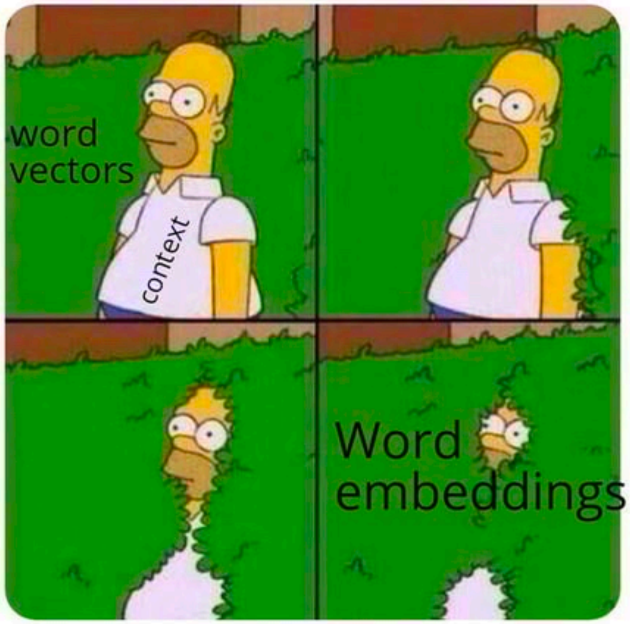
May 17, 2023.

DSML : NLP module.

BERT : Bidirectional Encoder Representation from Transformers.

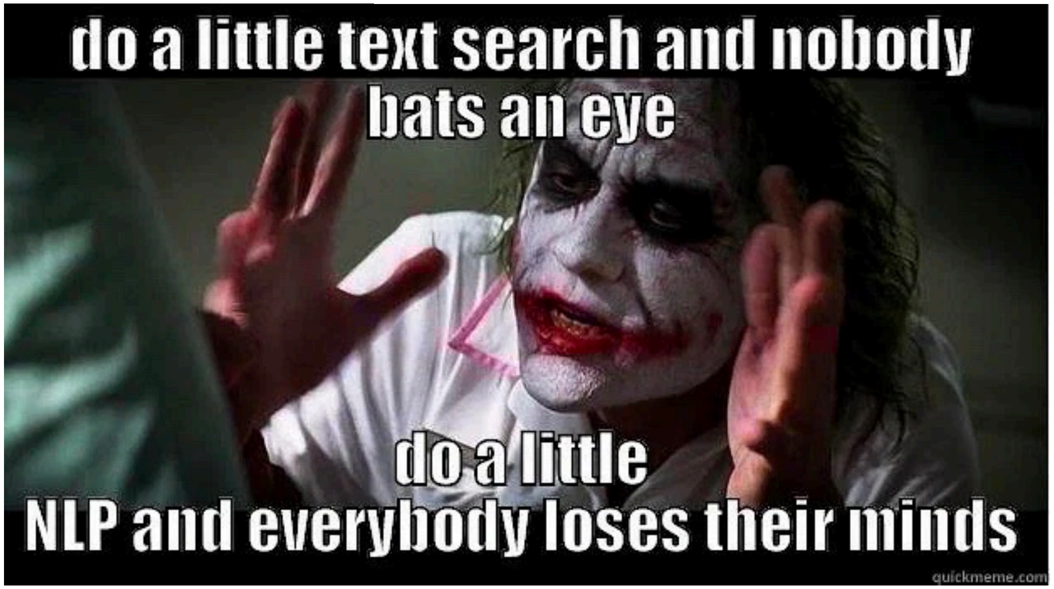
Class starts @ 9:05

Word embeddings in a nutshell



When you penalize your Natural Language Generation model for large sentence lengths

Me think, why waste time say lot word, when few word do trick.


do a little text search and nobody bats an eye
do a little NLP and everybody loses their minds
quickmeme.com

Recap:

- * Business Case : Neural Machine Translation.
  → given a sentence in English, translate the sentence to French.

- * Approach taken to solve the problem : Transformer Neural Network.

