AWS re:Invent

NOV. 29 – DEC. 3, 2021 | LAS VEGAS, NV

BOA312

Do computers dream of code? From human language to applications

Suman Debnath
Principal Developer Advocate
Amazon Web Services

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Let’s try this…

1. Take your mobile
2. Open any text editor/messenger
3. And type I <space>

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Let’s try this…

1. Take your mobile
2. Open any text editor/messenger
3. And type I <space>
4. You should see some prediction for the next word, select any of the words which you feel is most *appropriate*

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Let’s try this…

1. Take your mobile
2. Open any text editor/messenger
3. And type I <space>
4. You should see some prediction for the next word, select any of the words which you feel is most *appropriate*
5. Continue to select words from the predicted words for a few more times

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Natural language processing

- Natural language processing is a sub-field of {linguistics, computer science}
- Human-generated language is difficult for computers to understand and interpret
- In NLP, the goal is to make computers understand this complex language structure and retrieve meaningful pieces of information from it
- Some of the NLP use cases
  - Text classification
  - Speech recognition
  - Text summary
  - Topic modelling
  - Question answering

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Natural language processing

- Natural language processing is a sub-field of {linguistics, computer science}
- Human-generated language is difficult for computers to understand and interpret
- In NLP, the goal is to make computers understand this complex language structure and retrieve meaningful pieces of information from it
- Some of the NLP use cases
  - Text classification
  - Speech recognition
  - Text summary
  - Topic modelling
  - Question answering



Problem to solve?

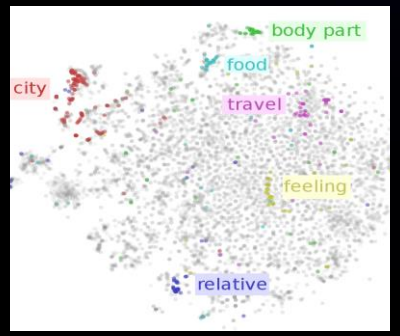
How can we build a mathematical representation of language that can help solve all these different use cases?

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

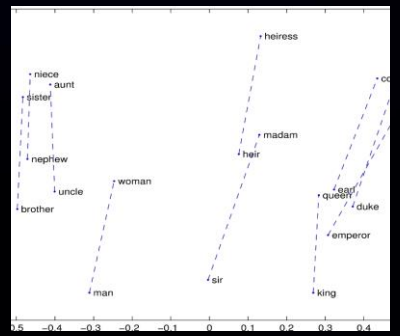
Evolution of NLP algorithms

Word2Vec (2013)



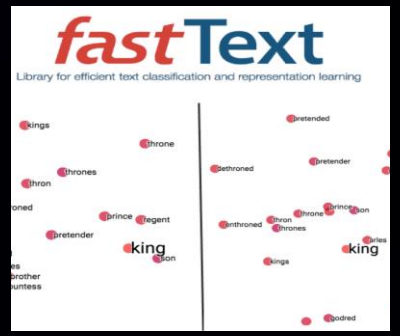
Simple NN predict the word based on the context window of other words in the sentence

GloVe (2014)



Global Vectors for Word Representation Matrix factorization

FastText (2015)

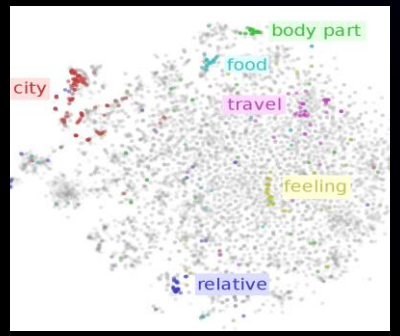


Extension of Word2Vec; each word is treated as a set of sub-words

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

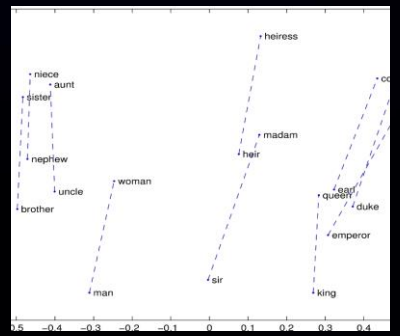
Evolution of NLP algorithms

Word2Vec (2013)



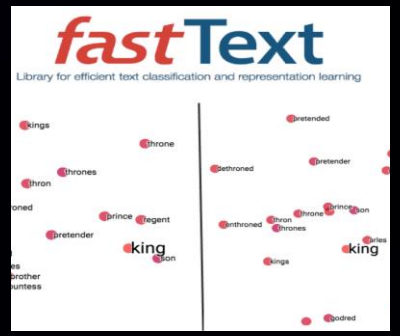
Simple NN predict the word based on the context window of other words in the sentence

GloVe (2014)



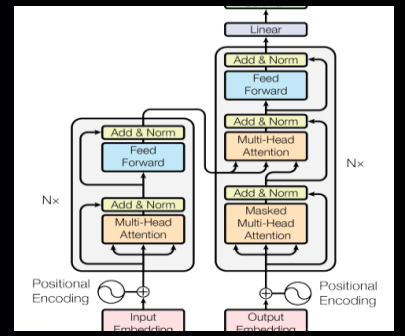
Global Vectors for Word Representation Matrix factorization

FastText (2015)

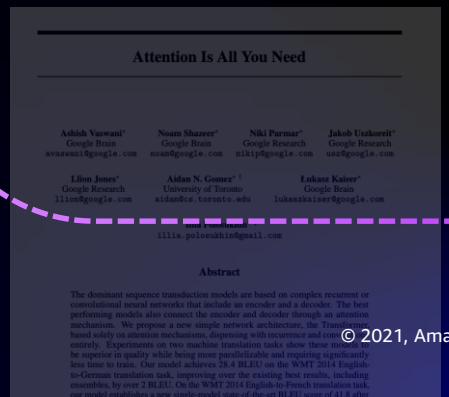


Extension of Word2Vec; each word is treated as a set of sub-words

Transformer (2017)



Attention is all you need

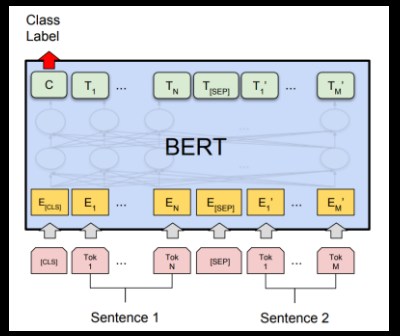

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Lukasz Kaiser* Google Brain lukaszkaiser@google.com
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after

