The architecture is a RNN language model. The model is formally defined as:

$$
\begin{aligned}
e^{(t)} &= x^{(t)} L \\
h^{(t)} &= \text{sigmoid}\left(h^{(t-1)} H + e^{(t)} I + b_1\right) \\
\hat{y}^{(t)} &= \text{softmax}\left(h^{(t)} U + b_2\right)
\end{aligned}
$$

$$p\left(x_{t+1} = v_j | x_t, \ldots, x_1\right) = \hat{y}_j^{(t)}$$

where L is the embedding matrix, I the input word representation matrix, H the hidden transformation matrix, and U is the output word representation matrix. b1 and b2 are biases. d is the embedding dimension, |V| is the vocabulary size, and Dh is the hidden layer dimension.

2. Compute the gradients for all the model parameters at a single point in time t:

$$\frac{\partial J^{(t)}}{\partial U} \quad \frac{\partial J^{(t)}}{\partial b_2} \quad \frac{\partial J^{(t)}}{\partial L_{x^{(t)}}} \quad \left.\frac{\partial J^{(t)}}{\partial I}\right|_{(t)} \quad \left.\frac{\partial J^{(t)}}{\partial H}\right|_{(t)} \quad \left.\frac{\partial J^{(t)}}{\partial b_1}\right|_{(t)}$$

The solution is provided as:

The partial derivatives:

$$
\begin{aligned}
\frac{\partial J^{(t)}}{\partial U} &= \left(\boldsymbol{h}^{(t)}\right)^T (\boldsymbol{y} - \hat{\boldsymbol{y}}) \\
\frac{\partial J^{(t)}}{\partial \boldsymbol{h}^{(t)}} &= (\boldsymbol{y} - \hat{\boldsymbol{y}}) \boldsymbol{U}^T \\
\frac{\partial J^{(t)}}{\partial \boldsymbol{b}_2} &= (\boldsymbol{y} - \hat{\boldsymbol{y}}) \\
\left.\frac{\partial J^{(t)}}{\partial \boldsymbol{b}_1}\right|_{(t)} &= \left(\frac{\partial J^{(t)}}{\partial \boldsymbol{h}^{(t)}} \odot \text{sigmoid}'\left(\boldsymbol{h}^{(t-1)} \boldsymbol{H} + e^{(t)} \boldsymbol{I} + \boldsymbol{b}_1\right)\right) \\
\frac{\partial J^{(t)}}{\partial \boldsymbol{L}_{x^{(t)}}} &= \frac{\partial J^{(t)}}{\partial e^{(t)}} = \left.\frac{\partial J^{(t)}}{\partial \boldsymbol{b}_1}\right|_{(t)} I^T \\
\left.\frac{\partial J^{(t)}}{\partial \boldsymbol{I}}\right|_{(t)} &= \left(e^{(t)}\right)^T \left.\frac{\partial J^{(t)}}{\partial \boldsymbol{b}_1}\right|_{(t)} \\
\left.\frac{\partial J^{(t)}}{\partial \boldsymbol{H}}\right|_{(t)} &= \left(h^{(t-1)}\right)^T \left.\frac{\partial J^{(t)}}{\partial \boldsymbol{b}_1}\right|_{(t)} \\
\frac{\partial J^{(t)}}{\partial \boldsymbol{h}^{(t-1)}} &= \boldsymbol{H}^T \left.\frac{\partial J^{(t)}}{\partial \boldsymbol{b}_1}\right|_{(t)}
\end{aligned}
$$

TODO: I was wondering (taking the first partial derivative as an example) where the h-transpose comes from in the calculation for the first partial derivative.