8.5

Understanding Self-Attention

Part 1: A Basic Attention Mechanism

Sebastian Raschka and the Lightning AI Team

Transformers are solely based on self-attention (no recurrence required)

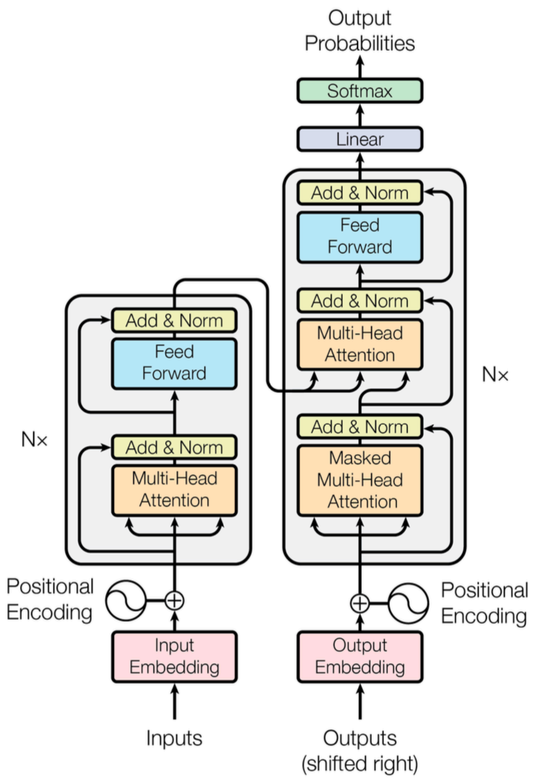


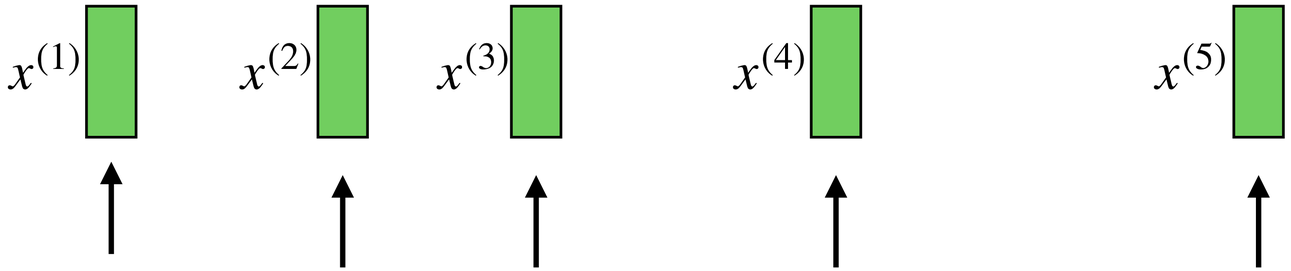
Figure 1: The Transformer - model architecture.

Attention is used to create context-aware embedding vectors

Context-aware means we model the dependency of the current input to all other inputs

Context-aware embedding vector at step (i)

