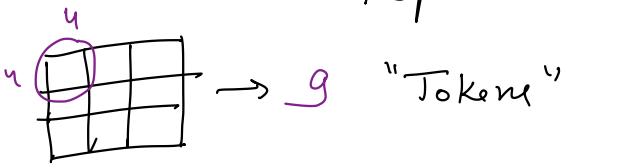
Agenda:

1. Transformer (✓)
2. Foundational / Multimodal
3. Foundational (Unimodal) (✓)
   - GPT
   - VIT
4. Recipes for Multimodality
5. Research Works

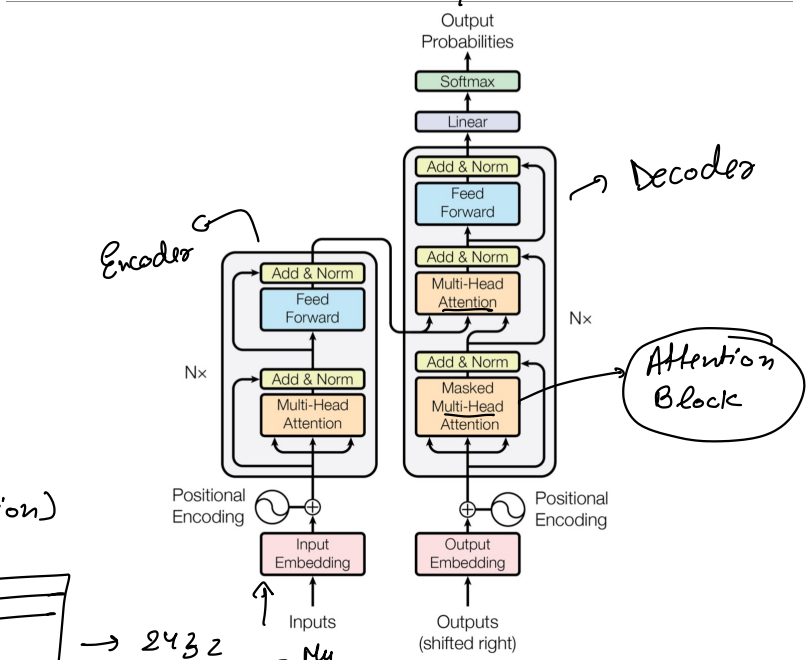
→ Transformer

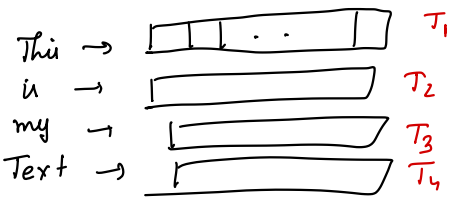