- * What we covered last class : Transformer

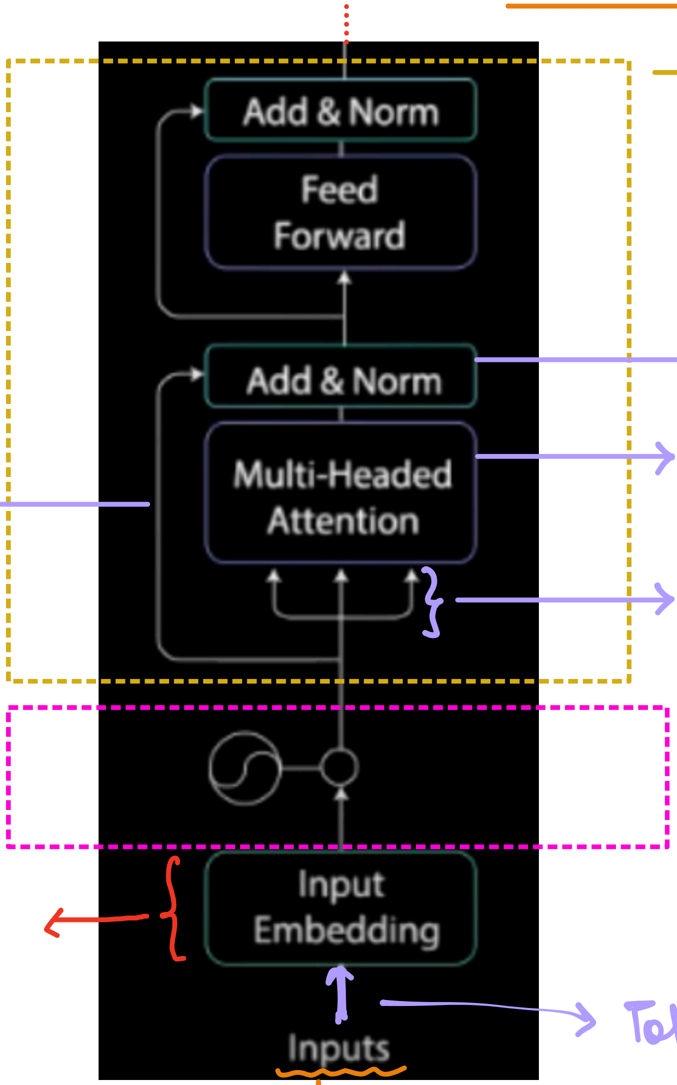
Transformer → Encoder, Decoder.

→ Encoder

- * Uses Attention only.
- * Requires positional encoding
- * Multiheaded self-attention.
- * Is trainable.

Detailed Recap:

Final Output: Query & Key



One block of the transformer encoder. This repeats 6 times.

3 trainable matrices for Q, K, V.

Self Attention: Q = K = V.

Skip connection.

Positional Encoder

Sinusoid vector + Embedding.

Word2Vec or GloVe. Usually fixed.

Tokenization.

Whole sentences.

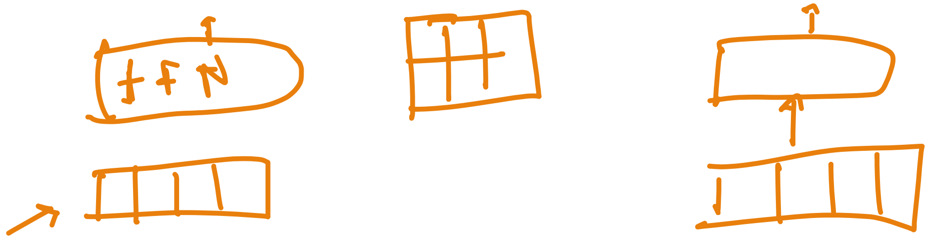
Agenda:

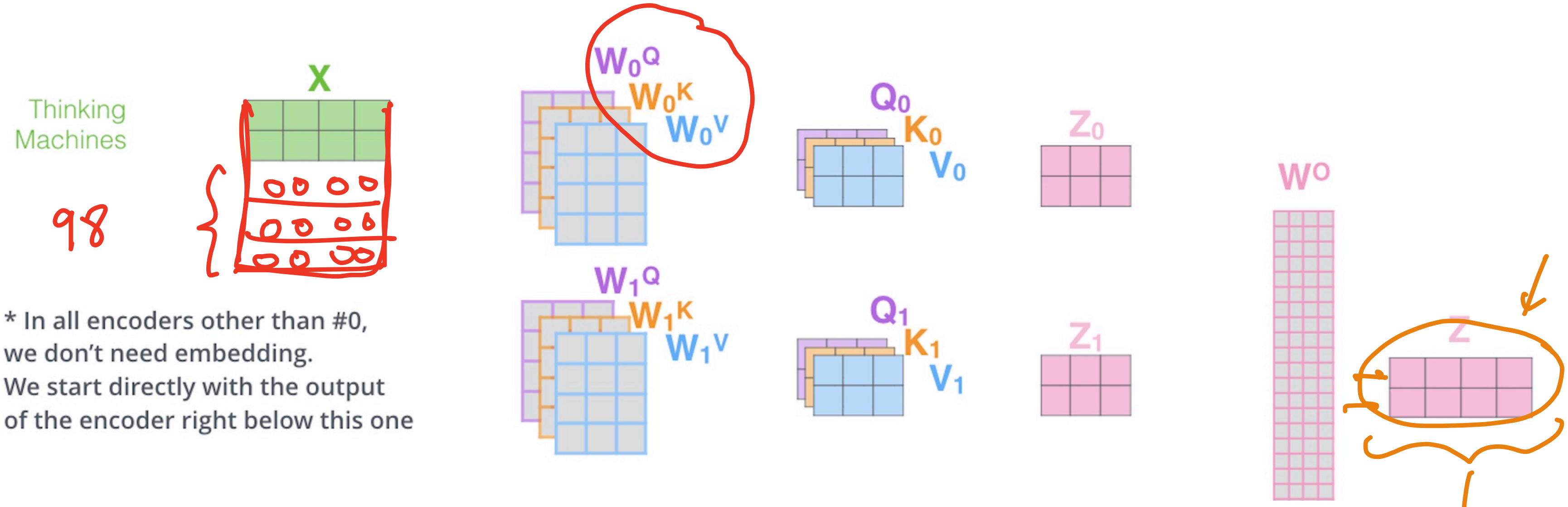
✓* Decoder architecture.

* BERT: An encoder only model.

* Transformer Implementation using Transfer Learning.

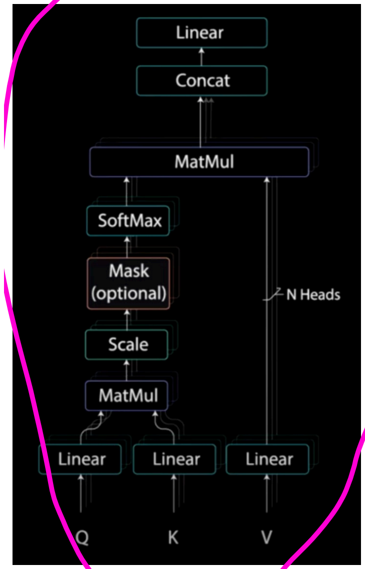
* Using BERT for an NER task.





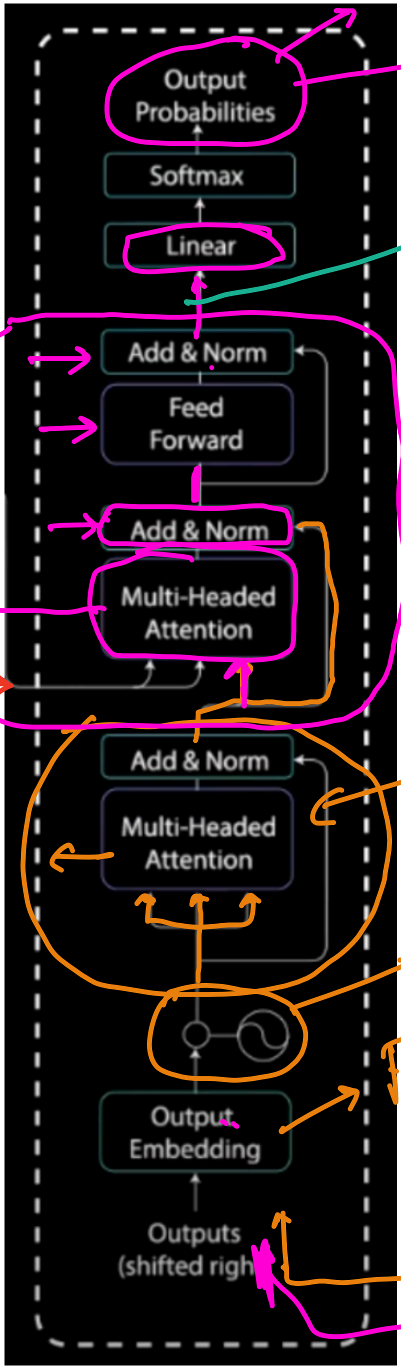
* In all encoders other than #0, we don't need embedding. We start directly with the output of the encoder right below this one

