3(b) $\dfrac{\partial(CE(\mathbf{y},\hat{\mathbf{y}}))}{\partial \mathbf{U}}$

$$= -\frac{\partial}{\partial \mathbf{U}} \sum_{w=1}^{W} \left(y_w \log(\widehat{y_w}) \right) \tag{1}$$

$$= \begin{bmatrix} -\dfrac{\partial}{\partial \mathbf{u}_1}\left(\left(y_1 \log \widehat{y_1}\right) + \ldots + \left(y_o \log \widehat{y_o}\right) + \ldots + \left(y_W \log \widehat{y_W}\right)\right) \\ \vdots \\ -\dfrac{\partial}{\partial \mathbf{u}_o}\left(\left(y_1 \log \widehat{y_o}\right) + \ldots + \left(y_o \log \widehat{y_o}\right) + \ldots + \left(y_W \log \widehat{y_W}\right)\right) \\ \vdots \\ -\dfrac{\partial}{\partial \mathbf{u}_W}\left(\left(y_1 \log \widehat{y_1}\right) + \ldots + \left(y_o \log \widehat{y_o}\right) + \ldots + \left(y_W \log \widehat{y_W}\right)\right) \\ \vdots \end{bmatrix} \tag{2}$$

$$= \begin{bmatrix} -\dfrac{\partial}{\partial \mathbf{u}_1}\left(\left(y_1 \log\left(\dfrac{\exp(\mathbf{u}_1{}^{\mathbf{T}}\mathbf{v}_c)}{\sum_{m=1}^{m=W}\exp(\mathbf{u}_m{}^{\mathbf{T}}\mathbf{v}_c)}\right)\right) + \ldots + \left(y_o \log\left(\dfrac{\exp(\mathbf{u}_o{}^{\mathbf{T}}\mathbf{v}_c)}{\sum_{m=1}^{m=W}\exp(\mathbf{u}_m{}^{\mathbf{T}}\mathbf{v}_c)}\right)\right) + \ldots + \left(y_W \log\left(\dfrac{\exp(\mathbf{u}_W{}^{\mathbf{T}}\mathbf{v}_c)}{\sum_{m=1}^{m=W}\exp(\mathbf{u}_m{}^{\mathbf{T}}\mathbf{v}_c)}\right)\right)\right) \\ \vdots \\ -\dfrac{\partial}{\partial \mathbf{u}_o}\left(\left(y_1 \log\left(\dfrac{\exp(\mathbf{u}_1{}^{\mathbf{T}}\mathbf{v}_c)}{\sum_{m=1}^{m=W}\exp(\mathbf{u}_m{}^{\mathbf{T}}\mathbf{v}_c)}\right)\right) + \ldots + \left(y_o \log\left(\dfrac{\exp(\mathbf{u}_o{}^{\mathbf{T}}\mathbf{v}_c)}{\sum_{m=1}^{m=W}\exp(\mathbf{u}_m{}^{\mathbf{T}}\mathbf{v}_c)}\right)\right) + \ldots + \left(y_W \log\left(\dfrac{\exp(\mathbf{u}_W{}^{\mathbf{T}}\mathbf{v}_c)}{\sum_{m=1}^{m=W}\exp(\mathbf{u}_m{}^{\mathbf{T}}\mathbf{v}_c)}\right)\right)\right) \\ \vdots \\ -\dfrac{\partial}{\partial \mathbf{u}_W}\left(\left(y_1 \log\left(\dfrac{\exp(\mathbf{u}_1{}^{\mathbf{T}}\mathbf{v}_c)}{\sum_{m=1}^{m=W}\exp(\mathbf{u}_m{}^{\mathbf{T}}\mathbf{v}_c)}\right)\right) + \ldots + \left(y_o \log\left(\dfrac{\exp(\mathbf{u}_o{}^{\mathbf{T}}\mathbf{v}_c)}{\sum_{m=1}^{m=W}\exp(\mathbf{u}_m{}^{\mathbf{T}}\mathbf{v}_c)}\right)\right) + \ldots + \left(y_W \log\left(\dfrac{\exp(\mathbf{u}_W{}^{\mathbf{T}}\mathbf{v}_c)}{\sum_{m=1}^{m=W}\exp(\mathbf{u}_m{}^{\mathbf{T}}\mathbf{v}_c)}\right)\right)\right) \\ \vdots \end{bmatrix} \tag{3}$$

