The solution is in Dr. Chris Manning hint/derivation:

$$J_{n\text{-}softmax}(v_c, o, U) = -u_o^T v_c + \log\left(\sum_{w \in V_{oc}} \exp(u_w^T v_c)\right).$$

then: $$\frac{\partial J}{\partial v_c} = \frac{\partial}{\partial v_c}\left[-u_o^T v_c + \log\left(\sum_{w \in V_{ob}} \exp(u_w^T v_c)\right)\right)$$

$$= -u_o + \frac{\partial}{\partial v_c} \log\left(\sum_{w \in V_{ob}} \exp(u_w^T v_c)\right)$$

chain-rule

$$= -u_o + \sum \frac{1}{\sum \exp(u_w^T v_c)} \left(\cancel{\#} \exp(u_w^T v_c \cdot u_w \right.$$

$$= -u_o + \sum \frac{\exp(u_w^T v_c)}{\sum \exp(u_w^T v_c)} \cdot u_w$$

$$= -u_o + \sum \hat{y}_w \cdot u_w$$ (4) from Not

$$= U(\hat{y} - y)$$ (3) from no

3 Extra Credit Challenge II (2.5 Points)

The partial derivatives of $J_{\text{naive-softmax}}(v_c, o, U)$ with respect to each of the 'outside' word vectors, u_w's is given below:

$$\frac{\partial J}{\partial U} = v_c(\hat{y} - y)^\top \tag{6}$$

or equivalently:

$$\frac{\partial J}{\partial u_w} = \begin{cases} (\hat{y}_w - 1)v_c & \text{if } w = o \\ \hat{y}_w v_c & \text{otherwise} \end{cases} \tag{7}$$

Write the steps required to arrive at the partial derivative of $J_{\text{naive-softmax}}(v_c, o, U)$ with respect to each of the 'outside' word vectors, u_w's. There are two cases you need to consider: when $w = o$, the true 'outside' word vector, and $w \neq o$, for all other words. Please write you answer in terms of y, $\hat{y}$, and v_c. The proof may take 4 or 5 steps. The loss function $J_{\text{naive-softmax}}(v_c, o, U)$ is:

$$J_{\text{naive-softmax}}(v_c, o, U) = -u_o^\top v_c + \log\left(\sum_{w' \in \text{Vocab}} \exp(u_{w'}^\top v_c)\right)$$

Similar as pevious derivation + Manning's Note.

$$\frac{\partial J}{\partial u_w} = \frac{\partial}{\partial u_w}\left[-u_o^T v_c + \log\left(\sum_{w' \in Voc} \exp(u_w^T v_c)\right)\right]$$

$$= \frac{\partial}{\partial u_w}(-u_o^T v_c) + \frac{\partial}{\partial u_w}\log\left(\sum_{w' \in V..}\exp(u_w^T v_c)\right)$$

(chain Rule)

when $w = o$

$$\frac{-u_o^T v_c}{\partial u_o} = -v_c$$

when $w \neq o$.

$$\frac{-u_o^T v_c}{\partial u_o} = 0.$$

chain Rule

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

$$= \frac{\exp(u_w^T v_c) \cdot v_c}{\sum \exp(u_w^T v_c)}$$

$$= \hat{y}_w \cdot v_c$$

We got (7): $\begin{cases} (\hat{y}_w - 1)v_c & \text{if } w=o \\ \hat{y}_w v_c & \text{if } w \neq o. \end{cases}$