Applied Machine Learning for Business Analytics

Lecture 3: Transformers

Lecturer: Zhao Rui

Logistics

- HW schedule has been updated.
  - 3 -> 2
- HW1 has been released

Fri 01/31	From Word2Vec to Transformers	Link	Form your team & Assignment I Out
Fri 02/07	LLM and its Practices I	Link	LangChain Tutorial & Assignment I Due
Fri 02/14	LLM and its Practices II	Link	N.A.
Fri 02/21	LLM and its Practices III	Link	Build your First RAG & Assignment II Out
Sun 03/02	Recess Week	N.A.	Proposal Due
Fri 03/07	Data Preparation	Link	Assignment II Due

Jim Fan ✔
@DrJimFan

Jensen Huang is the new Taylor Swift

0:13 / 0:22

4:07 AM · Mar 19, 2024 from San Jose, CA · **511.2K** Views

US Economy

Nvidia

Taylor Swift

3

Transforming AI Panel at GTC 2024

Jensen Huang will host a panel with the authors of "*Attention Is All You Need*", a seminal research paper that introduced the Transformer neural network architecture (NeurIPS, 2017)

Jensen Huang
Founder and CEO
NVIDIA

Ashish Vaswani
Co-Founder & CEO
Essential AI

Noam Shazeer
CEO and Co-Founder
Character.AI

Niki Parmar
Co-Founder
Essential AI

Jakob Uskoreit
CEO
Inceptive

Llion Jones
Co-Founder and CTO
Sakana AI

Aidan Gomez
Co-Founder and CEO
Cohere

Lukasz Kaiser
Member of Technical Staff
OpenAI
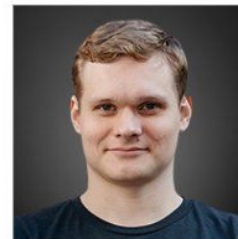
Illia Polosukhini
Co-Founder
NEAR Protocol

NVIDIA

4

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

Łukasz Kaiser[*]
Google Brain
lukaszkaiser@google.com

Illia Polosukhin[*] [‡]
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 training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

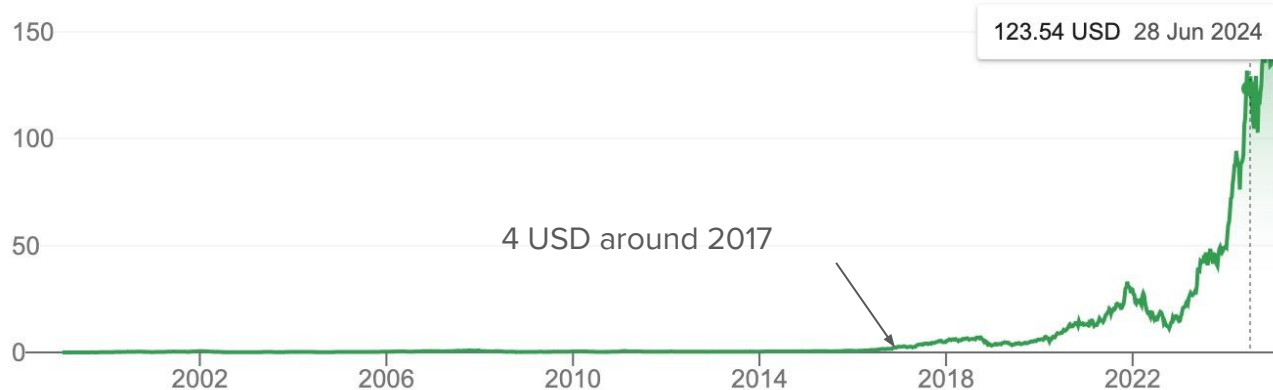Market Summary > NVIDIA Corp

128.86 USD

+128.82 (322,056.25%) ↑ all time

Closed: 28 Jan, 7:53 pm GMT-5 • Disclaimer
After hours 127.47 −1.39 (1.08%)

+ Follow

| 1D | 5D | 1M | 6M | YTD | 1Y | 5Y | **Max** |

123.54 USD 28 Jun 2024

4 USD around 2017

Open	121.81	Mkt cap	3.16T	52-wk high	153.13
High	129.00	P/E ratio	50.78	52-wk low	60.70
Low	116.25	Div yield	0.031%		

What happened to NVDA this WEEK?