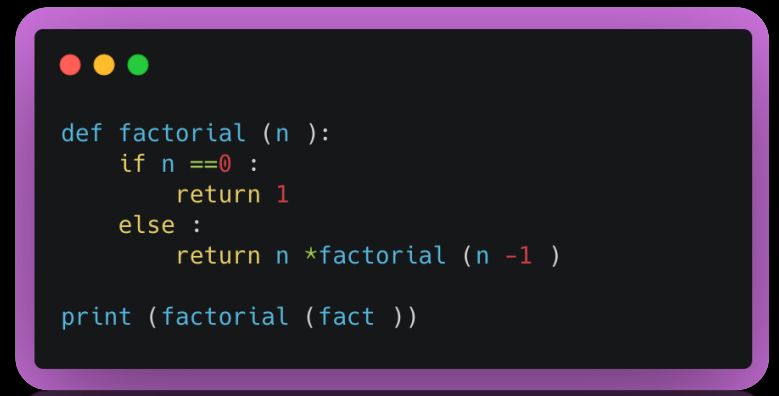
BERT (2018)



Pre-training of deep bidirectional transformers for language understanding

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

How the transformer works (code generator)



Encoder

Representation

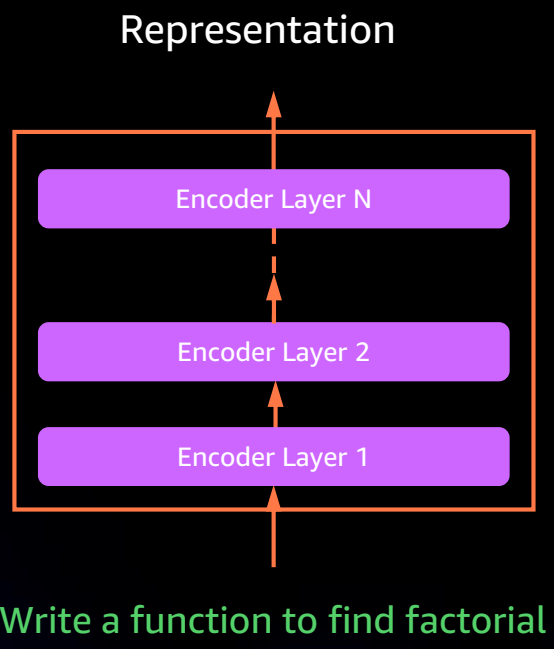
Decoder

Write a function to find factorial

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

The encoder of the transformer

- Stack of N number of encoders
- The output of one encoder is sent as input to the encoder above it



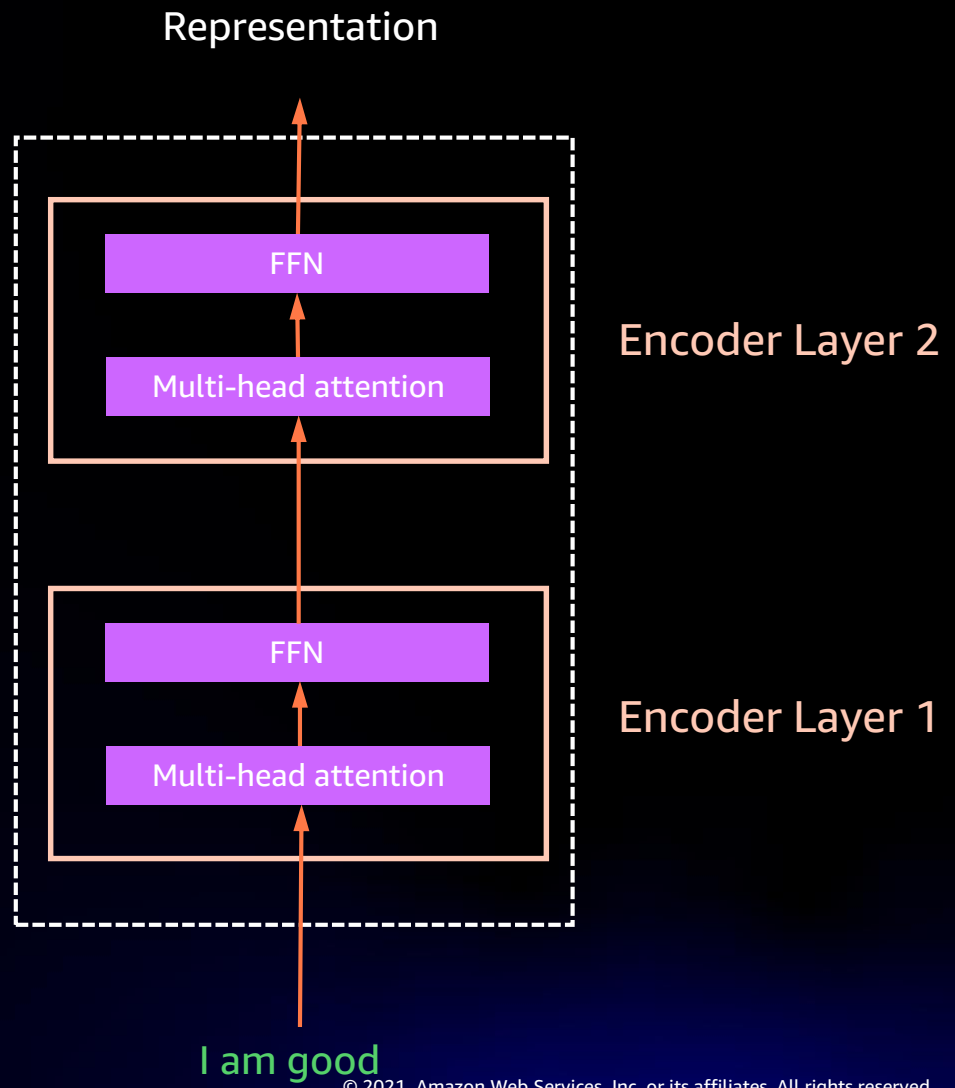
Questions?

- How exactly does the encoder work?
- How is it generating the representation for the given source sentence (input sentence)?

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

How exactly does the encoder work?

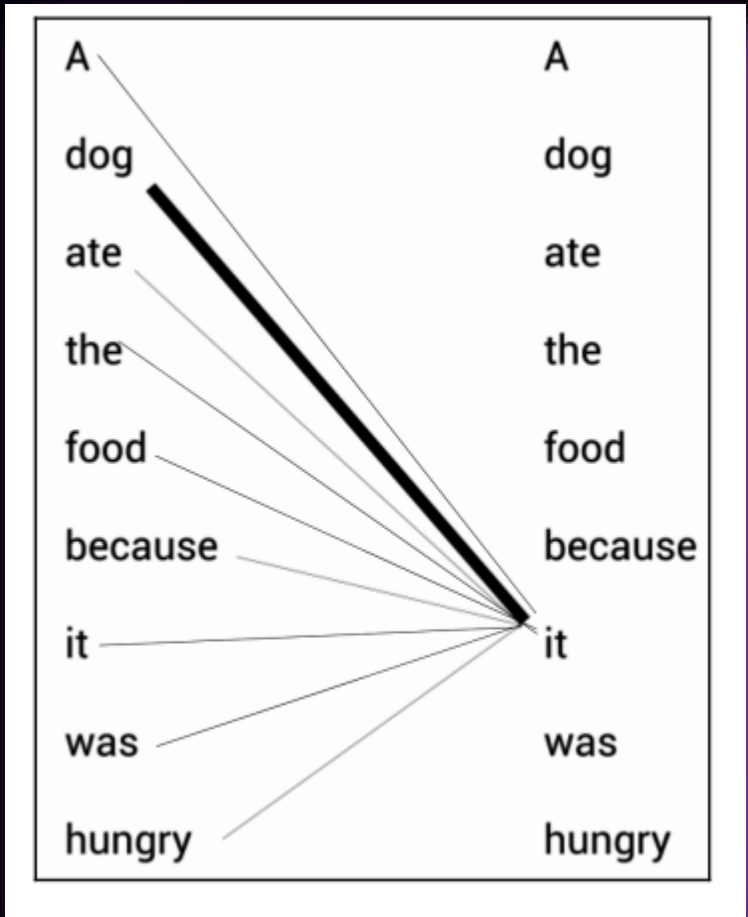
- All the encoder blocks are identical
- Each encoder block consists of 2 sublayers
  - Multi-head attention
  - Feedforward network (FNN)



