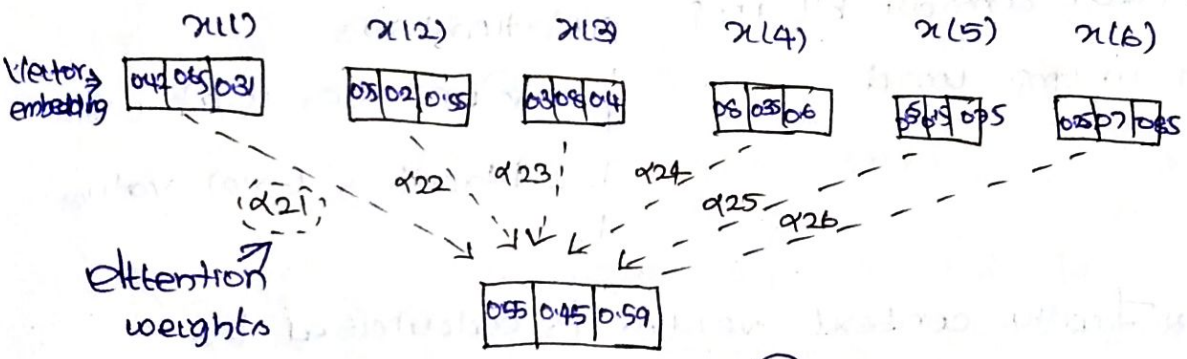
Lec-3

Introduction to Self Attention

* Self attention generalizes Bahdanau attention
* The goal is to create context vector for every word in the sequence by taking into consideration of other words in the sequence.
* Each word should know how much importance it has to attribute itself and other words.

Calculation of context vector for each words in the sequence:



z(2) ⟶ Context vector as attention weighted sum for x(2)

Attention score: (a metric to find similarity mathematically)

* It is computed using dot product of two vectors.

$$\omega = \vec{v}_1 \cdot \vec{v}_2$$

Interpretation,

- High dot product → high relevance
- Near zero → weak relevance.
- Negative → opposite direction.

* Query (Q) = Current word being processed.

* Key (K) = All words in the sequence.

* Attention score is calculated between query and every single input vector (key) by computing dot product bit it.

* Attention weights = Normalized form of attention score where the sum is 1.

* We can normalize by 2 methods,

Simple Normalization

$$\alpha_i = \frac{\omega_i}{\sum \omega}$$

* Fails with negative values
* Weak contrast bit imp and unimp. word

Softmax (Preferred)

$$\alpha_i = \frac{e^{\omega_i}}{\sum e^{\omega}}$$

* Emphasizes dominant relationships.
* Supresses weak ones.
* Handles (-ve) values.

* Finally context vector is calculated by the weight sum of attention weights and input embedding.

* Consider we have 6 words in the i/p sequence, then the matrix representation can be,

→ I/p embedding = 6x3
→ Attention = 6x6
→ O/p context vector = 6x3

Attention weights 6x6 × I/p embeddings 6x3 ⇓ Context Vector 6x3

—x—