] - exp(x) \cdot \frac{\partial}{\partial x}[exp(x)]}{(exp(x)+1)^2} = \frac{\frac{\partial}{\partial x} exp(x)}{(exp(x)+1)^2}$$

$\frac{\partial}{\partial x} \sigma(x)_{ij} =$

$$\frac{\partial}{\partial x_j} \sigma(x_i) = \frac{\frac{\partial}{\partial x_j} exp(x_i)}{(exp(x_i)+1)^2} = \begin{cases} 0, i \neq j \\ \frac{exp(x_i)}{(exp(x_i)+1)^2}, i = j \end{cases} = \begin{cases} 0, i \neq j \\ \sigma(x_i) \cdot \sigma(-x_i), i = j \end{cases}$$

$\implies \frac{\partial}{\partial x} \sigma(x) = Diag((\sigma(x_i) \cdot \sigma(-x_i))_i)$

Where $Diag((y_i)_i)$ is the diagonal matrix with x_i on it's diagonal line.

(1) $\frac{\partial}{\partial x} \frac{f(x)}{g(x)} = \frac{\frac{\partial}{\partial x}[f(x)] \cdot g(x) - \frac{\partial}{\partial x}[g(x)] \cdot f(x)}{g(x)^2}$

(2) $\frac{\partial}{\partial x} exp(x)_{ij} = \frac{\partial}{\partial x_j} exp(x_i) = \begin{cases} 0, i \neq j \\ exp(x_i), i = j \end{cases}$

(3) $\frac{exp(x_i)}{(exp(x_i)+1)^2} = \frac{exp(x_i)}{exp(x_i)+1} \cdot \frac{1}{exp(x_i)+1} = \sigma(x_i) \cdot \sigma(-x_i)$

(f)

$$J_{neg-sample}(v_c, o, U) = -log(\sigma(u_0^T v_c)) - \sum_{k=1}^{K} log(\sigma(-u_k^T v_c))$$

$\underline{\frac{\partial}{\partial v_c}[J_{neg-sample}(v_c, o, U)] =}$

$$\frac{\partial}{\partial v_c}[-log(\sigma(u_0^T v_c)) - \sum_{k=1}^{K} log(\sigma(-u_k^T v_c))] = -\frac{\partial}{\partial v_c}[log(\sigma(u_0^T v_c))] - \sum_{k=1}^{K} \frac{\partial}{\partial v_c}[log(\sigma(-u_k^T v_c))] \underset{(2)}{=}$$

$$-\frac{\frac{\partial}{\partial v_c}[\sigma(u_0^T v_c)]}{\sigma(u_0^T v_c)} - \sum_{k=1}^{K} \frac{\frac{\partial}{\partial v_c}[\sigma(-u_k^T v_c))]}{\sigma(-u_k^T v_c))} = (\cdot)$$

Define for every $w \in W$:
$x_w \underset{(1)}{=} u_w^T v_c, P_w = \sigma(x_w), N_w = \sigma(-x_w)\ \frac{\partial}{\partial v_c} x_w = u_w$
$\frac{\partial}{\partial v_c} P_w = \frac{\partial}{\partial x_w}\sigma(x_w) \cdot \frac{\partial x_w}{\partial v_c} \underset{(4)}{=} \sigma(x_w) \cdot \sigma(-x_w) \cdot u_w = P_w \cdot N_w \cdot u_o$
$\frac{\partial}{\partial v_c} N_w = \frac{\partial}{\partial x_w}\sigma(-x_w) \cdot \frac{\partial x_w}{\partial v_c} = \underset{(4)}{-}\sigma(x_w) \cdot \sigma(-x_w) \cdot u_w = -P_w \cdot N_w \cdot u_w$

$$(\cdot) = -\frac{\frac{\partial}{\partial v_c}[P_o]}{P_o} - \sum_{k=1}^{K} \frac{\frac{\partial}{\partial v_c}[N_k]}{N_k} = -\frac{P_o \cdot N_o \cdot u_o}{P_o} - \sum_{k=1}^{K} \frac{-P_k \cdot N_k \cdot u_k}{N_k} = -N_o \cdot u_o + \sum_{k=1}^{K} P_k \cdot u_k =$$

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

$\implies \frac{\partial}{\partial v_c}[J_{neg-sample}(v_c, o, U)] = -\sigma(-u_o^T v_c) \cdot u_o + \sum_{k=1}^{K} \sigma(u_k^T v_c) \cdot u_k$

