ELL881/AIL821

Large Language Models: Introduction and Recent Advances

Semester I, 2024-25

Quiz 1

Answer the questions in the spaces provided. No extra pages will be given. Write proper justifications for every answer.

Course Code: ____________________

Name: ____________________

Entry Number: ____________________

Total Marks: 20 **Time: 40 minutes**

Questions	**Marks**	**Score**
Variational Inference in Language Modeling	4	
Backpropagation Through Time	4	
Combining Word Embeddings Using Autoencoders	6	
A Sneak Peek into Transformers	6	
Total	**20**	

Question 1: Variational Inference in Language Modeling

Given two probability distributions $P(Y)$ and $Q(Y)$, the *Kullback-Leibler divergence* (popularly called *KL divergence*), between these two distributions can be expressed as:

$$D_{\mathrm{KL}}(P(Y) \,\|\, Q(Y)) = \mathbb{E}_{y \sim P} \log \left(\frac{P(Y = y)}{Q(Y = y)} \right).$$

Now, consider a language model where:

- $P(\text{output} \mid \text{context})$ represents the true posterior probability distribution for generating a specific output (e.g., a word or sequence of words) given an input context (e.g., a preceding sentence or phrase).
- $P(\text{output})$ is the prior probability distribution of the output before observing any context, which reflects the likelihood of generating the output without any conditioning information.
- $P(\text{context} \mid \text{output})$ is the likelihood of observing a given context given the output, which could be interpreted as the probability of the input context being consistent with the output.
- $Q(\text{output} \mid \text{context})$ is an approximate distribution used to estimate the true posterior distribution $P(\text{output} \mid \text{context})$.

1. Show that the *KL divergence* between the approximate distribution $Q(\text{output} \mid \text{context})$ and the true posterior $P(\text{output} \mid \text{context})$ can be expressed as:

$$\begin{aligned} D_{\text{KL}}(Q(\text{output} \mid \text{context}) \parallel P(\text{output} \mid \text{context})) = &\log P(\text{context}) \\ &- \mathbb{E}_{Q(\text{output}\mid\text{context})}\left[\log P(\text{context} \mid \text{output})\right] \\ &+ D_{\text{KL}}(Q(\text{output} \mid \text{context}) \parallel P(\text{output})). \end{aligned}$$

(4 marks)

Question 2: Backpropagation Through Time

In the process of backpropagation through time (BPTT) for training a Recurrent Neural Network (RNN), we compute various derivatives to update the network's parameters. For the scenarios given below, determine the dimensions of the derivatives.

1. **Derivative of the Loss with Respect to the Output Matrix**
   Assume that the output matrix $\mathbf{O}$ has dimensions $n \times m$. What are the dimensions of $\frac{\partial L}{\partial \mathbf{O}}$?

(1 mark)

2. **Derivative of the Output Matrix with Respect to the Hidden State Matrix**
   Assume that the output matrix $\mathbf{O}$ has dimensions $n \times m$ and the hidden state matrix $\mathbf{H}$ also has dimensions $n \times m$. What are the dimensions of $\frac{\partial \mathbf{O}}{\partial \mathbf{H}}$?

(1 mark)

3. **Derivative of the Hidden State Matrix with Respect to the Input Vector**
   The hidden state matrix $\mathbf{H}$ has dimensions $n \times m$ and the input vector $\mathbf{X}$ has p elements. What are the dimensions of $\frac{\partial \mathbf{H}}{\partial \mathbf{X}}$?

(1 mark)

4. **Second-Order Derivative of the Loss with Respect to the Hidden State Matrix**
   The loss function L depends on the hidden state matrix $\mathbf{H}$ with dimensions $n \times m$. What are the dimensions of the second-order derivative $\frac{\partial^2 L}{\partial \mathbf{H}^2}$?

(1 mark)

Question 3: Combining Word Embeddings Using Autoencoders

We discussed different word embedding methods in class. To combine the embeddings of two or more words, we can simply add or concatenate them - we saw this approach in CNN-based neural language models where embeddings are concatenated.
Another approach of combining word embeddings can be using an autoencoder.

In the autoencoder, two input word embeddings/vectors $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^{D_x \times 1}$ are first concatenated into a single vector $\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \in \mathbb{R}^{2D_x \times 1}$, and the parent vector $\mathbf{p}$ can be computed as:

$$\mathbf{p} = \text{ReLU}(W_1 \mathbf{x} + \mathbf{b}_1) \in \mathbb{R}^{D_p \times 1},$$

where $\text{ReLU}(x) = \max(0, x)$, and W_1 can be decomposed as:

$$W_1 = \begin{bmatrix} W_{11} & W_{12} \end{bmatrix}$$

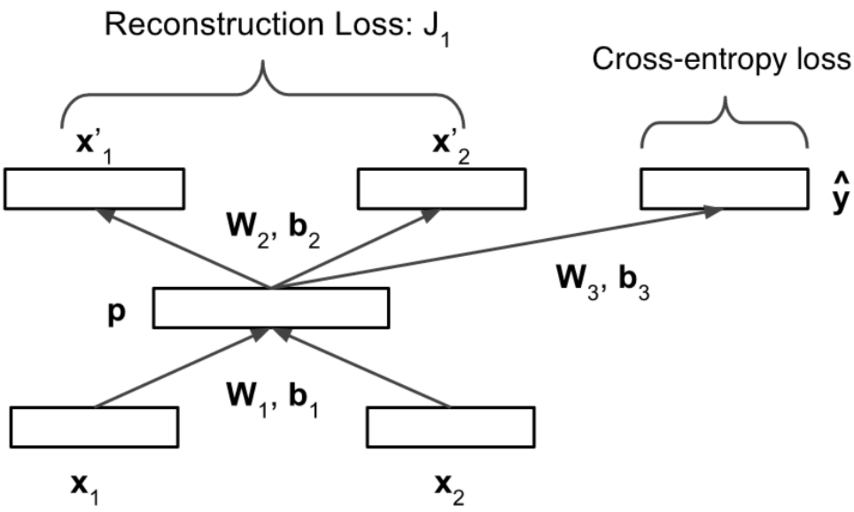

Figure 1: Using an autoencoder to combine embeddings of two words

thus $W_1\mathbf{x}$ becomes:

$$W_1\mathbf{x} = W_{11}\mathbf{x}_1 + W_{12}\mathbf{x}_2.$$

During training, we use the parent vector $\mathbf{p}$ to reconstruct the input vectors:

$$\mathbf{x}' = \begin{bmatrix}\mathbf{x}'_1 \\ \mathbf{x}'_2\end{bmatrix} = W_2\mathbf{p} + \mathbf{b}_2 \in \mathbb{R}^{2D_x \times 1}.$$

where $\mathbf{x}'_1, \mathbf{x}'_2 \in \mathbb{R}^{D_x \times 1}$ are the reconstructions. Correspondingly, a re-construction loss J_1 that computes the Euclidean distance between inputs and re-constructions is used during training:

$$J_1 = \frac{1}{2}\|\mathbf{x}' - \mathbf{x}\|^2 \in \mathbb{R}.$$

The network is trained using the total loss:

$$J = J_1 + J_2.$$

where, J_2 is a cross-entropy loss between actual label y and predicted label $\hat{y} = W_3\mathbf{p} + \mathbf{b}_3$:

$$J_2 = \mathrm{CE}(y, \hat{y}) \in \mathbb{R}$$

Now, compute the following gradients for the re-construction loss J_1.
You can use the following notation

$$\mathbb{1}\{x > 0\} = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{otherwise} \end{cases}$$

Using it on a matrix performs an element-wise operation, e.g.,

$$\mathbb{1}\left\{\begin{bmatrix} 5 & 0 \\ -3 & 7 \end{bmatrix} > 0\right\} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

1. $\delta_1 = \frac{\partial J_1}{\partial \mathbf{p}}$.

(2 marks)

2. $\delta_2 = \frac{\partial J_1}{\partial \mathbf{h}}$ (where, $\mathbf{h} = W_1\mathbf{x} + \mathbf{b}_1$) in terms of δ_1.

(2 marks)

3. $\frac{\partial J_1}{\partial W_1}$ in terms of δ_2.

(2 marks)

Question 4: A Sneak Peek into Transformers

Consider a transformer model used for natural language processing tasks. Given a sequence of input vectors $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$, where each $\mathbf{x}_i$ is in $\mathbb{R}^d$, the attention mechanism computes the attention scores using the following steps:

- **Query, Key, and Value Matrices:** The input vectors are linearly transformed into query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) matrices using weight matrices $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ respectively:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{X}\mathbf{W}_V$$

  where $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are in $\mathbb{R}^{d \times d}$.

- **Scaled Dot-Product Attention:** The attention scores are computed as *softmax* over the *scaled dot-product of the query and key matrices.* Then, the final output from each head is computed as:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}$$

Given the above, answer the following questions:

Suppose W_Q and W_K are **symmetric** and **diagonalizable** matrices with eigenvector matrices P_Q and P_K, respectively.

1. Prove that if the eigenvectors of W_Q and W_K are **aligned** (i.e., $P_Q = P_K$), then the attention scores will favor attention in the directions aligned with these common eigenvectors.

(2 marks)

2. Conversely, if every eigenvector of $\mathbf{W_Q}$ is orthogonal to the eigenvectors of $\mathbf{W_K}$, i.e.,

$$\forall i, \forall j, \mathbf{P}_Q[:i] \perp \mathbf{P}_K[:j]$$

   prove that the attention scores tend to be uniformly distributed.

(3 marks)

3. If the rank of the matrix $\mathbf{QK}^T$ is p and the rank of $\mathbf{V}$ is q, then what can you infer about the rank of the attention matrix $Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ in terms of p and q? Justify your answer.

(1 mark)