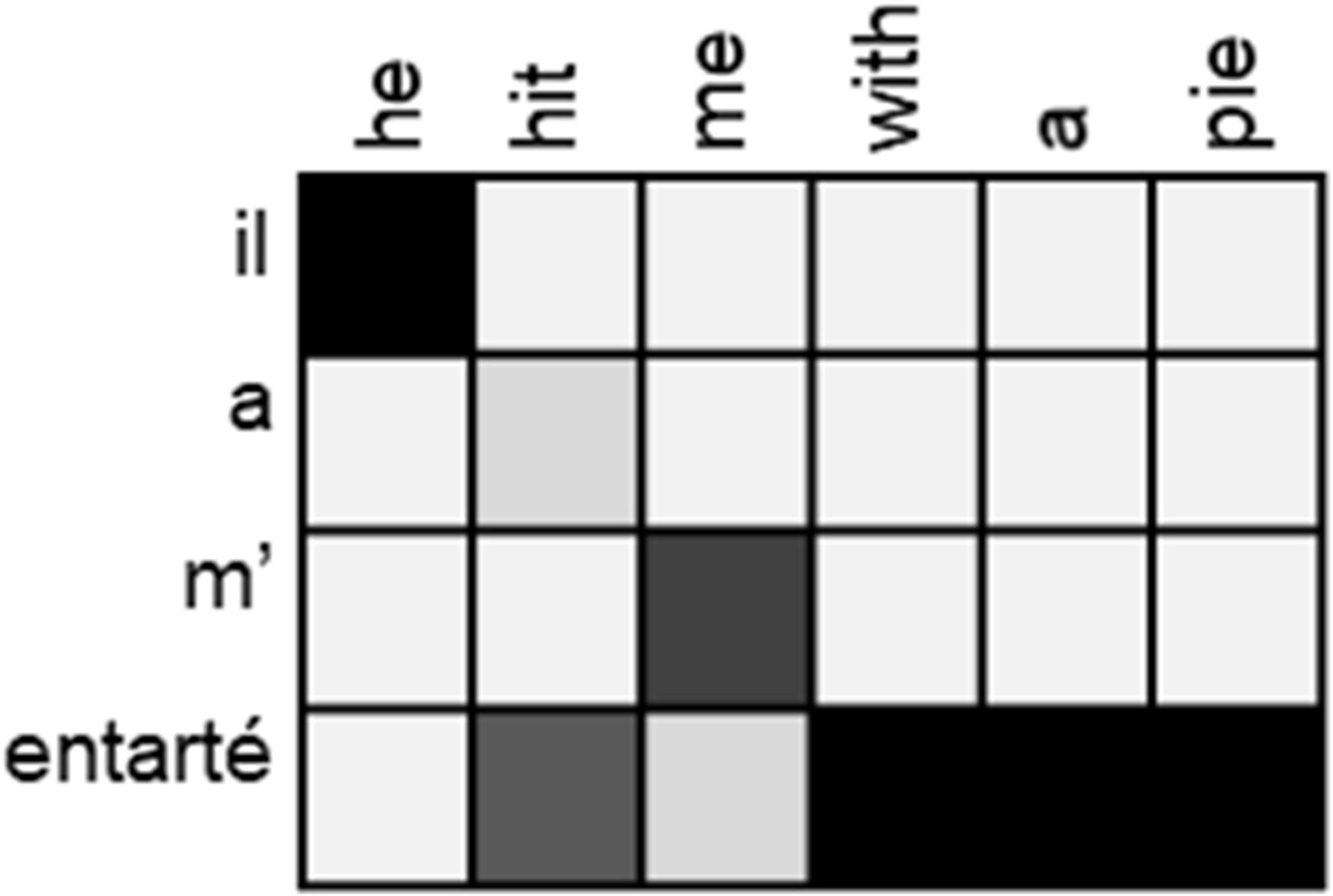
Admin

- Presentations on Thursday
  - Millie, Serafina, Braeden, Rhys, Pierce, Annie

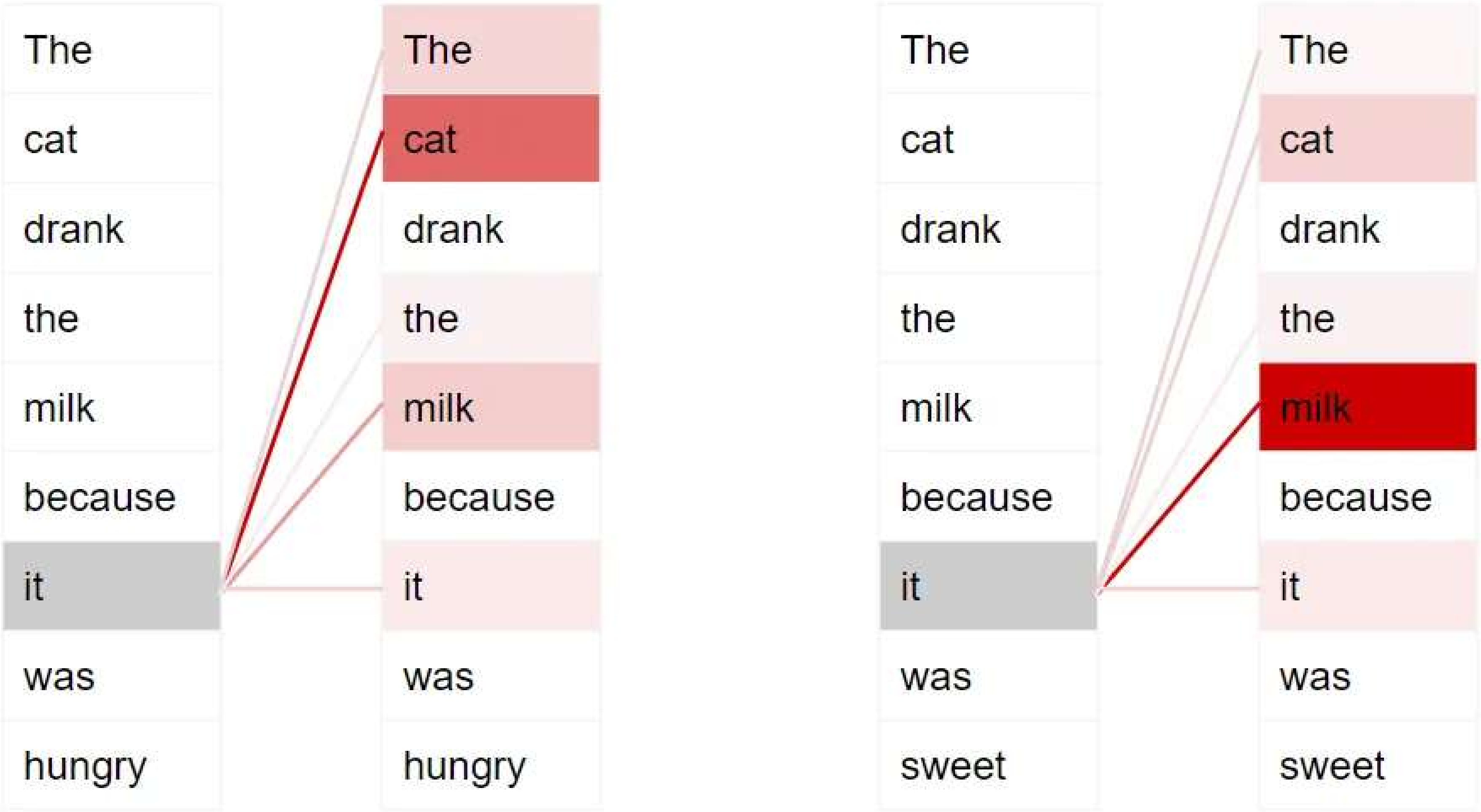
- Project code due Friday 5pm

