Self-Attention w/ trainable weights

Also, Scaled dot-product attention

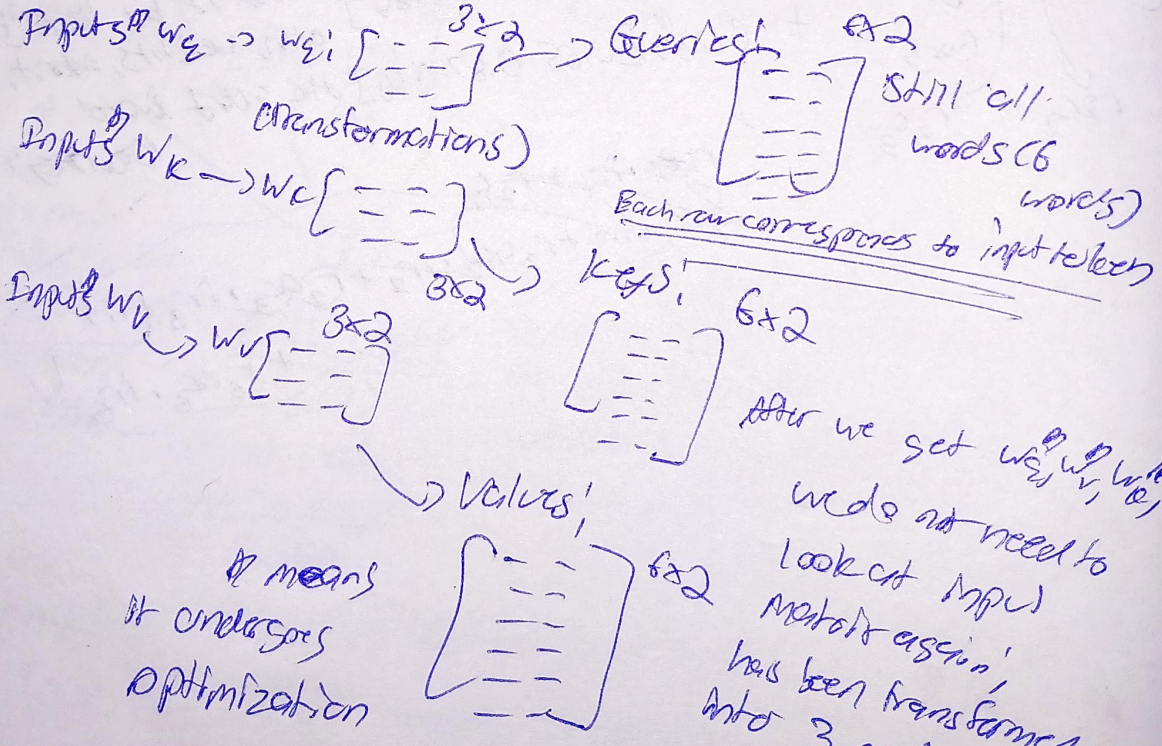
① compute context vectors as weighted sums over input vectors specific to a certain input element

