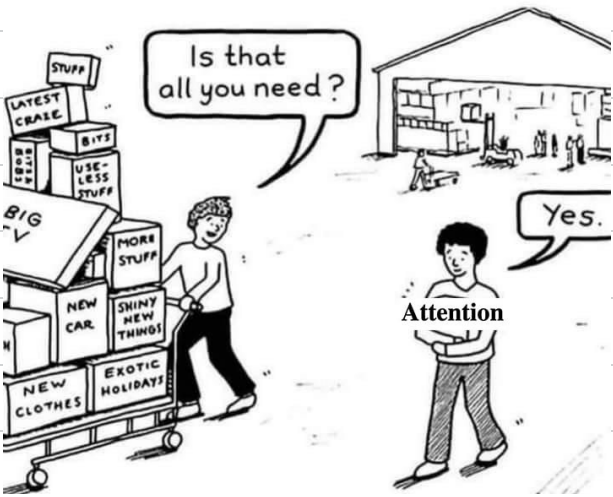
Session 09

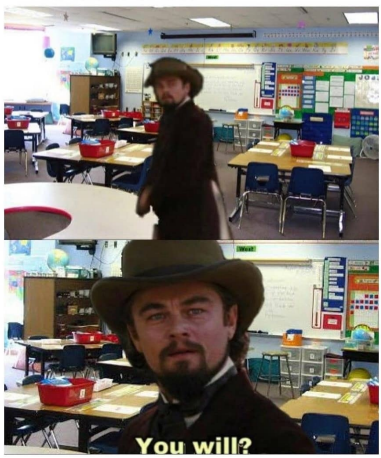
Attention

May 03, 2024



Teacher: If you don't want to pay attention I'll honestly end this class right now.

The Students:



Agenda

1. Case-study, how can we solve it?
2. What's wrong with RNNs?
3. How attention helps?
4. How attention works
5. Different kinds of attention
6. Self Attention and Multi-head attention
7. Code Walkthrough.

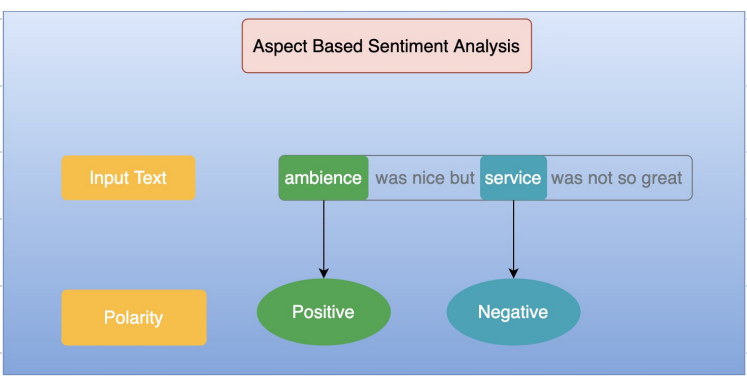
Aspect Based Sentiment Analysis

ABSA

Food is delicious but price is high → overall neutral

taste → V. Positive "taste"

