1 Preprocess

Alphabet Size: $|\mathcal{A}| = 4$
Character to One-hot Encoding:
Given a vocabulary $\mathcal{A} = \{h, e, l, o\}$, the one-hot encoding of a character $c \in \mathcal{A}$ is defined as:

$$\text{one\_hot}(c)_i = \begin{cases} 1 & \text{if } \mathcal{A}_i = c, \\ 0 & \text{otherwise.} \end{cases}$$

$$\begin{aligned} h &: [1\ 0\ 0\ 0] \\ e &: [0\ 1\ 0\ 0] \\ l &: [0\ 0\ 1\ 0] \\ o &: [0\ 0\ 0\ 1] \end{aligned}$$

Input and Target Sequences:

$$\begin{aligned} x_1 &= h & y_1 &= e \\ x_2 &= e & y_2 &= l \\ x_3 &= l & y_3 &= l \\ x_4 &= l & y_4 &= o \end{aligned}$$

Replacing characters with one-hot encoding:

$$\begin{aligned} x_1 &= [1\ 0\ 0\ 0] & y_1 &= [0\ 1\ 0\ 0] \\ x_2 &= [0\ 1\ 0\ 0] & y_2 &= [0\ 0\ 1\ 0] \\ x_3 &= [0\ 0\ 1\ 0] & y_3 &= [0\ 0\ 1\ 0] \\ x_4 &= [0\ 0\ 1\ 0] & y_4 &= [0\ 0\ 0\ 1] \end{aligned}$$

2 Forward Pass Formulas

For a sequence of characters $x_1, x_2, \ldots, x_T$, the network computes:

1. Hidden State at time t, h_t:

$$h_t = \tanh(Ux_t + Wh_{t-1} + b)$$

2. Output before softmax, Ω_t:

$$\Omega_t = Vh_t + c$$

3. Softmax Output, $\hat{y}_t$, for each character:

$$\text{softmax}(\Omega_t) = \hat{y}_t = \frac{e^{\Omega_t}}{\sum_{k=1}^{|\mathcal{A}|} e^{\Omega_{t,k}}} \text{ for } t = 1, \ldots, |\mathcal{A}|$$

4. Cross-Entropy Loss for the correct character y_t:

$$L_t = -\ y_t\ ln(\hat{y}_t)$$

3 Backpropagation Through Time Formulas

Gradients of the loss L with respect to the parameters U, W, V, b, c are computed as follows:

1. Gradient of Loss w.r.t. Output (Softmax Gradient):

$$\frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial \Omega_t} = (\hat{y}_t - y_t)$$

2. Updates for V and c:

$$\begin{aligned}
\frac{\partial L}{\partial V} &= \sum_{t=1}^{S} \frac{\partial L_t}{\partial V} \\
&= \sum_{t=1}^{S} \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial \Omega_t} \frac{\partial \Omega_t}{\partial V} \\
&= \sum_{t=1}^{S} (\hat{y}_t - y_t) \cdot h_t^\top \\
\frac{\partial L}{\partial c} &= \sum_{t=1}^{T} \frac{\partial L_t}{\partial c} \\
&= \sum_{t=1}^{T} \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial \Omega_t} \frac{\partial \Omega_t}{\partial c} \\
&= \sum_{t=1}^{T} (\hat{y}_t - y_t)
\end{aligned}$$