Representation



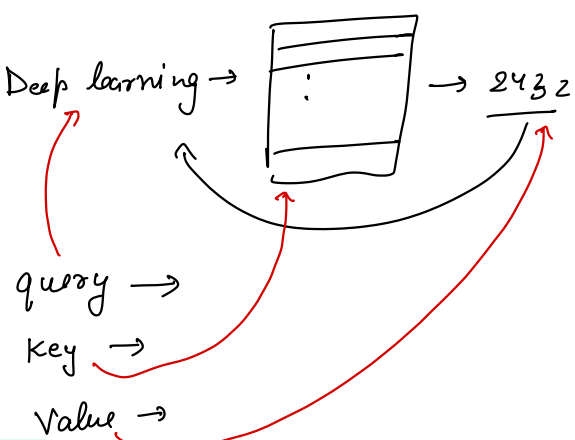
"This is a class" → 4 tokens

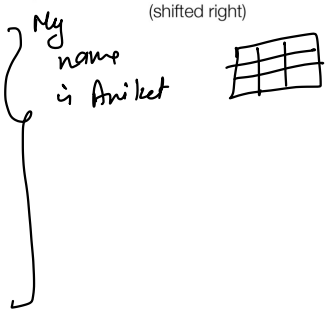
मेरा नाम अमीन है

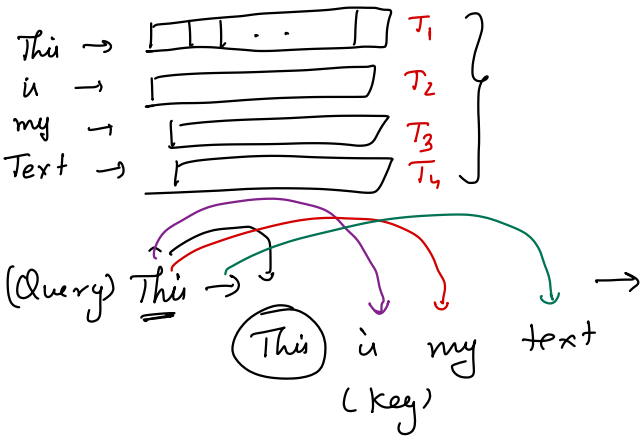


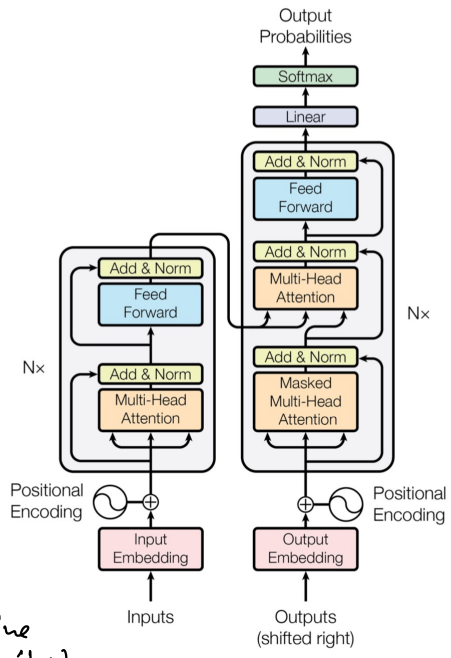


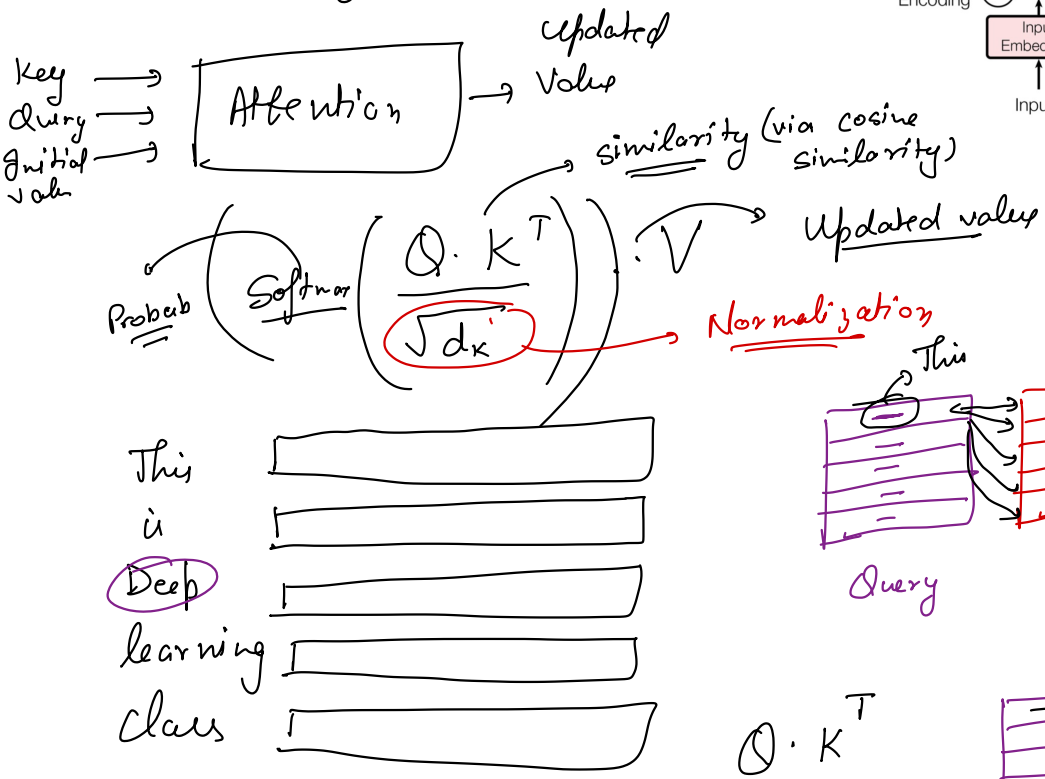
- Attention (Self-Attention)