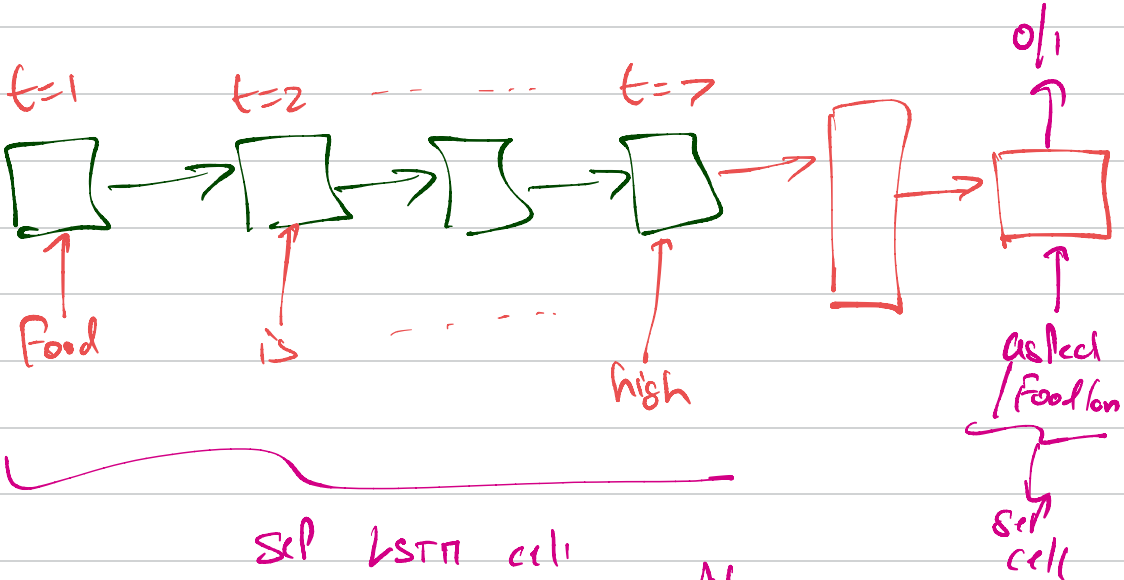
Cost → V. Negative.

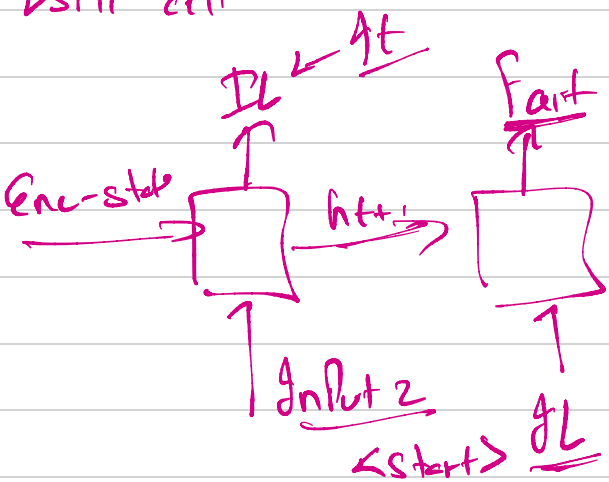


(1) Use NER to remove / select other entity

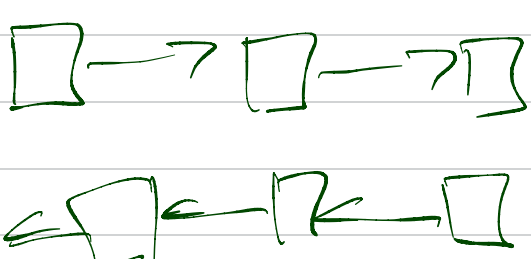
↳ Personalized NER model

Encoder - Decoder.





Bidirection



What are some problems with LSTM??

1. If I have a very long paragraph, let's say like 100 words, then the specific word like "taste" maybe present in the beginning of the word.

Now cell state vector takes care of that little bit, but still information is lost.

2. The weight matrix, for forget, input, and output gate are the same wt matrix, being multiplied to all the words.

3. Lack of parallel training for LSTM networks.

ATTENTION

(1) Key, Query, Value → matrices

(2) Attention function / Energy / Alignment

(3) Content Vector

Value matrix → [10, 512] ← W_V [512, 512]

1 Key matrix → $[w_1 \dots w_{10}]$

↓

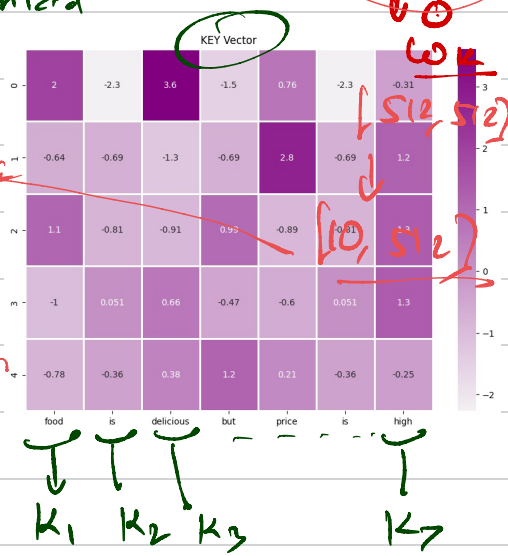
Representation of tokenized Sentence

→ [10, 512] → [10×512]

W_K [512, 512]

Food is delicious but price is high

Key matrix → [10, 512]



5d - !!. → For visualization

124d or 512d → PCA → 5d !!

