24/10/25

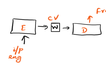
NLP - DSM 10 (Revision class)

by Dr Rajan Dabunda, IIT Delhi + Harvard Alumni

Transformer

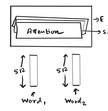
Seq2seq

application → language translation (English → French)



Self attention, multi-head attention

Figure 1: The Transformer - model architecture.

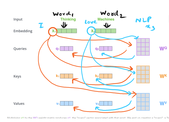


Step-1

$x_1 W_Q^T \Rightarrow q_1\ q_2\ q_3$

$K_1\ K_2\ K_3$

$V_1\ V_2\ V_3$



Attention is a way that maps a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is determined by the compatibility of the query with the corresponding key.

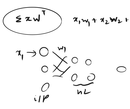
W^Q, W^K, W^V

adjustment weight matrices

↓

initialized randomly

$\sum x_i w^1 \quad x_1 w_1 + x_2 w_2 + \dots$



Step-2

$q \cdot K$

$K_1 q_1 = \checkmark$

$q_1\ K_2 = \checkmark$ } word1

$q_2\ K_1 = \checkmark$

$q_2\ K_2 = \checkmark$

Attention is a way that maps a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is determined by the compatibility of the query with the corresponding key.

$z = \frac{e^{x_i}}{\sum e^{x_i}} \quad \begin{bmatrix} 0.6 \\ 0.1 \\ 0.2 \\ 0.1 \end{bmatrix} \quad z[i] = \frac{e^{0.6}}{e^{0.6}+e^{0.1}+e^{0.2}+e^{0.1}}$

$\begin{bmatrix} 0.9 \\ 0.1 \end{bmatrix} \rightarrow$

$\frac{e^{x_1}}{\sum e^{x_i}}$

$\frac{e^{1.1}}{e^{1.1}+e^{0.1}}$

$q_1 k_1$

$q_1 k_2$

$z = f(\sum q_i x^T + b)$

Step-3,4

The softmax score determines how much each word will be expressed at this position. Clearly the word at this position will rate the highest softmax score, but sometimes it's useful to attend to another word that is relevant to the word.

W^Q

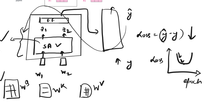
W^K

512 → embedding vector size

64 → dimension

hyper-parameter

softmax(x)



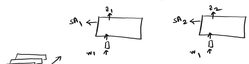
Pre-trained → customized.

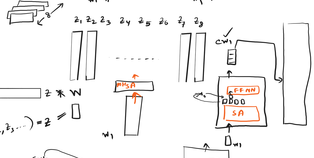
LLM

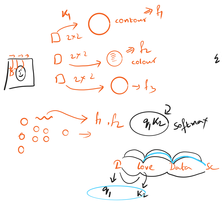
GPT

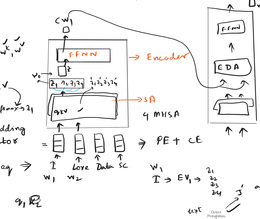
Enhance foundation

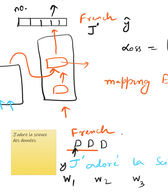
$$\text{softmax}\left(\frac{Q \times K^T}{\sqrt{d_k}}\right) \times V = Z$$











$q_1 k_1 \quad q_1 k_2$

Resnet

residual

Norm → Batch Normalization

or Layer ↓

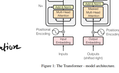
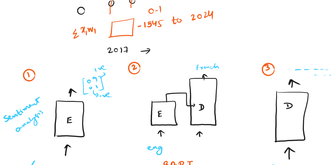


Figure 1: The Transformer - model architecture.

70 Eng

93 French



Applications

1. Name Entity Recognition
2. Text Summarization
3. Q&A
4. Seq2Seq Task
5. language generation
6. multi modal classification