Assignment 2

Hila Man 315212092

Gal Grudka 206960064

Nir Frid 322805151

1 Question 1

1.1 a

$$CE(y, \hat{y}) = -\sum_i y_i \cdot \log(\hat{y_i}) \qquad softmax(\theta)_i = \frac{\exp(\theta_i)}{\sum_j \exp(\theta_j)}$$

y is a one–hot vector, thus there exists a k s.t $y_j = 0$ for $j \neq k$ and $y_k = 1$. we get that:

$$CE(\theta) = -\log(\hat{y_k}) = -\log\left(\frac{\exp(\theta_k)}{\sum_j \exp(\theta_j)}\right) = -\theta_k + \log\left(\sum_j \exp(\theta_j)\right)$$

We can take the derivative of each term in the final equation w.r.t θ:

$$\frac{\partial}{\partial \theta_k}\theta_k = 1 \qquad \frac{\partial}{\partial \theta_j}\theta_k = 0 \qquad (\text{for } j \neq k)$$

and therefore $\frac{\partial}{\partial \theta}\theta_k = y$.

For the second term, we can use the chain rule:

$$\frac{\partial}{\partial \theta_i} \log\left(\sum_j \exp(\theta_j)\right) = \frac{\exp(\theta_i)}{\sum_j \exp(\theta_j)} = \hat{y_i}$$

Which gets us the final derivative:

$$\frac{\partial}{\partial \theta}\left(-\theta_k + \log\left(\sum_j \exp(\theta_j)\right)\right) = \hat{y} - y$$

1.2 b

Let $z = xW_1 + b_1$ and $\theta = hW_2 + b_2$. By the chain rule:

$$\frac{\partial J}{\partial x} = \frac{\partial J}{\partial z}\frac{\partial z}{\partial x}$$

To calculate the first derivative, we will use the chain rule once again:

$$\frac{\partial J}{\partial z} = \frac{\partial J}{\partial \theta}\frac{\partial \theta}{\partial h}\frac{\partial h}{\partial z}$$

From the previous clause we have that:

$$\frac{\partial J}{\partial \theta} = \hat{y} - y$$
$$\frac{\partial \theta}{\partial h} = W_2$$

since the derivative of a matrix multiplied by a vector w.r.t that vector is the matrix itself. For the third derivative we will use the derivative of the sigmoid function that we derived in HW0:

$$\frac{\partial h}{\partial z} = \sigma(z)(1 - \sigma(z))$$

Putting everything together:

$$\frac{\partial J}{\partial z} = (\hat{y} - y)W_2^t \cdot \sigma(z)(1 - \sigma(z))$$

Going back to the original derivative, we have:

$$\frac{\partial z}{\partial x} = W_1$$

as the derivative of a matrix multiplied by a vector w.r.t that vector. This gives us the final derivative:

$$\frac{\partial J}{\partial x} = \frac{\partial J}{\partial z}\frac{\partial z}{\partial x} = \left[(\hat{y} - y)W_2^t \cdot \sigma(z)(1 - \sigma(z))\right] W_1^t$$

1.3 d

We got dev perplexity of 112.86212240788677.

2 Question 2

2.1 a

$x^{(t)}$is a one-hot vector representing index of the current word. let denote this index by i.

$y^{(t)}$ is a one–hot vector , thus there exists a k s.t $y_j = 0$ for $j \neq k$ and $y_k = 1$.

$$J^{(t)}(\theta) = CE(y^{(t)}, \hat{y}^{(t)}) = -\sum_{i=1}^{|V|} y_i^{(t)} \log(\hat{y_i}^{(t)}) = -\log(\hat{y_k}^{(t)})$$

$$\boldsymbol{\frac{\partial J^{(t)}}{\partial U} = \frac{\partial J}{\partial \hat{y}^{(t)}} \cdot \frac{\partial \hat{y}^{(t)}}{\partial U}}$$

$$\frac{\partial J}{\partial \hat{y}^{(t)}} = -\frac{\frac{\partial \hat{y_k}^{(t)}}{\partial \hat{y}^{(t)}}\left(\hat{y_k}^{(t)}\right)}{\hat{y_k}^{(t)}} = -\frac{1}{\hat{y_k}^{(t)}}$$

$$\hat{y_i}^{(t)} = softmax\left(h^{(t)}U + b_2\right)_i = \frac{e^{(h^{(t)}U+b_2)_i}}{\sum_j e^{(h^{(t)}U+b_2)_j}}$$

$$\frac{\partial \hat{y_i}^{(t)}}{\partial U} = \frac{h^{(t)}e^{(h^{(t)}U+b_2)_i} \cdot \sum_j e^{(h^{(t)}U+b_2)_j} - e^{(h^{(t)}U+b_2)_i} \cdot h^{(t)}e^{(h^{(t)}U+b_2)_i}}{\left(\sum_j e^{(h^{(t)}U+b_2)_j}\right)^2} =$$

$$= \frac{h^{(t)}e^{(h^{(t)}U+b_2)_i}}{\sum_j e^{(h^{(t)}U+b_2)_j}} - \frac{e^{(h^{(t)}U+b_2)_i} \cdot h^{(t)}e^{(h^{(t)}U+b_2)_i}}{\left(\sum_j e^{(h^{(t)}U+b_2)_j}\right)^2} =$$

$$= h^{(t)}\left(\frac{e^{(h^{(t)}U_i+b_2)}}{\sum_j e^{(h^{(t)}U_j+b_2)}} - \frac{\left(e^{(h^{(t)}U+b_2)_i}\right)^2}{\left(\sum_j e^{(h^{(t)}U+b_2)_j}\right)^2}\right) = h^{(t)}\left(\hat{y_i}^{(t)} - \left(\frac{e^{(h^{(t)}U+b_2)_i}}{\sum_j e^{(h^{(t)}U+b_2)_j}}\right)^2\right) =$$

$$= h^{(t)}\left(\hat{y_i}^{(t)} - \left(\hat{y_i}^{(t)}\right)^2\right) = h^{(t)}\hat{y_i}^{(t)}\left(1 - \hat{y_i}^{(t)}\right)$$