Nvidia (NVDA) stock dropped nearly 17% Monday, leading a sell-off across chip stocks and the broader market after a new AI model from China's DeepSeek raised questions about AI investment and the rise of more cost-efficient artificial intelligence agents.

Nvidia's decline shaved $589 billion off the AI chipmaker's market cap, the largest single-day loss in stock market history.

Donald Trump: DeepSeek's AI should be 'wakeup call' to US tech industry

'The release of DeepSeek, AI from a Chinese company should be a wakeup call for our industries that we need to be laser-focused on competing to win'

DeepSeek is still on top of Transformer

performance comparable to OpenAI's o1 series models.

In this paper, we take the first step toward improving language model reasoning capabilities using pure reinforcement learning (RL). Our goal is to explore the potential of LLMs to develop reasoning capabilities without any supervised data, focusing on their self-evolution through a pure RL process. Specifically, we use DeepSeek-V3-Base as the base model and employ GRPO (Shao et al., 2024) as the RL framework to improve model performance in reasoning. During training, DeepSeek-R1-Zero naturally emerged with numerous powerful and interesting reasoning behaviors. After thousands of RL steps, DeepSeek-R1-Zero exhibits super performance on reasoning benchmarks. For instance, the pass@1 score on AIME 2024 increases from 15.6% to 71.0%, and with majority voting, the score further improves to 86.7%, matching the performance of OpenAI-o1-0912.

2.1 Basic Architecture

The basic architecture of DeepSeek-V3 is still within the Transformer (Vaswani et al., 2017) framework. For efficient inference and economical training, DeepSeek-V3 also adopts MLA and DeepSeekMoE, which have been thoroughly validated by DeepSeek-V2. Compared with DeepSeek-V2, an exception is that we additionally introduce an auxiliary-loss-free load balancing strategy (Wang et al., 2024a) for DeepSeekMoE to mitigate the performance degradation induced by the effort to ensure load balance. Figure 2 illustrates the basic architecture of DeepSeek-V3, and we will briefly review the details of MLA and DeepSeekMoE in this section.

DeepSeekR1:
https://arxiv.org/abs/2501.12948

DeepSeekv3:
https://arxiv.org/html/2412.19437v1

Agenda

1. Transformers
2. **Attention** is all you need
   a. Self-Attention
   b. Positional Embeddings
   c. Masked Self-Attention
   d. Encoder-Decoder Attention
3. Summary

1. Transformers

What is Transformer

Transformer is a sequence to sequence model (Encoder and Decoder)

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[*]
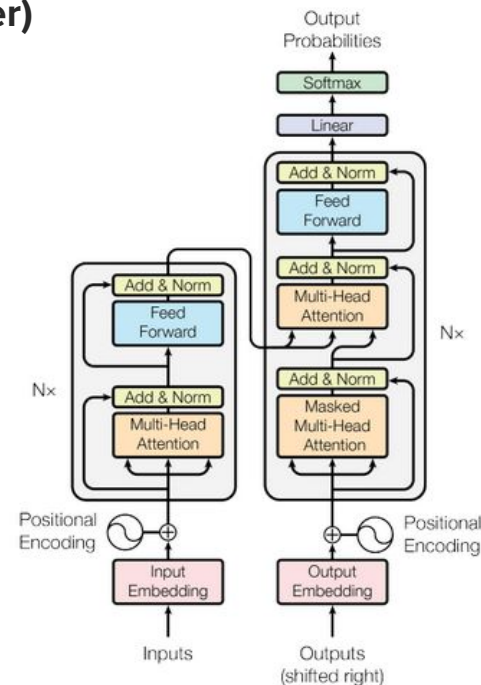Google Research
llion@google.com