3. Updates for U, W and b:

$$
\begin{aligned}
\frac{\partial L}{\partial W} &= \sum_{t=1}^{\mathcal{S}} \left(\sum_{k=1}^{t} \; \frac{\partial L_t}{\partial \hat{y}_t} \; \frac{\partial \hat{y}_t}{\partial \Omega_t} \; \frac{\partial \Omega_t}{\partial h_t} \prod_{j=k}^{t-1} \left(\frac{\partial h_{j+1}}{\partial h_j} \right) \frac{\partial h_k}{\partial W} \right) \\
&= \sum_{t=1}^{\mathcal{S}} \left(\sum_{k=1}^{t} \; (y_t - \hat{y}_t) \; V^\top \prod_{j=k}^{t-1} \left(W^\top \; (1 - h_{j+1}^2) \right) (1 - h_k^2) \; h_{k-1} \right) \\
\frac{\partial L}{\partial U} &= \sum_{t=1}^{\mathcal{S}} \left(\sum_{k=1}^{t} \; \frac{\partial L_t}{\partial \hat{y}_t} \; \frac{\partial \hat{y}_t}{\partial \Omega_t} \; \frac{\partial \Omega_t}{\partial h_t} \prod_{j=k}^{t-1} \left(\frac{\partial h_{j+1}}{\partial h_j} \right) \frac{\partial h_k}{\partial U} \right) \\
&= \sum_{t=1}^{\mathcal{S}} \left(\sum_{k=1}^{t} \; (y_t - \hat{y}_t) \; V^\top \prod_{j=k}^{t-1} \left(W^\top \; (1 - h_{j+1}^2) \right) (1 - h_k^2) \; x_k \right) \\
\frac{\partial L}{\partial b} &= \sum_{t=1}^{\mathcal{S}} \left(\sum_{k=1}^{t} \; \frac{\partial L_t}{\partial \hat{y}_t} \; \frac{\partial \hat{y}_t}{\partial \Omega_t} \; \frac{\partial \Omega_t}{\partial h_t} \prod_{j=k}^{t-1} \left(\frac{\partial h_{j+1}}{\partial h_j} \right) \frac{\partial h_k}{\partial b} \right) \\
&= \sum_{t=1}^{\mathcal{S}} \left(\sum_{k=1}^{t} \; (y_t - \hat{y}_t) \; V^\top \prod_{j=k}^{t-1} \left(W^\top \; (1 - h_{j+1}^2) \right) (1 - h_k^2) \right)
\end{aligned}
$$

Parameter Updates

The parameters are updated by subtracting the gradient scaled by a learning rate η:

$$
\begin{aligned}
V &= V - \eta \frac{\partial L}{\partial V} \\
c &= c - \eta \frac{\partial L}{\partial c} \\
W &= W - \eta \frac{\partial L}{\partial W} \\
U &= U - \eta \frac{\partial L}{\partial U} \\
b &= b - \eta \frac{\partial L}{\partial b}
\end{aligned}
$$

4 Parameters Initialized

The network parameters are initialized as follows:

$$
\begin{aligned}
U &= \begin{bmatrix} 0.1442 & -0.2315 & -0.6690 & 1.1585 \end{bmatrix} \\
W &= \begin{bmatrix} -0.5870 \end{bmatrix} \\
b &= \begin{bmatrix} 0 \end{bmatrix} \\
V &= \begin{bmatrix} -0.2246 \\ -0.3053 \\ 0.4905 \\ 0.2768 \end{bmatrix} \\
c &= \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}
\end{aligned}
$$

5 Forward Pass

5.1 Step 1

$$x_1 = [1\ 0\ 0\ 0]$$
$$h_1 = \tanh(U \cdot x_1 + W \cdot h_0 + b)$$
$$= \tanh\left(\begin{bmatrix}0.1442 & -0.2315 & -0.6690 & 1.1585\end{bmatrix} \cdot \begin{bmatrix}1\\0\\0\\0\end{bmatrix} + \begin{bmatrix}-0.5870\end{bmatrix} \cdot \begin{bmatrix}0\end{bmatrix} + \begin{bmatrix}0\end{bmatrix}\right)$$
$$h_1 = \begin{bmatrix}0.1432\end{bmatrix}$$
$$\Omega_1 = V \cdot h_1 + c$$
$$= \begin{bmatrix}-0.2246\\-0.3053\\0.4905\\0.2768\end{bmatrix} \cdot \begin{bmatrix}0.1432\end{bmatrix} + \begin{bmatrix}0\\0\\0\\0\end{bmatrix}$$
$$\Omega_1 = \begin{bmatrix}-0.0322\\-0.0437\\0.0703\\0.0396\end{bmatrix}$$
$$\hat{y}_1 = \text{softmax}(\Omega_1)$$
$$\hat{y}_1 = \begin{bmatrix}0.2398\\0.2370\\0.2656\\0.2576\end{bmatrix}$$