$$\frac{\partial J^{(t)}}{\partial U} = \frac{\partial J}{\partial \hat{y}^{(t)}} \cdot \frac{\partial \hat{y}^{(t)}}{\partial U} \underbrace{=}_{\substack{\text{in vector notation,} \\ y^{(t)} \text{ is a one} - \text{hot vector}}} -\frac{1}{\cancel{\hat{y}^{(t)}}} \cdot \cancel{\hat{y}^{(t)}}\left(h^{(t)}\right)^T\left(y^{(t)} - \hat{y}^{(t)}\right) = \boldsymbol{\left(h^{(t)}\right)^T\left(\hat{y}^{(t)} - y^{(t)}\right)}$$

$$\boldsymbol{\frac{\partial J^{(t)}}{\partial L_{x^{(t)}}} = \frac{\partial J^{(t)}}{\partial \hat{y}^{(t)}} \cdot \frac{\partial \hat{y}^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial \mathfrak{e}^{(t)}} \cdot \frac{\partial \mathfrak{e}^{(t)}}{\partial L_{x^{(t)}}}}$$

$$\frac{\partial J^{(t)}}{\partial \hat{y}^{(t)}} = -\frac{1}{\hat{y_k}^{(t)}} \underbrace{=}_{in\ vector\ notation} -\frac{y^{(t)}}{\hat{y}^{(t)}}$$

$$\hat{y_i}^{(t)} = softmax\left(h^{(t)}U + b_2\right)_i = \frac{e^{(h^{(t)}U+b_2)_i}}{\sum_j e^{(h^{(t)}U+b_2)_j}}$$

$$\frac{\partial \hat{y_i}^{(t)}}{\partial h^{(t)}} = \frac{U \cdot e^{(h^{(t)}U+b_2)_i} \cdot \sum_j e^{(h^{(t)}U+b_2)_j} - U \cdot e^{(h^{(t)}U+b_2)_i} \cdot e^{(h^{(t)}U+b_2)_i}}{\left(\sum_j e^{(h^{(t)}U+b_2)_j}\right)^2} =$$

$$= \frac{U \cdot e^{(h^{(t)}U+b_2)_i} \cdot \left(\sum_j e^{(h^{(t)}U+b_2)_j} - e^{(h^{(t)}U+b_2)_i}\right)}{\left(\sum_j e^{(h^{(t)}U+b_2)_j}\right)^2} =$$

$$= U \cdot \left[\frac{e^{(h^{(t)}U+b_2)_i}}{\sum_j e^{(h^{(t)}U+b_2)_j}} - \frac{\left(e^{(h^{(t)}U+b_2)_i}\right)^2}{\left(\sum_j e^{(h^{(t)}U+b_2)_j}\right)^2}\right] =$$

$$= U \cdot \left[\frac{e^{(h^{(t)}U+b_2)_i}}{\sum_j e^{(h^{(t)}U+b_2)_j}} - \left(\frac{e^{(h^{(t)}U+b_2)_i}}{\sum_j e^{(h^{(t)}U+b_2)_j}}\right)^2\right] =$$

$$= U \cdot \left[\hat{y}_i^{(t)} - \left(\hat{y}_i^{(t)}\right)^2\right] \underbrace{=}_{in\ vector\ notation} U \cdot \hat{y}^{(t)}\left[y^{(t)} - \hat{y}^{(t)}\right]$$

$$h^{(t)} = sigmoid\left(h^{(t-1)}H + \mathfrak{e}^{(t)}I + b_1\right) = \frac{1}{1 + e^{-\left(h^{(t-1)}H + \mathfrak{e}^{(t)}I + b_1\right)}}$$

$$\frac{\partial h^{(t)}}{\partial \mathfrak{e}^{(t)}} = \frac{I \cdot e^{-\left(h^{(t-1)}H + \mathfrak{e}^{(t)}I + b_1\right)}}{\left(1 + e^{-\left(h^{(t-1)}H + \mathfrak{e}^{(t)}I + b_1\right)}\right)^2} =$$

$$\underbrace{=}_{sigmoids'\ proprerty} I \cdot \sigma\left(h^{(t-1)}H + \mathfrak{e}^{(t)}I + b_1\right)\left(1 - \sigma\left(h^{(t-1)}H + \mathfrak{e}^{(t)}I + b_1\right)\right) =$$

$$\underbrace{=}_{h^{(t)} = \sigma\left(h^{(t-1)}H + \mathfrak{e}^{(t)}I + b_1\right)} I \cdot h^{(t)}(1 - h^{(t)})$$

$\mathfrak{e}^{(t)} = x^{(t)}L$, where $\mathrm{x^{(t)}}$ is one − hot vector representing index of the current word denoted by i .

L is the embedding matrix, $\mathrm{L_{x^{(t)}}}$ is the row of L corresponding to the current input word

$$\frac{\partial \mathfrak{e}^{(t)}}{\partial L_{x^{(t)}}} = x_i^{(t)} = 1$$

$$\frac{\partial J^{(t)}}{\partial L_{x^{(t)}}} = \frac{\partial J^{(t)}}{\partial \hat{y}^{(t)}} \cdot \frac{\partial \hat{y}^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial \mathfrak{e}^{(t)}} \cdot \frac{\partial \mathfrak{e}^{(t)}}{\partial L_{x^{(t)}}} = -\frac{1}{\cancel{\hat{y}_k^{(t)}}} \cdot U \cdot \cancel{\hat{y}_k^{(t)}} \cdot \left[y_k^{(t)} - \hat{y_k}^{(t)}\right] \cdot I \cdot h^{(t)}(1 - h^{(t)}) =$$

$$= \boldsymbol{U \cdot \left[y^{(t)} - \hat{y}^{(t)}\right] \cdot I \cdot h^{(t)}(1 - h^{(t)})}$$

$$\boldsymbol{\frac{\partial J^{(t)}}{\partial I}\,|_{(t)} = \frac{\partial J^{(t)}}{\partial \hat{y}^{(t)}}\,|_{(t)} \cdot \frac{\partial \hat{y}^{(t)}}{\partial h^{(t)}}\,|_{(t)} \cdot \frac{\partial h^{(t)}}{\partial I}\,|_{(t)}}$$