Aidan N. Gomez[* †]
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser[*]
Google Brain
lukaszkaiser@google.com

Illia Polosukhin[* ‡]
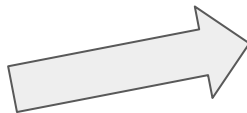illia.polosukhin@gmail.com

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 training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.
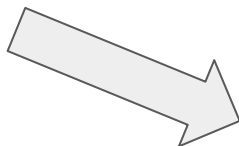
https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

11

All LLMs so far use transformer architecture
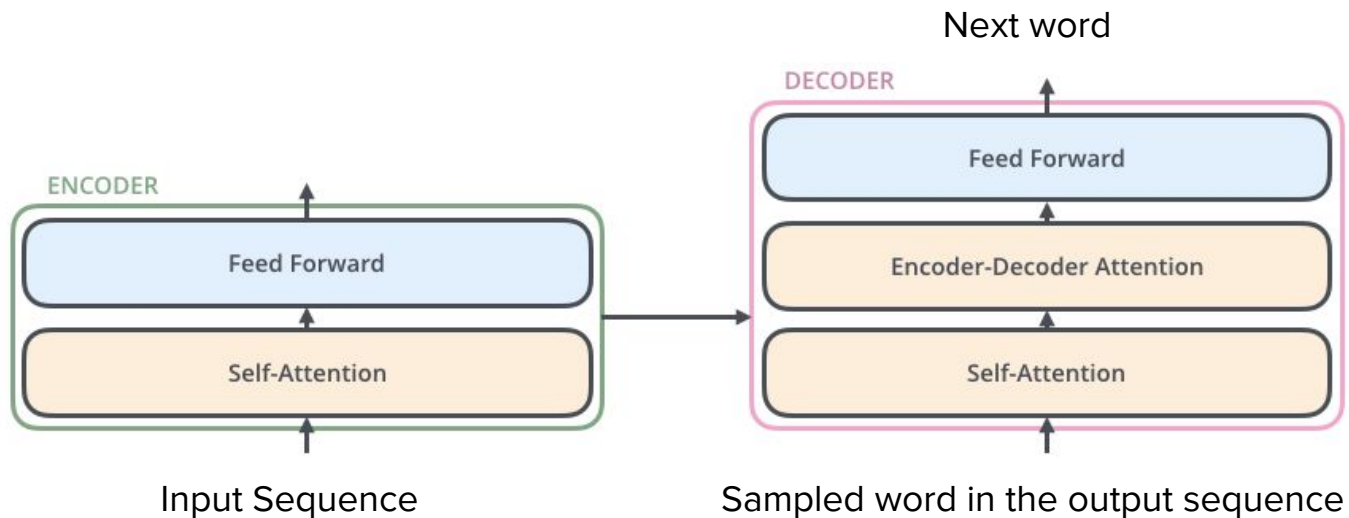
- BERT and GPT are the most representative ones

