Language Models

Q1

a)

$$CE(y,\hat{y}) = -\sum_i y_i \cdot \log(\hat{y}_i) \overset{\text{y is a one-hot vector } (y_i = 1)}{=} -y_i \cdot \log(\hat{y}_i) = -1 \cdot \log\left(\frac{\exp(\theta_i)}{\sum_j \exp(\theta_j)}\right) = -\,\theta_i + \log\left(\sum_j \exp(\theta_j)\right)$$

$$\frac{\partial CE(y,\hat{y})}{\partial \theta_i} = -1 + \frac{\exp(\theta_i)}{\sum_j \exp(\theta_j)}$$

(j≠i)

$$\frac{\partial CE(y,\hat{y})}{\partial \theta_j} = \frac{\exp(\theta_j)}{\sum_k \exp(\theta_k)}$$

Overall:

$$\frac{\partial CE(y,\hat{y})}{\partial \theta} = \hat{y} - y$$

b)

By the chain rule:

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

Where we define:

$$z = xw_1 + b_1, \theta = hw_2 + b_1$$

- $\frac{\partial J}{\partial \theta}$ = $\hat{y} - y$ (from part (a))
- $\frac{\partial \theta}{\partial h} = \ w_2 \ (w_2 \in \mathbb{R}^{D_h \times D_y})$
- $\frac{\partial h}{\partial z} = \sigma(z)^T \cdot \big(1 - \sigma(z)\big)\ (hence\ \frac{\partial h}{\partial z} \in \mathbb{R}^{D_h \times D_h})$
- $\frac{\partial z}{\partial x} = w_1 \ (w_1 \in \mathbb{R}^{D_x \times D_h})$

Overall:

$\frac{\partial J}{\partial x} = (\hat{y} - y) \cdot w_2 \cdot \sigma(xw_1 + b_1)^T \cdot \big(1 - \sigma(xw_1 + b_1)\big) \cdot {w_1}^T \ (hence\ \frac{\partial J}{\partial x} \in \mathbb{R}^{1 \times D_X})$

Theoretical Inquiry of a Simple RNN Language Model

For any timestep t, the model is defined as follows:

$$J^{(t)}(\theta) = CE(y^{(t)}, \hat{y}^{(t)}) = -\textstyle\sum_{i=1}^{|V|} {y_i}^{(t)} log(\hat{y}_i^{(t)}), \quad \hat{y}^{(t)} = softmax(h^{(t)}U + b_2)$$

$$z^{(t)} := h^{(t)}U + b_2 \implies \hat{y}_i^{(t)} = softmax(z^{(t)})_i = \frac{exp(z_i^{(t)})}{\sum_{k=1}^{|V|} exp(z_k^{(t)})}$$

$$J^{(t)}(\theta) = CE(y^{(t)}, \hat{y}^{(t)}) = -\sum_{i=1}^{|V|} {y_i}^{(t)} log(\hat{y}_i^{(t)}) = -\sum_{i=1}^{|V|} {y_i}^{(t)} log(\frac{exp(z_i^{(t)})}{\sum_{k=1}^{|V|} exp(z_k^{(t)})}) =$$

$$-\sum_{i=1}^{|V|} {y_i}^{(t)}(z_i^{(t)} - log(\sum_{k=1}^{|V|} exp(z_k^{(t)}))) = -\sum_{i=1}^{|V|} {y_i}^{(t)} z_i^{(t)} + \sum_{i=1}^{|V|} {y_i}^{(t)} log(\sum_{k=1}^{|V|} exp(z_k^{(t)})) =$$

$$-y^{(t)} \cdot z^{(t)} + log(\sum_{k=1}^{|V|} exp(z_k^{(t)})) \sum_{i=1}^{|V|} {y_i}^{(t)} = -y^{(t)} \cdot z^{(t)} + log(\sum_{k=1}^{|V|} exp(z_k^{(t)}))$$

$$e^{(t)} = x^{(t)}L, \quad h^{(t)} = \sigma(h^{(t-1)}H + e^{(t)}I + b_1)$$

a

Compute the gradients for all model parameters at a single point in time (timestep) t:

$$\frac{\partial J^{(t)}}{\partial U} = \frac{\partial J^{(t)}}{\partial z^{(t)}} \cdot \frac{\partial z^{(t)}}{\partial U}$$

$$\frac{\partial z^{(t)}}{\partial U} = \frac{\partial}{\partial U}[h^{(t)}U + b_2] = h^{(t)}$$

$$\frac{\partial J^{(t)}}{\partial z^{(t)}} = \frac{\partial}{\partial z^{(t)}}[-y^{(t)} \cdot z^{(t)} + log(\sum_{k=1}^{|V|} exp(z_k^{(t)}))] = -y^{(t)} + \frac{\partial}{\partial z^{(t)}}[log(\sum_{k=1}^{|V|} exp(z_k^{(t)}))]$$

$$\frac{\partial}{\partial z_i^{(t)}}[log(\sum_{k=1}^{|V|} exp(z_k^{(t)}))] = \frac{\frac{\partial}{\partial z_i^{(t)}}[\sum_{k=1}^{|V|} exp(z_k^{(t)})]}{\sum_{k=1}^{|V|} exp(z_k^{(t)})} = \frac{\frac{\partial}{\partial z_i^{(t)}}[exp(z_i^{(t)})]}{\sum_{k=1}^{|V|} exp(z_k^{(t)})} =$$

$$= \frac{exp(z_i^{(t)})}{\sum_{k=1}^{|V|} exp(z_k^{(t)})} = \hat{y}_i^{(t)}$$

$$\implies \frac{\partial}{\partial z}[log(\sum_{k=1}^{|V|} exp(z_k^{(t)}))] = \hat{y}^{(t)} \implies \frac{\partial J^{(t)}}{\partial z} = -y^{(t)} + \hat{y}^{(t)}$$

$$\implies \frac{\partial J^{(t)}}{\partial U} = (-y^{(t)} + \hat{y}^{(t)})^T \cdot h^{(t)}$$

$$\frac{\partial J^{(t)}}{\partial b_2} = \frac{\partial J^{(t)}}{\partial z^{(t)}} \cdot \frac{\partial z^{(t)}}{\partial b_2} = \frac{\partial J^{(t)}}{\partial z^{(t)}} \cdot \frac{\partial [h^{(t)}U + b_2]}{\partial b_2} = (-y^{(t)} + \hat{y}^{(t)}) \cdot \mathbb{I}_{|V|}$$

(Where $\mathbb{I}_a$ is the unit matrix of dimensions $a \times a$.)

$$\frac{\partial J^{(t)}}{\partial L_{x^{(t)}}} = \frac{\partial J^{(t)}}{\partial z^{(t)}} \cdot \frac{\partial z^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial e^{(t)}} \cdot \frac{\partial e^{(t)}}{\partial L_{x^{(t)}}}$$

$$\frac{\partial e^{(t)}}{\partial L_{x^{(t)}}} = \frac{\partial [x^{(t)} \cdot L]}{\partial L_{x^{(t)}}} = \frac{\partial [L_{x^{(t)}}]}{\partial L_{x^{(t)}}} = \mathbb{I}_d$$

$$w^{(t)} = h^{(t-1)} H + e^{(t)} I + b_1 \implies h^{(t)} = \sigma(w^{(t)})$$

$$\frac{\partial h^{(t)}}{\partial e^{(t)}} = \frac{\partial [\sigma(w^{(t)})]}{\partial w^{(t)}} \cdot \frac{\partial w^{(t)}}{\partial e^{(t)}} = \sigma(w^{(t)})^T (1 - \sigma(w^{(t)})) \cdot I^T = (h^{(t)})^T \cdot (1 - h^{(t)}) \cdot I^T$$

$$\frac{\partial z^{(t)}}{\partial h^{(t)}} = U$$

$$\implies \frac{\partial J^{(t)}}{\partial L_{x^{(t)}}} = (-y^{(t)} + \hat{y}^{(t)}) \cdot U^T \cdot (h^{(t)})^T \cdot (1 - h^{(t)}) \cdot I^T \cdot \mathbb{I}_d$$