$$\frac{\partial h^{(t)}}{\partial I}\,|_{(t)} = \frac{\mathfrak{e}^{(t)} \cdot e^{-\left(h^{(t-1)}H + \mathfrak{e}^{(t)}I + b_1\right)}}{\left(1 + e^{-\left(h^{(t-1)}H + \mathfrak{e}^{(t)}I + b_1\right)}\right)^2}\,|_{(t)} =$$

$$\underbrace{=}_{sigmoids'\ proprerty} \left[\mathfrak{e}^{(t)} \cdot \sigma\left(h^{(t-1)}H + \mathfrak{e}^{(t)}I + b_1\right)\left(1 - \sigma\left(h^{(t-1)}H + \mathfrak{e}^{(t)}I + b_1\right)\right)\right]\,|_{(t)} =$$

$$\underbrace{=}_{h^{(t)} = \sigma\left(h^{(t-1)}H + \mathfrak{e}^{(t)}I + b_1\right)} \left[\mathfrak{e}^{(t)} \cdot h^{(t)}(1 - h^{(t)})\right]\,|_{(t)}$$

We calculated earlier:

$$\frac{\partial \hat{y_i}^{(t)}}{\partial h^{(t)}} = U \cdot \hat{y}^{(t)} \left[y^{(t)} - \hat{y}^{(t)} \right]$$
$$\frac{\partial J^{(t)}}{\partial \hat{y}^{(t)}} = -\frac{1}{\hat{y_k}^{(t)}} \underbrace{=}_{in\ vector\ notation} -\frac{y^{(t)}}{\hat{y}^{(t)}}$$

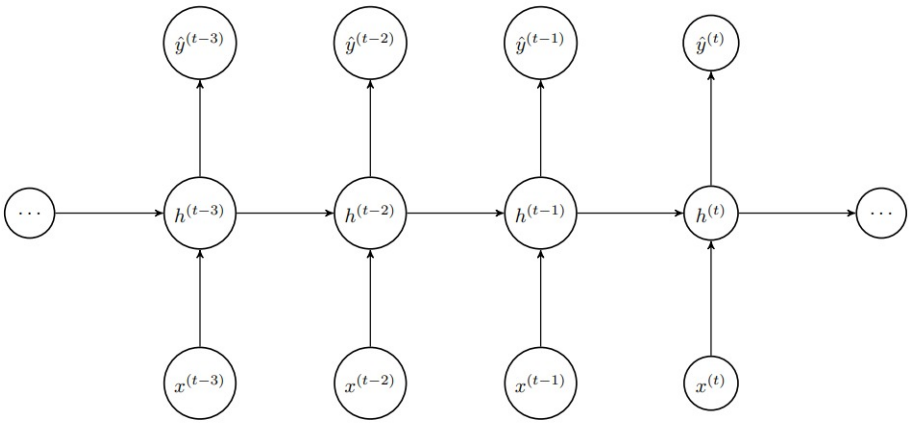
Then when applying the gradients we calculated above, we get that:

$$\frac{\partial J^{(t)}}{\partial I} \mid_{(t)} = -\frac{1}{\cancel{\hat{y_k}^{(t)}}} \cdot U \left[y_k^{(t)} - \hat{y_k}^{(t)} \right] \cdot \cancel{\hat{y}_k^{(t)}} \cdot \mid_{(t)} \left[\mathfrak{e}^{(t)} \cdot h^{(t)} (1 - h^{(t)}) \right] \mid_{(t)} =$$
$$= \left[\boldsymbol{y^{(t)} - \hat{y}^{(t)}} \right] \boldsymbol{U} \left[\boldsymbol{\mathfrak{e}^{(t)} \cdot h^{(t)} (1 - h^{(t)})} \right] \mid_{\boldsymbol{(t)}}$$

$$\boldsymbol{\frac{\partial J^{(t)}}{\partial H} \mid_{(t)} = \frac{\partial J^{(t)}}{\partial \hat{y}^{(t)}} \mid_{(t)} \cdot \frac{\partial \hat{y}^{(t)}}{\partial h^{(t)}} \mid_{(t)} \cdot \frac{\partial h^{(t)}}{\partial H} \mid_{(t)}}$$
$$\frac{\partial h^{(t)}}{\partial H} \mid_{(t)} = \frac{h^{(t-1)} \cdot e^{-\left(h^{(t-1)} H + \mathfrak{e}^{(t)} I + b_1 \right)}}{\left(1 + e^{-\left(h^{(t-1)} H + \mathfrak{e}^{(t)} I + b_1 \right)} \right)^2} \mid_{(t)} =$$
$$\underbrace{=}_{sigmoids'\ proprerty} \left[h^{(t-1)} \cdot \sigma \left(h^{(t-1)} H + \mathfrak{e}^{(t)} I + b_1 \right) \left(1 - \sigma \left(h^{(t-1)} H + \mathfrak{e}^{(t)} I + b_1 \right) \right) \right] \mid_{(t)} =$$
$$\underbrace{=}_{h^{(t)} = \sigma \left(h^{(t-1)} H + \mathfrak{e}^{(t)} I + b_1 \right)} \left[h^{(t-1)} \cdot h^{(t)} (1 - h^{(t)}) \right] \mid_{(t)}$$
$$\frac{\partial J^{(t)}}{\partial H} \mid_{(t)} = -\frac{1}{\cancel{\hat{y_k}^{(t)}}} \cdot U \left[y_k^{(t)} - \hat{y_k}^{(t)} \right] \cdot \cancel{\hat{y}_k^{(t)}} \cdot \mid_{(t)} \left[h^{(t-1)} \cdot h^{(t)} (1 - h^{(t)}) \right] \mid_{(t)} =$$
$$= \left[\boldsymbol{y^{(t)} - \hat{y}^{(t)}} \right] \boldsymbol{U} \left[\boldsymbol{h^{(t-1)} \cdot h^{(t)} (1 - h^{(t)})} \right] \mid_{\boldsymbol{(t)}}$$