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Self-attention mechanism

“A dog ate the food because it was hungry.”



How exactly does this work?

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

I am good

The embedding of the word *I* is :

x_1 = [1.76, 2.22, … ,6.66]

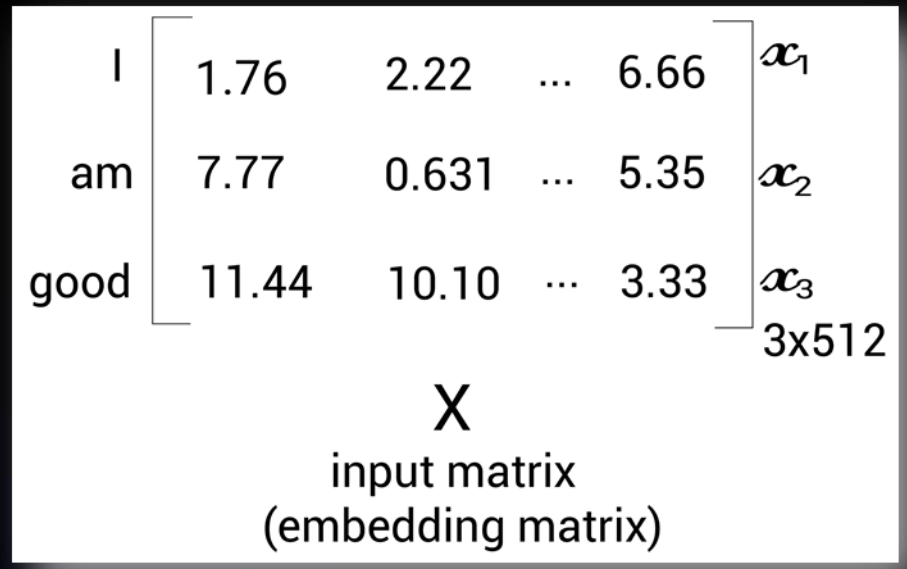
The embedding of the word *am* is :

x_2 = [7.77, 0.631, … ,5.35]

The embedding of the word *good* is :

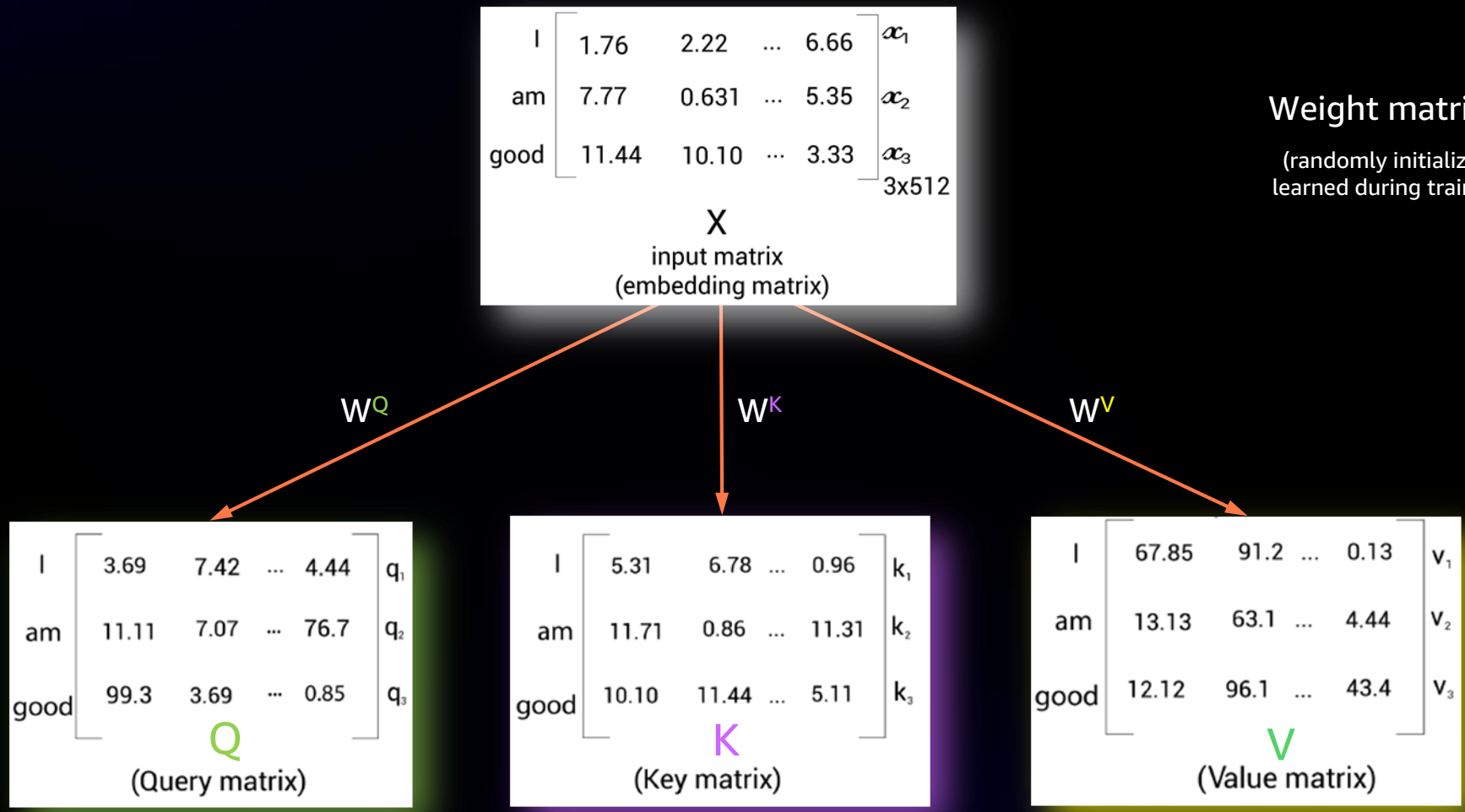
x_3 = [11.44, 10.10, … ,3.33]

