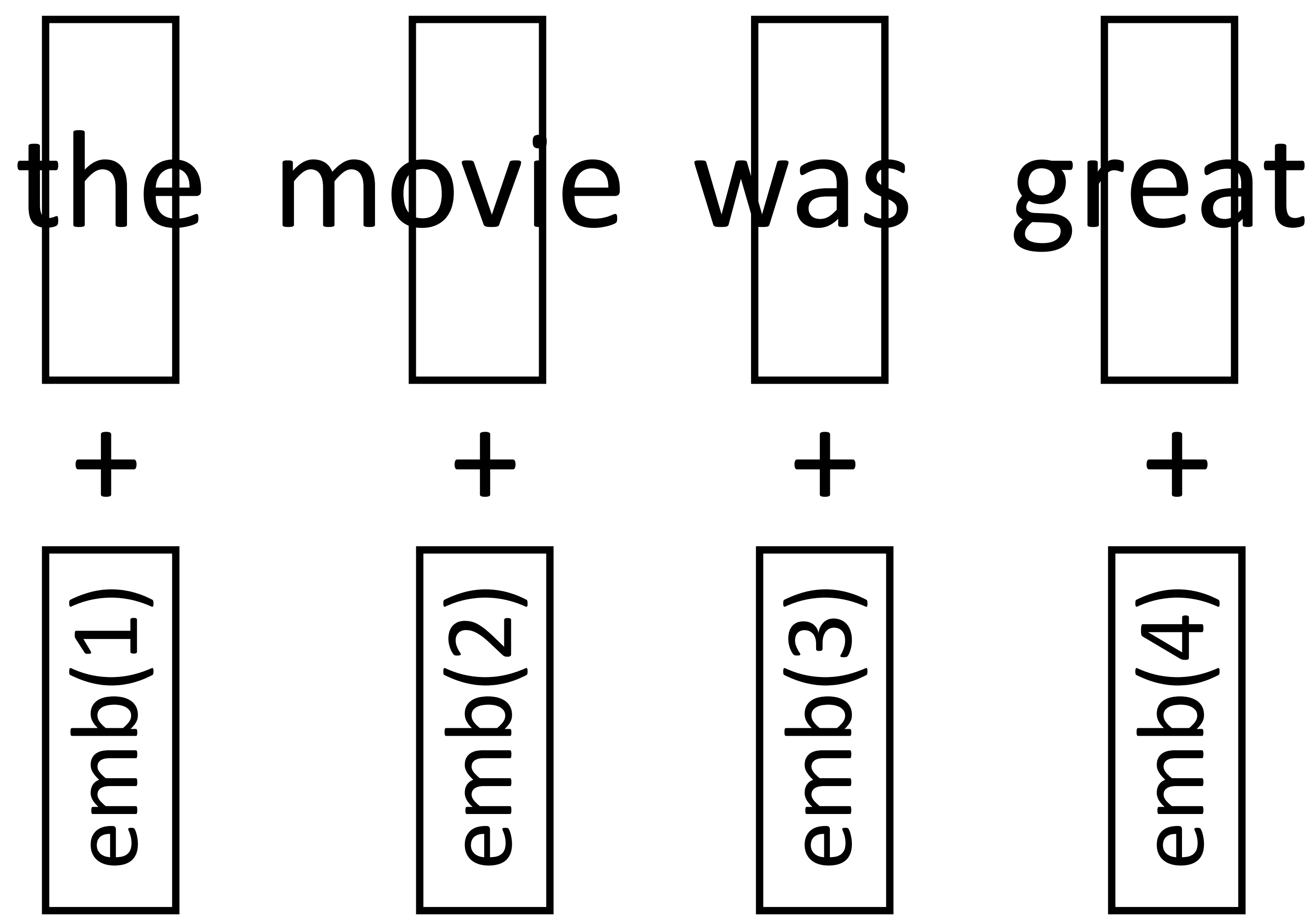
Positional Encoding

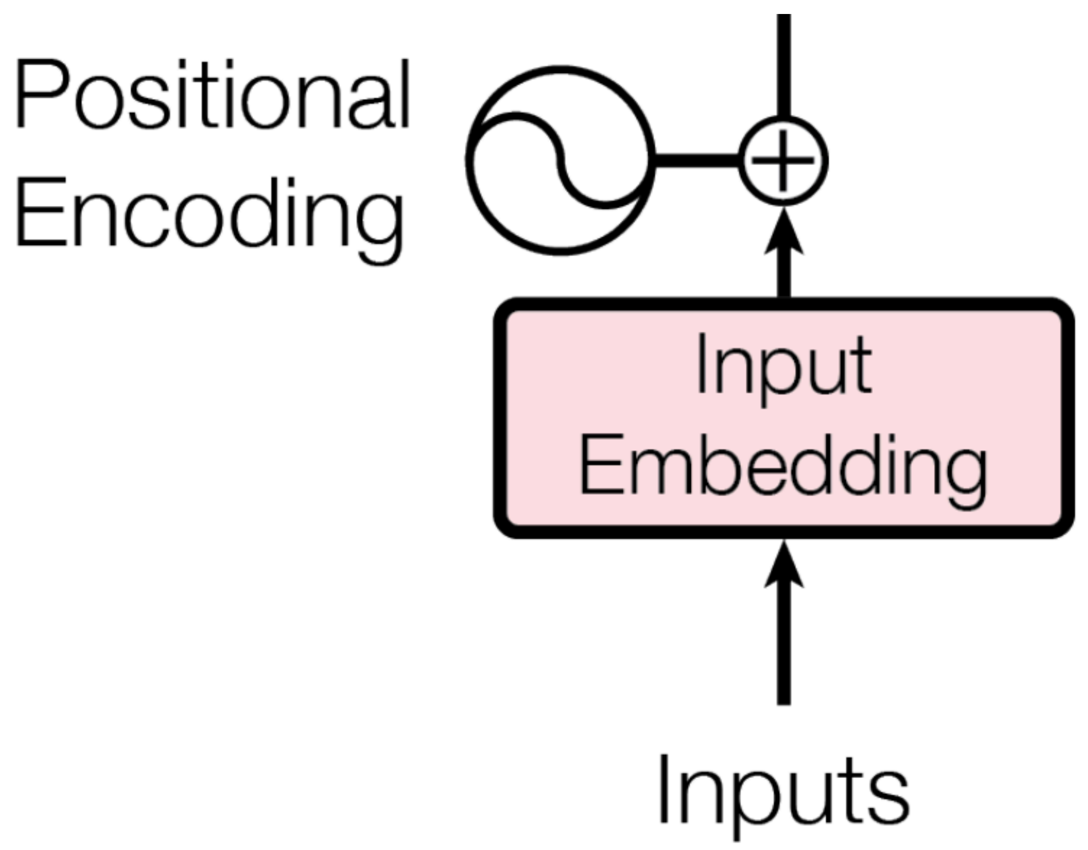
I visited New York and ate ______

- *visited* and *ate* actually look the same from the perspective of self-attention, unless we provide position information

- *Positional encoding* refers to a family of schemes to provide this information to Transformer models