$$\boldsymbol{\frac{\partial J^{(t)}}{\partial h^{(t-1)}} = \frac{\partial J^{(t)}}{\partial \hat{y}^{(t)}} \cdot \frac{\partial \hat{y}^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial h^{(t-1)}}}$$
$$\frac{\partial h^{(t)}}{\partial h^{(t-1)}} \mid_{(t)} = \frac{H \cdot e^{-\left(h^{(t-1)} H + \mathfrak{e}^{(t)} I + b_1 \right)}}{\left(1 + e^{-\left(h^{(t-1)} H + \mathfrak{e}^{(t)} I + b_1 \right)} \right)^2} \mid_{(t)} =$$
$$\underbrace{=}_{sigmoids'\ proprerty} \left[H^T \cdot \sigma \left(h^{(t-1)} H + \mathfrak{e}^{(t)} I + b_1 \right) \left(1 - \sigma \left(h^{(t-1)} H + \mathfrak{e}^{(t)} I + b_1 \right) \right) \right] \mid_{(t)} =$$
$$\underbrace{=}_{h^{(t)} = \sigma \left(h^{(t-1)} H + \mathfrak{e}^{(t)} I + b_1 \right)} \left[H^T \cdot h^{(t)} (1 - h^{(t)}) \right] \mid_{(t)}$$
$$\frac{\partial J^{(t)}}{\partial h^{(t-1)}} \mid_{(t)} = -\frac{1}{\cancel{\hat{y_k}^{(t)}}} \cdot U \left[y_k^{(t)} - \hat{y_k}^{(t)} \right] \cdot \cancel{\hat{y}_k^{(t)}} \cdot \mid_{(t)} \left[H^T \cdot h^{(t)} (1 - h^{(t)}) \right] \mid_{(t)} =$$
$$= \left[\boldsymbol{y^{(t)} - \hat{y}^{(t)}} \right] \boldsymbol{U} \left[\boldsymbol{H^T \cdot h^{(t)} (1 - h^{(t)})} \right] \mid_{\boldsymbol{(t)}}$$

2.2 b



Unrolled network for 3 time-steps :1 איור

$$\frac{\partial \boldsymbol{J^{(t)}}}{\partial \boldsymbol{L_{x^{(t)}}}} = \frac{\partial \boldsymbol{J^{(t)}}}{\partial \boldsymbol{\hat{y}^{(t)}}} \cdot \frac{\partial \boldsymbol{\hat{y}^{(t)}}}{\partial \boldsymbol{h^{(t)}}} \cdot \frac{\partial \boldsymbol{h^{(t)}}}{\partial \boldsymbol{h^{(t-1)}}} \cdot \frac{\partial \boldsymbol{h^{(t-1)}}}{\partial \boldsymbol{\mathfrak{e}^{(t-1)}}} \cdot \frac{\partial \boldsymbol{\mathfrak{e}^{(t-1)}}}{\partial \boldsymbol{L_{x^{(t-1)}}}}$$

We already calculated in part (a):

$$\frac{\partial J^{(t)}}{\partial \hat{y}^{(t)}} = -\frac{1}{\hat{y_k}^{(t)}} \underbrace{=}_{in\ vector\ notation} -\frac{y^{(t)}}{\hat{y}^{(t)}}$$

$$\frac{\partial \hat{y}^{(t)}}{\partial h^{(t)}} = U \cdot \hat{y}^{(t)} \left[y^{(t)} - \hat{y}^{(t)} \right]$$

$$\frac{\partial h^{(t)}}{\partial h^{(t-1)}} = \left[H^T \cdot h^{(t)} (1 - h^{(t)}) \right]$$

$$h^{(t-1)} = sigmoid\left(h^{(t-2)} H + \mathfrak{e}^{(t-1)} I + b_1 \right) = \frac{1}{1 + e^{-\left(h^{(t-2)} H + \mathfrak{e}^{(t-1)} I + b_1 \right)}}$$

$$\frac{\partial h^{(t-1)}}{\partial \mathfrak{e}^{(t-1)}} = \frac{I \cdot e^{-\left(h^{(t-2)} H + \mathfrak{e}^{(t-1)} I + b_1 \right)}}{\left(1 + e^{-\left(h^{(t-2)} H + \mathfrak{e}^{(t-1)} I + b_1 \right)} \right)^2} =$$

$$\underbrace{=}_{sigmoids'\ proprerty} I \cdot \sigma\left(h^{(t-2)} H + \mathfrak{e}^{(t-1)} I + b_1 \right) \left(1 - \sigma\left(h^{(t-2)} H + \mathfrak{e}^{(t-1)} I + b_1 \right) \right) =$$

$$\underbrace{=}_{h^{(t-1)} = \sigma\left(h^{(t-2)} H + \mathfrak{e}^{(t-1)} I + b_1 \right)} I \cdot h^{(t-1)} (1 - h^{(t-1)})$$

$\mathfrak{e}^{(t-1)} = x^{(t-1)} L$, where $\mathrm{x}^{(t-1)}$ is one − hot vector representing index of the current word denoted by i .
L is the embedding matrix, $\mathrm{L}_{\mathrm{x}^{(t)}}$ is the row of L corresponding to the current input word

$$\frac{\partial \mathfrak{e}^{(t-1)}}{\partial L_{x^{(t-1)}}} = x_i^{(t-1)} = 1$$

