1 Word-Level Neural Bi-gram Language Model

1.a

Starting from our fundamental cross-entropy definition:

$$\text{CE}(y, \hat{y}) = -\sum_i y_i \cdot \log(\hat{y}_i)$$

For a word-level model with softmax output, we can express $\hat{y}_i$ as:

$$\hat{y}_i = \text{softmax}(\theta)_i = \frac{\exp(\theta_i)}{\sum_j \exp(\theta_j)}$$

Let's derive the gradient. For any output k (where y is the one-hot vector):

$$\frac{\partial \text{CE}}{\partial \theta_k} = -\sum_i y_i \cdot \frac{\partial}{\partial \theta_k} \log(\hat{y}_i) \tag{1}$$

$$= -\sum_i y_i \cdot \frac{\partial}{\partial \theta_k} \log\left(\frac{\exp(\theta_i)}{\sum_j \exp(\theta_j)}\right) \tag{2}$$

$$= -\sum_i y_i \cdot (\delta_{ik} - \hat{y}_k) \tag{3}$$

$$= \hat{y}_k - y_k \tag{4}$$

Therefore, we can concisely express the gradient as:

$$\frac{\partial \text{CE}}{\partial \theta} = \hat{y} - y$$

where y is the one-hot target vector and $\hat{y}$ is the predicted probability distribution.

1.b

Following the chain rule of differentiation, we can express the gradient as:

$$\frac{\partial J}{\partial x} = \frac{\partial J}{\partial (hW_2 + b_2)} \cdot \frac{\partial (hW_2 + b_2)}{\partial h} \cdot \frac{\partial \sigma}{\partial (xW_1 + b_1)} \cdot \frac{\partial (xW_1 + b_1)}{\partial x}$$

$$= (\hat{y} - y) \cdot W_2^\top \odot (h(1-h)) \cdot W_1^\top$$

This derivation shows the complete backward propagation of the gradient through the neural network layers, accounting for the weights (W_1, W_2), biases (b_1, b_2), and the activation function's derivative.

1.c

Code is attached.

1.d

After training the network for 40K iterations, we got 113.699 dev perplexity.

2 Generating Shakespeare Using a Character-level Language Model

2.a

Advantages of Character-Based Language Models:

Character-based models can handle any word, even unseen words or words that are not in the vocabulary. They have a smaller vocabulary compared to word-based models, which makes them more efficient in terms of memory. Additionally, they can capture subtleties like prefixes, suffixes, or spelling variations better than word-based models.

Advantages of Word-Based Language Models:

Word-based models are faster to train because they process fewer tokens for the same text length. They are also able to achieve better syntactic and semantic understanding, resulting in more coherent and meaningful text generation.

2.b Graph Plot

Figure 1: Average loss over epochs

3 Perplexity

3.a

$$
\begin{aligned}
2^{-\frac{1}{M}\sum_{i=1}^{M}\log_2 P(s_i|s_1,\ldots,s_{i-1})} &= \left(2^{\sum_{i=1}^{M}\log_2 P(s_i|s_1,\ldots,s_{i-1})}\right)^{-\frac{1}{M}} \\
&= \left(2^{\log_2 P(s_1)+\log_2 P(s_2|s_1)+\cdots+\log_2 P(s_M|s_1,\ldots,s_{M-1})}\right)^{-\frac{1}{M}} \\
&= \left(P(s_1)\cdot P(s_2 \mid s_1)\cdot\cdots\cdot P(s_M \mid s_1,\ldots,s_{M-1})\right)^{-\frac{1}{M}} \\
&= \left(e^{\ln P(s_1)+\ln P(s_2|s_1)+\cdots+\ln P(s_M|s_1,\ldots,s_{M-1})}\right)^{-\frac{1}{M}} \\
&= \left(e^{\sum_{i=1}^{M}\ln P(s_i|s_1,\ldots,s_{i-1})}\right)^{-\frac{1}{M}} \\
&= e^{-\frac{1}{M}\sum_{i=1}^{M}\ln P(s_i|s_1,\ldots,s_{i-1})}.
\end{aligned}
$$