Positional Encoding

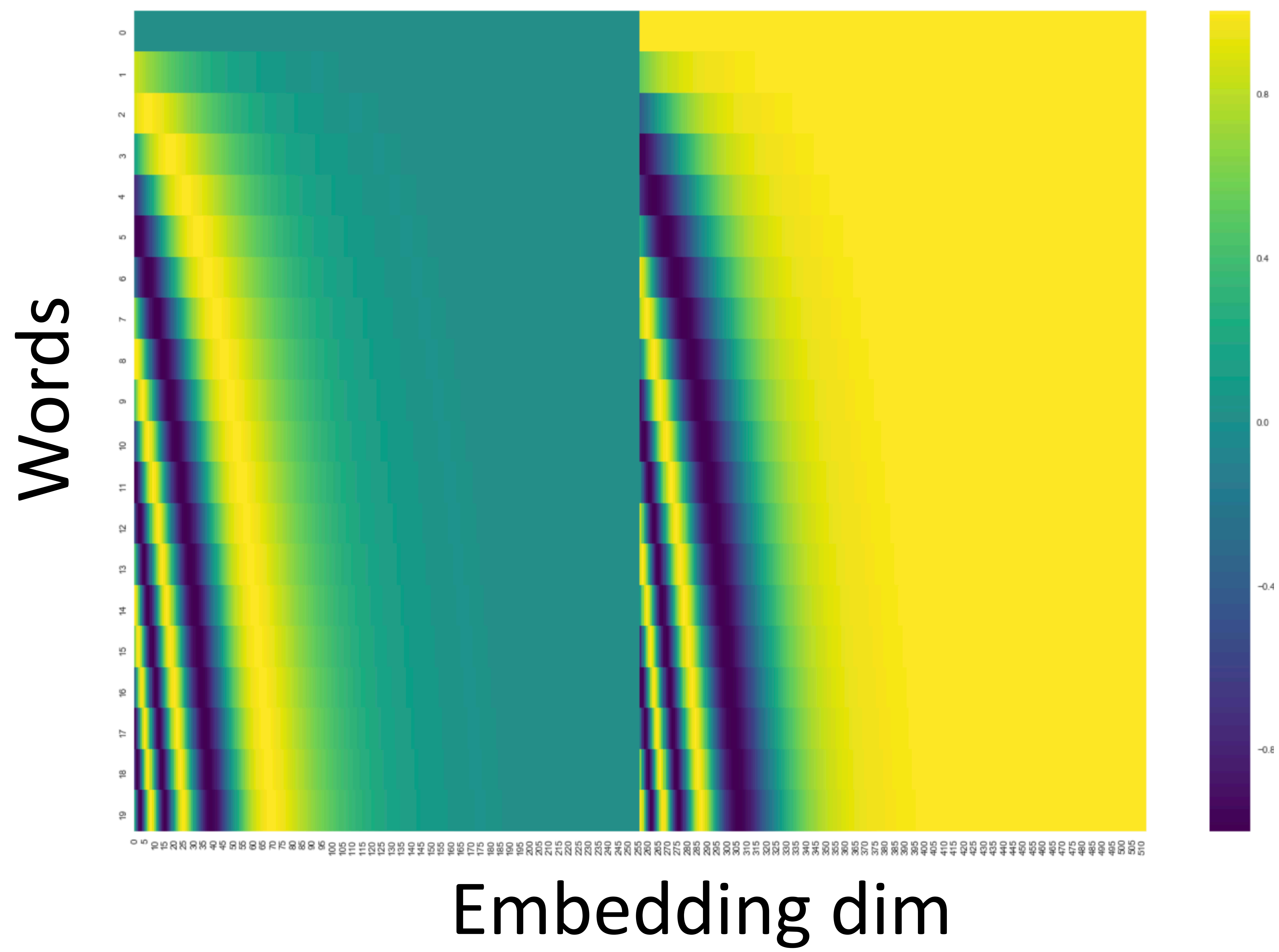




- Encode each sequence position as an integer, add it to the word embedding vector
- What are some drawbacks of this?

Multi-head Self-Attention

- Alternative from Vaswani et al.: sines/cosines of different frequencies (closer words get higher dot products by default)



$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

Positional Encoding: Variants

- Relative positional encoding (used in T5): self-attention computation depends on the distance between two tokens

- ALiBi (Press et al., 2022): $\text{softmax}(\mathbf{q}_i\mathbf{K}^\top + m \cdot [-(i-1), ..., -2, -1, 0])$
  - Adds a linear bias to disprefer attending farther back, use different slope *m* for each attention head

- No positional encoding (NoPE) actually works too! (Kazemnejad et al., 2023)
  - Transformers with self-attention into the past can learn to “count” tokens; each token can determine what position it is after a single layer.