Bi-directional Encoder Representations from Transformers
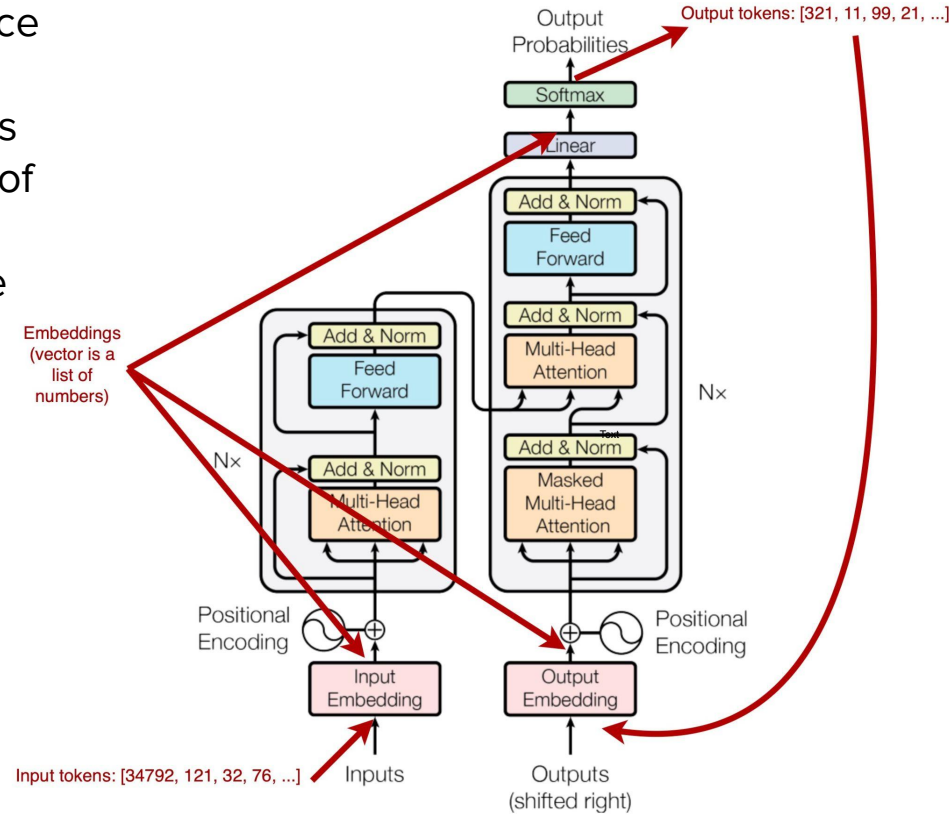
Generative Pre-trained Transformers

Transformer is solving Seq2Seq

Next word

DECODER

Feed Forward

Encoder-Decoder Attention

Self-Attention

ENCODER

Feed Forward

Self-Attention

Input Sequence

Sampled word in the output sequence

- Input text is encoded with tokenizers to sequence of integers
- Input tokens are mapped to sequence of vectors
- Output vectors can be classified to a sequence of tokens
- Output tokens can then be decoded back to the text

Output tokens: [321, 11, 99, 21, ...]

Output Probabilities

Softmax

Linear

Add & Norm

Feed Forward

Add & Norm

Multi-Head Attention

Add & Norm

Masked Multi-Head Attention

N×

Text

Embeddings (vector is a list of numbers)

Add & Norm

Feed Forward

Add & Norm

Multi-Head Attention

N×

Positional Encoding

Input Embedding

Positional Encoding

Output Embedding

Input tokens: [34792, 121, 32, 76, ...]
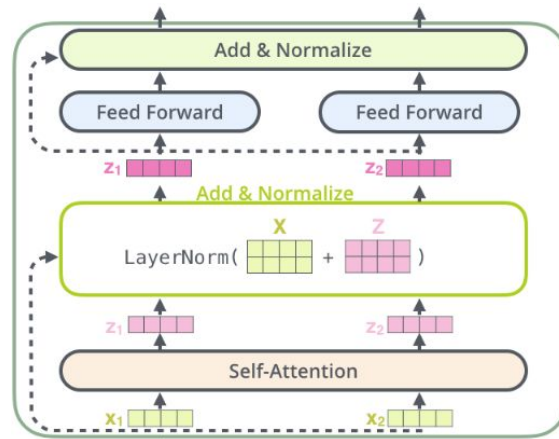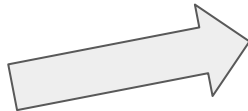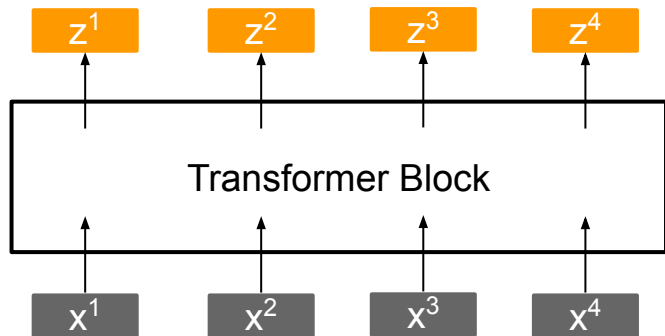
Inputs