3.b

Neural Bi-gram Language Model:

Shakespeare Perplexity: 7.122318650322853

Wikipedia Perplexity:25.75261330172207

Character-level Language Model:

Shakespeare Perplexity: 7.459401319332429

Wikipedia Perplexity: 19.97325117242642

3.c

For the Character-level Language Model we can assume those result are due to the fact that he learns at the charachter level and also was trained on shakespear data which can give the advantage to more similiar text who is also contains non standard spelling and word forms but on the other hand the result on wikipedia was worse to the more complex and structured word level patterns. For the Bi-gram Model we would assume it struggled with Wikipedia and got even a worse score than the character-level language model because it is limited to a 2 word context, where in Wikipedia text usually contains long sentences with dependencies that span multiple words. Our assumption that the score is even lower then the Character-level Language Model to the the later ability to capture meanings of suffixes, prefixes even in complex modern words while not relying on find word pairs. Also we think the the Bi-gram model performs better on shakespear because the word pairs and sentence structures in shakespeares text are relatively predictable and consistent .

3.d

Neural Bi-gram Language Model: (After preprocessing)

Shakespeare Perplexity: 4.475288916573775

Wikipedia Perplexity: 2.536595085380771

Character-level Language Model: (After preprocessing)

Shakespeare Perplexity: 6.740292861033496

Wikipedia Perplexity: 16.24853261546454

During preprocessing, we primarily focused on "cleaning" the data. This included converting all text to lowercase for better generalization, removing all non-printable characters (which are likely irrelevant to the context), and eliminating extra spaces.

4 Deep Averaging Networks

4.a

The plot of the accuracy vs the number of epochs:

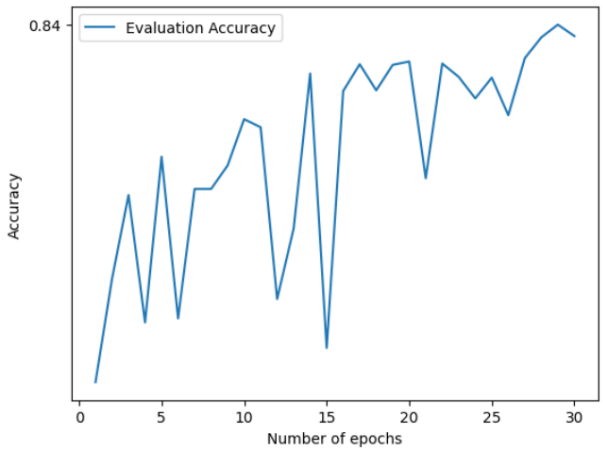

Figure 2: The plot of the accuracy vs the number of epochs

The accuracy reaches a maximal value of 0.84

4.b

The accuracy as a function of the dropout rate:

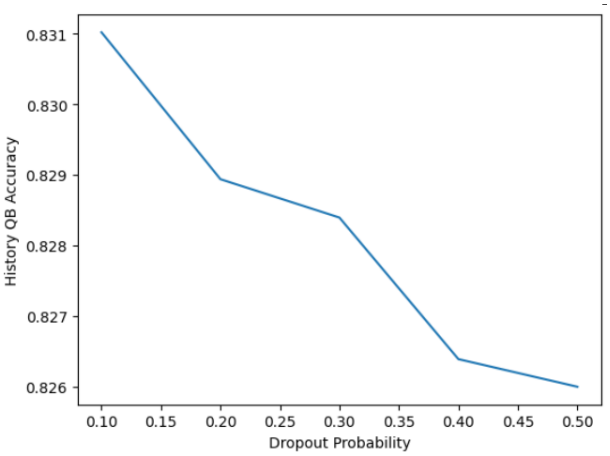

Figure 3: The accuracy as a function of the dropout rate

4.c

The accuracy as function of layer numbers:

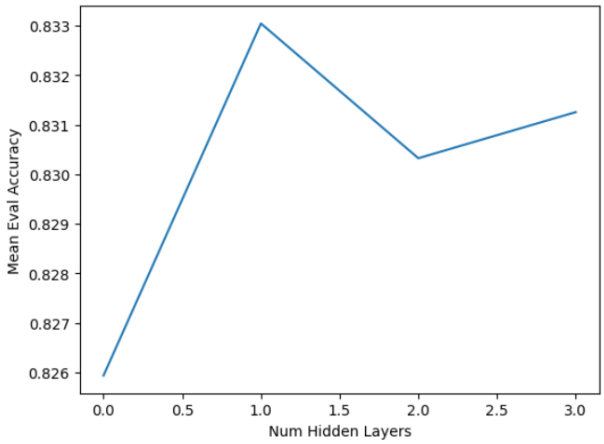

Figure 4: mean value of accuracy as function of the number of layers

Diminishing returns occur when the model's capacity exceeds what is necessary for the given task. On average, the model will perform poorly on 0 layers, since it will be linear and have trouble fitting itself into nonlinear data. Between 1 and 2, we expect the highest accuracy because the deeper network can model more complex functions. Above that, we will expect a decrease in accuracy with the potential for overfitting because the model will be too complex.

Each training included 30 epochs. We can see that the slope decreases at 1, indicating the reduction of accuracy due to diminishing returns as we expected. We can also see a small increase from 2 to 3 which doesn't surpass the peak at 1 hidden layer. Our assumption is that the transition from 1 to 2 layers may introduce optimization difficulties such as vanishing/exploding gradients and poor initialization, that degrade the performance.

4.d

.

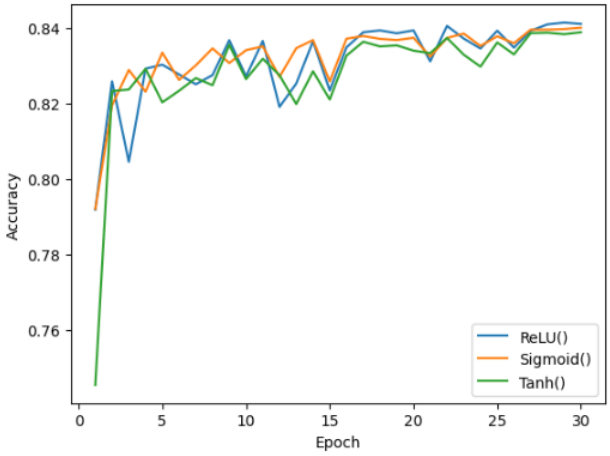

Figure 5: Plot of the accuracy across epochs

The activation functions we used are Relu, sigmoid, and Tanh. We can see that the accuracy values are similar, with some differences. The graph of the Tanh function starts with a steep increase, which stabilizes similarly to the two other functions. The Relu graph shows a steady improvement, and the graph of the Sigmoid function is smoother. All three stabilize on a value close to 0.84. We can learn that the different functions may achieve similar levels of accuracy even with different initial behaviors.

4.e

Example 1:

Prediction: 1, Label: 0

Written by, directed by and starring the champ of camp Bruce Campbell. Easy on its easy to tell this is a budget on a shoestring affair; filmed independently in Bulgaria. All I can really say for sure is that silly is not always funny. Campbell plays an affluent American business man with a cheating wife(Antoinette Byron)and trying to close a business transaction before he is murdered. He hires a cabbie to drive him around a strange little town; not knowing that his wife is 'carrying on' with the taxi driver. Within moments of Campbell being bludgeoned; the cabbie is killed in the same location. A mad scientist(Stacy Keach)proceeds with an experiment putting the cabbie's brain inside the American's head. With massive stitches on his forehead, Campbell breaks free and roams the streets looking for his wife; all the while he is arguing with a strange voice inside his over-sized head. Campbell contorts his rubbery face making silly expressions as he argues with himself. Thus, Bruce is doing what he does best and no doubt his many fans will be pleased. I get the impression this must have been written as a straight comedy. Rounding out the cast are Ted Raimi, Tamara Gorski, and Vladmir Kolev. Watch for this on the Sci-Fi Channel.

