Attention and Transformers

COSC 410: Applied Machine Learning
Fall 2025
Prof. Forrest Davis

November 6, 2025

Warm-up

1. Discuss with your neighbor the course you are most excited about taking next semester
2. For a recurrent neural network with one hidden layer processing a time sequence of t steps, what is the receptive field for each hidden representation as a function of t?

Logistics

- Codelet 5 is due Nov 14
- Midterm Exam II is now Nov. 19
- Schedule a feedback meeting with me (include all your group members) here

Learning Objectives

- Describe the relationship between databases and attention
- Calculate attention weights and the output of applying attention
- Describe the matrix representation of attention
- Describe one application of transformer LLMs, in-context learning

Summary: We tackle the last neural network architecture we will consider in this class, the transformer. In particular, we dive into the attention mechanism which is the main advance underlying the transformer architecture. Along the way, you demonstrate you can calculate attention for a small sample, can identify some interesting limitations of the approach, and can engage with one active area in the deployment of transformer language models.

Databases as a Way to Motivate Attention

DATABASES ARE ONE WAY TO CONCEPTUALIZE what the attention mechanism in a transformer does. We go through that abstraction together on the slides and highlight a key limitation in RNNs that motivates the development of attention.

Attention Mechanism: Details and Complications

THE CONNECTION BETWEEN ATTENTION AND DATABASE QUERY motivates two key desiderata:

1. We want a way to compute the similarity between a *key* (the database) and the *query* (the input)
2. We want a way to generate an output (how do we weigh a set of similar items in our database)

First comes first, a natural way of representing things in a neural network is vectors.

Practice Problems

1. Compute the similarity between a query vector: $\begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}$ and the key vectors $\begin{bmatrix} -1 \\ 0 \\ 0.5 \end{bmatrix}$ and $\begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}$
2. What function could you use to normalize these similarities so you had a sense of the proportion of total similarity each key represented? We might want to know that one key is 30% more similar than another, for example.
3. Give the normalized output of the similarity between the query and the two keys in 1.

Second comes second, given attention weights (e.g., a normalized measure of how similar a query is to a set of keys), we need to generate a response (i.e., retrieve something from our database). Each key is associated with a value (e.g., a last name as a key might be associated with a first name as a value).

Practice Problems

1. Say you want to return to the user a representation from your database that mixes the values in your database in proportion to the attention weight associated with them. Generate a representation given the attention weights $\begin{bmatrix} 0.8 \\ 0.2 \end{bmatrix}$ and the values $\begin{bmatrix} 4 \\ 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} -3 \\ 2 \\ -6 \end{bmatrix}$

Where does these queries, keys, and values come from? They come from weight matrices (they are parameters that are learned during training). If my input had 5 features and I wanted a query with 3 features, W_Q would be 5×3.

Practice Problems

Calculate the output of attention for the inputs (with 2 samples and 5 features) below assuming

- $W_K = \begin{bmatrix} 1 & 2 & 2 \\ 2 & 1 & 2 \\ 1 & 2 & 2 \\ 2 & 2 & 2 \\ 2 & 2 & 2 \end{bmatrix}$
- $W_Q = \begin{bmatrix} 1 & 2 & 1 \\ 1 & 1 & 2 \\ 1 & 1 & 1 \\ 2 & 1 & 1 \\ 1 & 1 & 2 \end{bmatrix}$
- $W_V = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 2 & 1 \\ 2 & 1 & 1 \\ 1 & 1 & 1 \\ 2 & 2 & 2 \end{bmatrix}$

1. The input is $\begin{bmatrix} 1 & 2 & 1 & 2 & 1 \\ 2 & 2 & 1 & 2 & 2 \end{bmatrix}$
2. The input is $\begin{bmatrix} 2 & 2 & 1 & 2 & 2 \\ 1 & 2 & 1 & 2 & 1 \end{bmatrix}$

Practice Problems

Give an attention matrix for an input that has 6 samples (time steps) and every feature attends fully to the time period to it's right.

Before Next Class

- Complete final project proposal
- Work on Codelet 5