→ o/p of multiheaded attention.



Decoder block.

x N.



Generates 1 word at a time.

New value vectors.

<start> This is a
0 1 2 3

cat <end>
4 5

Encoder Outputs:

✓ Key Matrix

✓ Query Matrix.

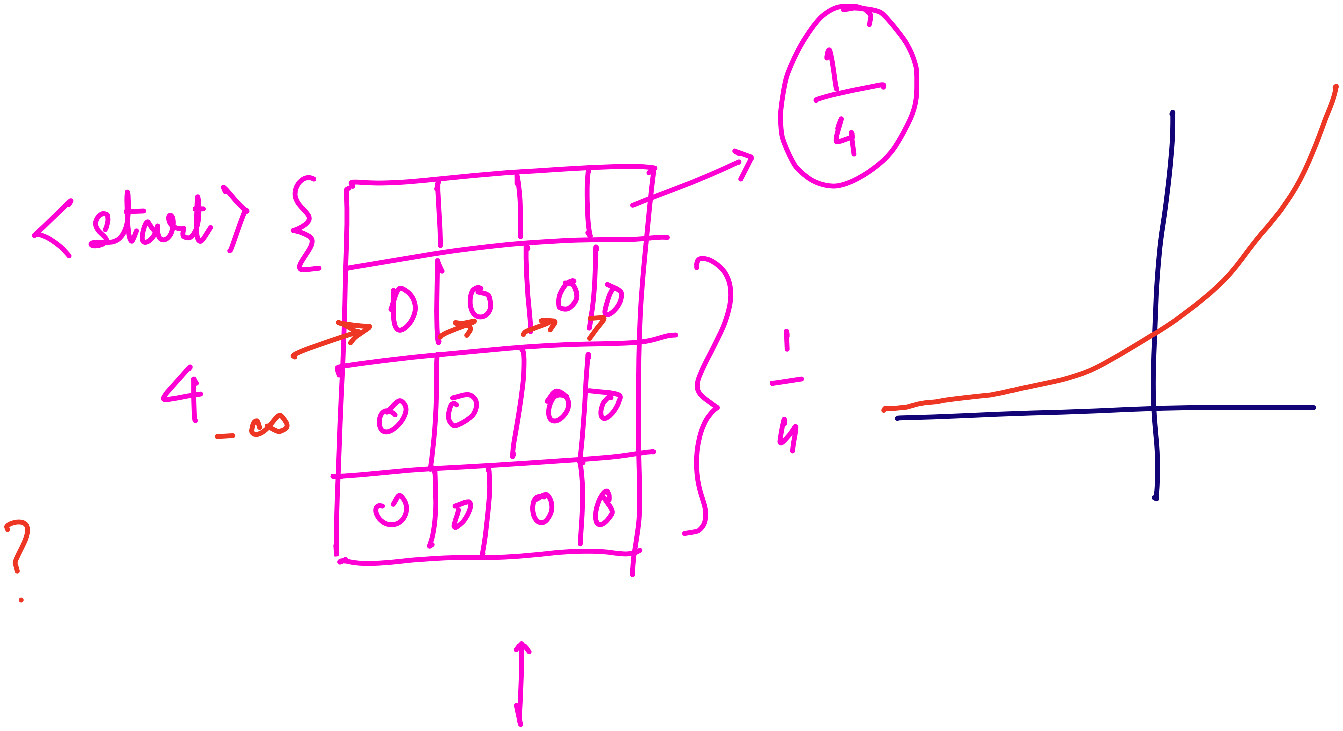
Value vectors.