Explanation: This review might have been predicted positive because of positive comments about the actor: "Written by, directed by and starring the champ of camp Bruce Campbell", "Bruce is doing what he does best and no doubt his many fans will be pleased".

Example 2:

Prediction: 1, Label: 0

It would be wrong and reprehensible of me to advise you to watch Killjoy 2, you must have better things to do, washing the car, throwing stones in a stream, but at the same time it's nowhere near as awful as you probably think it is. It's almost a proper film, which a lot more than most straight-to-DVD sludge can manage. Killjoy 2 is helped a great deal by Trent Haaga's manic turn as the eponymous clown, he throws himself into the role with such fevered abandonment that he almost tips the scales in the movie's favour, but, of course, it takes more than one man in big shoes. Tammi Sutton gives the most entertaining director cameo since Roger Corman in Creature from the Haunted Sea and the whole thing is nearly destroyed by the rushed, sugary ending. All over the place and almost good fun.

Explanation: This review might have been predicted positive because of phrases like "it's nowhere near as awful as you probably think it is" and "it's almost a proper film" which might seem to the model as if the movie succeeded. The reviewer also compliments Trent Haaga' performance which could also be interpreted as positive.

Example 3:

Prediction: 0, Label: 1

I am not a footie fan by any means but watched this with a friend as there wasn't anything else on the box at the time.(thank goodness). Not only did we laugh from start to finish but about a week later in the pub, when we started discussing it, we made a right spectacle of ourselves with uncontrollable laughter. Does that sloping pitch actually exist??? I have released my e-mail here so if anyone hears about it's future availability or a repeat on the telly please let me know. Definitely the funniest thing I've seen on television!!! King Leek was good too!! another Tim Healy classic

Explanation: This review might have been predicted negative because of the opening sentence "I am not a footie fan by any means but watched this with a friend as there wasn't anything else on the box at the time" which could be interpreted as disinterest in the game.

Example 4:

Prediction: 1, Label: 0

Okay, I haven't read the book yet but I have to say that the lead character was miscast. How can I say such a thing haven't read the book you ask? It's simple. As a viewer of this miniseries, I grew irritated by the mannerisms, gestures and look of the lead character, Fannie Price. It's one thing to be a good person but it's quite another to be a stick in the mud creature who disdains from looking anyone in the face or otherwise meeting their gaze. Apart from the overdone "Susan B. Anthony" profile, she seemed resolute in refusing to look at another person. The scene where Edmond is pouring his heart out to her, she is looking straight ahead the whole time, forcing him to do the same. As a result, it was just awkward and I just couldn't fathom anyone being in love with her let alone both Henry and Edmond. Many have said it was true to the book, if that is the case I find it hard to believe that Jane Austen would create such a character as her lead heroine. It's possible to create a character who has been put upon by others and succeeds in earning their trust and endearment but the portrayal of this character in this miniseries just didn't do it for me.

Explanation: The review might have been predicted positive due to it's starting word "Okay", perhaps the model connected it to the movie and not as a part of a sentence.

Example 5:

Prediction: 0, Label: 1

I had the good fortune to be at Perris Island in the fall of 1959. The DI showed one evening at the outdoor theater directly in front of our barracks, Plt 162, B Co, 1st Bn, 1st ITR. Although we hadn't been there long enough to even think about seeing a movie, we could hear those that were laughing. It's one of the many indelible memories of my thirteen weeks at PI. At some later date, I got to actually see it in a theater. I'm still convinced that, to date, it remains the most realistic portrayal of the experience in the late 1950's ever done. No one has done it better than Jack Webb...