- Today
  - Attention
  - Transformers

Attention in machine translation



Self Attention



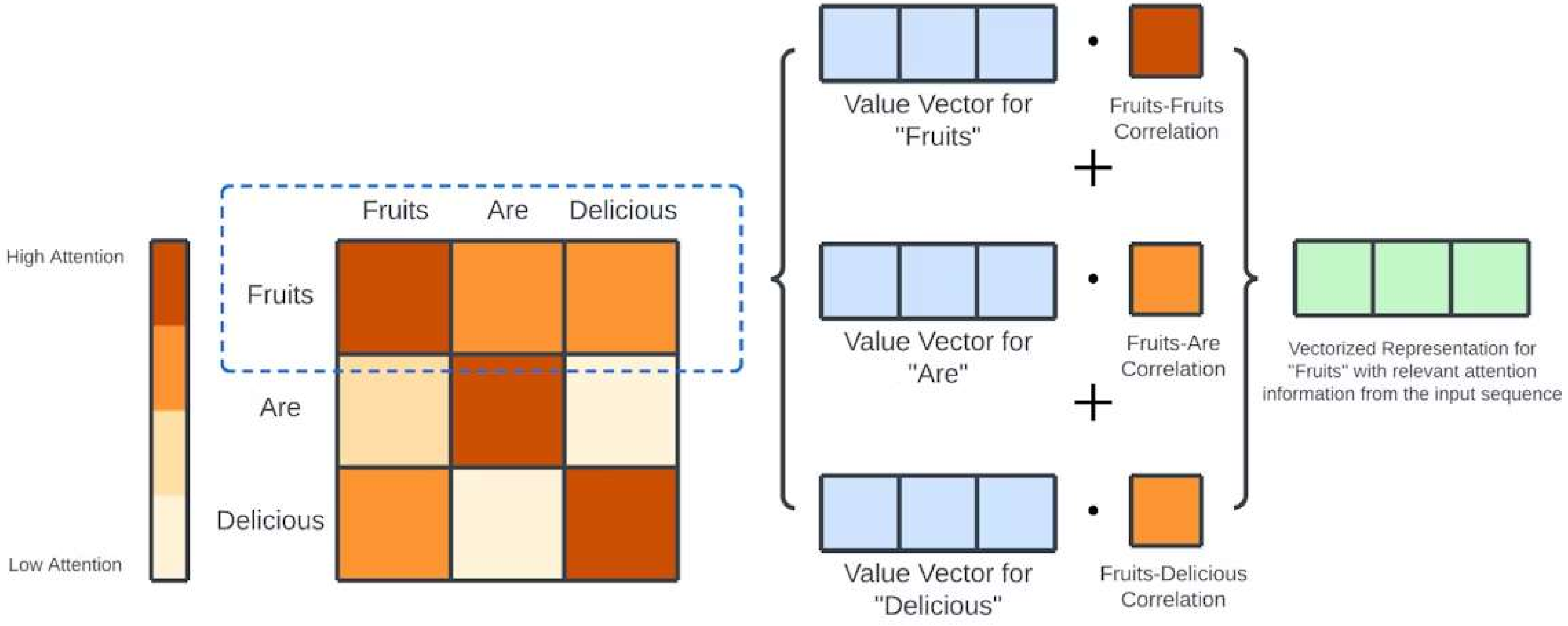
The
cat
drank
the
milk
because
it
was
hungry