Based on the maximum value of the softmax function, our model predicts it as 'l', but we actually want it to predict 'e'.

$$\begin{aligned}\text{L}_1 &= -y_1 \cdot \ln(\hat{y}_1)\\ &= -\begin{bmatrix}0\\1\\0\\0\end{bmatrix} \cdot \ln\begin{bmatrix}0.2398\\0.2370\\0.2656\\0.2576\end{bmatrix}\\ &= 1.4397\end{aligned}$$

5.2 Step 2

$$x_2 = [0\ 1\ 0\ 0]$$
$$h_2 = \tanh(U \cdot x_2 + W \cdot h_1 + b)$$
$$= \tanh\left(\begin{bmatrix}0.1442 & -0.2315 & -0.6690 & 1.1585\end{bmatrix} \cdot \begin{bmatrix}0\\1\\0\\0\end{bmatrix} + \begin{bmatrix}-0.5870\end{bmatrix} \cdot \begin{bmatrix}0.1432\end{bmatrix} + \begin{bmatrix}0\end{bmatrix}\right)$$
$$h_2 = \begin{bmatrix}-0.3055\end{bmatrix}$$
$$\Omega_2 = V \cdot h_2 + c$$
$$= \begin{bmatrix}-0.2246\\-0.3053\\0.4905\\0.2768\end{bmatrix} \cdot \begin{bmatrix}-0.3055\end{bmatrix} + \begin{bmatrix}0\\0\\0\\0\end{bmatrix}$$
$$\Omega_2 = \begin{bmatrix}0.0686\\0.0933\\-0.1498\\-0.0845\end{bmatrix}$$
$$\hat{y}_2 = \text{softmax}(\Omega_2)$$
$$\hat{y}_2 = \begin{bmatrix}0.2712\\0.2780\\0.2180\\0.2327\end{bmatrix}$$

Based on the maximum value of the softmax function, our model predicts it as 'e', but we actually want it to predict 'l'.

$$\begin{aligned}
\mathrm{L}_2 &= -y_2 \cdot \ln(\hat{y}_2) \\
&= -\begin{bmatrix}0\\0\\1\\0\end{bmatrix} \cdot \ln \begin{bmatrix}0.2712\\0.2780\\0.2180\\0.2327\end{bmatrix} \\
&= 1.5232
\end{aligned}$$

5.3 Step 3

$$
\begin{aligned}
x_3 &= [0\ 0\ 1\ 0] \\
h_3 &= \tanh(U \cdot x_3 + W \cdot h_2 + b) \\
&= \tanh\left(\begin{bmatrix}0.1442 & -0.2315 & -0.6690 & 1.1585\end{bmatrix} \cdot \begin{bmatrix}0\\0\\1\\0\end{bmatrix} + \begin{bmatrix}-0.5870\end{bmatrix} \cdot \begin{bmatrix}-0.3055\end{bmatrix} + \begin{bmatrix}0\end{bmatrix}\right) \\
h_3 &= \begin{bmatrix}-0.4540\end{bmatrix} \\
\Omega_3 &= V \cdot h_3 + c \\
&= \begin{bmatrix}-0.2246\\-0.3053\\0.4905\\0.2768\end{bmatrix} \cdot \begin{bmatrix}-0.4540\end{bmatrix} + \begin{bmatrix}0\\0\\0\\0\end{bmatrix} \\
&= \begin{bmatrix}0.1019\\0.1386\\-0.2227\\-0.1256\end{bmatrix} \\
\hat{y}_3 &= \mathrm{softmax}(\Omega_3) \\
\hat{y}_3 &= \begin{bmatrix}0.2812\\ \color{red}{0.2917}\\0.2032\\0.2239\end{bmatrix}
\end{aligned}
$$

Based on the maximum value of the softmax function, our model predicts it as 'e', but we actually want it to predict 'l'.