At the end we get that:

$$\frac{\partial \boldsymbol{J^{(t)}}}{\partial \boldsymbol{L_{x^{(t)}}}} = \frac{\partial \boldsymbol{J^{(t)}}}{\partial \boldsymbol{\hat{y}^{(t)}}} \cdot \frac{\partial \boldsymbol{\hat{y}^{(t)}}}{\partial \boldsymbol{h^{(t)}}} \cdot \frac{\partial \boldsymbol{h^{(t)}}}{\partial \boldsymbol{h^{(t-1)}}} \cdot \frac{\partial \boldsymbol{h^{(t-1)}}}{\partial \boldsymbol{\mathfrak{e}^{(t-1)}}} \cdot \frac{\partial \boldsymbol{\mathfrak{e}^{(t-1)}}}{\partial \boldsymbol{L_{x^{(t-1)}}}}$$

$$\frac{\partial J^{(t)}}{\partial L_{x^{(t-1)}}} = \frac{\partial J^{(t)}}{\partial \hat{y}^{(t)}} \cdot \frac{\partial \hat{y}^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial h^{(t-1)}} \cdot I \cdot h^{(t-1)}(1-h^{(t-1)}) \cdot 1 =$$
$$= \frac{\partial J^{(t)}}{\partial \hat{y}^{(t)}} \cdot \frac{\partial \hat{y}^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial h^{(t-1)}} I \cdot h^{(t-1)}(1-h^{(t-1)})$$

$$\boldsymbol{\frac{\partial J^{(t)}}{\partial L_{x^{(t)}}}|_{(t-1)} = \frac{\partial J^{(t)}}{\partial \hat{y}^{(t)}} \cdot \frac{\partial \hat{y}^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial h^{(t-1)}} \cdot \frac{\partial h^{(t-1)}}{\partial \mathfrak{e}^{(t-1)}} \cdot \frac{\partial \mathfrak{e}^{(t-1)}}{\partial L_{x^{(t-1)}}}}$$

We already calculated in part (a):

$$h^{(t-1)} = sigmoid\left(h^{(t-2)}H + \mathfrak{e}^{(t-1)}I + b_1\right) = \frac{1}{1+e^{-\left(h^{(t-2)}H + \mathfrak{e}^{(t-1)}I + b_1\right)}}$$

$$\frac{\partial h^{(t-1)}}{\partial \mathfrak{e}^{(t-1)}} = \frac{I \cdot e^{-\left(h^{(t-2)}H + \mathfrak{e}^{(t-1)}I + b_1\right)}}{\left(1+e^{-\left(h^{(t-2)}H + \mathfrak{e}^{(t-1)}I + b_1\right)}\right)^2} =$$

$$\underbrace{=}_{sigmoids'\ proprerty} I \cdot \sigma\left(h^{(t-2)}H + \mathfrak{e}^{(t-1)}I + b_1\right)\left(1-\sigma\left(h^{(t-2)}H + \mathfrak{e}^{(t-1)}I + b_1\right)\right) =$$

$$\underbrace{=}_{h^{(t-1)}=\sigma\left(h^{(t-2)}H + \mathfrak{e}^{(t-1)}I + b_1\right)} I \cdot h^{(t-1)}(1-h^{(t-1)})$$

$\mathfrak{e}^{(t-1)} = x^{(t-1)}L$, where $\mathrm{x}^{(t-1)}$ is one − hot vector representing index of the current word denoted by i .

L is the embedding matrix, $\mathrm{L}_{\mathrm{x}^{(t)}}$ is the row of L corresponding to the current input word

$$\frac{\partial \mathfrak{e}^{(t-1)}}{\partial L_{x^{(t-1)}}} = x_i^{(t-1)} = 1$$

At the end we get that:

$$\boldsymbol{\frac{\partial J^{(t)}}{\partial L_{x^{(t)}}}|_{(t-1)} = \frac{\partial J^{(t)}}{\partial \hat{y}^{(t)}} \cdot \frac{\partial \hat{y}^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial h^{(t-1)}} \cdot \frac{\partial h^{(t-1)}}{\partial \mathfrak{e}^{(t-1)}} \cdot \frac{\partial \mathfrak{e}^{(t-1)}}{\partial L_{x^{(t-1)}}}}$$

$$\frac{\partial J^{(t)}}{\partial L_{x^{(t-1)}}} = \frac{\partial J^{(t)}}{\partial \hat{y}^{(t)}} \cdot \frac{\partial \hat{y}^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial h^{(t-1)}} \cdot I \cdot h^{(t-1)}(1-h^{(t-1)})$$

$$\boldsymbol{\frac{\partial J^{(t)}}{\partial I}\ |_{(t-1)} = \frac{\partial J^{(t)}}{\partial \hat{y}^{(t)}} \cdot \frac{\partial \hat{y}^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial h^{(t-1)}} \cdot \frac{\partial h^{(t-1)}}{\partial I}}$$

$$\frac{\partial J^{(t)}}{\partial I} = \frac{\partial J^{(t)}}{\partial \hat{y}^{(t)}} \cdot \frac{\partial \hat{y}^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial h^{(t-1)}} \cdot \left[\mathfrak{e}^{(t-1)} \cdot h^{(t-1)}(1-h^{(t-1)})\right]\ |_{(t-1)}$$

$$\frac{\partial J^{(t)}}{\partial b_1}|_{(t-1)} = \frac{\partial J^{(t)}}{\partial \hat{y}^{(t)}} \cdot \frac{\partial \hat{y}^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial h^{(t-1)}} \cdot \frac{\partial h^{(t-1)}}{\partial b_1}$$

$$h^{(t-1)} = sigmoid\left(h^{(t-2)}H + \mathfrak{e}^{(t-1)}I + b_1\right) = \frac{1}{1+e^{-\left(h^{(t-2)}H + \mathfrak{e}^{(t-1)}I + b_1\right)}}$$

$$\frac{\partial h^{(t-1)}}{\partial b_1} = \frac{e^{-\left(h^{(t-2)}H + \mathfrak{e}^{(t-1)}I + b_1\right)}}{\left(1+e^{-\left(h^{(t-2)}H + \mathfrak{e}^{(t-1)}I + b_1\right)}\right)^2}$$

$$\underbrace{=}_{sigmoids'\ proprerty} \sigma\left(h^{(t-2)}H + \mathfrak{e}^{(t-1)}I + b_1\right)\left(1-\sigma\left(h^{(t-2)}H + \mathfrak{e}^{(t-1)}I + b_1\right)\right) =$$

$$\underbrace{=}_{h^{(t-1)}=\sigma\left(h^{(t-2)}H + \mathfrak{e}^{(t-1)}I + b_1\right)} h^{(t-1)}(1-h^{(t-1)})$$

$$\frac{\partial J^{(t)}}{\partial I}\ |_{(t-1)} = \frac{\partial J^{(t)}}{\partial \hat{y}^{(t)}} \cdot \frac{\partial \hat{y}^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial h^{(t-1)}} \cdot h^{(t-1)}(1-h^{(t-1)})$$

3 Question 3

3.1 a

One advantage of a character-based language model over a word-based language model could be learning more delicate rules of language and, which enables a more detailed analysis of linguistic patterns such as rare words and different spelling.