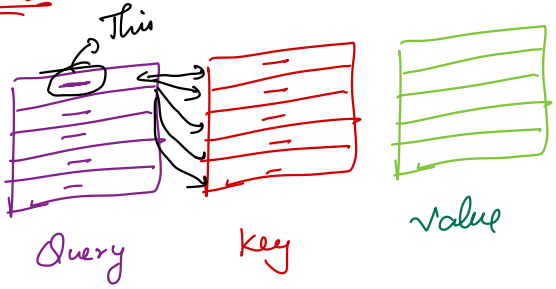




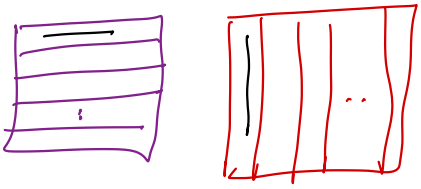


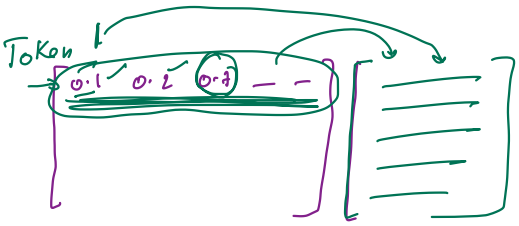


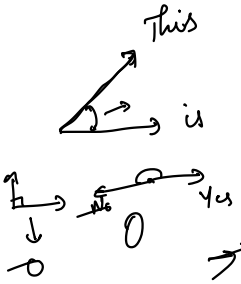




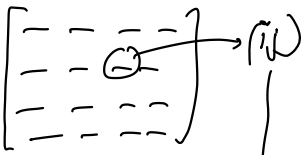
$Q \cdot K^T$







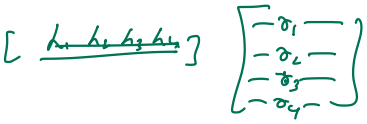
How similar Token i is with token j



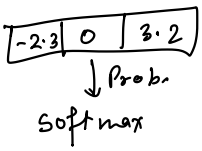
How similar token i is with token j

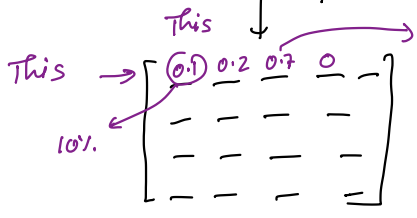
Happy, fun → Probability

Softmax



$h_1 v_1 + h_2 v_2 + h_3 v_3 + h_4 v_4$





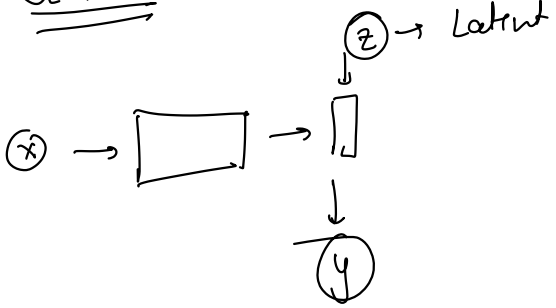
Attention Matrix

- Foundation Model

(1) Trained on a large-scale data

(2) It is able to perform multiple task

⇒ GPT

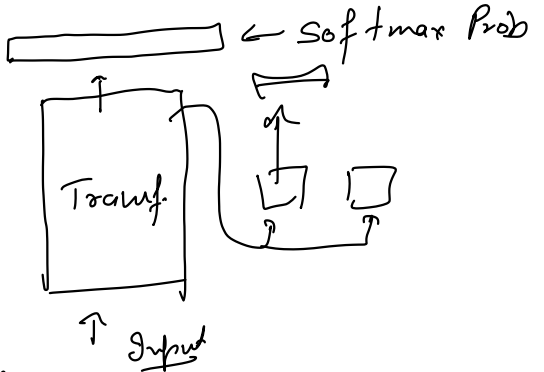


- Self-supervised
  - Labels are generated from data itself.

→ Next word prediction

"Labels are generated itself, then it is called" Input

Self-supervision → output.



Prediction: Generative

Stackoverflow → Q: ___ ? → Answer

GMDB ↓

Labels are gen ___ then it is called

Mixture of Experts.

Yes!