Never do this in Practice

2 Query Matrix → [1, 512] ⊙ W_Q [512, 512]

↳ Representation of tokenized aspects

↳ 1 word → "cost"

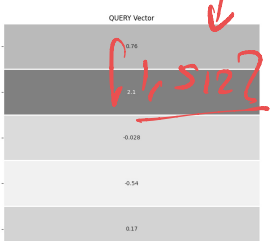
↓

1 word

2 word

⋮

n word



[1, 512]

(3) Value Vector

↓

In our Platform

Store separately in RAM

Value Vector = Key Vector

Initially same

Key matrix: 5×7 | 10, 512

Query matrix: 5×1 | 1, 512

Value matrix: 5×7 | 10, 512

Attention → Google

↓

Secret Engine

Key : Value

↓

1 trillion Keys

Query

most relevant Key

Key matrix: 5x7 (10, 512

Query matrix: 5x1 (1, 512

Value matrix: 5x7 (10, 512

$K \cdot Q^T$

(10, 512) 0 (512, 1)

= 10, 1

7x5 · 5x1 = 7x1

Dot Product of $(K^T \cdot Q)$

What do Key retrieve

Vec A · Vec B → A·B → similarity of how similar A → B

Similarity of Query w.r.t. each token in Key

APPly softmax on

Softmax $(K^T \cdot Q)$

0-1

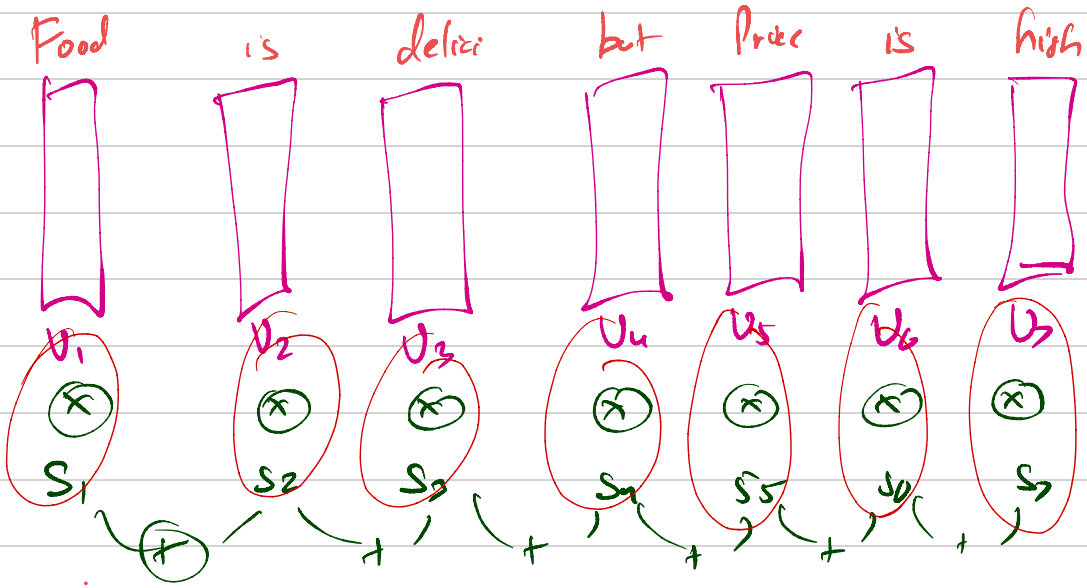
7x1

3 Context Vector

↳ Here Enters the Value matrix

Element wise multiplication of softmax output with Value matrix

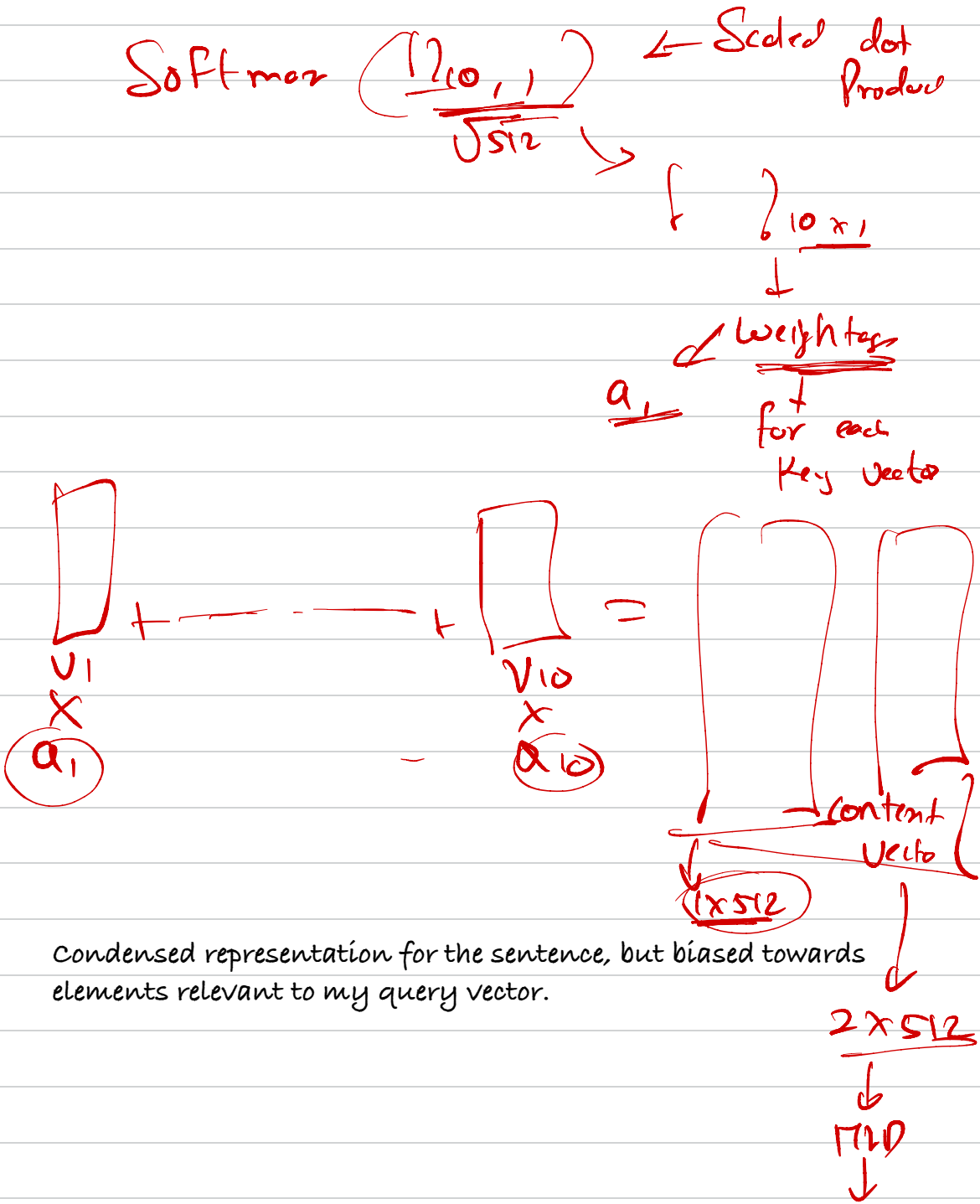
↳ 5×7



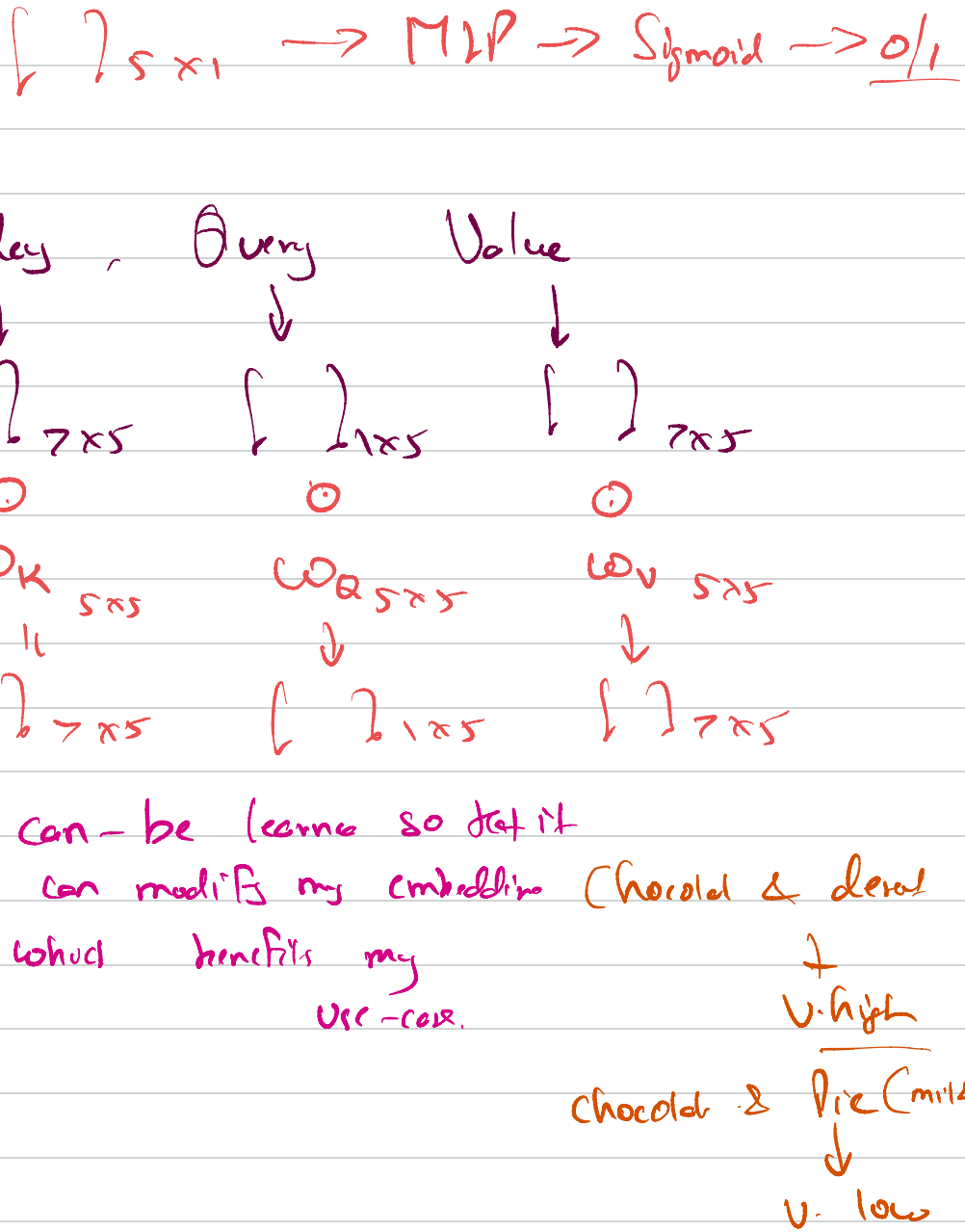