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

$$\frac{\partial h^{(t)}}{\partial I} = \frac{\partial h^{(t)}}{\partial w^{(t)}} \cdot \frac{\partial w^{(t)}}{\partial I} = \frac{\partial [\sigma(w^{(t)})]}{\partial w^{(t)}} \cdot \frac{\partial w^{(t)}}{\partial I} = \sigma(w^{(t)})^T (1 - \sigma(w^{(t)})) \cdot e^{(t)}$$

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

$$\frac{\partial J^{(t)}}{\partial H}|_{(t)} = \frac{\partial J^{(t)}}{\partial z^{(t)}} \cdot \frac{\partial z^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial w^{(t)}} \cdot \frac{\partial w^{(t)}}{\partial H}|_{(t)} = (-y^{(t)} + \hat{y}^{(t)}) \cdot U^T \cdot {h^{(t)}}^T (1 - h^{(t)}) \cdot \frac{\partial w^{(t)}}{\partial H}|_{(t)}$$

$$\frac{\partial w^{(t)}}{\partial H}|_{(t)} = h^{(t-1)}$$

$$\implies \frac{\partial J^{(t)}}{\partial H}|_{(t)} = (-y^{(t)} + \hat{y}^{(t)}) \cdot U^T \cdot {h^{(t)}}^T (1 - h^{(t)}) \cdot h^{(t-1)}$$

$$\frac{\partial J^{(t)}}{\partial b_1}|_{(t)} = \frac{\partial J^{(t)}}{\partial z^{(t)}} \cdot \frac{\partial z^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial b_1}|_{(t)}$$

$$\frac{\partial h^{(t)}}{\partial b_1}|_{(t)} = \mathbb{I}_{D_n}$$

$$\implies \frac{\partial J^{(t)}}{\partial b_1}|_{(t)} = (-y^{(t)} + \hat{y}^{(t)}) \cdot U^T \cdot {h^{(t)}}^T (1 - h^{(t)}) \cdot \mathbb{I}_{D_n}$$

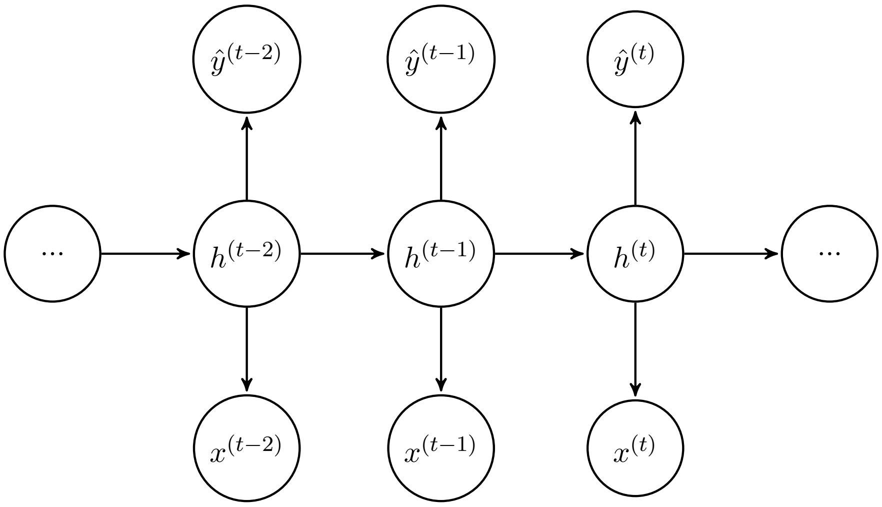
$$\frac{\partial J^{(t)}}{\partial h^{(t-1)}} = \frac{\partial J^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial h^{(t-1)}}$$

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

$$\implies \frac{\partial J^{(t)}}{\partial h^{(t-1)}} = (-y^{(t)} + \hat{y}^{(t)}) \cdot U^T \cdot {h^{(t)}}^T (1 - h^{(t)}) \cdot H^T$$

b

Draw the unrolled network for 3 timesteps and compute the "backpropagation-through-time"gradients