The
cat
drank
the
milk
because
it
was
hungry

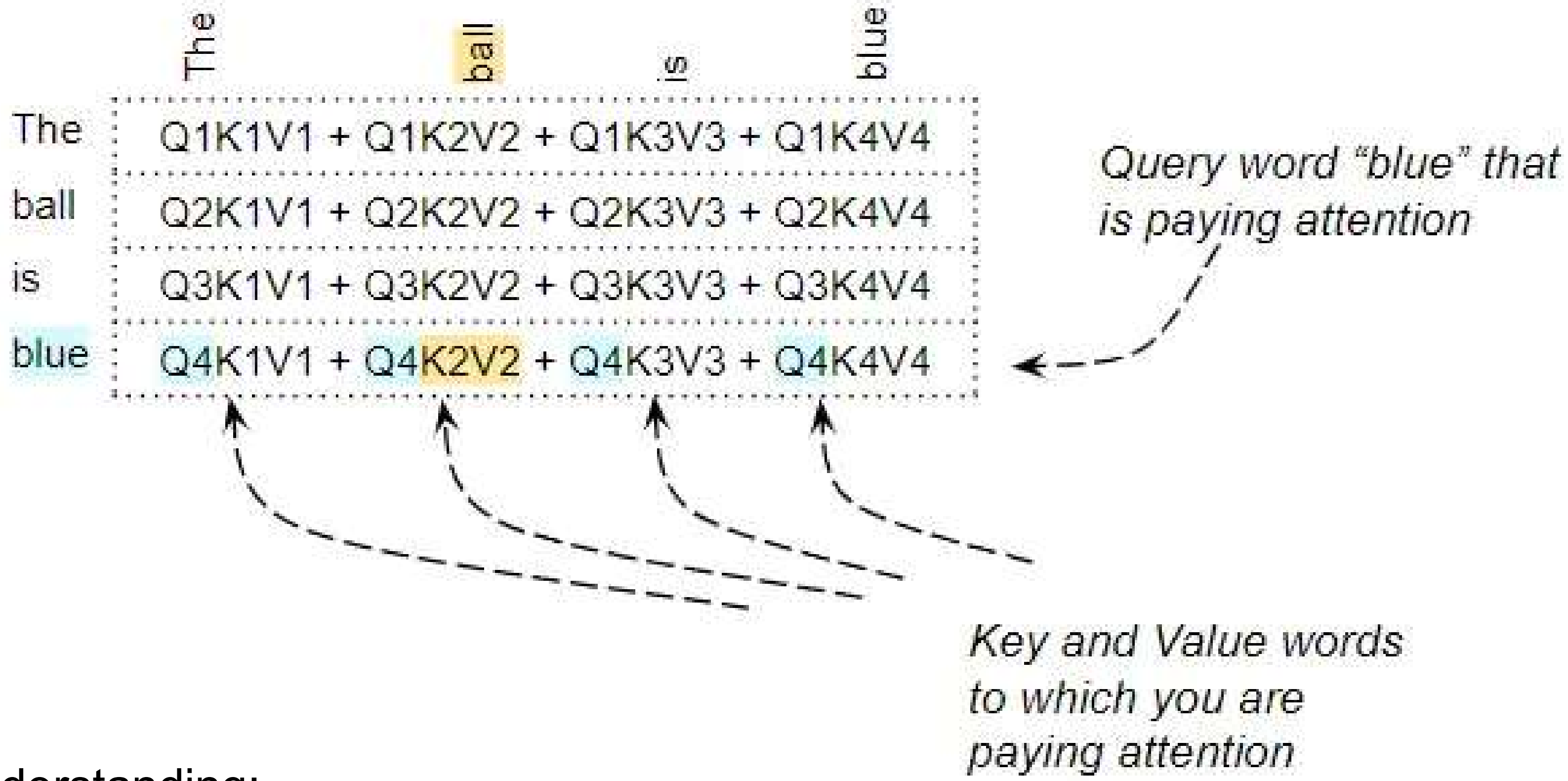
The
cat
drank
the
milk
because
it
was
sweet