A word-based language model, on the other hand, can reduce the number of possibilities for sequences it has to generate as well as the size of the vocabulary it must consider, thus reducing computational complexity compared to character-based models.

3.2 b

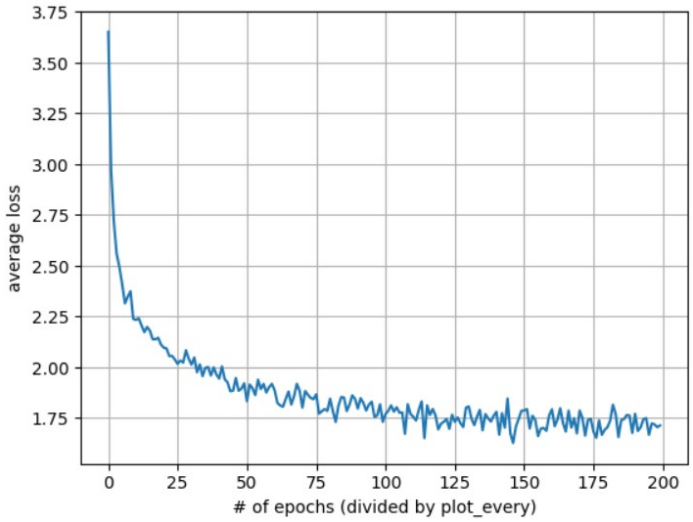


Figure :2 Training Loss

4 Question 4

4.1 a

$$2^{-\frac{1}{M}\sum_{i=1}^{M}\log_2 p(s_i|s_1,\ldots,s_{i-1})} \underbrace{=}_{base\ change} 2^{-\frac{1}{M}\sum_{i=1}^{M}\frac{\log_e p(s_i|s_1,\ldots,s_{i-1})}{\log_e 2}} = 2^{-\frac{1}{M}\sum_{i=1}^{M}\frac{\ln p(s_i|s_1,\ldots,s_{i-1})}{\ln 2}} =$$

$$= 2^{-\frac{1}{M\cdot\ln 2}\sum_{i=1}^{M}\ln p(s_i|s_1,\ldots,s_{i-1})} =$$

$$\underbrace{=}_{powers\ rules} \left(2^{\log_2 e}\right)^{\left(-\frac{1}{M}\sum_{i=1}^{M}\ln p(s_i|s_1,\ldots,s_{i-1})\right)} = e^{-\frac{1}{M}\sum_{i=1}^{M}\ln p(s_i|s_1,\ldots,s_{i-1})}$$

5 Question 5

5.1 (a) The Evaluation Accurancy as a function of the Number of Epochs

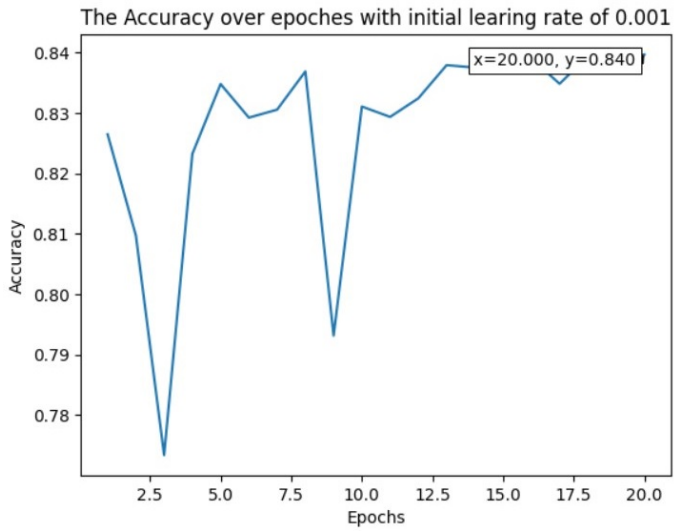


Figure :3 The evaluation accuracy as a function of the number of epochs when running the DAN model with the following parameters: Number of epoches = ,20 Both train and evel batch size = 8,Learning rate = .0.001 Using $ReLU$ activation function between adjecent hidden linear layers (execpt for the last layer), whose dimentions are: first hidden layer $[450, 100]$second hidden layer $[100, 50]$third hidden layer $[50, 10]$last and fifth hidden layer $[10, 2]$

5.2 (b) The Evaluation Accurancy as a Function of the Dropout Probability

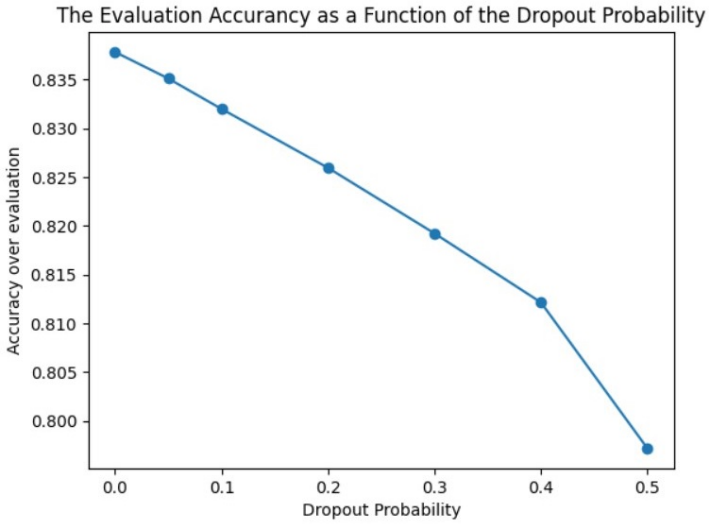


Figure :4 The evaluation accuracy as a function of the dropout probability when running the DAN model with the following parameters: Number of epoches = ,12 Both train and evel batch size = 8,Learning rate = .0.001 Using $ReLU$ activation function between adjecent hidden linear layers (execpt for the last layer), whose dimentions are: first hidden layer $[450, 100]$second hidden layer $[100, 50]$third hidden layer $[50, 10]$last and fifth hidden layer $[10, 2]$

As for using word dropout as a regularization factor in the model, we can observe that there wasn't an improvement of the accuracy as the dropout probability increases. Furthermore , when increasing the dropout probability we can see that our model's accuracy decreases. One possible explanation of the phenomena can be that when droppoutting words, we drop words with significant

sentiment attached to them leading the sentiment of the sentence to be indistinguishable to our model.