$$z^{(i)} = \sum_{j=1}^{T} \alpha_{ij} \cdot x^{(j)}$$

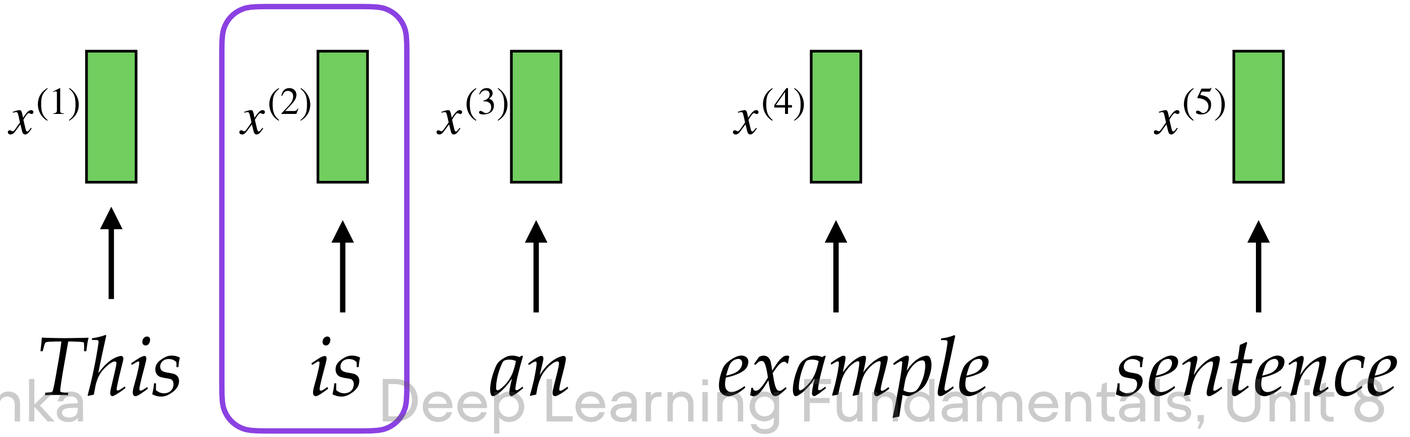


This *is* *an* *example* *sentence* $T = 5$

Context-aware embedding vector at step (i)

$$z^{(i)} = \sum_{j=1}^{T} \alpha_{ij} \cdot x^{(j)}$$

Suppose $i = 2$ (We create the context-vector for the second input)

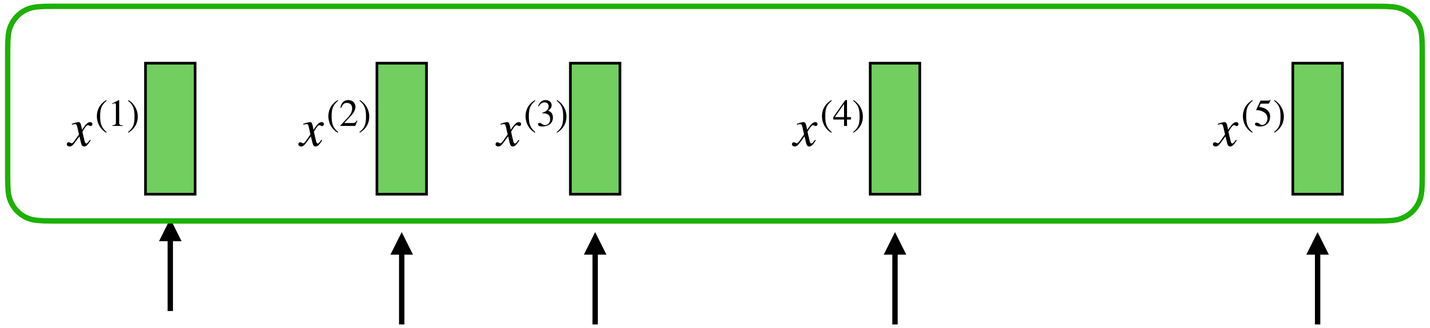


$T = 5$

Context-aware embedding vector at step (i)

$$z^{(i)} = \sum_{j=1}^{T} \alpha_{ij} \cdot \boxed{x^{(j)}}$$

There is an interaction with
all other inputs
(weighted by $\alpha's$ for the i-th context)