Some Kind of Condense representation

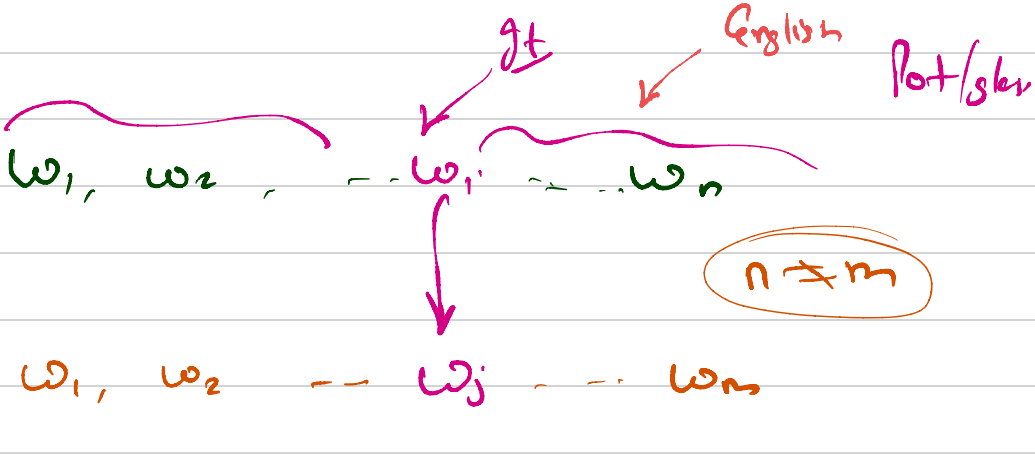
Aggregated information from all parts of data but biased towards elements which have more weightage/important.

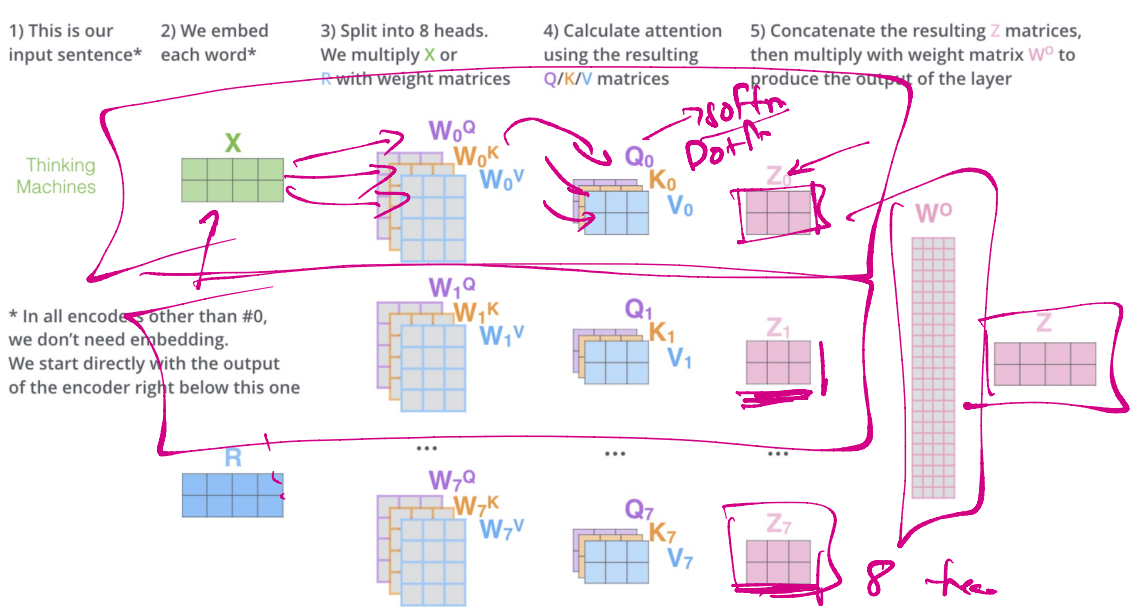
[ ]5×1



Condensed representation for the sentence, but biased towards elements relevant to my query vector.


