Lec-5

Causal Attention Mechanism

(aka Masked Self Attention)

* Causal attention is a special form of self attention where,

→ each token is allowed to attend to itself and past tokens.

→ future tokens are blocked.

* This is the exact attention mechanism used in GPT and autoregressive language models.

Why Causal Attention ?

* Large Language Models works autoregressively,

→ predicts next token

→ append it to the input

→ predict next token again.

* At time t, the model must not know tokens from t+1 onward.

* Consider the sentence "dream big and work for it,"

→ If we are computing the context vector for "big", the model should not attend to "and", "work", "for", "it".

→ because these tokens are in the future.

Solution to implement Causal Attention :-

* Replace the attention scores for future positions with $-\infty$

*While applying softmax after scaling it will become 0.

$e^{-\infty} = 0$

Dropout in Causal Attention:

*Dropout is applied to prevent over-reliance on specific token-to-token paths. It improves robustness and reduces overfitting.

*It is applied by randomly zero out some attention weights

*Scaling remaining weights by : $\frac{1}{1-p}$

Final Attention Pipeline:

1) Input Embeddings → Q, K, V
2) Compute Attention Scores → QK^T
3) Apply Causal Mask (−∞ to future)
4) Scaling by $1/\sqrt{d_k}$
5) Apply Softmax
6) Apply Dropout.
7) Multiply by V.

— x —