*embedding dimension be 512



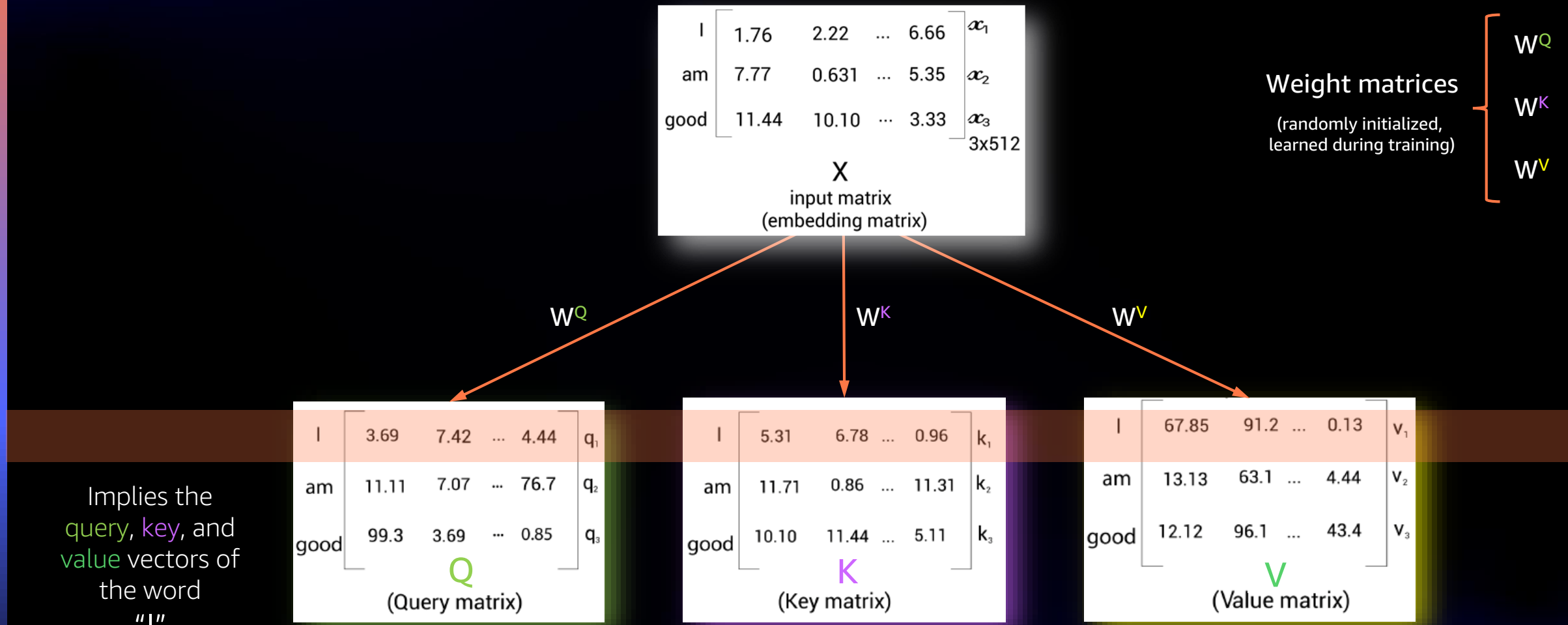
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

3 new matrices: {query, key, value}



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

3 new matrices: {query, key, value}



Implies the query, key, and value vectors of the word "I"

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Why are we computing this?
What is the use of query, key, and value matrices?

How is this going to help us?

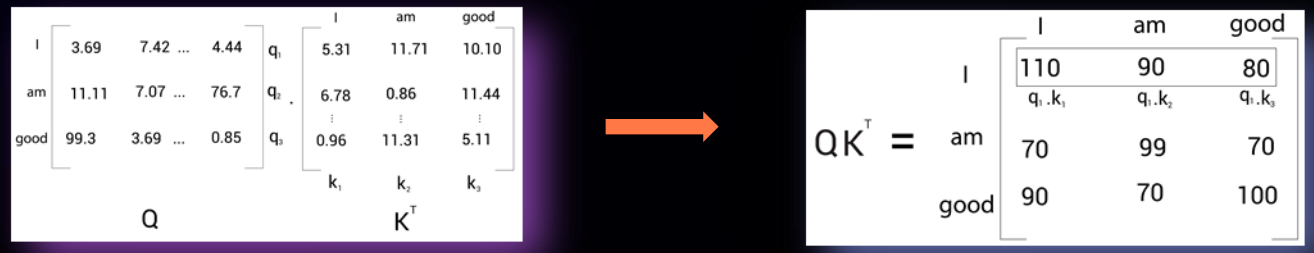
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

4-step process

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

4-step process

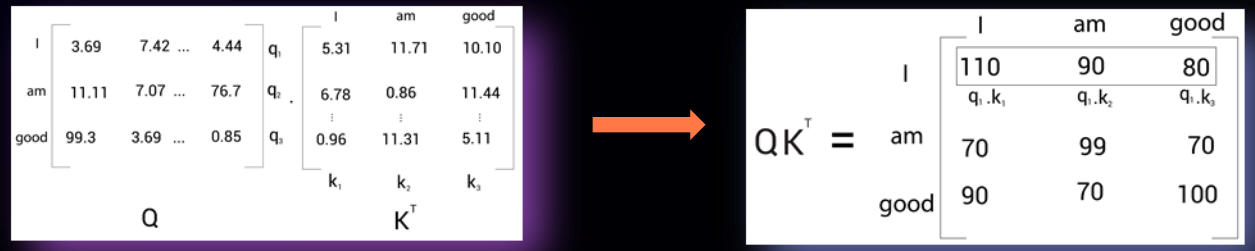
- Step 1: "Dot product" between the query matrix, Q, and the key matrix, K^T



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

4-step process

- Step 1: "Dot product" between the query matrix, Q, and the key matrix, K^T



- Step 2: "Divide" the matrix by the square root of the dimension of the key vector

$$\frac{QK^T}{\sqrt{d_K}} = \frac{QK^T}{8} =$$

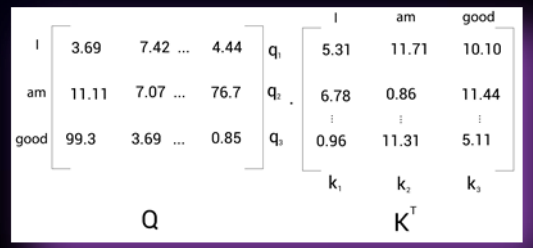
	I	am	good
I	13.75	11.25	10
am	8.75	12.375	8.75
good	11.25	8.75	12.5

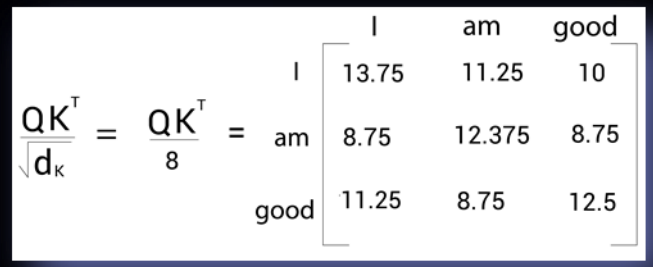
*the dimension of the key vector(d_k) is 64.