This changes for every new word which is generated.

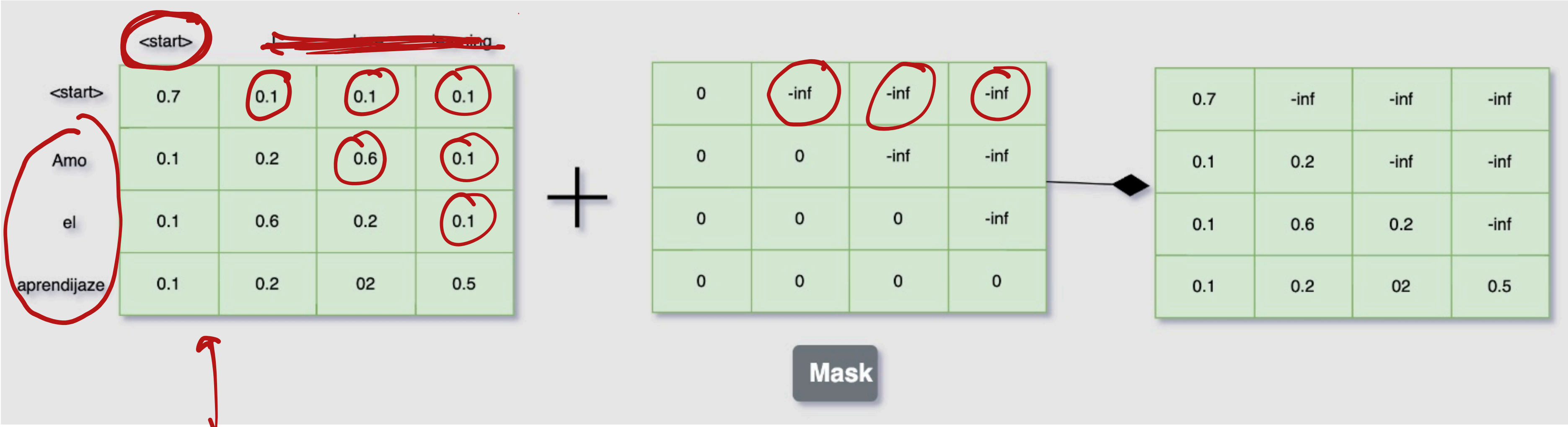
Positional encoding.

From Glove

$e^{\boxed{-\infty}} = 0$?



Issue: softmax will artificially put $\frac{1}{4}$ as the attention weight when we use padded inputs.



	<start>			
<start>	0.7	0.1	0.1	0.1
Amo	0.1	0.2	0.6	0.1
el	0.1	0.6	0.2	0.1
aprendijaze	0.1	0.2	02	0.5