The
cat
drank
the
milk
because
it
was
sweet

New representation modified by attentior


High Attention
Low Attention
Fruits
Are
Delicious
Fruits
Are
Delicious
Value Vector for "Fruits"
Fruits-Fruits Correlation
+
Value Vector for "Are"
Fruits-Are Correlation
+
Value Vector for "Delicious"
Fruits-Delicious Correlation
Vectorized Representation for "Fruits" with relevant attention information from the input sequence

Use Attention Score(QK) to update values



	The	ball	is	blue
The	Q1K1V1 +	Q1K2V2 +	Q1K3V3 +	Q1K4V4
ball	Q2K1V1 +	Q2K2V2 +	Q2K3V3 +	Q2K4V4
is	Q3K1V1 +	Q3K2V2 +	Q3K3V3 +	Q3K4V4
blue	Q4K1V1 +	Q4K2V2 +	Q4K3V3 +	Q4K4V4

Query word "blue" that is paying attention

Key and Value words to which you are paying attention

My understanding:

Query word: the target word (the word is paying attention)
Key-value pair: the source word, to which a target is paying attention to

- Value word: the value vector generated from a source word,
- Key word: the source word

QK: the similarity (relationship) between Query and Key and used as a weight to update Value

Attention is just a weighted average!

- Attention: weights, weighted voting
- Apply attention: get a weighted average

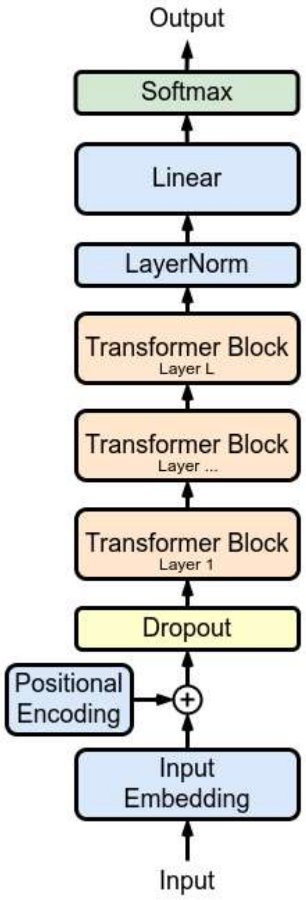
Attention is a *general* Deep Learning technique

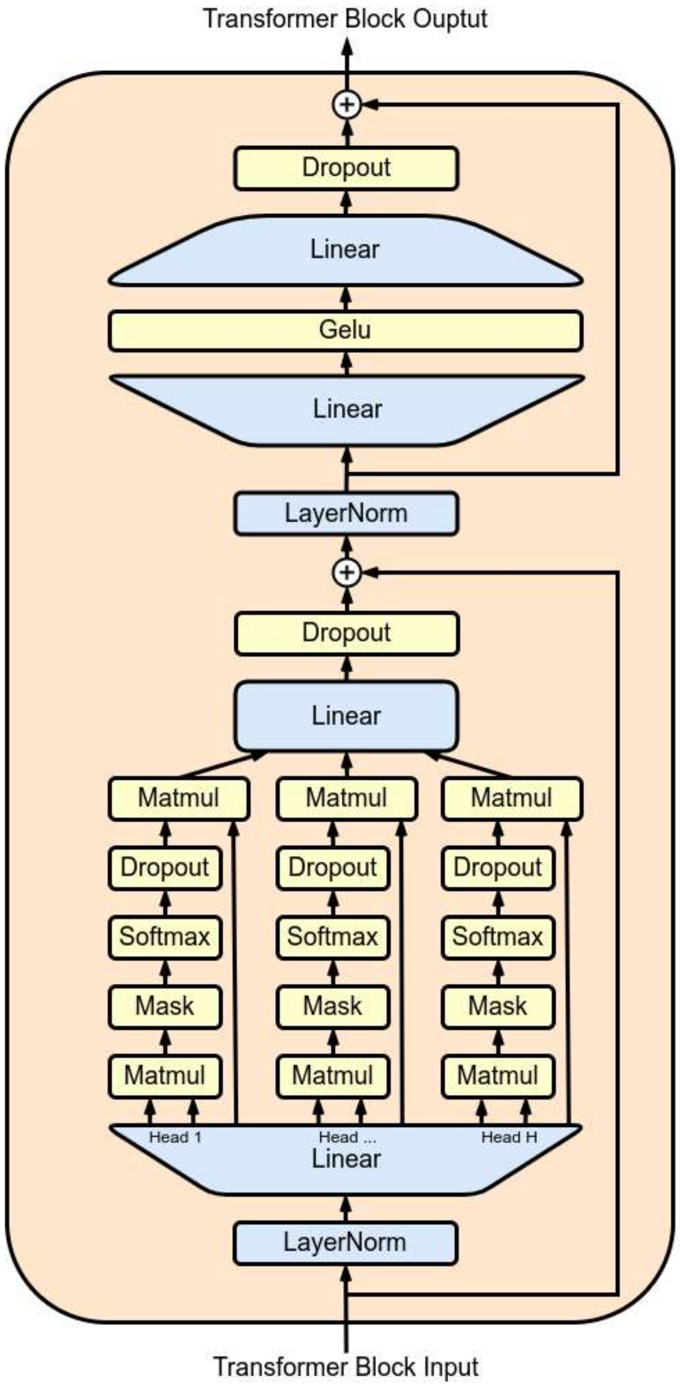
More general definition of attention:

Given a set of vector *values*, and a vector *query*, attention is a technique to compute a weighted sum of the values, dependent on the query.

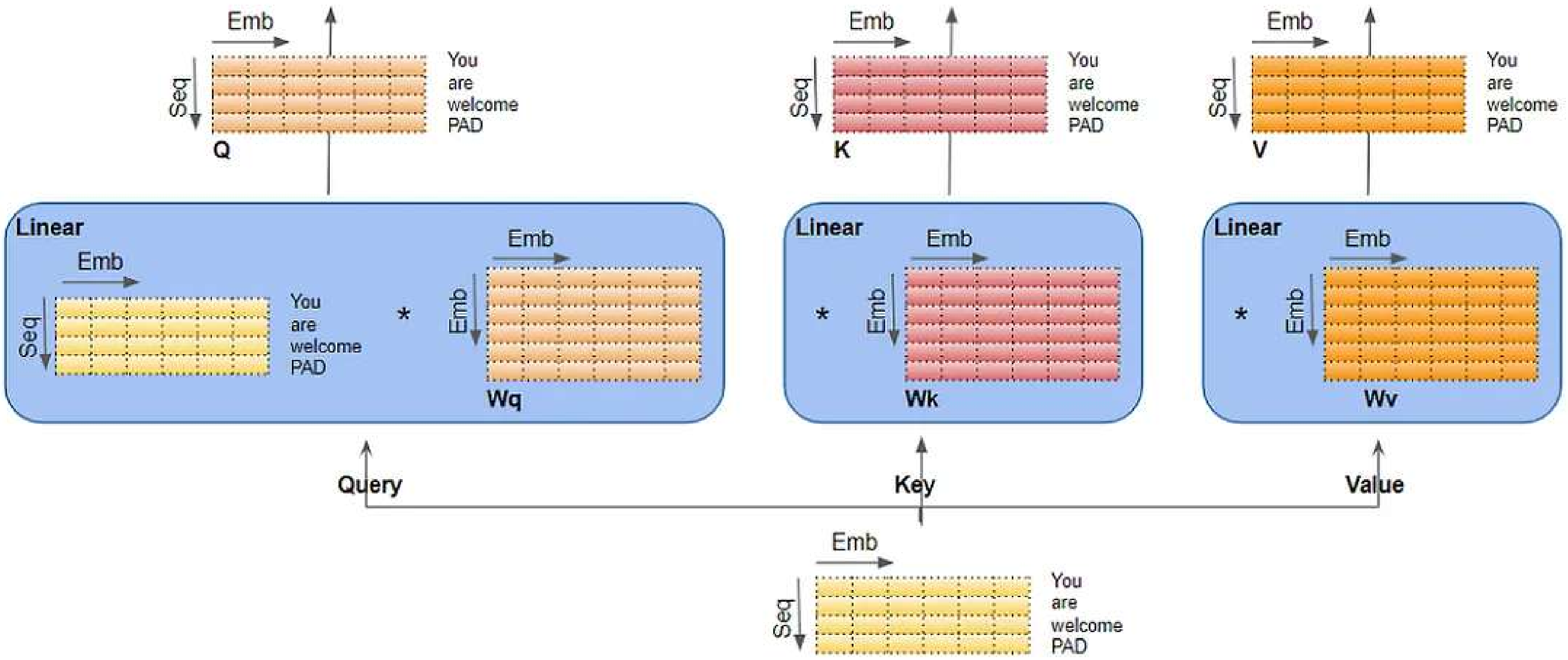
Intuition:

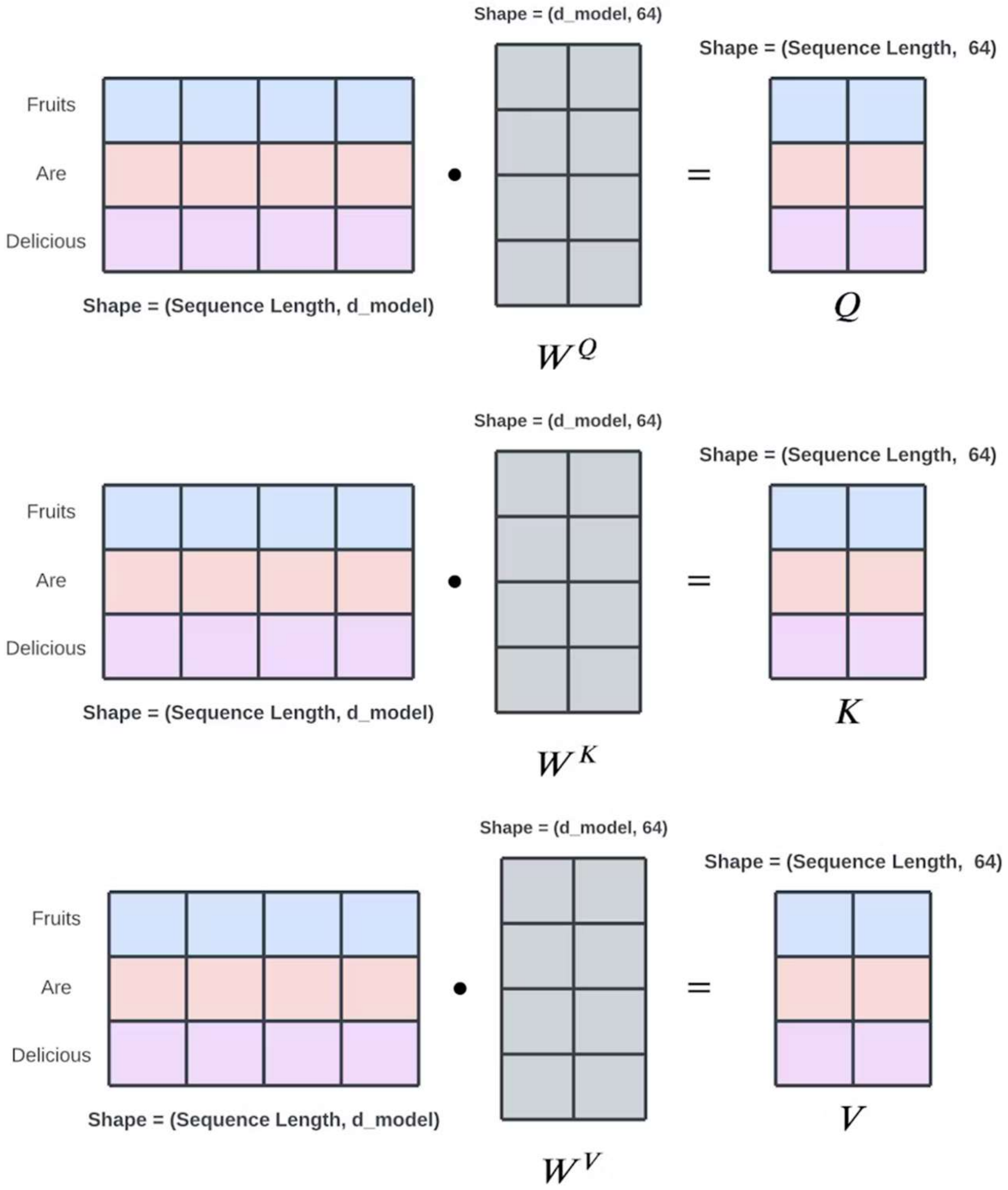
- The weighted sum is a *selective summary* of the information contained in the values, where the query determines which values to focus on.
- Attention is a way to obtain a *fixed-size representation of an arbitrary set of representations* (the values), dependent on some other representation (the query).


Output
Softmax
Linear
LayerNorm
Transformer Block
Layer L
Transformer Block
Layer ...
Transformer Block
Layer 1
Dropout
Positional Encoding
Input Embedding
Input