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

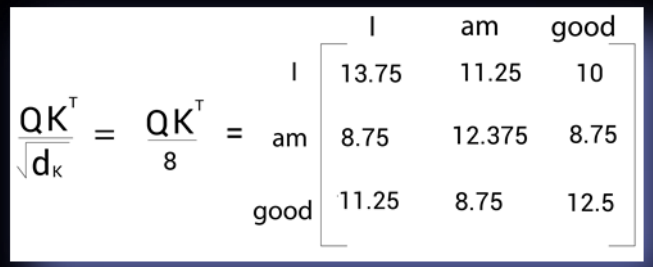
4-step process

- Step 1: "Dot product" between the query matrix, Q, and the key matrix, K^T



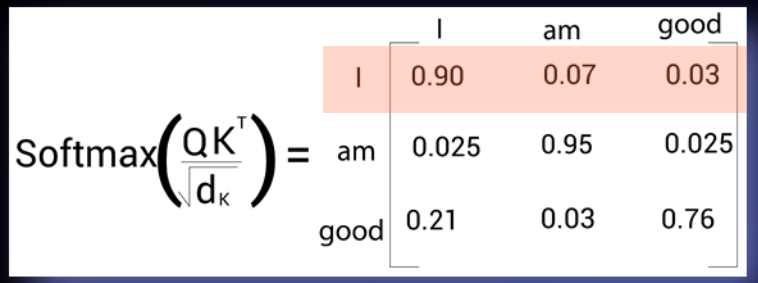


- Step 2: "Divide" the matrix by the square root of the dimension of the key vector



*the dimension of the key vector(d_k) is 64.

- Step 3: "Normalize" the matrix by the square root of the dimension of the key vector

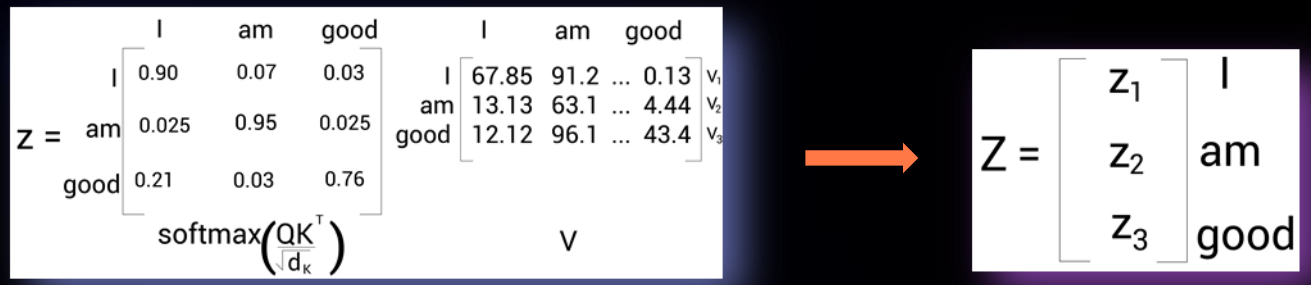


Word "I" is related to:
- itself by 90%
- am by 7%
- good by 3%

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

4-step process

- Step 4: final step in the self-attention mechanism is to compute the "attention matrix"



Self-attention of the word "I" is computed as the sum of the value vectors weighted by the scores

$$z_1 = 0.90\,[67.85\ 91.2\ \ldots]\,(v_1(I)) + 0.07\,[13.13\ 63.1\ \ldots]\,(v_2(am)) + 0.03\,[12.12\ 96.1\ \ldots]\,(v_3(good))$$

$$z_2 = 0.025\,[67.85\ 91.2\ \ldots]\,(v_1(I)) + 0.95\,[13.13\ 63.1\ \ldots]\,(v_2(am)) + 0.025\,[12.12\ 96.1\ \ldots]\,(v_3(good))$$

$$z_3 = 0.21\,[67.85\ 91.2\ \ldots]\,(v_1(I)) + 0.03\,[13.13\ 63.1\ \ldots]\,(v_2(am)) + 0.76\,[12.12\ 96.1\ \ldots]\,(v_3(good))$$

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Multi-head attention mechanism

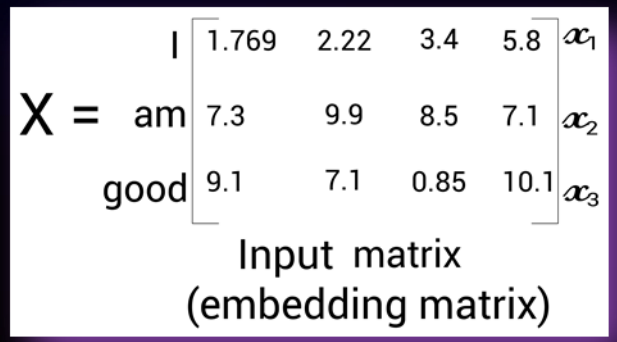
- Instead of having a *single* attention head, we can use *multiple* attention heads
- This will be useful only in circumstances where the meaning of the actual word is ambiguous, e.g.

"A dog ate the food because *it* was hungry."

$$\textit{Multi-head attention} = \textit{Concatenation}(Z_1, Z_2, Z_3, \ldots Z_i, \ldots, Z_8)\, W_0$$

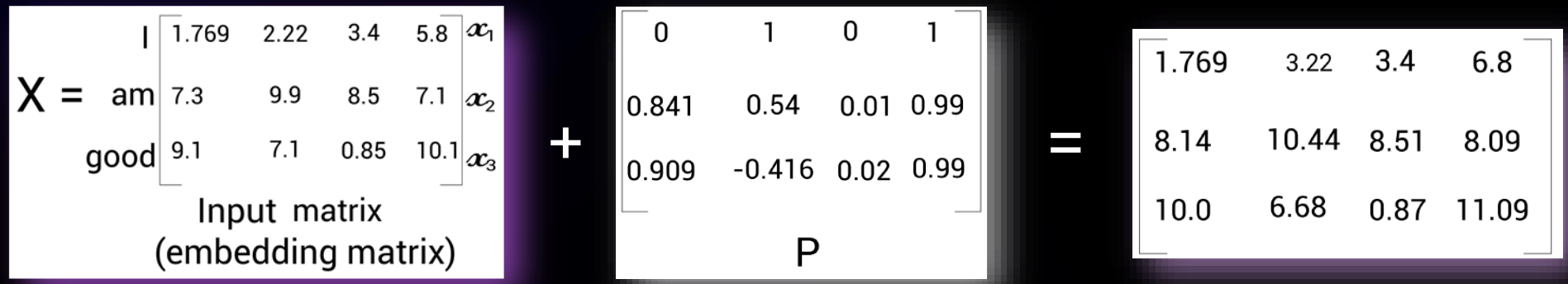
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Positional encoding



