Understanding LLMs
University of Tuebingen
SoSe 2024
Karahan Sarıtaş - 6661689

Homework I

Exercise 1: Understanding language modeling

(a) **Q1**: Consider the corpus C with the following sentences: C = "The cat sleeps", "The mouse sings", "The cat sleeps", "A dog sings". (a) Define the vocabulary V of this corpus (assuming by-word tokenization). (b) Pick one of the four sentences in C. Formulate the probability of that sentence in the form of the chain rule. Given the corpus, calculate the probability of each term in the chain rule.
A1:
V = {"The", "cat", "sleeps", "mouse", "sings", "A", "dog"}
Freq: D = {The: 3, cat: 2, sleeps: 2, mouse: 1, sings: 2, A: 1, dog: 1}

Let's pick the sentence "The cat sleeps".
The probability of the sentence can be formulated as:

$$P(\text{"The cat sleeps"}) = P(\text{"sleeps"} \mid \text{"The cat"}) \cdot P(\text{"cat"} \mid \text{"The"}) \cdot P(\text{"The"})$$

$$P(\text{"The"}) = \frac{1}{4}$$

$$P(\text{"cat"} \mid \text{"The"}) = \frac{2}{3}$$

$$P(\text{"sleeps"} \mid \text{"The cat"}) = \frac{2}{2} = 1$$

$$P(\text{"The cat sleeps"}) = \frac{1}{4} \cdot \frac{2}{3} \cdot \frac{2}{2} = \frac{1}{6}$$

(b) **Q2**: We want to train a neural network that takes as input two numbers x_1, x_2, passes them through three hidden linear layers, each with 13 neurons, each followed by the ReLU activation function, and outputs three numbers y_1, y_2, y_3. Write down all weight matrices of this network with their dimensions. (Example: if one weight matrix has the dimensions 3x5, write $M_1 \in R^{3\times5}$)
A2:
There will be 4 weight matrices in total with corresponding bias vectors. For simplicity, I'm assuming a batch size of 1 for the following equations. There are different formats in the literature for the ordering of the dimensions of the weight matrices. I'm complying with the one provided in the slides "02-PyTorch-ANNs-RNNs" on page 18.

$$
\begin{aligned}
x &= [x_1, x_2]^T \in R^{2\times 1} \\
W_1 &\in R^{13\times 2} \\
b_1 &\in R^{13\times 1} \\
a_1 &= ReLU(W_1 x + b_1)
\end{aligned}
$$

$$
\begin{aligned}
W_2 &\in R^{13\times 13} \\
b_2 &\in R^{13\times 1} \\
a_2 &= ReLU(W_2 a_1 + b_2)
\end{aligned}
$$

$$
\begin{aligned}
W_3 &\in R^{13\times 13} \\
b_3 &\in R^{13\times 1} \\
a_3 &= ReLU(W_3 a_2 + b_3)
\end{aligned}
$$

$$
\begin{aligned}
W_4 &\in R^{3\times 13} \\
b &\in R^{3\times 1} \\
y &= g(W_4 a_3 + b)
\end{aligned}
$$

g can be the softmax function for a classification task or identity/linear activations for regression tasks.

(c) **Q3**: Consider the sequence: "Input: Some students trained each language model". Assuming that each word+space/punctuation corresponds to one token, consider the following token probabilities of this sequence under some trained language model: $p = [0.67, 0.91, 0.83, 0.40, 0.29, 0.58, 0.75]$. Compute the average surprisal of this sequence under that language model. [Note: in this class we always assume the base e for log, unless indicated otherwise. This is also usually the case throughout NLP.]

A3:

Negative log likelihood $= -\log(P_{\text{LM}}(w_{1:n}))$ where $w_{1:n}$ is the target sentence

Average surprisal $= -\frac{1}{n}\log(P_{\text{LM}}(w_{1:n}))$

$P_{\text{LM}}(w_{1:n}) = \prod_{i=1}^{n} P_{\text{LM}}(w_i \mid w_{1:i-1})$

$$
\begin{aligned}
P_{\text{LM}}(w_{1:n}) &= -\frac{1}{7}\log(0.67 \cdot 0.91 \cdot 0.83 \cdot 0.40 \cdot 0.29 \cdot 0.58 \cdot 0.75) \\
&= -\frac{1}{7}\log(0.0255) \\
&= -\frac{1}{7} \cdot -3.67 \\
&= 0.52
\end{aligned}
$$