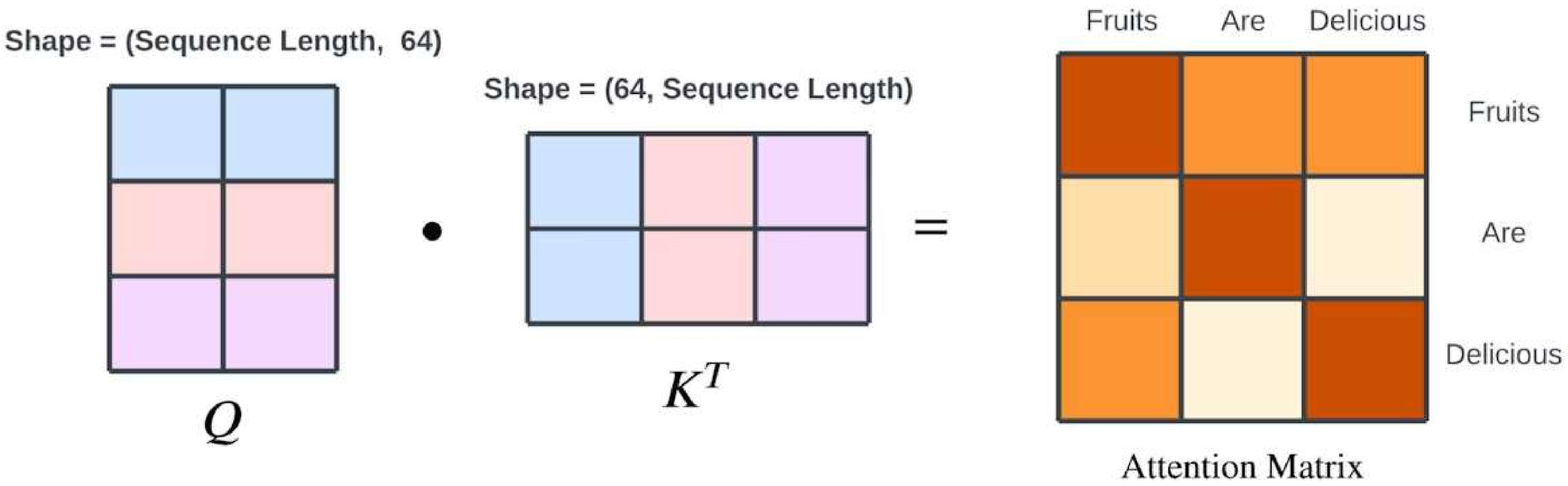

Transformer Block Ouptut
Dropout
Linear
Gelu
Linear
LayerNorm
Dropout
Linear
Matmul
Matmul
Matmul
Dropout
Dropout
Dropout
Softmax
Softmax
Softmax
Mask
Mask
Mask
Matmul
Matmul
Matmul
Head 1
Head ...
Head H
Linear
LayerNorm
Transformer Block Input

Self-Attention: How to get Q, K, V

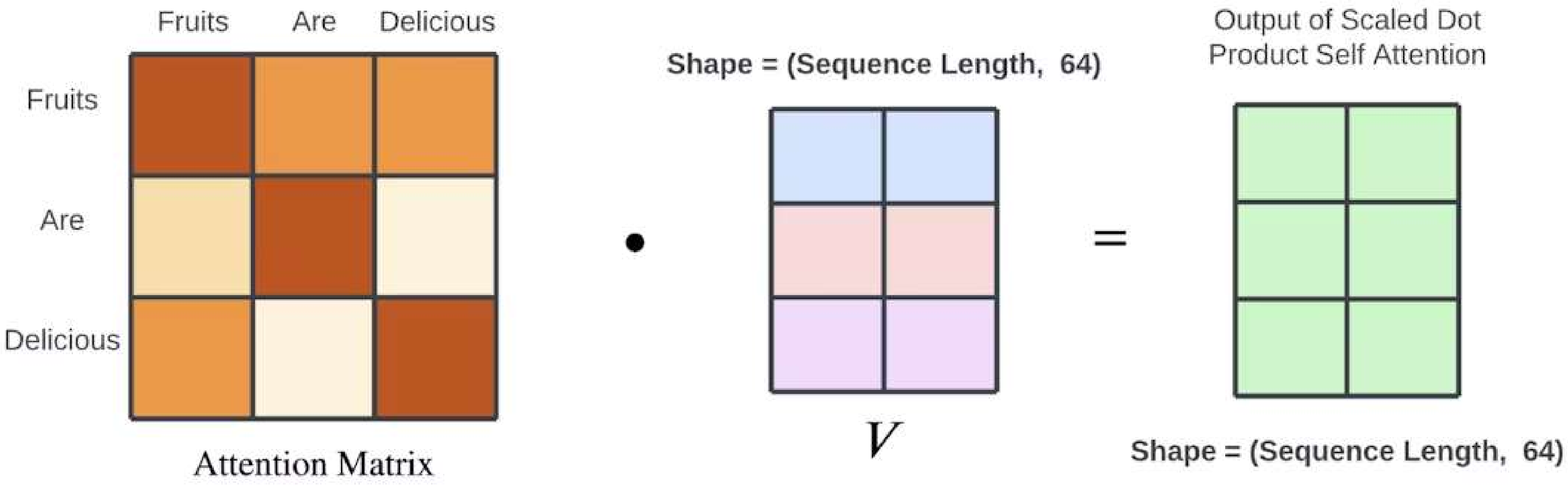



Shape = (d_model, 64)
Shape = (Sequence Length, 64)
Fruits
Are
Delicious
•
=
Shape = (Sequence Length, d_model)
W^Q
Q
Shape = (d_model, 64)
Shape = (Sequence Length, 64)
Fruits
Are
Delicious
•
=
Shape = (Sequence Length, d_model)
W^K
K
Shape = (d_model, 64)
Shape = (Sequence Length, 64)
Fruits
Are
Delicious
•
=
Shape = (Sequence Length, d_model)
W^V
V

