Exercise 3: Deep Learning

Ehsan Mousavi, Kasra Alishahi

December 26, 2023

Problem 1: Neural Chatbot

Neural Dialog Model are Sequence-to-Sequence (Seq2Seq) models that produce conversational response given the dialog history. In this assignment we will build a simple single turn conversations to make based on the data
In this assignment you will implement,

- Seq2Seq encoder-decoder model base on LSTM model
- Seq2Seq model with attention mechanism
- Greedy search decoding algorithms

In this exercise, we use Cornell Movie Dialog Corpus to train the model. However, it is optional to train the model on a larger corpus to get a better performance.

See the colab file for more details.

Problem 2: Multi-Head Attention

Let $X \in \mathbb{R}^{N\times d}$ be the input of attention layer. For example, you can assume that there are N token and each of them has a d dimension embedding. Suppose that we have H heads indexed by $h = 1, \ldots, H$ of the form

$$H_h = Attention(Q_h, K_h.V_h) = Softmax\Big[\frac{Q_h K_h^T}{\sqrt{d_h}}\Big]V_h$$

where $Q_h, K_h, V_h \in \mathbb{R}^{N\times d_h}$. Here, we have defined separate query, key, and value matrices for each head using

$$Q_h = XW_h^{(q)}, K_h = XW_h^{(k)}, V_h = XW_h^{(v)}$$

The heads are first concatenated into a single matrix, and the result is then linearly transformed using a matrix $W^{(o)}$ to give a combined output in the form

$$O = Concat[H_1, \ldots, H_H]W^{(o)}$$

i) Consider matrices $W_h^{(q)}$ and $W_h^{(k)}$, where all elements are independent and identically distributed (i.i.d.) random variables with a mean of 0 and a variance of σ^2. Additionally, let x_i and x_j in $\mathbb{R}^{1\times d}$ represent the embeddings of tokens i and j, satisfying $\|x_i\|^2 = \|x_j\|^2 = 1$.

   Define $q_i = x_i W_h^{(q)}$ and $k_j = x_j W_h^{(k)}$ as the query for token i and the key for token j. Introduce the similarity measure between query i and key j as

$$\alpha_{i.j} = \frac{q_i k_j^T}{\sqrt{d_h}}$$

   where d_h is the dimensionality of the hidden space.

   Now, compute the expected value $E[\alpha_{i.j}]$ and the variance $\Sigma_{i,j} = \mathrm{var}[\alpha_{i.j}]$. Finally, elucidate the rationale behind incorporating the scaling factor $\sqrt{d_h}$ by using central limit theorem.

ii) Let $d_h = d/H$ for all $h = 1, \ldots, H$. How many multiplication is required to compute the output O? You can assume that running time of computing e^x is $O(1)$.

   Hint: The time complexity of matrix multiplication for matrices A of size $m \times n$ and B of size $n \times p$ is given by $O(m \cdot n \cdot p)$. The resulting product matrix $C = A \cdot B$ will have dimensions $m \times p$.

iii) Show that there exists matrices $\tilde{H}_h \in \mathbb{R}^{N\times N}$ and $W^{(h)} \in \mathbb{R}^{d\times d}$ such that

$$O = \sum_{h=1}^{H} \tilde{H}_h X W^{(h)}$$

   Argue that these two representations are not equivalent.
   Hint: What is rank$(W^{(h)})$?

Problem 3: Coming Soon!!!