5 x 1
MLP
Sigmoid
0/1
Key
Query
Value
7x5
1x5
7x5
W_K 5x5
W_Q 5x5
W_V 5x5
7x5
1x5
7x5
can be learned so that it can modify my embedding (chocolate & dessert + v. high) chocolate & pie (mixed) v. low
beau
beautiful





8 heads

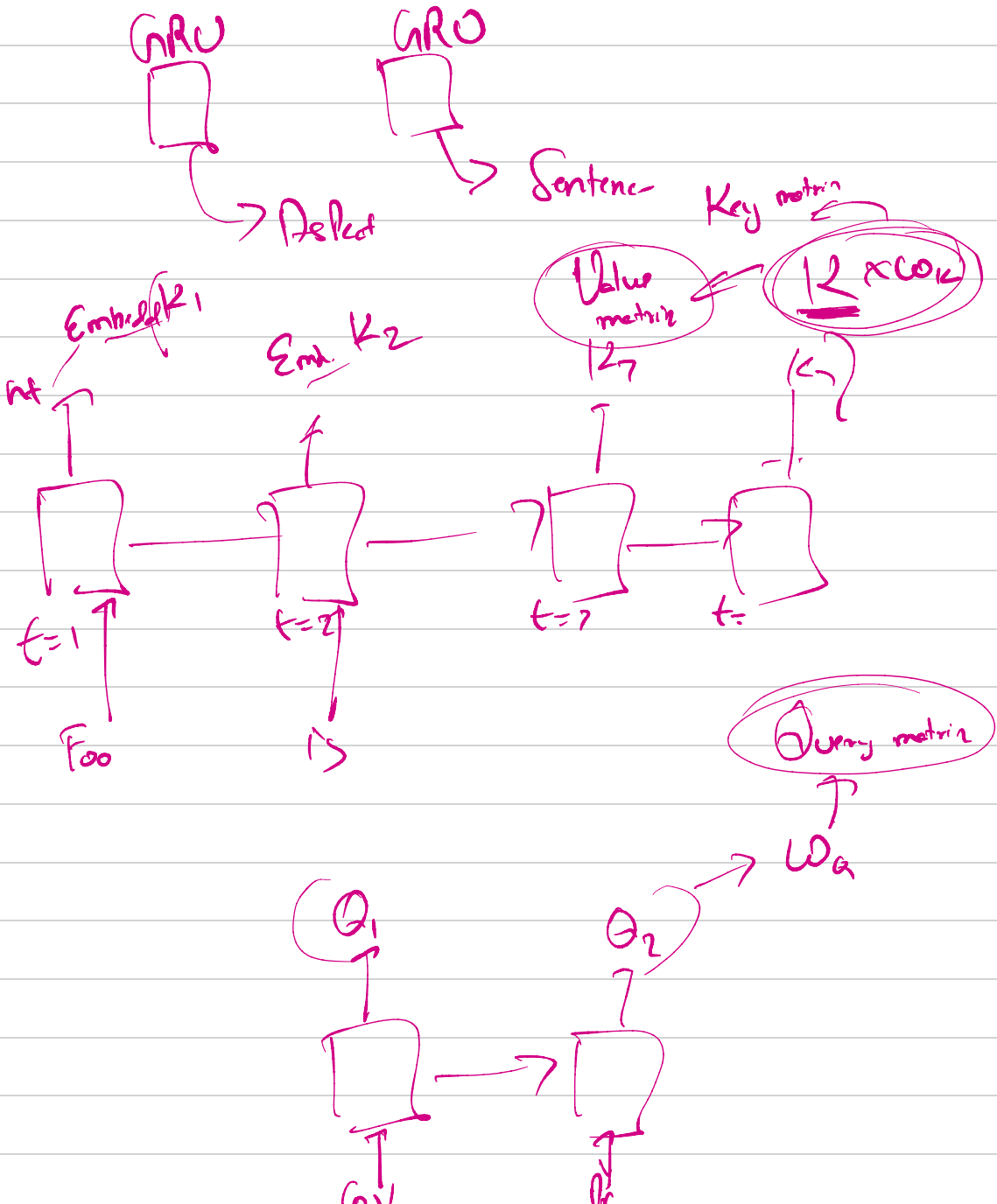
1st → [ ] 5×1

2nd → [ ] 5×1

8th → [ ] 5×1

⇒ [5, 8] × [8, 1]

⇒ [5, 1]


GRU
GRU
Sentenc
Key matrix
Value matrix
Embedd K1
Emb. K2
Foo
Query matrix

I want food

I want paratha

aspect hunger