Q K to Attention Matrix



Output of self-attention



Fruits Are Delicious

Fruits

Are

Delicious

Attention Matrix

•

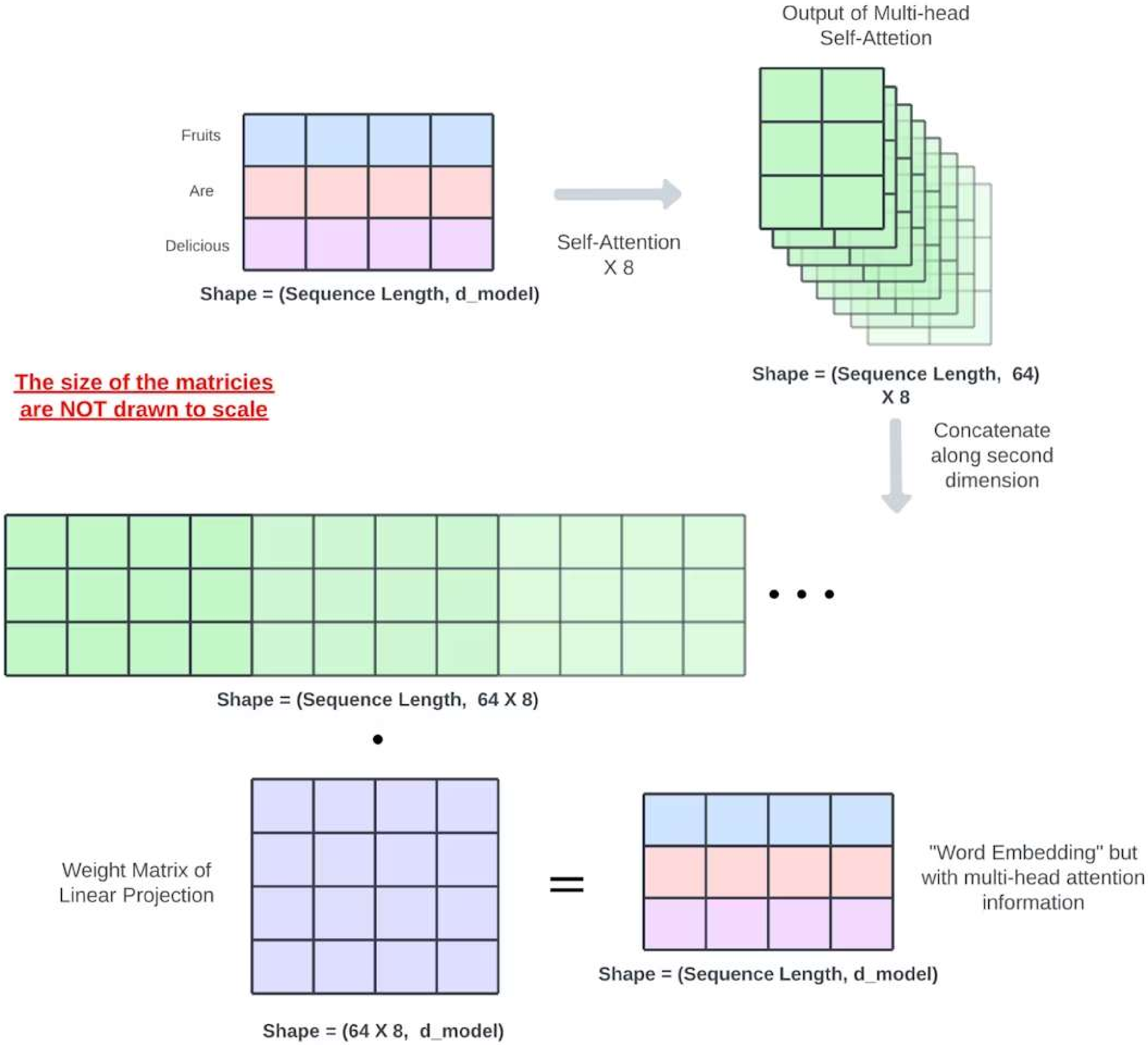
Shape = (Sequence Length, 64)

V

=

Output of Scaled Dot Product Self Attention

Shape = (Sequence Length, 64)



Output of Multi-head Self-Attetion

Fruits

Are

Delicious

Shape = (Sequence Length, d_model)

Self-Attention X 8

Shape = (Sequence Length, 64) X 8

The size of the matricies are NOT drawn to scale

Concatenate along second dimension

Shape = (Sequence Length, 64 X 8)

Weight Matrix of Linear Projection

=

"Word Embedding" but with multi-head attention information

Shape = (64 X 8, d_model)

Shape = (Sequence Length, d_model)

BertViz

- https://colab.research.google.com/drive/1hXIQ77A4TYS4y3UthWF-Ci7V7vVUoxmQ?usp=sharing#scrollTo=-QnRteSLP0Hm

Attention mechanism: breakthrough in NLP

- To learn those embeddings and weights
    - Two words in a sentence are relevant to each other,
        - word vectors will be aligned.
        - And hence produce a higher attention score.
    - For words that are not relevant to each other,
        - the word vectors will not be aligned
        - and will produce a lower attention score.

What patterns does BERT learn

- Attention to next word
- Attention to previous word
- Attention to identical/related words
- Attention to identical/related words in another sentence
- Attention to other words predictive of word
  - Straw-----berries
- Attention to delimiters
  - word to SEP