This *is* *an* *example* *sentence*

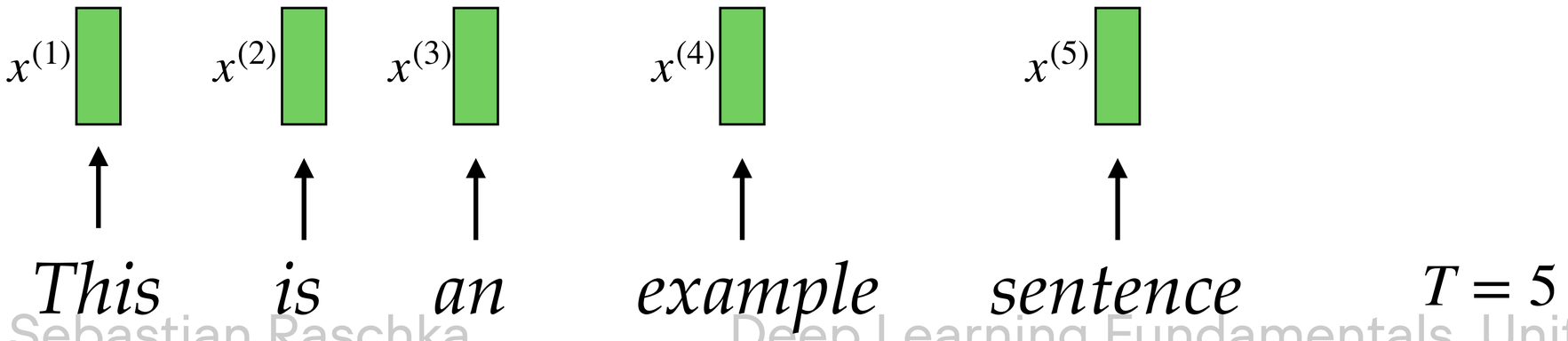
$T = 5$

Let’s start with a basic form of self-attention

$$z^{(i)} = \sum_{j=1}^{T} \boxed{\alpha_{ij}} \cdot x^{(j)}$$

1. Similarity between i-th element all inputs $j = 1 \ldots T$

$$\omega_{ij} = x^{(i)\top} \cdot x^{(j)}$$



$T = 5$

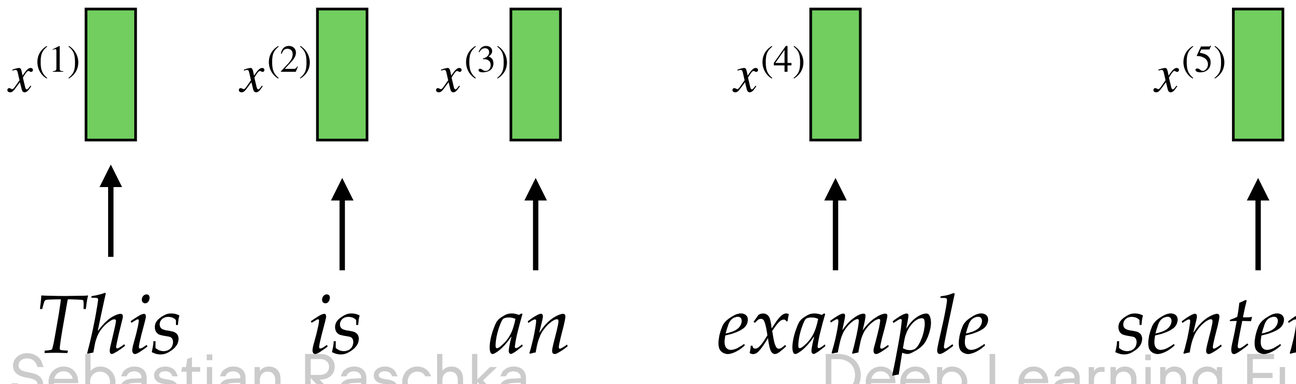
$$z^{(i)} = \sum_{j=1}^{T} \boxed{\alpha_{ij}} \cdot x^{(j)}$$

1. Similarity between i-th element all inputs $j = 1 \ldots T$

$$\omega_{ij} = x^{(i)\top} \cdot x^{(j)}$$

2. Normalize ω values to obtain attention scores α

$$\alpha_{ij} = \frac{\exp\left(\omega_{ij}\right)}{\sum_{j=1}^{T} \exp\left(\omega_{ij}\right)} = \text{softmax}\left(\left[\omega_{ij}\right]_{j=1\ldots T}\right)$$



$T = 5$

$$z^{(i)} = \sum_{j=1}^{T} \boxed{\alpha_{ij}} \cdot x^{(j)}$$

1. Similarity between i-th element all inputs $j = 1 \ldots T$

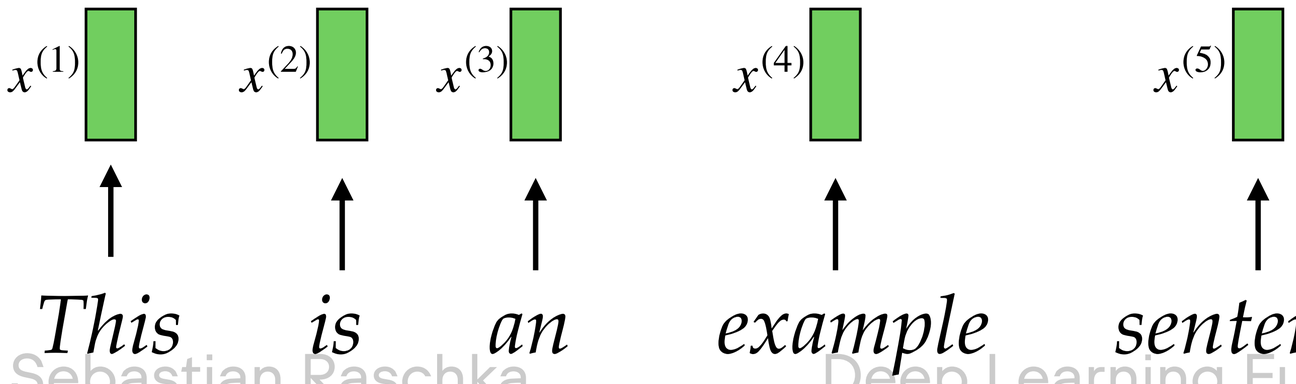
$$\omega_{ij} = x^{(i)\top} \cdot x^{(j)}$$