5.3 (c) The Evaluation Accuracy as the Number of Layers Increases

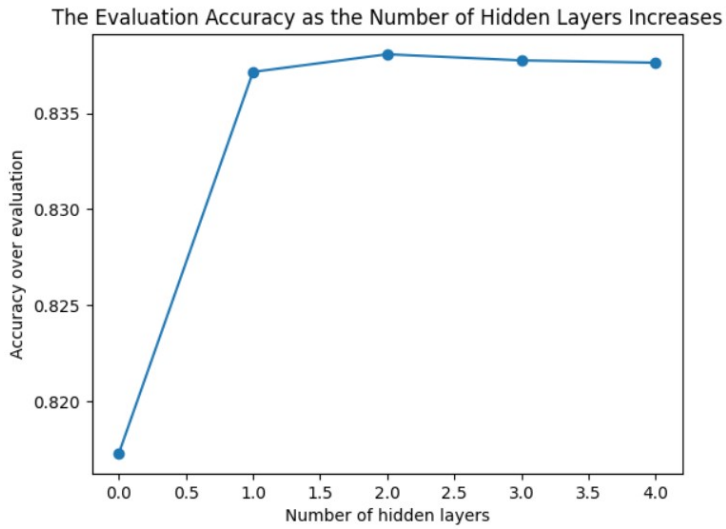


Figure :5 The accuracy as the number of layers increases. Running five DAN models : Model -(0) without hidden layers. The only layer dimensions are $[300, 2]$. Model (1) - with **one** hidden layer with dimentions $[450, 2]$. Model (2) - with **two** hidden layers with dimentions $[450, 100], [100, 2]$. Model (3) - with **three** hidden layers with dimentions $[450, 100], [100, 50], [50, 2]$. Model (4) - with **four** hidden layers with dimentions $[450, 100], [100, 50], [50, 10], [10, 2]$.with the following commmon parameters: Number of epoches = ,12 Both train and evel batch size = ,8 Learning rate = 0.001 and Using $ReLU$ activation function between adjecent hidden linear layers execpt for the last layer.

We can observe that the accuracy of the model increases overall when the number of hidden layer grows larger but the improvement of the accuracy isn't stiff with over two layers. There might be a diminishing return which starts with over three hidden layers, but its unclear to determine from the experiment we did.

We expected that there should be a significant improvement of the accuracy when adding more than two hidden layers but the results of the experiment shows otherwise.

Furthermore, the linear model without hidden layers did not outperformed the four layer model.

5.4 (d) experimenting with different activation functions over the hidden layers

Choosing the following activation functions: $nn.ReLU$, $nn.LeakyReLU$, $nn.Hardswish()$:

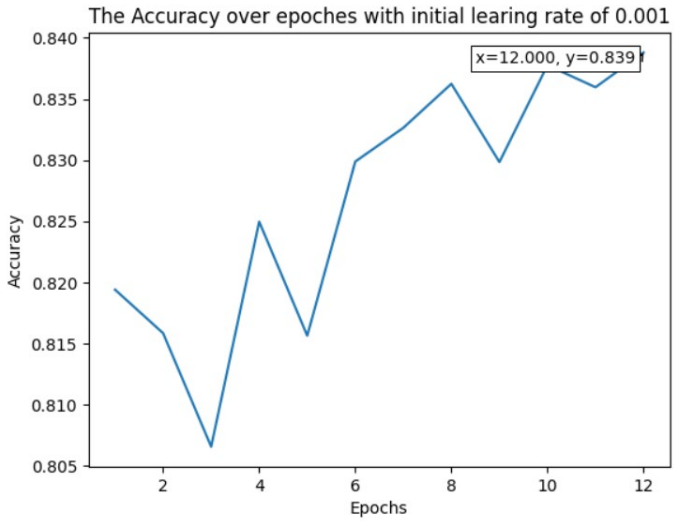


Figure :6 The accuracy over epochs when experimenting with different activation functions, $nn.ReLU$, $nn.LeakyReLU$, $nn.Hardswish()$.Using 4 hidden layers with the same dimentions mentioned in the previous part. The order of the activation functions over the hidden layer is as follows: $[nn.LeakyReLU(), nn.ReLU(), nn.Hardswish(), nn.ReLU()]$

We can observe that adding different activation functions over the linear hidden layers leads to a less stable improvement of the accuracy, with more ups and downs in the accuracy. But overall, the maximal accuracy achieved is almost the same as with applying the same activation function over all the layers.

5.5 e - Interpreting a Sample of Wrongly Predicted Inputs

1. "This is the latest entry in the long series of films with the French agent, O.S.S. 117 (the French answer to James Bond). The series was launched in the early 1950's, and spawned at least eight films (none of which was ever released in the U.S.). 'O.S.S.117:Cairo,Nest Of Spies' is a breezy little comedy that should not...repeat NOT, be taken too seriously. Our protagonist finds himself in the middle of a spy chase in Egypt (with Morroco doing stand in for Egypt) to find out about a long lost friend. What follows is the standard James Bond/Inspector Cloussou kind of antics. Although our man is something of an overt xenophobe,sexist,homophobe, it's treated as pure farce (as I said, don't take it too seriously). Although there is a bit of **rough language** & cartoon **violence**, it's basically $okay\ for\ older\ kids$ (ages 12 & up). As previously stated in the subject line, just sit back,pass the popcorn &$just\ enjoy$."