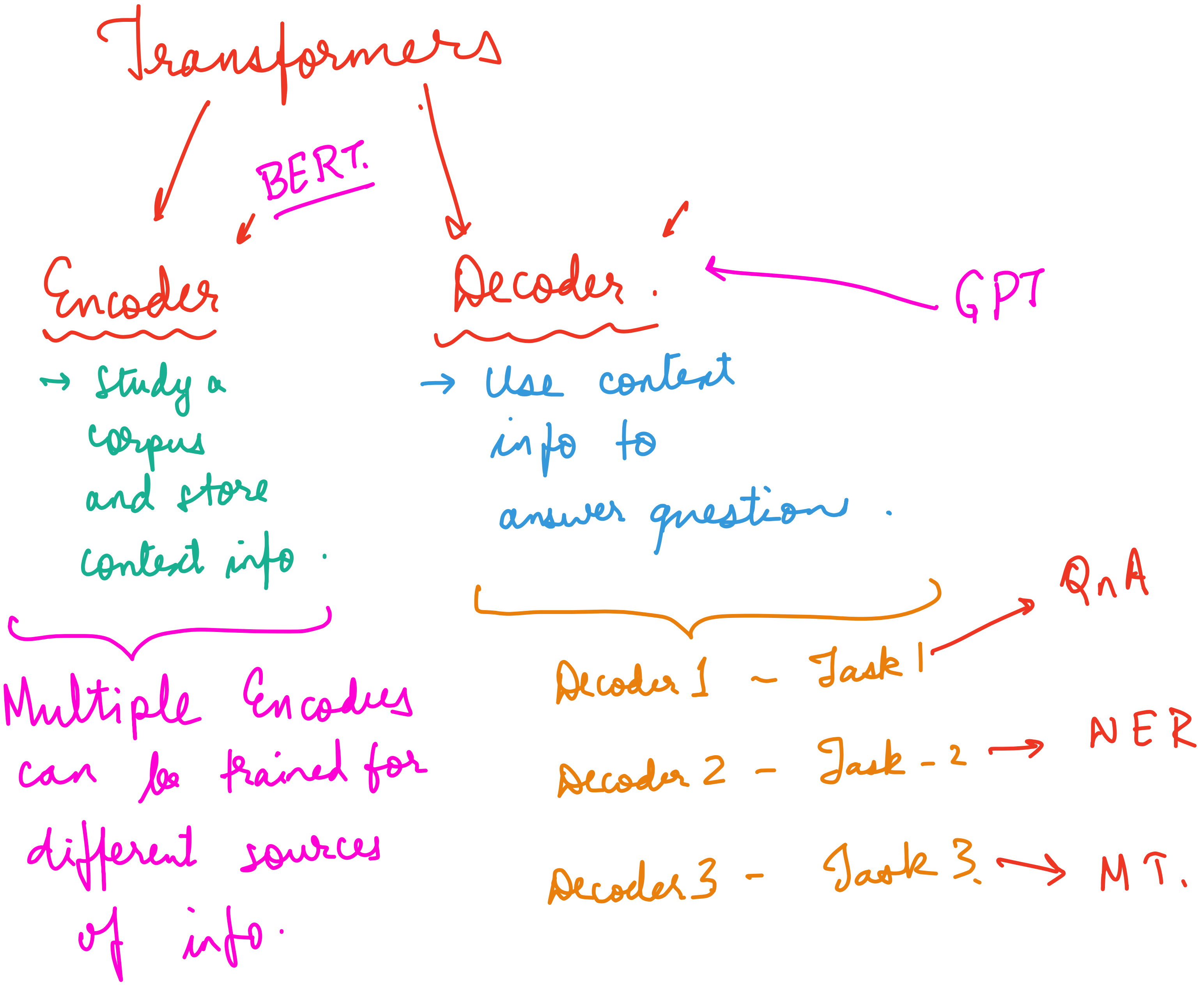
+

0	-inf	-inf	-inf
0	0	-inf	-inf
0	0	0	-inf
0	0	0	0

Mask

0.7	-inf	-inf	-inf
0.1	0.2	-inf	-inf
0.1	0.6	0.2	-inf
0.1	0.2	02	0.5

$$\left(\frac{Q^T K}{\sqrt{d}}\right)$$



Transformers

BERT.

Encoder

→ Study a corpus and store context info.

Multiple Encoders can be trained for different sources of info.

Decoder.