$$\begin{aligned}
\frac{\partial J^{(t)}}{\partial L_{x^{(t-1)}}} &= \frac{\partial J^{(t)}}{\partial h^{(t-1)}} \cdot \frac{\partial h^{(t-1)}}{\partial e^{(t-1)}} \cdot \frac{\partial e^{(t-1)}}{\partial L_{x^{(t-1)}}} = \\
&= (-y^{(t)} + \hat{y}^{(t)}) \cdot U^T \cdot {h^{(t)}}^T (1 - h^{(t)}) \cdot H^T \cdot (h^{(t-1)})^T \cdot (1 - h^{(t-1)}) \cdot I^T \cdot \mathbb{I}_d \\
\frac{\partial J^{(t)}}{\partial H}|_{(t-1)} &= \frac{\partial J^{(t)}}{\partial h^{(t-1)}} \cdot \frac{\partial h^{(t-1)}}{\partial H} = \\
&= (-y^{(t)} + \hat{y}^{(t)}) \cdot U^T \cdot {h^{(t)}}^T (1 - h^{(t)}) \cdot H^T \cdot (h^{(t-1)})^T \cdot (1 - h^{(t-1)}) \cdot h^{(t-2)} \\
\frac{\partial J^{(t)}}{\partial I}|_{(t-1)} &= \frac{\partial J^{(t)}}{\partial h^{(t-1)}} \cdot \frac{\partial h^{(t-1)}}{\partial I} = \\
&= (-y^{(t)} + \hat{y}^{(t)}) \cdot U^T \cdot {h^{(t)}}^T (1 - h^{(t)}) \cdot H^T \cdot (h^{(t-1)})^T \cdot (1 - h^{(t-1)}) \cdot e^{(t-1)} \\
\frac{\partial J^{(t)}}{\partial b_1}|_{(t-1)} &= \frac{\partial J^{(t)}}{\partial h^{(t-1)}} \cdot \frac{\partial h^{(t-1)}}{\partial b_1} = \\
&= (-y^{(t)} + \hat{y}^{(t)}) \cdot U^T \cdot {h^{(t)}}^T (1 - h^{(t)}) \cdot H^T \cdot (h^{(t-1)})^T \cdot (1 - h^{(t-1)}) \cdot \mathbb{I}_{D_n}
\end{aligned}$$

We already computed these derivatives in (a):

$$\begin{aligned}
\frac{\partial J^{(t)}}{\partial h^{(t-1)}} &= (-y^{(t)} + \hat{y}^{(t)}) \cdot U^T \cdot {h^{(t)}}^T (1 - h^{(t)}) \cdot H^T \\
\frac{\partial h^{(t-1)}}{\partial L_{x^{(t-1)}}} &= \frac{\partial h^{(t-1)}}{\partial e^{(t-1)}} \cdot \frac{\partial e^{(t-1)}}{\partial L_{x^{(t-1)}}} = (h^{(t-1)})^T \cdot (1 - h^{(t-1)}) \cdot I^T \cdot \mathbb{I}_d \\
\frac{\partial h^{(t-1)}}{\partial H}|_{(t-1)} &= (h^{(t-1)})^T \cdot (1 - h^{(t-1)}) \cdot h^{(t-2)} \\
\frac{\partial h^{(t-1)}}{\partial I} &= (h^{(t-1)})^T \cdot (1 - h^{(t-1)}) \cdot e^{(t-1)} \\
\frac{\partial h^{(t-1)}}{\partial b_1} &= (h^{(t-1)})^T \cdot (1 - h^{(t-1)}) \cdot \mathbb{I}_{D_n}
\end{aligned}$$

Q3

a)