$\underline{\frac{\partial}{\partial u_w}[J_{neg-sample}(v_c, o, U)] =}$

$$\frac{\partial}{\partial u_w}[-log(\sigma(u_0^T v_c)) - \sum_{k=1}^{K} log(\sigma(-u_k^T v_c))] = -\frac{\partial}{\partial u_w}[log(\sigma(u_0^T v_c))] - \sum_{k=1}^{K} \frac{\partial}{\partial u_w}[log(\sigma(-u_k^T v_c))] \underset{(2)}{=}$$

$$-\frac{\frac{\partial}{\partial u_w}[\sigma(u_0^T v_c)]}{\sigma(u_0^T v_c)} - \sum_{k=1}^{K} \frac{\frac{\partial}{\partial u_w}[\sigma(-u_k^T v_c))]}{\sigma(-u_k^T v_c))} = -\frac{\frac{\partial}{\partial u_w}[P_o]}{P_o} - \sum_{k=1}^{K} \frac{\frac{\partial}{\partial u_w}[N_k]}{N_k} = (\cdot)$$

$\frac{\partial}{\partial u_w} x_w = v_c$
$\frac{\partial}{\partial u_w} P_w = \frac{\partial}{\partial x_w}\sigma(x_w) \cdot \frac{\partial x_w}{\partial u_w} \underset{(4)}{=} \sigma(x_w) \cdot \sigma(-x_w) \cdot v_c = P_w \cdot N_w \cdot v_c$
$\frac{\partial}{\partial u_w} N_w = \frac{\partial}{\partial x_w}\sigma(-x_w) \cdot \frac{\partial x_w}{\partial u_w} = \underset{(4)}{-}\sigma(x_w) \cdot \sigma(-x_w) \cdot v_c = -P_w \cdot N_w \cdot v_c$

<u>if w=o:</u>

$$(\cdot) = -\frac{\frac{\partial}{\partial u_o}[P_o]}{P_o} - \sum_{k=1}^{K} \frac{\frac{\partial}{\partial u_o}[N_k]}{N_k} = -\frac{P_o \cdot N_o \cdot v_c}{P_o} = -N_o \cdot v_c = -\sigma(u_o^T v_c) \cdot v_c$$

(As $w_o \notin \{w_k\}_{k \in [1,K]}$ and $\frac{\partial}{\partial u_w} f(u_{w'})$=0 if $w \neq w'$).

<u>if w=k∈[1,K]:</u>

$$(\cdot) = -\frac{\frac{\partial}{\partial u_k}[P_o]}{P_o} - \sum_{k'=1}^{K} \frac{\frac{\partial}{\partial u_k}[N_{k'}]}{N_{k'}} = -\frac{\frac{\partial}{\partial u_k}[P_o]}{P_o} - \frac{\frac{\partial}{\partial u_k}[N_k]}{N_k} - \sum_{k' \in [1,K], k' \neq k} \frac{\frac{\partial}{\partial u_k}[N_{k'}]}{N_{k'}} = \frac{P_k \cdot N_k \cdot v_c}{P_k} =$$

$N_k \cdot v_c = \sigma(-v_k^T v_c) \cdot v_c$

$$\implies \frac{\partial}{\partial u_w}[J_{neg-sample}(v_c, o, U)] = \begin{cases} -\sigma(u_o^T v_c) \cdot v_c, w = o \\ \sigma(-v_k^T v_c) \cdot v_c, w = k \in [1, K] \end{cases}$$

(1) $\frac{\partial}{\partial a} b^T a = b$
(2) $\frac{\partial}{\partial x} log(f(x)) = \frac{\frac{\partial}{\partial x} f(x)}{f(x)}$
(3) $\frac{\partial}{\partial x}\sigma(x) = \sigma(x) \cdot \sigma(-x)$ as seen in part (e) where x is a scalar

(g)

$$J_{skip-gram}(v_c, w_{t-m}, ..., w_{t+m}, U) = \sum_{-m<=j<=m, j\neq 0} J(v_c, w_{t+j}, U)$$

$(i)\partial J_{skip-gram}(v_c, w_{t-m}, ..., w_{t+m}, U)/\partial U =$

$$\frac{\partial}{\partial U} \sum_{-m<=j<=m, j\neq 0} J(v_c, w_{t+j}, U) = \sum_{-m<=j<=m, j\neq 0} \frac{\partial}{\partial U} J(v_c, w_{t+j}, U)$$

$(ii)\partial J_{skip-gram}(v_c, w_{t-m}, ..., w_{t+m}, U)/\partial v_c =$

$$\frac{\partial}{\partial v_c} \sum_{-m<=j<=m, j\neq 0} J(v_c, w_{t+j}, U) = \sum_{-m<=j<=m, j\neq 0} \frac{\partial}{\partial v_c} J(v_c, w_{t+j}, U)$$

$(iii) w \neq c : \partial J_{skip-gram}(v_c, w_{t-m}, ..., w_{t+m}, U)/\partial v_w =$

$$\frac{\partial}{\partial v_w} \sum_{-m<=j<=m, j\neq 0} J(v_c, w_{t+j}, U) = \sum_{-m<=j<=m, j\neq 0} \frac{\partial}{\partial v_w} J(v_c, w_{t+j}, U) = 0$$

(since $J(v_c, w_{t+j}, U)$ is independent by v_w if $w \neq c$).