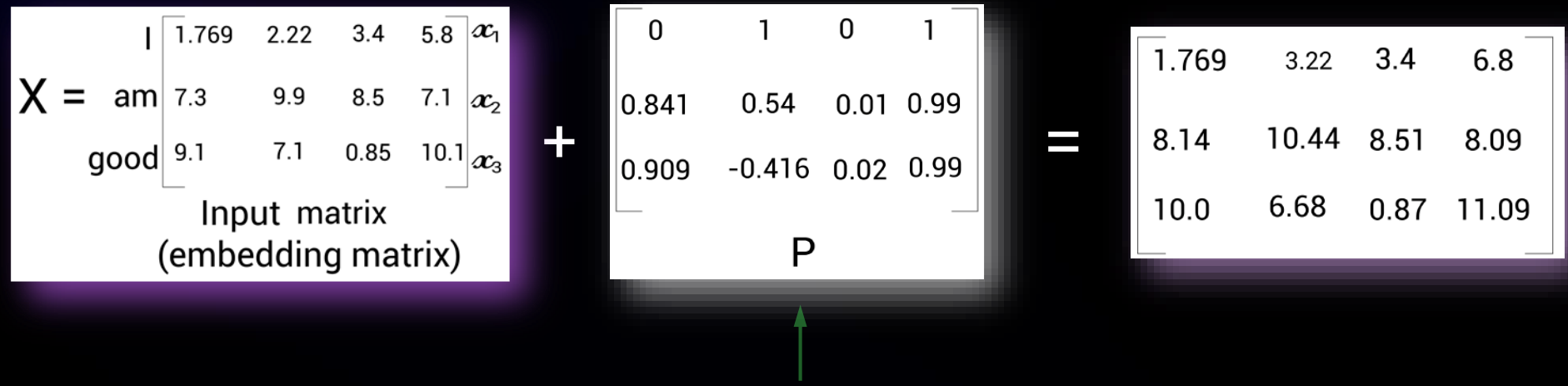
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Positional encoding



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Positional encoding



But how does it generate this?

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Positional encoding

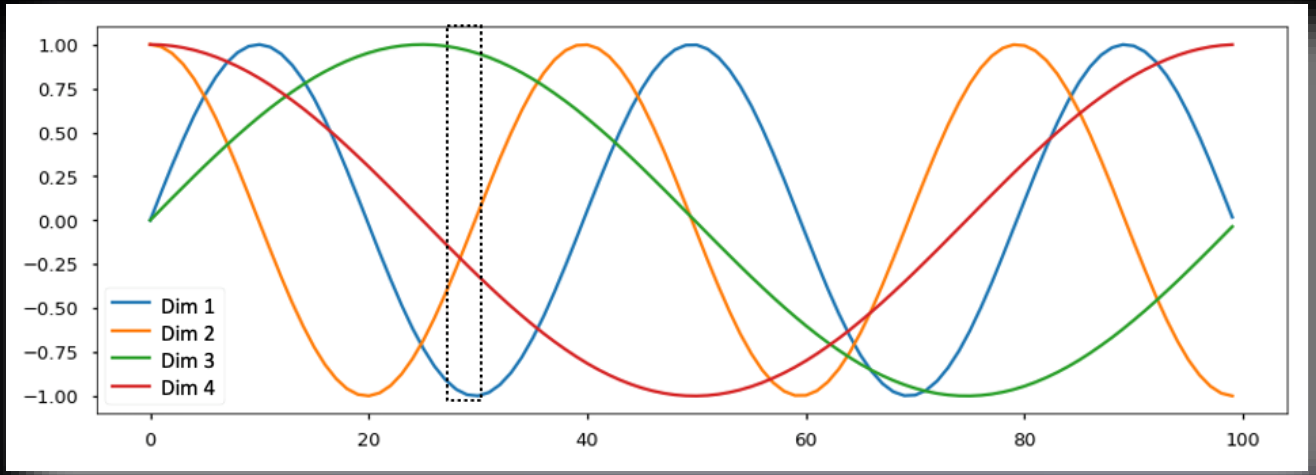
$$X = \begin{matrix} \\ I \\ am \\ good \end{matrix} \begin{bmatrix} 1.769 & 2.22 & 3.4 & 5.8 \\ 7.3 & 9.9 & 8.5 & 7.1 \\ 9.1 & 7.1 & 0.85 & 10.1 \end{bmatrix} \begin{matrix} x_1 \\ x_2 \\ x_3 \end{matrix}$$

Input matrix
(embedding matrix)

$$+ \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0.841 & 0.54 & 0.01 & 0.99 \\ 0.909 & -0.416 & 0.02 & 0.99 \end{bmatrix}$$

P

$$= \begin{bmatrix} 1.769 & 3.22 & 3.4 & 6.8 \\ 8.14 & 10.44 & 8.51 & 8.09 \\ 10.0 & 6.68 & 0.87 & 11.09 \end{bmatrix}$$



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Positional encoding

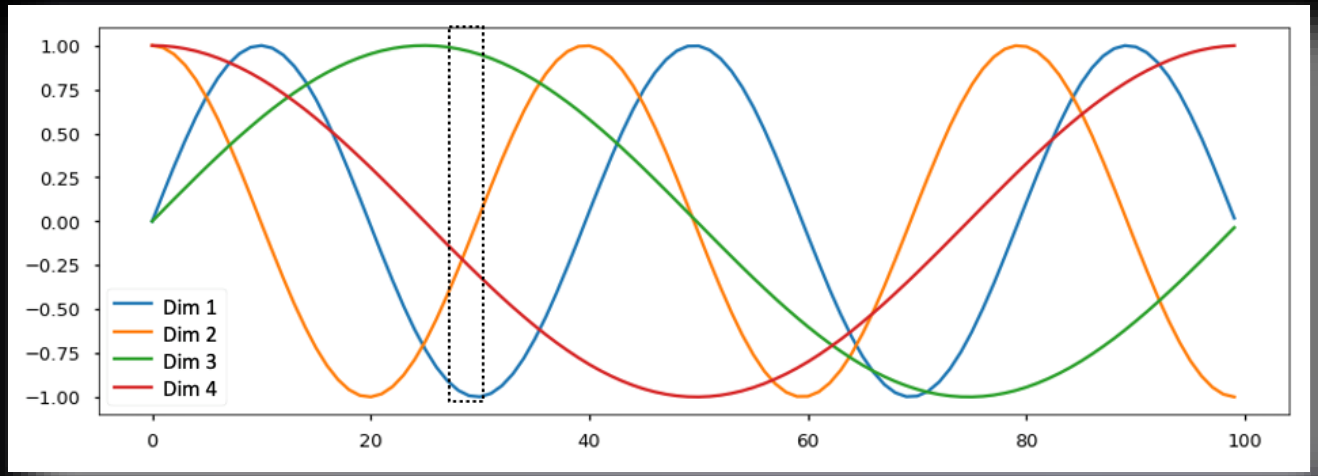
$$X = \begin{matrix} \\ I \\ am \\ good \end{matrix} \begin{bmatrix} 1.769 & 2.22 & 3.4 & 5.8 \\ 7.3 & 9.9 & 8.5 & 7.1 \\ 9.1 & 7.1 & 0.85 & 10.1 \end{bmatrix} \begin{matrix} \\ x_1 \\ x_2 \\ x_3 \end{matrix}$$