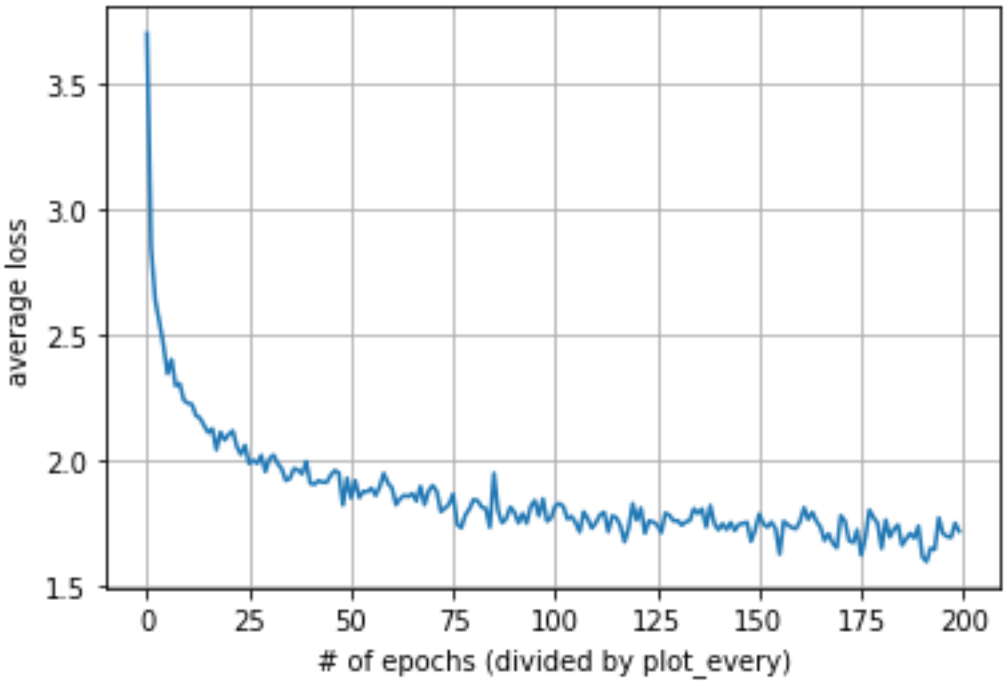
(i) Advantages of character-based language model over word-based language model:

One advantage of using a character-based language model is that the size of the matrix of all possible text generations is equal to the number of characters appear in the text, which is much lower than the size of all words appear in the same text. Another advantage is that a character-based language model can deal with spelling mistakes, while a word-based language model will see this spelling mistake as a different word in the text (which may cause lots of errors in the generating phase).

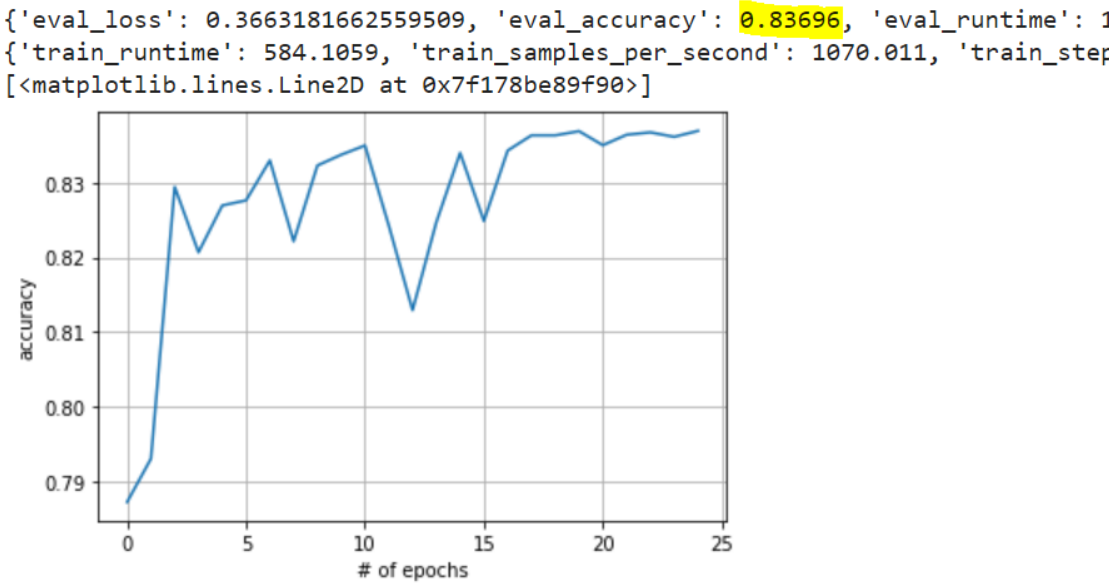
(ii) Advantages of word-based language model over character-based language model:

A word-based language model will surely make more sense in semantics and syntax, as it learns by taking whole words and not characters, as we can see it the printed test in our Shakespeare model. Also it improves significantly the option of generating words that are not in the vocabulary, which a character-based language model generate incessantly.