$$
\begin{aligned}
\mathrm{L}_3 &= -y_3 \cdot \ln(\hat{y}_3) \\
&= -\begin{bmatrix}0\\0\\1\\0\end{bmatrix} \cdot \ln \begin{bmatrix}0.2812\\ \color{red}{0.2917}\\0.2032\\0.2239\end{bmatrix} \\
&= 1.5934
\end{aligned}
$$

5.4 Step 4

$$
\begin{aligned}
x_4 &= [0\ 0\ 1\ 0] \\
h_4 &= \tanh(U \cdot x_4 + W \cdot h_3 + b) \\
&= \tanh\left(\begin{bmatrix}0.1442 & -0.2315 & -0.6690 & 1.1585\end{bmatrix} \cdot \begin{bmatrix}0\\0\\1\\0\end{bmatrix} + \begin{bmatrix}-0.5870\end{bmatrix} \cdot \begin{bmatrix}-0.4540\end{bmatrix} + \begin{bmatrix}0\end{bmatrix}\right) \\
h_4 &= \begin{bmatrix}-0.3821\end{bmatrix} \\
\Omega_4 &= V \cdot h_4 + c \\
&= \begin{bmatrix}-0.2246\\-0.3053\\0.4905\\0.2768\end{bmatrix} \cdot \begin{bmatrix}-0.3821\end{bmatrix} + \begin{bmatrix}0\\0\\0\\0\end{bmatrix} \\
\Omega_4 &= \begin{bmatrix}0.0858\\0.1166\\-0.1874\\-0.1058\end{bmatrix} \\
\hat{y}_4 &= \text{softmax}(\Omega_4) \\
\hat{y}_4 &= \begin{bmatrix}0.2764\\ \color{red}{0.2851}\\0.2103\\0.2282\end{bmatrix}
\end{aligned}
$$

Based on the maximum value of the softmax function, our model predicts it as 'e', but we actually want it to predict 'o'.

$$
\begin{aligned}
\text{L}_4 &= -y_4 \cdot \ln(\hat{y}_4) \\
&= -\begin{bmatrix}0\\0\\0\\1\end{bmatrix} \cdot \ln \begin{bmatrix}0.2764\\ \color{red}{0.2851}\\0.2103\\0.2282\end{bmatrix} \\
&= 1.4775
\end{aligned}
$$

$$
\sum_{t=1}^{k} L_t = 1.4397 + 1.5232 + 1.5934 + 1.4775 = \color{red}{6.0338}
$$

6 Backpropagation Through Time

6.1 Gradient of L w.r.t. Output

$$\begin{aligned}
\frac{\partial L_1}{\partial \hat{y}_1}\frac{\partial \hat{y}_1}{\partial \Omega_1} &= (\hat{y}_1 - y_1) \\
&= \begin{bmatrix} 0.2398 \\ 0.2370 \\ 0.2656 \\ 0.2576 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} 0.2398 \\ -0.7630 \\ 0.2656 \\ 0.2576 \end{bmatrix} \\
\frac{\partial L_2}{\partial \hat{y}_2}\frac{\partial \hat{y}_2}{\partial \Omega_2} &= (\hat{y}_2 - y_2) \\
&= \begin{bmatrix} 0.2712 \\ 0.2780 \\ 0.2180 \\ 0.2327 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} 0.2712 \\ 0.2780 \\ -0.7820 \\ 0.2327 \end{bmatrix} \\
\frac{\partial L_3}{\partial \hat{y}_3}\frac{\partial \hat{y}_3}{\partial \Omega_3} &= (\hat{y}_3 - y_3) \\
&= \begin{bmatrix} 0.2812 \\ 0.2917 \\ 0.2032 \\ 0.2239 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} 0.2812 \\ 0.2917 \\ -0.7968 \\ 0.2239 \end{bmatrix} \\
\frac{\partial L_4}{\partial \hat{y}_4}\frac{\partial \hat{y}_4}{\partial \Omega_4} &= (\hat{y}_4 - y_4) \\
&= \begin{bmatrix} 0.2764 \\ 0.2851 \\ 0.2103 \\ 0.2282 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \\
&= \begin{bmatrix} 0.2764 \\ 0.2851 \\ 0.2103 \\ -0.7718 \end{bmatrix}
\end{aligned}$$