If we look at the w^{th} row of Equation (3), it is of the form below:

=	$-\frac{\partial}{\partial \mathbf{u}_w}\left(\sum_{i=1}^{i=W} y_i \log\left(\frac{\exp(\mathbf{u}_i{}^{\mathbf{T}}\mathbf{v}_c)}{\sum_{m=1}^{m=W}\exp(\mathbf{u}_m{}^{\mathbf{T}}\mathbf{v}_c)}\right)\right)$	(4)
=	$-\sum_{i=1}^{i=W}\frac{\partial}{\partial \mathbf{u}_w}\left(y_i \log\left(\frac{\exp(\mathbf{u}_i{}^{\mathbf{T}}\mathbf{v}_c)}{\sum_{m=1}^{m=W}\exp(\mathbf{u}_m{}^{\mathbf{T}}\mathbf{v}_c)}\right)\right)$	(5)
=	$-\sum_{i=1}^{i=W}\frac{\partial}{\partial \mathbf{u}_w}\left(y_i \log\left(\exp(\mathbf{u}_i{}^{\mathbf{T}}\mathbf{v}_c)\right) - \log\sum_{m=1}^{m=W}\exp(\mathbf{u}_m{}^{\mathbf{T}}\mathbf{v}_c)\right)$	(6)
=	$-\sum_{i=1}^{i=W} y_i\left(\frac{\partial\left(\mathbf{u}_i{}^{\mathbf{T}}\mathbf{v}_c\right)}{\partial \mathbf{u}_w} - \frac{\partial}{\partial \mathbf{u}_w}\left(\log\sum_{m=1}^{m=W}\exp(\mathbf{u}_m{}^{\mathbf{T}}\mathbf{v}_c)\right)\right)$	(7)
=	$-\sum_{i=1}^{i=W} y_i\left(\frac{\partial\left(\mathbf{u}_i{}^{\mathbf{T}}\mathbf{v}_c\right)}{\partial \mathbf{u}_w} - \left(\frac{1}{\sum_{m=1}^{m=W}\exp(\mathbf{u}_m{}^{\mathbf{T}}\mathbf{v}_c)}\right)\frac{\partial}{\partial \mathbf{u}_w}\left(\sum_{x=1}^{x=W}\exp(\mathbf{u}_x{}^{\mathbf{T}}\mathbf{v}_c)\right)\right)$	(8)
=	$-\sum_{i=1}^{i=W} y_i\left(\frac{\partial\left(\mathbf{u}_i{}^{\mathbf{T}}\mathbf{v}_c\right)}{\partial \mathbf{u}_w} - \sum_{x=1}^{x=W}\frac{\frac{\partial}{\partial \mathbf{u}_w}\left(\exp(\mathbf{u}_x{}^{\mathbf{T}}\mathbf{v}_c)\right)}{\sum_{m=1}^{m=W}\exp(\mathbf{u}_m{}^{\mathbf{T}}\mathbf{v}_c)}\right)$	(9)

When *i* = *w* in Equation (9):

=	$-\left(\ldots + y_w\left(\frac{\partial\left(\mathbf{u}_w{}^{\mathbf{T}}\mathbf{v}_c\right)}{\partial \mathbf{u}_w} - \frac{\frac{\partial}{\partial \mathbf{u}_w}\left(\exp(\mathbf{u}_1{}^{\mathbf{T}}\mathbf{v}_c) + \ldots. + \exp(\mathbf{u}_w{}^{\mathbf{T}}\mathbf{v}_c) + \ldots. + \exp(\mathbf{u}_W{}^{\mathbf{T}}\mathbf{v}_c)\right)}{\sum_{m=1}^{m=W}\exp(\mathbf{u}_m{}^{\mathbf{T}}\mathbf{v}_c)}\right) + \ldots\right)$	(10)
=	$-\left(\ldots + y_w\left(\mathbf{v}_c - \frac{\exp(\mathbf{u}_w{}^{\mathbf{T}}\mathbf{v}_c)\mathbf{v}_c}{\sum_{m=1}^{m=W}\exp(\mathbf{u}_m{}^{\mathbf{T}}\mathbf{v}_c)}\right) + \ldots\right)$	(11)
=	$-\left(\ldots + y_w\left(1-\hat{y}_w\right)\mathbf{v}_c + \ldots\right)$	(12)
=	$-\left(\ldots + y_w\mathbf{v}_c - y_w\hat{y}_w\mathbf{v}_c + \ldots\right)$	(13)