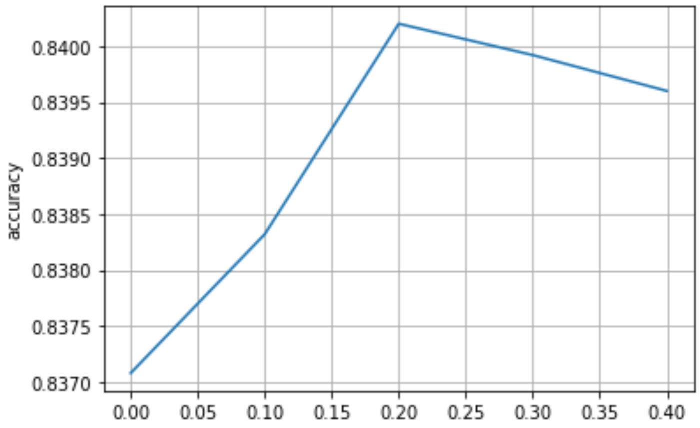
b) Dev perplexity: 112.817



5. a. Using the following architecture, we have achieved ~ 83.7% accuracy.



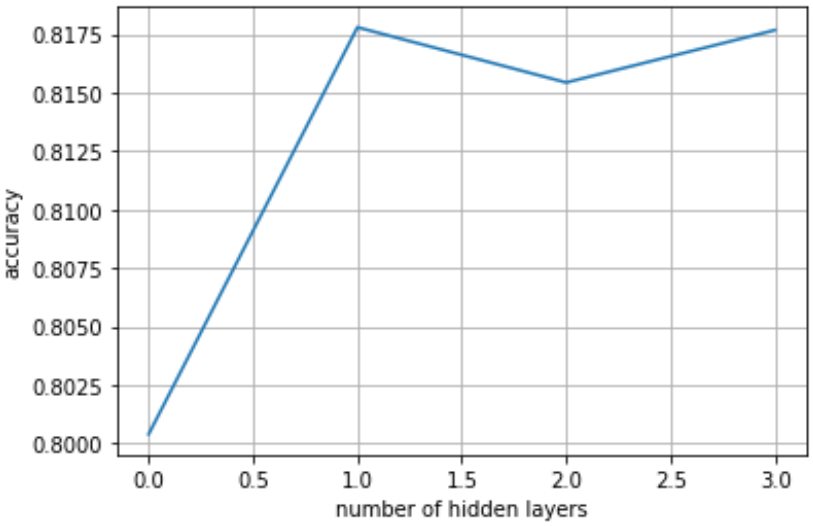
```
{'eval_loss': 0.3663181662559509, 'eval_accuracy': 0.83696, 'eval_runtime': 1
{'train_runtime': 584.1059, 'train_samples_per_second': 1070.011, 'train_step
[<matplotlib.lines.Line2D at 0x7f178be89f90>]
```



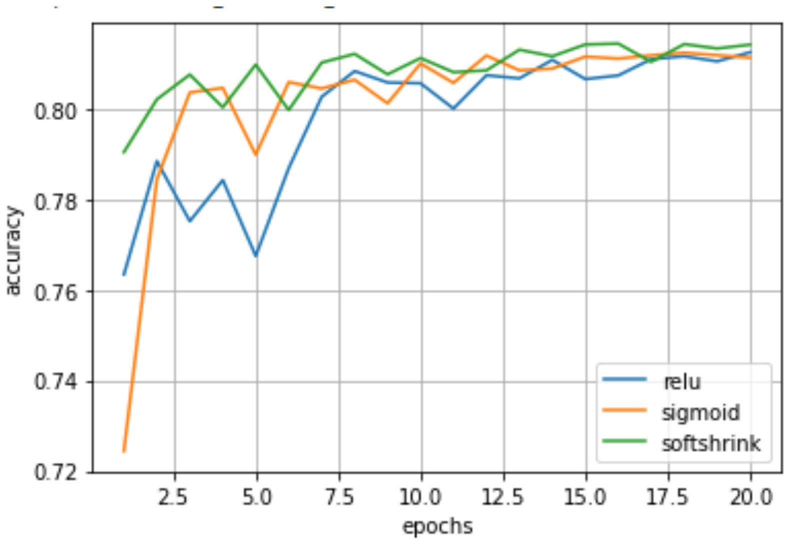
b.