Explanation: The review can be labeled as negative maybe due to the fact that the model misses the positive tone because it's mostly descriptive (except for the words good and indelible which are positive adjectives).

5 Attention Exploration

5.a 1.a

5.a.i

α can be interpreted as categorial probability distribution because:

1. $\alpha_i > 0$ for each $i \in [n]$ due to the $\exp(x)$ function properties.

2.
$$\sum_{i=1}^{n} \alpha_i = \sum_{i=1}^{n} \frac{\exp(k_i^\top q)}{\sum_{j=1}^{n} \exp(k_j^\top q)} = \frac{\sum_{i=1}^{n} \exp(k_i^\top q)}{\sum_{j=1}^{n} \exp(k_j^\top q)} = 1.$$

since α hold those two conditions it can be interpreted as categorial probability distribution.

5.a.ii

α_j is dependent on k_j and q, so in order to achieve $\alpha_j \gg \sum_{i \neq j} \alpha_i$, we would want $k_j \gg k_i$ for each $i \neq j$ in $[n]$.

5.a.iii

Since $\alpha_j \gg \sum_{i \neq j} \alpha_i$, we can presume $\alpha_j \approx 1$.

And since $c = \sum_{i=1}^{n} v_i \alpha_i \approx 1 \cdot v_j = v_j$, it means we will get c that is very close to v_j.

5.a.iv

It means that if the product between one of the key vectors and the query is very large compared to the other keys, then that query and the key are similar. This is because the output of the attention will be very close to the value associated with that key.

5.b

5.b.i

First, let's take a look at Ms:

$$\text{Ms} = M(v_a + v_b) = Mv_a + Mv_b = M(c_1 a_1 + c_2 a_2 + \cdots + c_n a_n) + M(d_1 b_1 + d_2 b_2 + \cdots + d_m b_m)$$

$$= MAc + MBd$$

where

$$c = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix}, \quad d = \begin{pmatrix} d_1 \\ \vdots \\ d_m \end{pmatrix}$$

Since $Ac = v_a$ and $Bd = v_b$, we can use the orthogonal properties of both bases and choose $M = AA^T$, yielding:

$$MAc + MBd = AA^TAc + AA^TBd = AIc + A^*0d = Ac = v_a$$

5.b.ii

First, since $c \approx \frac{1}{2}(v_a + v_b)$, we get $\alpha_a = \alpha_b = \frac{1}{2}$. From equation (1a), we can deduce that this implies $k_a^T q + k_b^T q \gg k_i^T q$, for each $i \neq a, b \in [n]$.

Now, if we choose $q = \beta(k_a + k_b)$, with $\beta \gg 0$, we will get:

$$\alpha_i = \begin{cases} \frac{\exp(\beta)}{n-2+2\exp(b)} & \text{if } i = a \text{ or } i = b, \\ \frac{\exp(0)}{n-2+2\exp(\beta)} & \text{if } i \neq a, b. \end{cases}$$

Since $\beta \gg 0$, we have $\exp(\beta) \to \infty$, which means:

$$\alpha_i = \begin{cases} \frac{1}{2} & \text{if } i = a \text{ or } i = b, \\ 0 & \text{if } i \neq a, b. \end{cases}$$

as wanted.

5.c

5.c.i

The covariance matrices are diagonal and it is given that the values of *alpha* on the diagonal are vanishingly small. That means the random variables are uncorrelated and their variances are approaching 0. Therefore, since $k_i \sim N(\mu_i, \Sigma_i)$ we can assume that $k_i \approx \mu_i$.

$\mu_a^T \mu_b = 0$ if $a \neq b$ and $\|\mu_a\|^2$, therefore we can use the result of the previous section:

$$q = \beta(\mu_a + \mu_b)$$

5.c.ii

Let's calculate the new covariance function:

$\sum_a = \alpha I + \frac{1}{2}(\mu_\alpha^T \mu_\alpha) = \alpha I + \frac{1}{2} I$ for $i \neq a : \sum_i = \alpha I$

Since α is vanishingly small:

$$i = a \to \sum_a \approx \frac{1}{2} I$$

$$i \neq a \to \sum_i \approx 0$$

Therefore, all variables in k_a are approximately uncorrelated, their variances are 0.5 and $k_a \approx \lambda \mu_a$ with $\lambda \sim N(1, 0.5)$. $k_{i \neq a} \approx \mu_i$ as before.