6.2 Update V

$$
\begin{aligned}
\frac{\partial L}{\partial V} &= \sum_{t=1}^{S} \frac{\partial L_t}{\partial V} \\
&= \sum_{t=1}^{S} (\hat{y}_t - y_t) \cdot h_t^{\top} \\
&= \begin{bmatrix} -0.2818 \\ -0.4356 \\ 0.5583 \\ 0.1591 \end{bmatrix} \\
\eta &= 0.1 \\
V_{new} &= V - \eta \frac{\partial L}{\partial V} \\
&= \begin{bmatrix} -0.2246 \\ -0.3053 \\ 0.4905 \\ 0.2768 \end{bmatrix} - 0.1 \cdot \begin{bmatrix} -0.2818 \\ -0.4356 \\ 0.5583 \\ 0.1591 \end{bmatrix} \\
V_{new} &= \begin{bmatrix} -0.1964 \\ -0.2617 \\ 0.4347 \\ 0.2609 \end{bmatrix}
\end{aligned}
$$

6.3 Update c:

$$\begin{aligned}
\frac{\partial L}{\partial c} &= \sum_{t=1}^{\mathcal{S}} \frac{\partial L_t}{\partial c} \\
&= \sum_{t=1}^{\mathcal{S}} (\hat{y}_t - y_t) \\
&= \begin{bmatrix} 1.0686 \\ 0.0917 \\ -1.1028 \\ -0.0575 \end{bmatrix} \\
\eta &= 0.1 \\
c_{new} &= c - \eta \frac{\partial L}{\partial c} \\
&= \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} - 0.1 \cdot \begin{bmatrix} 1.0686 \\ 0.0917 \\ -1.1028 \\ -0.0575 \end{bmatrix} \\
c_{new} &= \begin{bmatrix} -0.1069 \\ -0.0092 \\ 0.1103 \\ 0.0058 \end{bmatrix}
\end{aligned}$$

6.4 Update W:

$$\begin{aligned}
\frac{\partial L}{\partial W} &= \sum_{t=1}^{\mathcal{S}} \left(\sum_{k=1}^{t} \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial \Omega_t} \frac{\partial \Omega_t}{\partial h_t} \prod_{j=k}^{t-1} \left(\frac{\partial h_{j+1}}{\partial h_j} \right) \frac{\partial h_k}{\partial W} \right) \\
&= \sum_{t=1}^{\mathcal{S}} \left(\sum_{k=1}^{t} (y_t - \hat{y}_t) \; V^\top \prod_{j=k}^{t-1} \left(W^\top \; (1 - h_{j+1}^2) \right) (1 - h_k^2) \; h_{k-1} \right) \\
&= \begin{bmatrix} 0.1466 \end{bmatrix} \\
\eta &= 0.1 \\
W_{new} &= W - \eta \frac{\partial L}{\partial W} \\
&= \begin{bmatrix} -0.5870 \end{bmatrix} - 0.1 \cdot \begin{bmatrix} 0.1466 \end{bmatrix} \\
W_{new} &= \begin{bmatrix} -0.6016 \end{bmatrix}
\end{aligned}$$