   (a) The true label 1 (positive)

   (b) The predicted Label: 0 (negative)

      i. We think that although the reviewer has some complains about the movie, he chooses to take them in an easy manner and was able to enjoy. Our model probably miss understood this point of view of the reviewer as the model's input is the average embedding it encountered. This might occur because it containes strongly negative related words such as **violence** and **rough language.**

2. "This is a really **sad**, and $touching$ movie! It deals with the subject of child **abuse**. It's really **sad**, but mostly a true story, because it happens everyday. Elijah Wood and Joseph Mazzello play the two children or Lorraine Bracco, a single mother who just tries to make a home for them. While living with her parents, a man, who likes to be called "The King" comes into their life. He **hits** the youngest boy, Bobby, but the two brothers vow not to tell their mother. But finally she finds out, after the Bobby is hurt **badly**. The end kind of **ruined** it for me, because it is so totally unbelievable. But, except for that, I $love$ the movie.".

   (a) The true label 1 (positive)

   (b) The predicted Label: 0 (negative)

i. The words appear in this quote are taken from the movies' terminology, which probably was a sad movie disscussing about a sad phenonema, child abuse . Therefore it has more negativly sentiment words related to it although the overall review convey that the reviewer loved the movie. Probably the embeddings of those words reflected in the resulting average embedding.

3. "These days, writers, directors and producers are relying more and more on the "$surprise$" ending. The old art of bringing a movie to closure, taking all of the information we have learned through out the movie and bringing it to a $nice$ complete ending, has been **lost**. Now what we have is a movie that, no matter how complex, detailed, or frivolous, can be wrapped up in 5 minutes. It was all in his/her head. That explanation is the director's safety net. If all else **fails**, or if the writing wasn't that $good$, or if we ran out of money to complete the movie, we can always say "it was all in his/her head" and end the movie that way. The audience will buy it because, well, none of us are psychologists, and none of us are suffering from schizophrenia (not that we know about) so we take the story and $believe$ it. After all, the mind is a $powerful$ thing. Some movies have $pulled\ it\ off$. But those movies are the reason why we are getting more and more of these **crap** endings. Every director/writer now thinks they can $pull\ it\ off$ because, $well$, Fight Club did it and it made a lot of money. So we get movies like The Machinist, Secret Window, Identity, and this movie (just to name a few)".

   (a) The true label 0 (negative)

   (b) The predicted Label: 1 (positive)

      i. This quote has more words which have a stronger semantically positive sentiment to them rather than a negative which probably reflected in the **average embedding**. We assume that the writer of this review wanted to express his disappointment of the movie by telling a story about movies in general in his point of view. Our model probably wasn't able to understand it as the main idea is spelled with fewer words with negative sentiment to them.

4. "It **could be easy to complain** about the quality of this movie (you don't have to throw cartloads of money at a movie to make it $good$, nor will it guarantee that it is $worth\ watching$) but I think that is totally **missing the point.** If your expecting fast cars, T&A or a movie that will spell itself out for you then don't watch this, you'll be **disappointed** and **dumbfounded**.

This movie was $thoroughly\ enjoyable$, kept us on the edge of our seats and $made\ us\ really\ think$. The writer obviously put a lot of thought and research behind this movie and it shows through the end, just remember to keep an open mind.

Note: the school scenes were all filmed at McMaster University and most of the rest was done in Toronto."

   (a) The true label 1 (positive)

   (b) The predicted Label: 0 (negative)

      i. Similarly to previously explanations we suggested , this quote has more words which with **stronger** semantically negative sentiment to them rather than a positive sentiment.We assume this was probably reflected in the **average embedding**. Although the massage of the reviewer passed using words which might interpreted as negative or neutral sentiment, the overall sentiment is positive.

5. "**A not so good** action thriller because it unsuccessfully trends the same water as early Steven Seagal films because there **is not a very good** set piece. Steven Seagal plays the same kind of character that he has played since Above the Law. In my opinion the performance of Keenen Ivory Wayans is **wasted** in such an **average film** and belongs in a much better film. Bob Gunton is okay as the main heavy.$The\ best\ acting$ in the entire film belongs to Brian Cox who is $very\ frightening$ in the role of the murderer. My $favorite\ scenes$ are the fight scenes with the Russian mafia. One of the film reasons to see The Glimmer Man(1996) is for the brief appearence of the beautiful and voluptupus Nikki Cox. Its **too bad** that there were not more scenes with her in them.

   (a) The true label 0 (negative)

   (b) The predicted Label: 1 (positive)

i. As we can observe , this review has both positive and negative sentiment to it but the overall review is more negative. As for the model's architecture, it averages the embeddings of all the sequence and then the average embedding is fed is input. We can see that both the negative or the positive aspects of the review are not expressed with strongly sentiment words associated with them. The overall note might got missed by the averaging.

6 Question 6 - Right-to-left vs left-to-right Estimation

Let $x_0, x_1, ..., x_n$ be any sentence, where x_0 is the start symbol and x_n is the end symbol.

$$
\begin{aligned}
P(x_0 x_1 ... x_n) &= P(x_n)P(x_{n-1}|x_n) \cdot ... \cdot P(x_0|x_1) = \\
&\underbrace{=}_{bayes\ rule} P(x_n)\frac{P(x_n|x_{n-1})P(x_{n-1})}{P(x_n)} \cdot \frac{P(x_{n-1}|x_{n-2})P(x_{n-2})}{P(x_{n-1})} \cdot ... \cdot \frac{P(x_1|x_0)P(x_0)}{P(x_1)} = \\
&= P(x_n|x_{n-1}) \cdot P(x_{n-1}|x_{n-2}) \cdot ... \cdot P(x_1|x_0)P(x_0) = \\
&\underbrace{=}_{Commutative\ property} P(x_0)P(x_1|x_0)...P(x_n|x_{n-1})
\end{aligned}
$$