We used the following values for dropout: 0.1,0.2,0.3,0.4. we got the best results for 20% dropout.

c.



This model didn't outperform the one we found in a. The effect of diminishing returns states that as we are adding more layers the value, they add decreases – in the graph we can see that 1 and 3 layers gave pretty much the same accuracy. Important to mention that in a&b we used more types of activation functions, hence, the models in a&b do better. In c we used only ReLU and linear layers.



d.

We can see that while using different activations, eventually they all converge. Note: with softshrink we had to remove the last layer of activation to see a learning process.

e.

1.

This movie turned out to be better than I had expected it to be. Some parts were pretty funny. It was nice to have a movie with a new plot.
Prediction = negative, real = positive.
It's not clear why the model misclassified this sample, it has "positive vibes".

2.

"The screen-play is very bad, but there are some action sequences that i really liked. I think the image is good, better than other romanian movies. I liked also how the actors did their jobs."
Prediction = positive, real = negative.
The first sentence indicates for the human reader that the whole experience was "bad" but it follows some compliments and words that imply good things. Hence, it was misclassified.

3.

If it wasn't for the terrific music, I would not hesitate to give this cinematic underachievement 2/10. But the music actually makes me like certain passages, and so I give it 5/10.
Prediction = positive, real = negative.
Some words like 'terrific', 'like' can really push the weights to the positive direction.

4.

A really realistic, sensible movie by Ramgopal Verma . No stupidity like songs as in other Hindi movies. Class acting by Nana Patekar. <br /><br />Much similarities to real 'encounters'.
Prediction = negative, real = positive.
Some words like 'stupidity' can really push the weights to the negative direction.

5.

im sure he doesnt need the money for a life saving operation or transplant. in all honesty i think this review qualifies as a better movie than 'bulletproof'. thanks for listening

Prediction = positive, real = negative.
The text doesn't refer to the movie but some general sentence, plus the words 'better movie' takes the sentiment to the positive direction.

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. Prove that:

$$P(x_0 x_1 ... x_n) = P(x_n) \cdot P(x_{n-1} \mid x_n) \cdot ... \cdot P(x_0 \mid x_1) = P(x_0) \cdot P(x_1 \mid x_0) \cdot ... \cdot P(x_n \mid x_{n-1})$$

Proof. Count based bi-gram model is defined by using the probabilities:

$$P(w_i \mid w_{i-1}) = \frac{c(w_i, w_{i-1})}{c(w_{i-1})}$$

And it actually follows the Bayse Theorem:

$$P(w_i \mid w_{i-1}) \cdot \frac{P(w_{i-1})}{P(w_i)} = \frac{c(w_i, w_{i-1})}{c(w_{i-1})} \frac{c(w_{i-1})}{c(w_i)} = \frac{c(w_i, w_{i-1})}{c(w_i)} = P(w_{i-1} \mid w_i)$$

Using this we get:

$$\begin{aligned}
P(x_n) \cdot P(x_{n-1} \mid x_n) \cdot ... \cdot P(x_0 \mid x_1) &= P(x_n) \cdot \prod_{i=1}^{n} P(x_{i-1} \mid x_i) = \\
&= P(x_n) \cdot \prod_{i=1}^{n} P(x_i \mid x_{i-1}) \cdot \frac{P(x_{i-1})}{P(x_i)} = \\
&= P(x_n) \cdot \prod_{i=1}^{n} P(x_i \mid x_{i-1}) \cdot \prod_{i=1}^{n} \frac{P(x_{i-1})}{P(x_i)} = \\
&= \prod_{i=1}^{n} P(x_i \mid x_{i-1}) \cdot \frac{P(x_n) \cdot P(x_{n-1}) \cdot ... \cdot P(x_1) \cdot P(x_0)}{P(x_n) \cdot P(x_{n-1}) \cdot ... \cdot P(x_1)} = \\
&= P(x_0) \cdot \prod_{i=1}^{n} P(x_i \mid x_{i-1}) = \\
&= P(x_0) \cdot P(x_1 \mid x_0) \cdot ... \cdot P(x_n \mid x_{n-1})
\end{aligned}$$