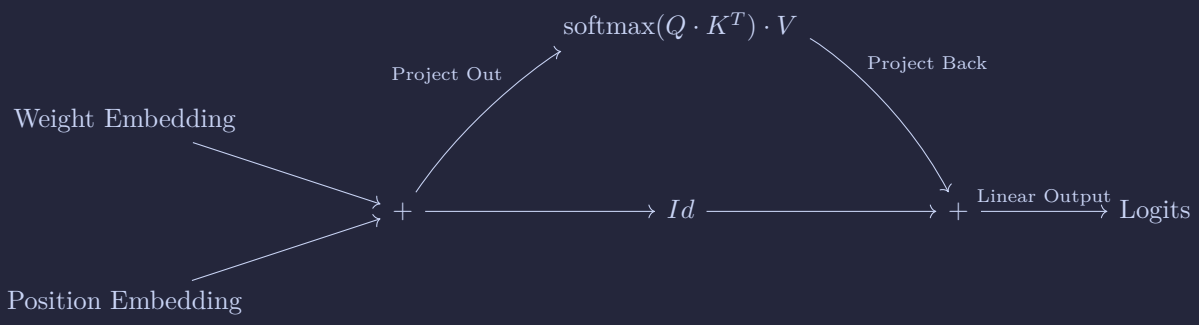
Walking Through a Tiny Transformer

Zach Virgilio

November 5, 2025

The data was a synthetic dataset consisting of the pattern $'ABC'$ repeating over and over. The model architecture is visualized in the following diagram.



The embedding weights are:

$$A \sim 0 \to \begin{bmatrix} 1.1988 \\ 1.3633 \end{bmatrix}, \quad B \sim 1 \to \begin{bmatrix} -1.8032 \\ -0.7401 \end{bmatrix}, \quad C \sim 2 \to \begin{bmatrix} -1.0425 \\ 1.4757 \end{bmatrix}$$

Where A, B and C are encoded numerically as 0, 1 and 2 respectively. The context size is 2 pieces of information, so the position weights are:

$$\begin{bmatrix} -0.3160 \\ 0.9374 \end{bmatrix}$$

So the first piece of context is has a negative weight, while the second piece of context (the value directly preceeding the token we wish to generate) has a positive weight. This makes sense since ultimately, this model only needs to learn 3 facts:

- If I see 'A' as the last term, generate 'B'
- If I see 'B' as the last term, generate 'C'
- If I see 'C' as the last term, generate 'A'

We will walk through the linear algebra of the simple one layer transformer to see how the model encoded these facts.