2. Normalize ω values to obtain attention scores α

$$\alpha_{ij} = \frac{\exp\left(\omega_{ij}\right)}{\sum_{j=1}^{T} \exp\left(\omega_{ij}\right)} = \text{softmax}\left(\left[\omega_{ij}\right]_{j=1\ldots T}\right)$$

So that attention value sum to 1

$$\sum_{j=1}^{T} \alpha_{ij} = 1$$



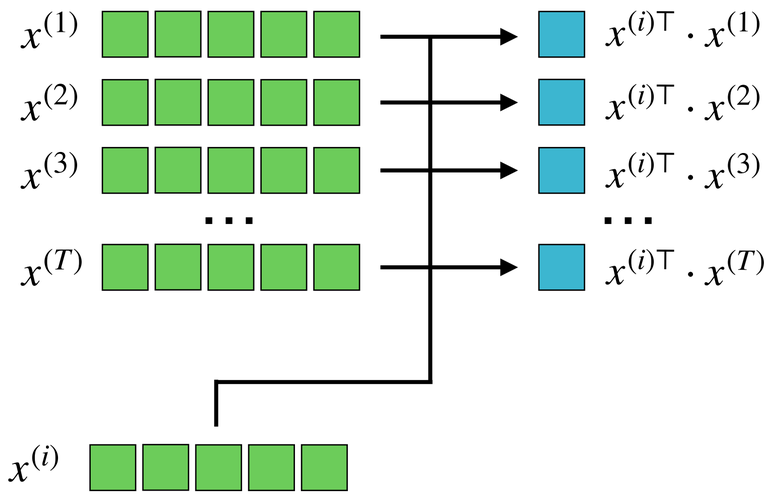
$T = 5$

Let’s summarize this with an illustration

1. Similarity between i-th element all inputs $j = 1 \ldots T$

$$\omega_{ij} = x^{(i)\top} \cdot x^{(j)}$$

Input sequence:

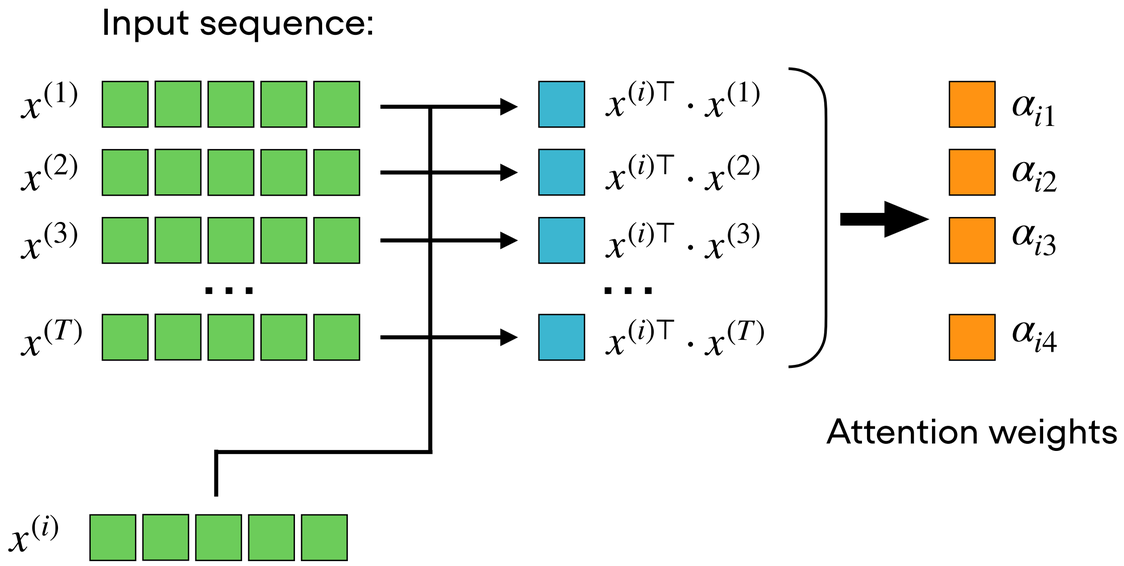


1. Similarity between i-th element all inputs $j = 1 \ldots T$

$$\omega_{ij} = x^{(i)\top} \cdot x^{(j)}$$

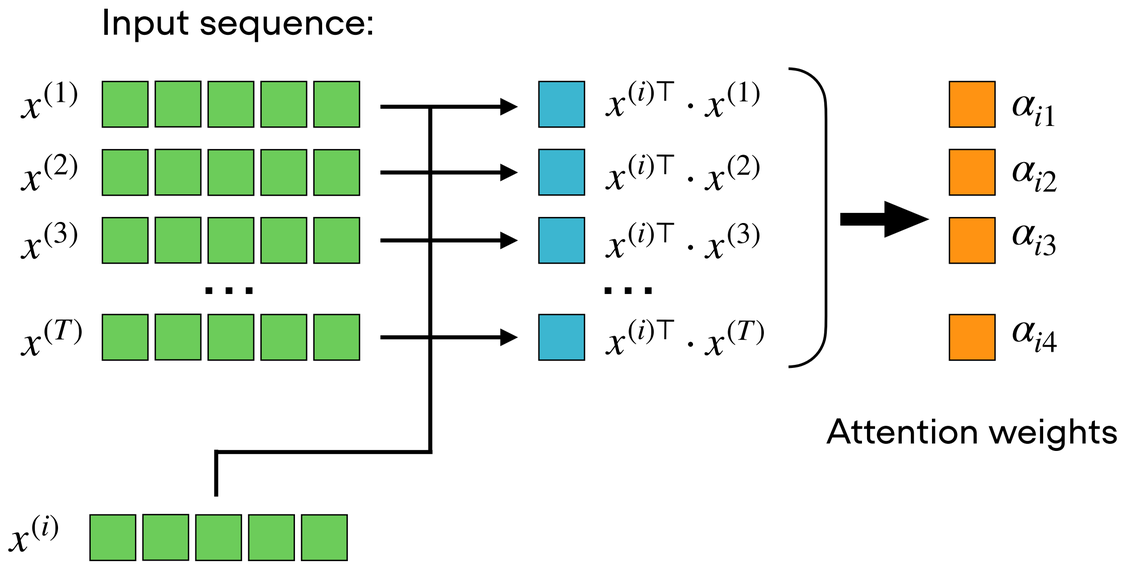
2. Normalize ω values to obtain attention scores α

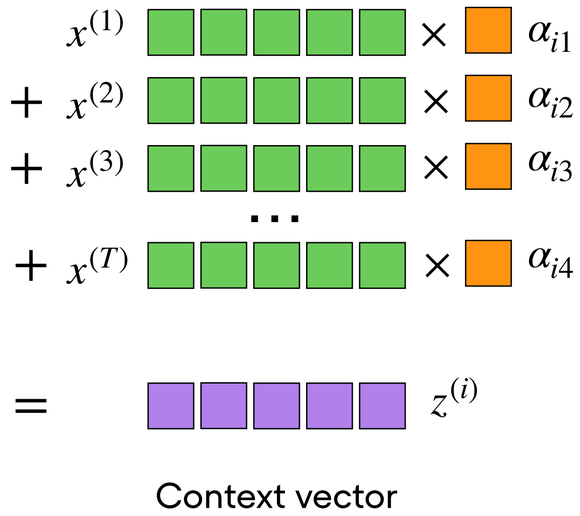
$$\alpha_{ij} = \frac{\exp\left(\omega_{ij}\right)}{\sum_{j=1}^{T} \exp\left(\omega_{ij}\right)} = \text{softmax}\left(\left[\omega_{ij}\right]_{j=1\ldots T}\right)$$



3. Compute context vector $z^{(i)}$

$$z^{(i)} = \sum_{j=1}^{T} \alpha_{ij} \cdot x^{(j)}$$





Next: Self-attention with learnable weights