Now let's consider all cases where $i \neq w$ in Equation (9):

=	$-\left(\ldots + y_{i \neq w} \left(\dfrac{\partial \left(\mathbf{u}_{i \neq w}{}^{\mathbf{T}} \mathbf{v}_{\mathbf{c}} \right)}{\partial \mathbf{u}_w} - \dfrac{\dfrac{\partial}{\partial \mathbf{u}_w} \left(\exp(\mathbf{u}_1{}^{\mathbf{T}} \mathbf{v}_{\mathbf{c}}) + \ldots. + \exp(\mathbf{u}_w{}^{\mathbf{T}} \mathbf{v}_{\mathbf{c}}) + \ldots. + \exp(\mathbf{u}_W{}^{\mathbf{T}} \mathbf{v}_{\mathbf{c}}) \right)}{\sum_{m=1}^{m=W} \exp(\mathbf{u}_m{}^{\mathbf{T}} \mathbf{v}_{\mathbf{c}})} \right) + \ldots \right)$	(14)
=	$-\left(\ldots + y_{i \neq w} \left(0 - \dfrac{\exp(\mathbf{u}_w{}^{\mathbf{T}} \mathbf{v}_{\mathbf{c}}) \mathbf{v}_{\mathbf{c}}}{\sum_{m=1}^{m=W} \exp(\mathbf{u}_m{}^{\mathbf{T}} \mathbf{v}_{\mathbf{c}})} \right) + \ldots \right)$	(15)
=	$-\left(\ldots + y_{i \neq w} \left(0 - \hat{y}_w \right) \mathbf{v}_{\mathbf{c}} + \ldots \right)$	(16)
=	$-\left(\ldots + - y_{i \neq w} \hat{y}_w \mathbf{v}_{\mathbf{c}} + \ldots \right)$	(17)

Combining Equations (13) and (17):

=	$-\left(\ldots + y_w \mathbf{v}_{\mathbf{c}} - y_w \hat{y}_w \mathbf{v}_{\mathbf{c}} - y_{i \neq w} \hat{y}_w \mathbf{v}_{\mathbf{c}} + \ldots \right)$	(18)
=	$-y_w \mathbf{v}_{\mathbf{c}} + y_w \hat{y}_w \mathbf{v}_{\mathbf{c}} + y_{i \neq w} \hat{y}_w \mathbf{v}_{\mathbf{c}} + \ldots$	(19)
=	$-y_w \mathbf{v}_{\mathbf{c}} + \hat{y}_w \mathbf{v}_{\mathbf{c}} \sum_{n=1}^{n=W} y_n$	(20)

Now, $\sum_n y_n = 1$ because $\mathbf{y}$ is a one-hot label vector. Substituting in Equation (20):

=	$-y_w \mathbf{v}_{\mathbf{c}} + \hat{y}_w \mathbf{v}_{\mathbf{c}} (1)$	(21)
=	$\left(\hat{y}_w - y_w \right) \mathbf{v}_{\mathbf{c}}$	(22)

Substituting in Equation (3):

$\dfrac{\partial(CE(\mathbf{y}, \hat{\mathbf{y}}))}{\partial \mathbf{U}} =$	$\begin{bmatrix} \left(\hat{y}_1 - y_1 \right) \mathbf{v}_{\mathbf{c}} \\ \vdots \\ \left(\hat{y}_o - y_o \right) \mathbf{v}_{\mathbf{c}} \\ \vdots \\ \left(\hat{y}_W - y_W \right) \mathbf{v}_{\mathbf{c}} \end{bmatrix}$	(23)

Since only **o** is the expected word ($y_o = 1$):

$$\frac{\partial(CE(\mathbf{y},\hat{\mathbf{y}}))}{\partial \mathbf{U}} = \begin{Bmatrix} (\hat{y}_w - 1)\mathbf{v_c}, & w = o \\ \hat{y}_w \mathbf{v_c}, & \text{otherwise} \end{Bmatrix} \tag{24}$$