② Introduce weight matrices that are updated during model training

③ Crucial that model can learn to produce "good" context vectors

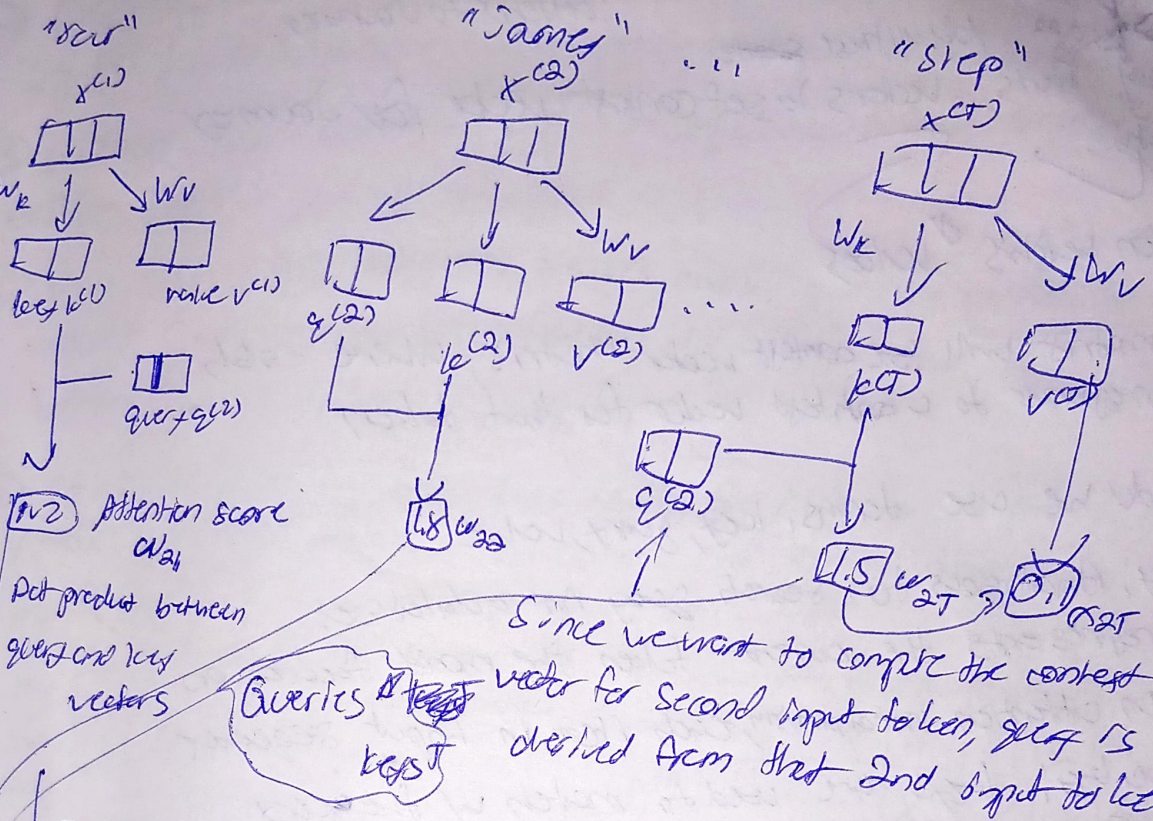
④ W_q, W_k and W_v — query, key, value

↳ these 3 matrices are used to project embedded input tokens $x^{(i)}$ into query, key and value vectors



(5) Compute only one context vector $z^{(2)}$ for illustration purposes

↳ computing attention scores:



(1.2) Attention score ω_{24}

Dot product between query and key vectors

Since we want to compute the context vector for second input token, query is derived from that 2nd input token

↳ Then we compute Attention weights

(3) Attention weights; (scale by $\sqrt{d\_keys}$) (2) in this case

Scale attention scores by $\sqrt{d\_keys}$ before softmax

Attn_score → Scale by $\sqrt{d\_keys}$ → softmax → Attention weights

(4) Context vector

↓

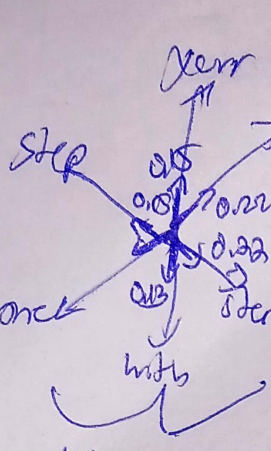
context vector

Last step is multiplying each value vector w/ its respective attention weight and then summing them to obtain the context vector

Multiply value vector w/ attention weight, which is then sent to create context vector

0.3 | 0.8

context vector $z^{(2)}$



6×6 X 6×2 → 6×2

enriched vector, contains semantic meaning of Journey and through other words attend to Journey.

add all these vectors to get context vector for Journey

Attention weights * values

6×2 matrix will be context vector matrix, where each row corresponds to a context vector for that token

* Why do we use terms: key, query, value

Query: Analogous to search query in a database. It represents the current token the model focuses on

key: In attention mechanism, each item in input sequence has a key. keys are used to match w/ the query

value: It represents the actual content or representation of the input items. Once model determines which keys (which parts of the input) are most relevant to the query (current focus item), it retrieves the corresponding values.