Tuesday Mar 25, 2025

In-class Handout

COSC 410A Applied Machine Learning

Prof. Forrest Davis

Name:

Discuss and complete the following questions with the person nearest you. You **may** be asked to share your thoughts with the class.

1. Calculate the output of attention given the following values.

- Embeddings are $\begin{bmatrix} 0 & 2 & 2 \\ 1 & 1 & 4 \\ 3 & 1 & 0 \\ 1 & 0 & 2 \end{bmatrix}$
- WQ is $\begin{bmatrix} 2 & 1 \\ 2 & 2 \\ 1 & 1 \end{bmatrix}$
- WV is $\begin{bmatrix} 2 & 2 \\ 2 & 1 \\ 2 & 0 \end{bmatrix}$
- WK is $\begin{bmatrix} 2 & 0 \\ 0 & 2 \\ 2 & 0 \end{bmatrix}$
- Q is $\begin{bmatrix} 6 & 6 \\ 8 & 7 \\ 8 & 5 \\ 4 & 3 \end{bmatrix}$
- V is $\begin{bmatrix} 8 & 2 \\ 12 & 3 \\ 8 & 7 \\ 6 & 2 \end{bmatrix}$
- K is $\begin{bmatrix} 4 & 4 \\ 10 & 2 \\ 6 & 2 \\ 6 & 0 \end{bmatrix}$
- Output is $\begin{bmatrix} 12 & 3 \\ 12 & 3 \\ 12 & 3 \\ 0 & 0 \end{bmatrix}$

2. Would reordering the embeddings in the sequence change the output? If not, how could you modify this mechanism to add back in sequential order?

3. When is attention slow? That is, what drives the runtime? Consider, the sequence length, dimensionality of the embeddings, etc.

Attention is quadratic in sequence length, so it doesn't scale well with longer context sizes.

4. Consider the case of language modeling, where you only condition predictions on prior words. How would you modify the attention mechanism to work in this domain.