Input matrix
(embedding matrix)

$$+ \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0.841 & 0.54 & 0.01 & 0.99 \\ 0.909 & -0.416 & 0.02 & 0.99 \end{bmatrix}$$

P

$$= \begin{bmatrix} 1.769 & 3.22 & 3.4 & 6.8 \\ 8.14 & 10.44 & 8.51 & 8.09 \\ 10.0 & 6.68 & 0.87 & 11.09 \end{bmatrix}$$

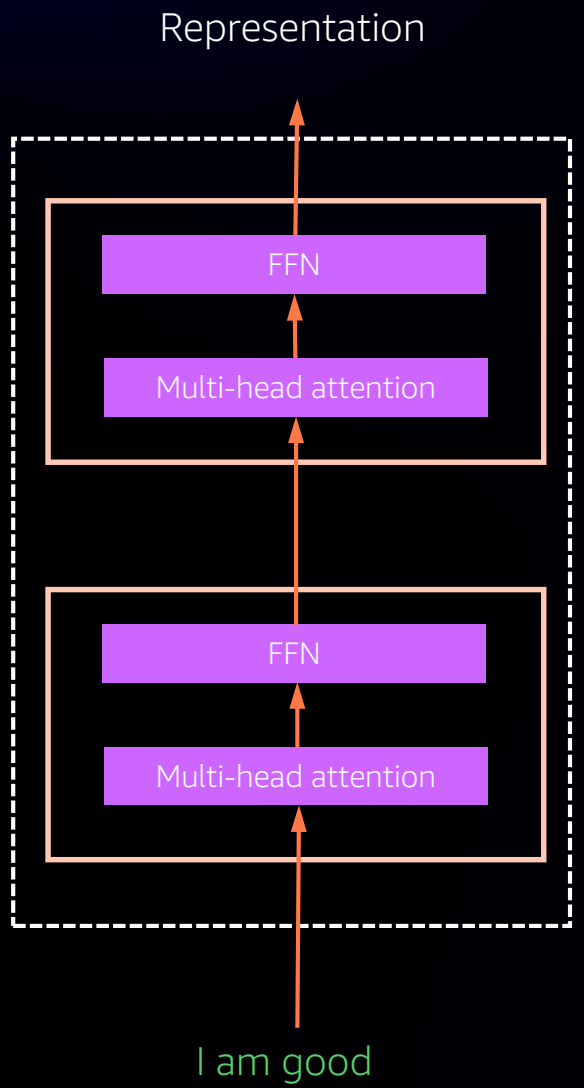


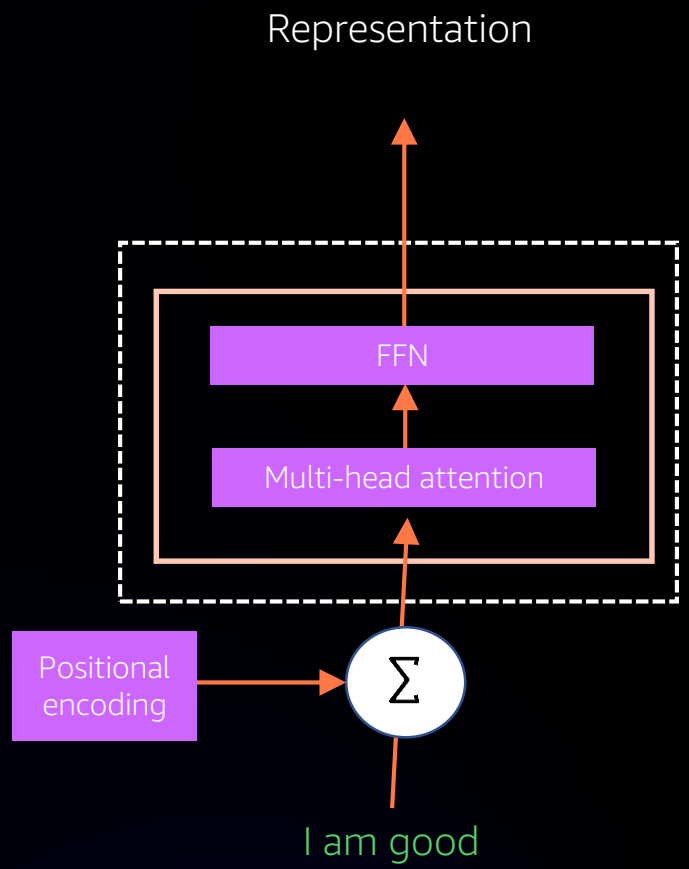
-1
-0.25
0.91
-0.25

Positional encoding for the 30th word

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

