Consider an H-head attention, w/ k,q,v size d_1, x size d.

How many parameters are introduced in 1 attention layer?

1. Projection of x to K, Q, V
2. Compute score + softmax
3. Weighted avg of value vecs.
4. W^O projection

1. W^Q, W^K, W^V for 1 head
$= [d \times d_1]$ — $d d_1$

$3 \times H \times d d_1$

2. 0
3. 0
4. [illegible]

(4) W^O ← input is

Concat of H v vectors

Input dim — $H \cdot d_1$

Output dim → a_3 is

same dim as $x_3 \rightarrow d$

$H d_1 \cdot d$

Overall → $3 H d_1 d +$

$H d_1 d$

$= 4 H d_1 d$

Suppose (prev question)

+ L transformer layers

Vocabulary size V,

FFNN hidden layer size

d_2.

Total no. of parameters??

- Para. in Attention Block
- " " layer norm
- " " FFNN
- " " E

$\times L$

QA →

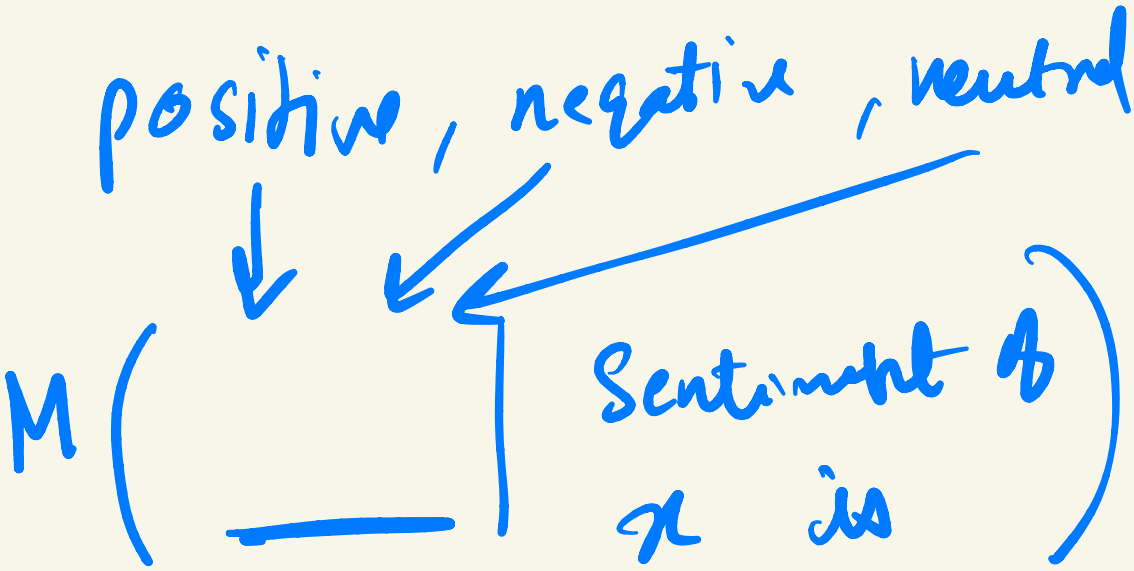
Input q "Who did ... ?"

Generate output token from M(.| q)

SA →

Input x → Movie Rev—

Generate output tokens from M(. | What is the sentiment of x ?)


positive, negative, neutral
M(___ Sentiment of x is)

Bidirectional RNN