6.5 Update U:

$$
\begin{aligned}
\frac{\partial L}{\partial U} &= \sum_{t=1}^{\mathcal{S}} \left(\sum_{k=1}^{t} \ \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial \Omega_t} \frac{\partial \Omega_t}{\partial h_t} \prod_{j=k}^{t-1} \left(\frac{\partial h_{j+1}}{\partial h_j} \right) \frac{\partial h_k}{\partial U} \ \right) \\
&= \sum_{t=1}^{\mathcal{S}} \left(\sum_{k=1}^{t} \ (y_t - \hat{y}_t) \ V^\top \prod_{j=k}^{t-1} \left(\ W^\top \ (1 - h_{j+1}^2) \right) (1 - h_k^2) \ x_k \ \right) \\
&= \begin{bmatrix} 0.5300 & -0.2733 & -0.5002 & 0 \end{bmatrix} \\
\eta &= 0.1 \\
U_{new} &= U - \eta \frac{\partial L}{\partial U} \\
&= \begin{bmatrix} 0.1442 & -0.2315 & -0.6690 & 1.1585 \end{bmatrix} - 0.1 \cdot \begin{bmatrix} 0.5300 & -0.2733 & -0.5002 & 0 \end{bmatrix} \\
U_{new} &= \begin{bmatrix} 0.0912 & -0.2041 & -0.6190 & 1.1585 \end{bmatrix}
\end{aligned}
$$

6.6 Update b:

$$
\begin{aligned}
\frac{\partial L}{\partial b} &= \sum_{t=1}^{\mathcal{S}} \left(\sum_{k=1}^{t} \ \frac{\partial L_t}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial \Omega_t} \frac{\partial \Omega_t}{\partial h_t} \prod_{j=k}^{t-1} \left(\frac{\partial h_{j+1}}{\partial h_j} \right) \frac{\partial h_k}{\partial b} \ \right) \\
&= \sum_{t=1}^{\mathcal{S}} \left(\sum_{k=1}^{t} \ (y_t - \hat{y}_t) \ V^\top \prod_{j=k}^{t-1} \left(\ W^\top \ (1 - h_{j+1}^2) \right) (1 - h_k^2) \ \right) \\
&= \begin{bmatrix} -0.2436 \end{bmatrix} \\
\eta &= 0.1 \\
b_{new} &= b - \eta \frac{\partial L}{\partial b} \\
&= \begin{bmatrix} 0 \end{bmatrix} - 0.1 \cdot \begin{bmatrix} -0.2436 \end{bmatrix} \\
b_{new} &= \begin{bmatrix} 0.0244 \end{bmatrix}
\end{aligned}
$$

6.7 Following Epochs Loss Values

$$
\begin{aligned}
L_1 &= -y_1 \cdot \ln(\hat{y}_1) \\
&= -\begin{bmatrix} 0. \\ 1. \\ 0. \\ 0. \end{bmatrix} \cdot \ln \left(\begin{bmatrix} 0.2169 \\ 0.2374 \\ 0.2898 \\ 0.2559 \end{bmatrix} \right) \\
L_1 &= 1.4381
\end{aligned}
$$

$$
\begin{aligned}
L_2 &= -y_2 \cdot \ln(\hat{y}_2) \\
&= -\begin{bmatrix} 0. \\ 0. \\ 1. \\ 0. \end{bmatrix} \cdot \ln\left(\begin{bmatrix} 0.2389 \\ 0.2676 \\ 0.2544 \\ 0.2391 \end{bmatrix}\right) \\
L_2 &= 1.3687
\end{aligned}
$$

$$
\begin{aligned}
L_3 &= -y_3 \cdot \ln(\hat{y}_3) \\
&= -\begin{bmatrix} 0. \\ 0. \\ 1. \\ 0. \end{bmatrix} \cdot \ln\left(\begin{bmatrix} 0.2494 \\ 0.2826 \\ 0.2377 \\ 0.2303 \end{bmatrix}\right) \\
L_3 &= 1.4368
\end{aligned}
$$

$$
\begin{aligned}
L_4 &= -y_4 \cdot \ln(\hat{y}_4) \\
&= -\begin{bmatrix} 0. \\ 0. \\ 0. \\ 1. \end{bmatrix} \cdot \ln\left(\begin{bmatrix} 0.2440 \\ 0.2749 \\ 0.2463 \\ 0.2349 \end{bmatrix}\right) \\
L_4 &= 1.4486
\end{aligned}
$$

$$
\sum_{t=1}^{k} L_t = 1.4381 + 1.3687 + 1.4368 + 1.4486 = 5.6922
$$

in Epoch #100 Total Loss Will be 2.0603
in Epoch #1000 Total Loss Will be 1.3455