Problem Set 4

Please submit a typed PDF addressing all problems below. This problem set contains 3 questions and is worth 25 points. **All responses MUST be in *your* own words.** Justification must be provided for **all** written answers. Statements made without any supporting explanation/justification will receive **no credit**. For mathematical derivations and plots, you may insert pictures of handwritten work if you find this easier. The required weekly readings and lecture slides should be helpful in completing the assignment. You can find these on our course website.

1. **Transformers vs. RNNs and MLPs [6 points]:** For each of the following, provide *two distinct examples* to support the requested architecture comparison.

   (a) How are transformers similar to RNNs?

   (b) How are transformers different from RNNs?

   (c) How are transformers similar to MLPs?

   (d) How are transformers different from MLPs?

2. **Transformer Architectures [8 points]:** Recall the self attention equation:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

   where Q, K, V are the query, key, and value matrices.

   (a) Explain in your words the importance of each of the following:

      i. The denominator term: $\sqrt{d_k}$

      ii. The use of the softmax function.

   (b) The softmax term in equation (1), $softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)$, is often referred to as the *attention map* or *attention matrix*. Why is this term referred to in this way?

   (c) What is the motivation for using multi-headed attention layers vs. single-headed attention layers?

   (d) Describe the difference between transformer encoder and transformer decoder architectures. Provide an example real world use case for each.

3. **Computing Attention Maps [11 points]:**

(a) Compute the attention map that results when using self-attention with the input tokens and weights provided below. Use the dot product to calculate the query-key similarity as outlined in equation (1). For simplicity, you do **NOT** need to scale the dot product result by $\sqrt{d_k}$. Mathematical steps must be shown for full credit. Round all steps to 2 decimal points.

(b) Using the attention map from the previous step, compute the resulting output tokens from the self-attention operation. Mathematical steps must be shown for full credit. Round all steps to 2 decimal points.

(c) Visualize the computed attention map from part (a) by creating a plot (or hand drawing) showing the connections between input and output tokens. Plot the tokens as dots organized vertically, with one vertical line for the inputs and another vertical line parallel to the first for the outputs. For each output token, plot a solid line connecting it to the input token which influenced it the most (i.e., largest attention map value for that row). When applicable, use dashed lines to connect output tokens to their corresponding second most influential input tokens.

(d) Discuss the patterns you observe in the attention map visualization.

1	-1	0	2
0	-1	2	1
-2	2	-1	0

Input Tokens

0	1
1	0
-0.5	0
1	1

Query Weights

1	-0.5
0	1
0.5	-0.5
1	0.5

Key Weights

1	0.5
-0.5	0
-1	2
0.5	-1

Value Weights

4. **Extra Credit [2.5 Points]:** Recall the cross attention equation for input token sets X_1 and X_2:

$$\begin{aligned} Q &= X_1 W_Q \\ K &= X_2 W_K \\ V &= X_2 W_V \\ CrossAttention(Q, K, V) &= softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) V \end{aligned} \tag{2}$$

   (a) Compute the cross attention map using the data provided below. For simplicity, you do **NOT** need to scale the dot product result by $\sqrt{d_k}$. Mathematical steps must be shown for full credit. Round all steps to 2 decimal points.

   (b) Using the attention map from the previous step, compute the resulting output tokens from the cross-attention operation. Mathematical steps must be shown for full credit. Round all steps to 2 decimal points.

   (c) Visualize the attention map by providing a plot in a similar style to question 3 part (c). Use X_2 as the input tokens.

   (d) Discuss the patterns you observe in the attention map visualization.

-1	1	0
0	1	0

X_1

-1	1
0	1
1	1
2	0

X_2

0	-1
1	1
-1	0

W_Q

1	0
-1	1

W_K

-1	0	1	2
1	-1	0	-2

W_V

Collaboration versus Academic Misconduct: Collaboration with other students (or AI) is permitted, but the work you submit must be your own. Copying/plagiarizing work from another student (or AI) is not permitted and is considered academic misconduct. For more information about University of Colorado Boulder's Honor Code and academic misconduct, please visit the course syllabus.