With $q = \beta(\mu_a + \mu_b)$:

$$k_a^T q \approx \lambda \mu_a^T \beta(\mu_a + \mu_b)$$
$$k_a^T q \approx \lambda \mu_b^T \beta(\mu_a + \mu_b)$$

From orthonormality:

$$k_a^T q \approx \lambda\beta$$
$$k_b^T q \approx \beta$$

We will calculate α_a and α_b:

$$\alpha_a \approx \frac{exp(\lambda\beta)}{exp(\lambda\beta) + exp(\beta) + \sum(exp(0))_{i \neq a,b}} = \frac{exp(\lambda\beta)}{exp(\lambda\beta) + exp(\beta) + n - 2} \approx \frac{1}{exp(\beta - \lambda\beta) + 1}$$

$$\alpha_b \approx \frac{exp(\beta)}{exp(\lambda\beta) + exp(\beta) + \sum(exp(0))_{i \neq a,b}} \approx \frac{1}{exp(\lambda\beta - \beta) + 1 + n - 2} \approx \frac{1}{exp(\lambda\beta - \beta) + 1}$$

λ is approximately between 0.5 and 1.5, therefore:

for $\lambda = 0.5$:

$$\alpha_a \approx \frac{1}{exp(0.5\beta) + 1} \approx 0$$

$$\alpha_b \approx \frac{1}{exp(-0.5\beta) + 1} \approx 1$$

$$c = \alpha_a v_a + \alpha_b v_b \approx v_b$$

for $\lambda = 1.5$:

$$\alpha_a \approx \frac{1}{exp(-0.5\beta) + 1} \approx 1$$

$$\alpha_b \approx \frac{1}{exp(0.5\beta) + 1} \approx 0$$
$$c = \alpha_a v_a + \alpha_b v_b \approx v_a$$

The expected value of c is indeed approximately $\frac{1}{2}(v_a + v_b)$

5.d

5.d.i

As in the previous section $\mu_i \approx k_i$.

Let's pick $q_1 = \beta(\mu_a + \mu_b)$ that will give $c_1 = 0.5(v_a + v_b)$ $q_2 = \beta(\mu_a + \mu_b)$ that will give $c_2 = 0.5(v_a + v_b)$. Thus the average is:

$$c = 0.5(c_1 + c_2) = 0.5(0.5 \cdot 2v_a + 0.5 \cdot 2v_b) = 0.5(v_a + v_b)$$

This will yield:

$$\alpha_a \approx 1, c_1 \approx v_a$$

$$\alpha_b \approx 1, c_2 \approx v_b$$

In total:

$$k_a^T = \lambda$$

$$c \approx 0.5(v_a + v_b)$$

5.d.ii

Since we add more heads we expect less oscillation of c between v_a and v_b because the factor λ will get closer to the mean 1. As in 5.c.ii:

$$k_a \approx \lambda \mu_a$$

$$\lambda \approx N(1, 0.5)$$

$$k_b \approx \mu_b$$

Thus:

$$k_a^T q_1 = \lambda \mu_a^T \mu_a \beta = \lambda\beta \rightarrow \alpha_a \approx \frac{1}{exp(\beta - \lambda\beta) + 1}$$

$$k_a^T q_2 = \mu_b^T \mu_b \beta = \beta \rightarrow \alpha_b \approx \frac{1}{exp(-\beta + \lambda\beta) + 1}$$

$$c \approx \frac{1}{exp(-\beta + \beta) + 1} + \frac{1}{exp(\beta - \beta) + 1} = 0.5(v_a + v_b)$$

as λ tends to 1.