GPT

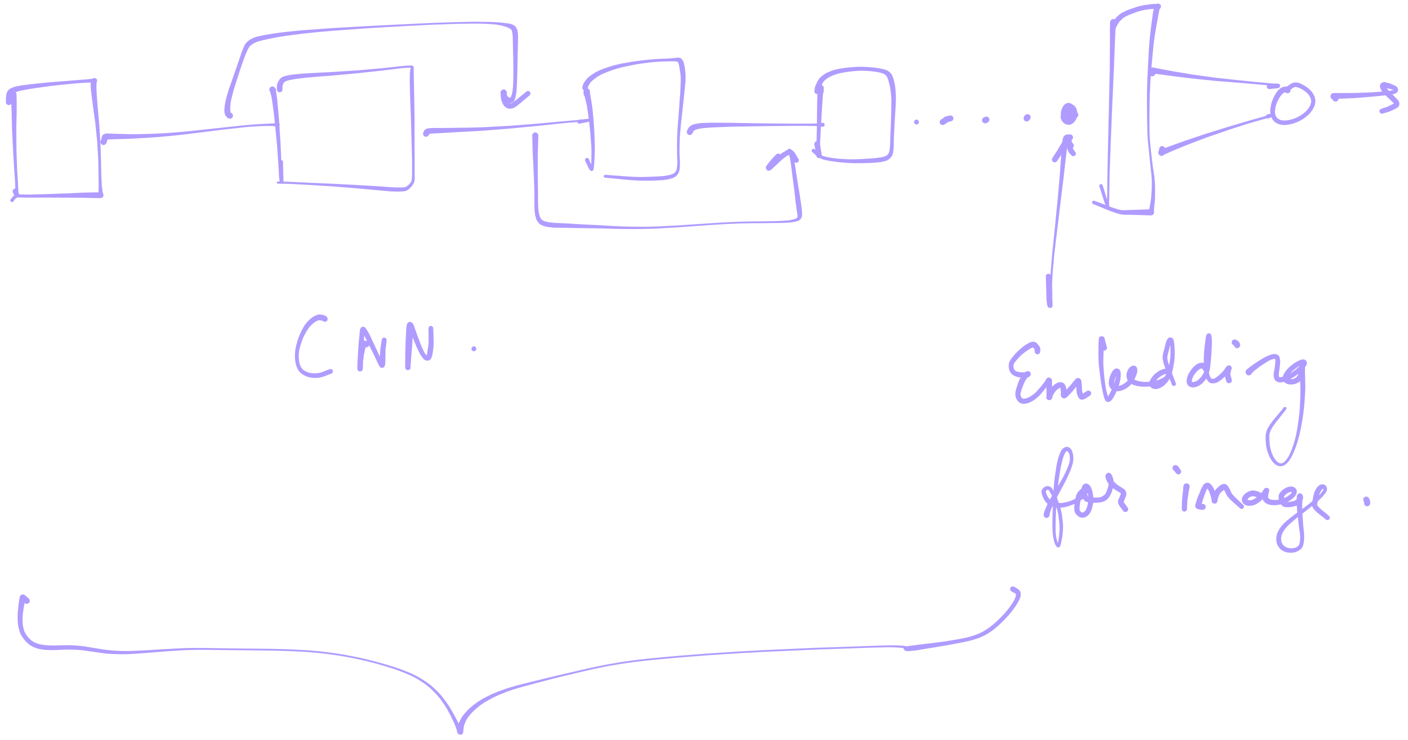
→ Use context info to answer question.