(a)

$$\mathcal{L}(\theta) = \Pi_{t=1}^{T}\Pi_{-m<=j<=m,j\neq 0}p_\theta(w_{t+j} \mid w_t)$$

$$\mathcal{J}(\theta) = log(\mathcal{L}(\theta)) = \sum_{t=1}^{T}\sum_{-m<=j<=m,j\neq 0} log(p_\theta(w_{t+j} \mid w_t))$$

$$\theta^* = argmax_\theta \mathcal{L}(\theta)$$

For a fixed c,o, the number of occurrences of $p_\theta(o \mid c)$ in $\mathcal{L}(\theta)$ is exactly the number of occurrences of c,o in the courpus with c as a center word and o as outside word, that is, $\#(c, o)$. This means that

$$\mathcal{L}(\theta) = \Pi_{(c,o)}p_\theta(o \mid c)^{\#(c,o)}$$

Let's fix a c and define

$$\mathcal{L}_c(\theta) = \Pi_o p_\theta(o \mid c)^{\#(c,o)}$$

$$J_c(\theta) = log(\mathcal{L}_c(\theta)) = \sum_o \#(c,o) log(p_\theta(o \mid c))$$

Notice that for this fixed c, $\sum_o p_\theta(o|c) = 1$, $p_\theta(o \mid c) >= 0$ since this is a distribution. Using Lagrange Multipliers we can find the optimum of θ under this constraint:

$$L(\theta, \lambda) = \sum_o \#(c,o) log(p_\theta(o \mid c)) - \lambda(\sum_o p_\theta(o|c) - 1)$$

Let us derive by $p_\theta(o|c)$ for some fixed o and find the optimum:

$$\nabla_{p_\theta(o|c)} = \#(c,o) \cdot \frac{1}{p_\theta(o|c)} - \lambda = 0$$

$$\frac{\#(c,o)}{p_\theta(o|c)} = \lambda$$

$$p_\theta(o|c) = \frac{\#(c,o)}{\lambda}$$

Now we get from the constraint that:

$$1 = \sum_o p_\theta(o|c) = \sum_o \frac{\#(c,o)}{\lambda} = \frac{\sum_o \#(c,o)}{\lambda} \implies \lambda = \sum_o \#(c,o)$$

Finally we get

$$\frac{\#(c,o)}{p_\theta(o|c)} = \lambda = \sum_o \#(c,o) \implies p_\theta(o|c) = \frac{\#(c,o)}{\sum_o \#(c,o)}$$

as needed.

(b)

Let there be a vocabulary of 4 words: $W = \{a, b, c, d\}$ and a corpus $C = \{\ddot{a}a", "bb", "cc", "dd"\}$.
Let $\theta^* = argmax_\theta \Pi_{t=1}^{T} \Pi_{-m<=j<=m, j\neq 0} p_\theta(w_t + j \mid w_t)$.
From (a) we know that $p_{\theta^*} = \frac{\#(c,o)}{\sum_{o'} \#(c,o')}$ so for example $p_{\theta^*}(a|c) = 0$.
Let $\theta = (U = (u_a, u_b, u_c, u_d), V = (v_a, v_b, v_c, v_d))$ where u_i, v_i are scalars. If we would use the Skip Gram algorithm on this vocabulary and corpus with the log maximum likelihood as loss function, we know that p_θ is actually the softmax function which is a positive function: for every choice of scalars u_a, v_c

$$p_\theta(a|c) = \frac{\cancel{exp(u_a \cdot v_c)}}{1 + exp(u_a \cdot v_c)} > 0$$

So we will never obtain $p_\theta(a|c) = 0$ for any θ.

$p(\text{the pair is a paraphrase}|\ x_1, x_2) = \sigma(relu(\mathbf{x}_1)^T relu(\mathbf{x}_2))$ where $relu(x) = max(0, x)$

(a)

In this model, since $relu(\mathbf{x}_i) >= 0$ then $(relu(\mathbf{x}_1)^T relu(\mathbf{x}_2)) >= 0$ for every choice of $\mathbf{x}_1$, $\mathbf{x}_2$. This means that $\sigma(relu(\mathbf{x}_1)^T relu(\mathbf{x}_2)) >= 0.5$ so the model will return true for each choice of 2 sentences. If the ratio of positive to negative samples is 1:3 then the model will give an incorrect answer for all the negative samples, which means that the maximal accuracy will be 0.25.

(b)

In order to fix this I would suggest to not use relu when computing the prediction. This will allow the σ function to return any value in [0,1] and thus not all predictions will be True:$p(\text{the pair is a paraphrase}|\ x_1, x_2) = \sigma(\mathbf{x}_1^T \mathbf{x}_2)$