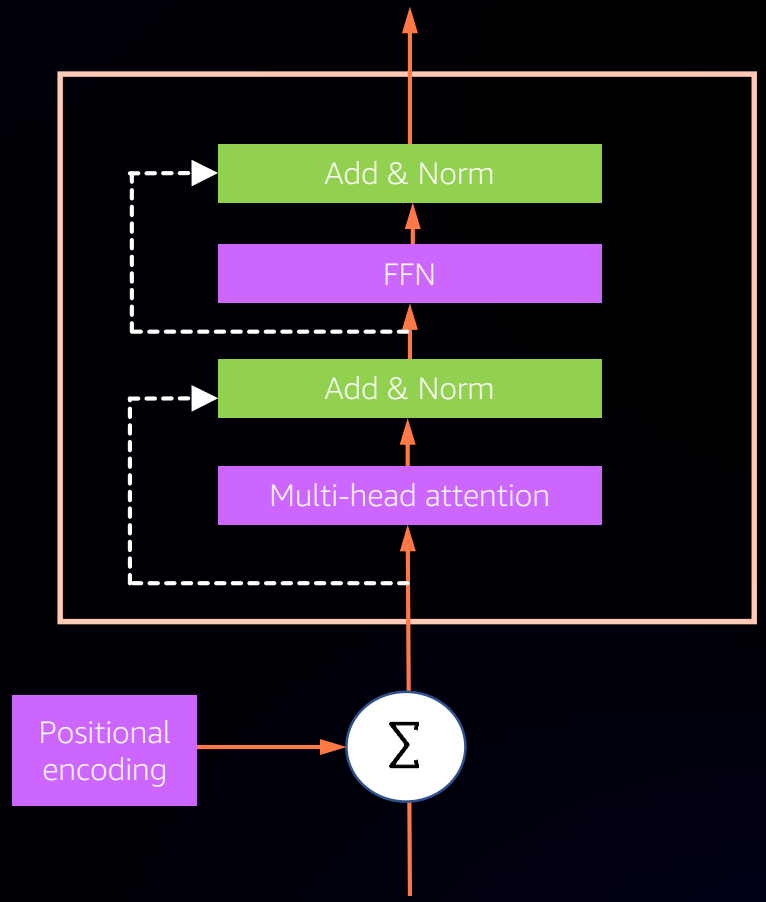
Let’s recall





© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Add and norm component



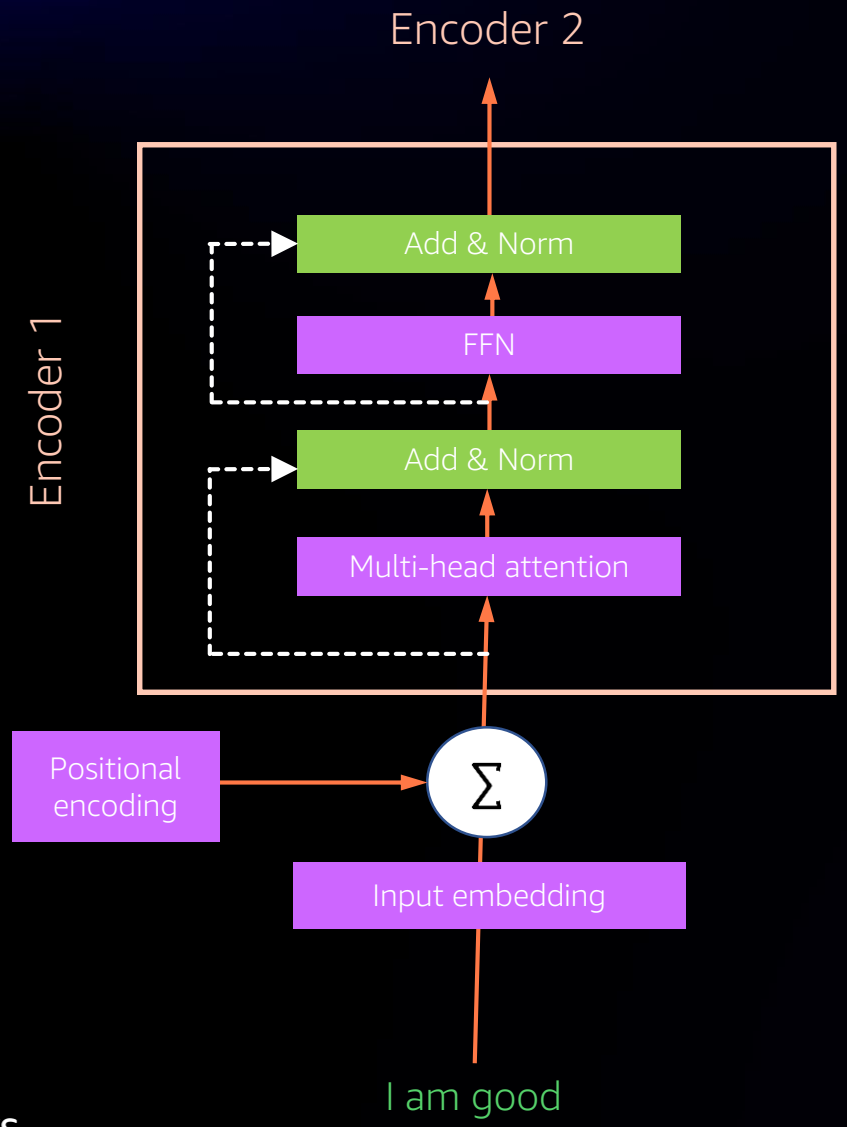
Encoder block

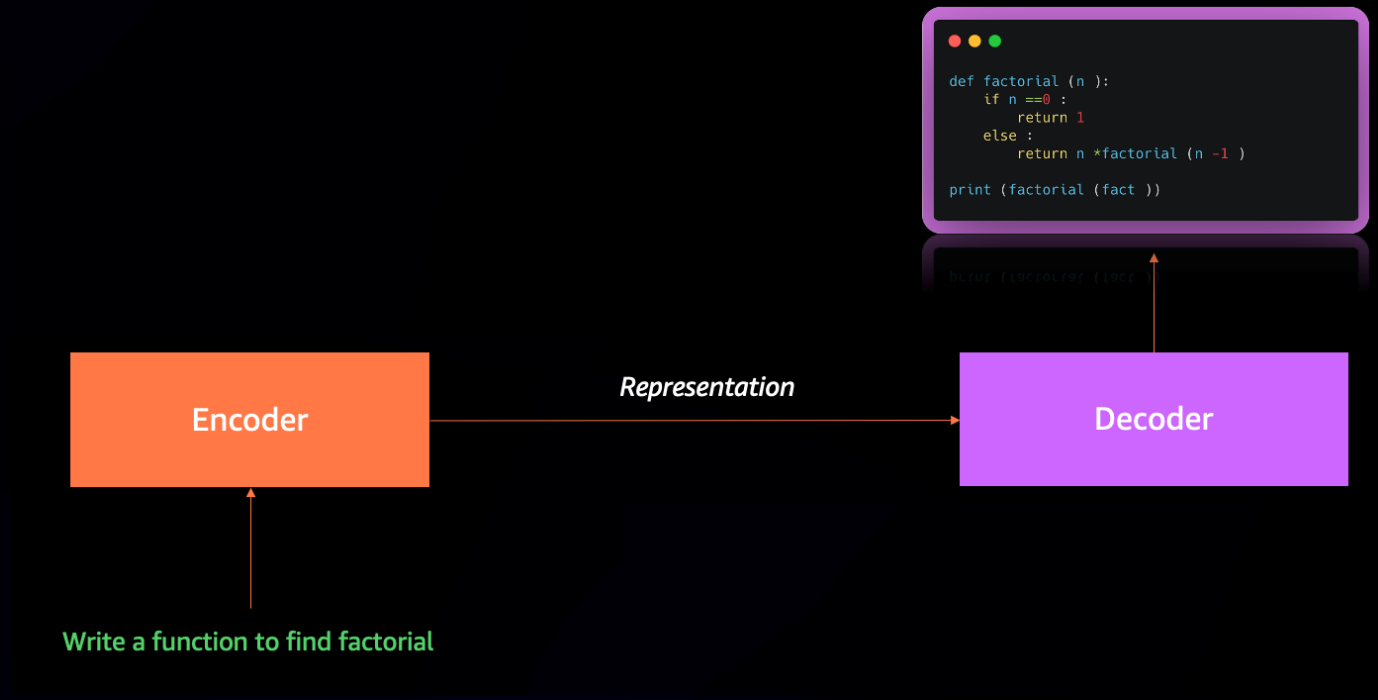
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Putting it all together



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.





© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Resources

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.