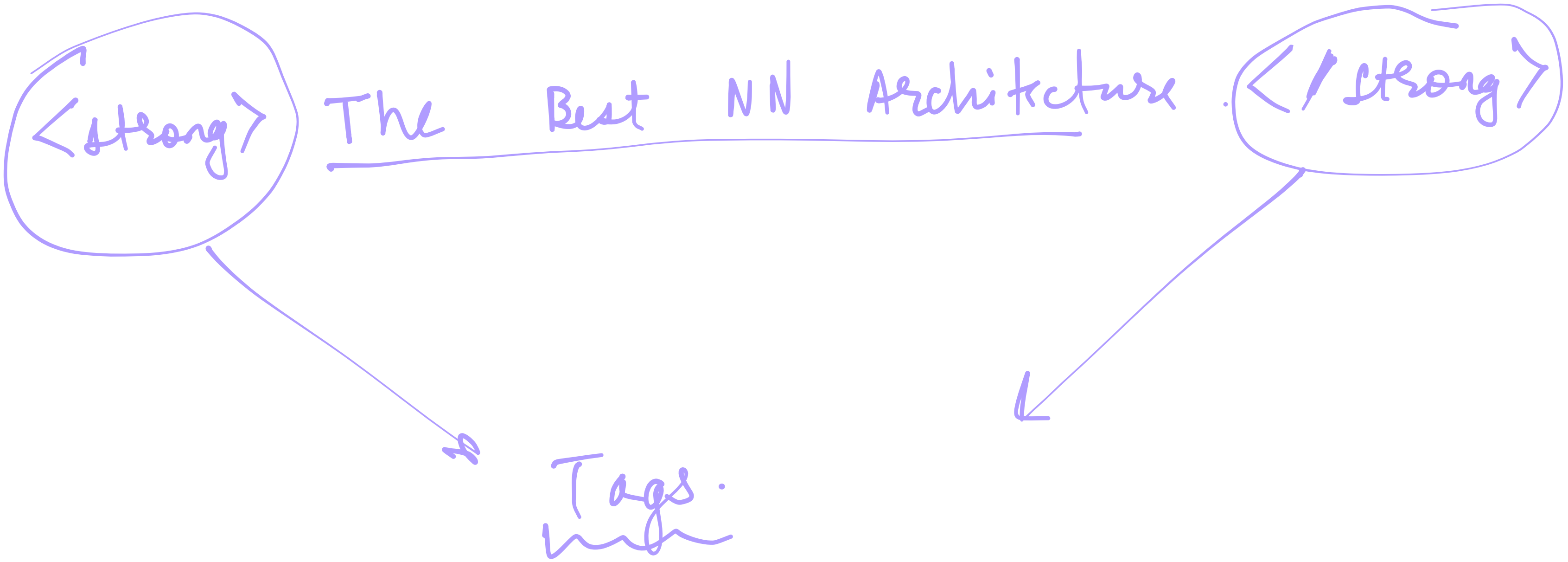
Decoder 1 - Task 1 → QnA

Decoder 2 - Task - 2 → NER

Decoder 3 - Task 3 → MT.

Recommender Systems → Embeddings using Matrix factorization.



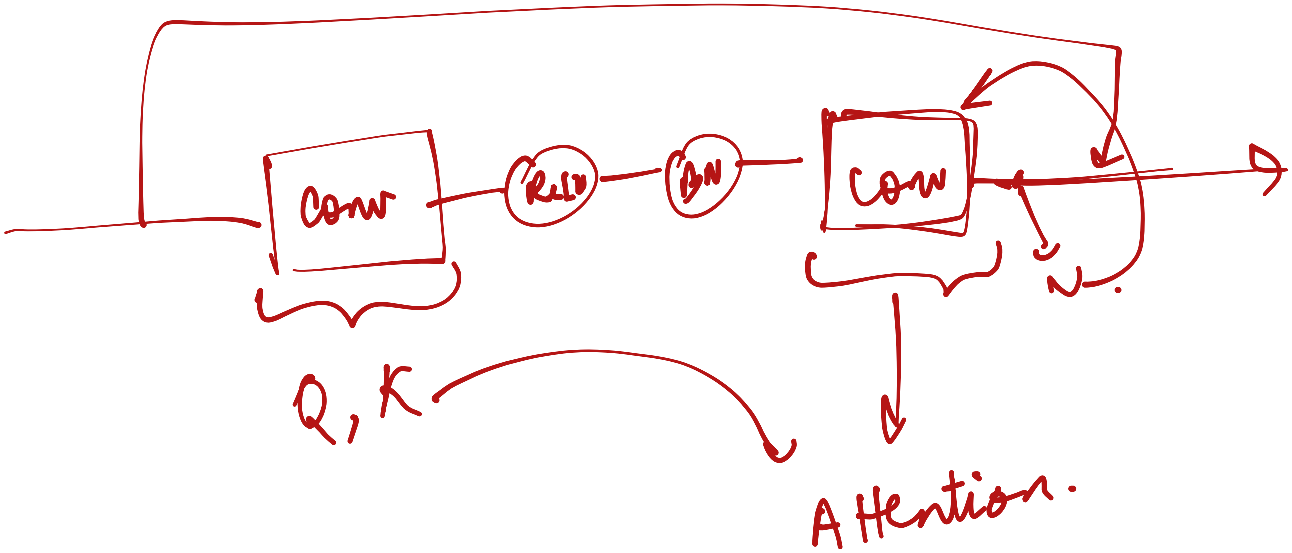

<strong> The Best NN Architecture </strong>
Tags.

Closing Remarks:

- Heuristic methods : Human intervention for useful info from text .

- Probabilistic methods; CRF (Conditional Random fields)
  Naïve Bayes .

- Neural network approaches :


Conv
ReLU
BN
Conv
N.
Q, K
Attention.