Outputs (shifted right)

Transformer

1. Transformer block: operation unit
   a. Consists of multiple computations
   b. A sequence of embeddings in
   c. A sequence of embeddings out
2. Encoder:
   a. Stack 6 transformer blocks
   b. Learn representations for the input sequence
3. Decoder:
   a. Stack another 6 transformer blocks.
   b. Generate output sequences conditioned on the learned representations from encoder.
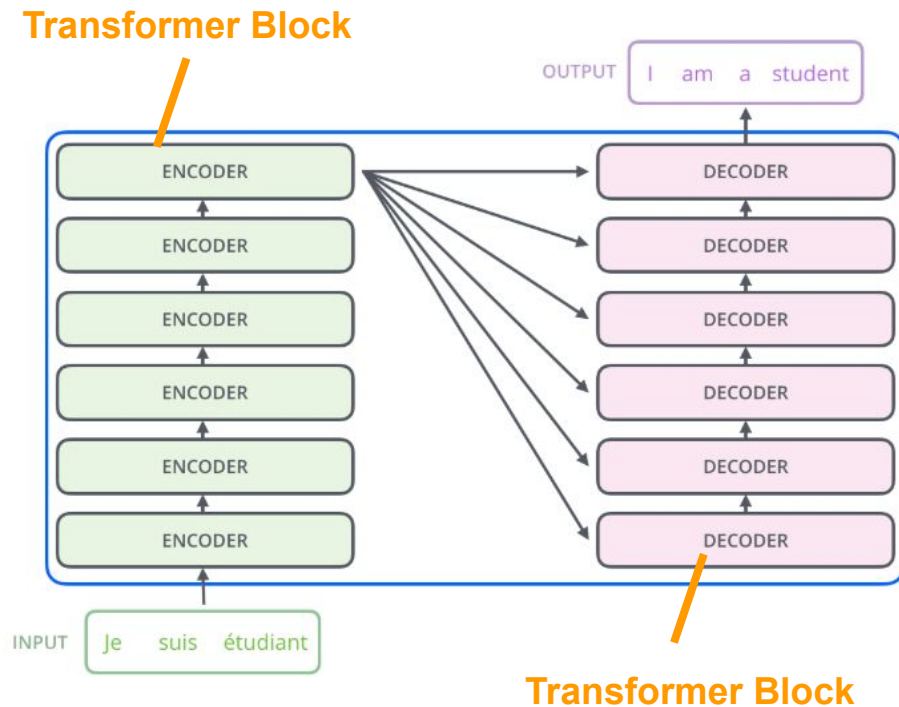
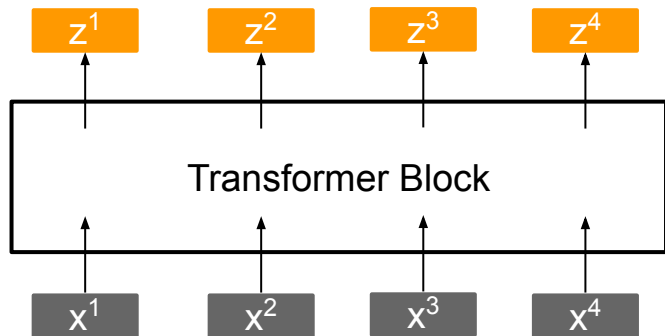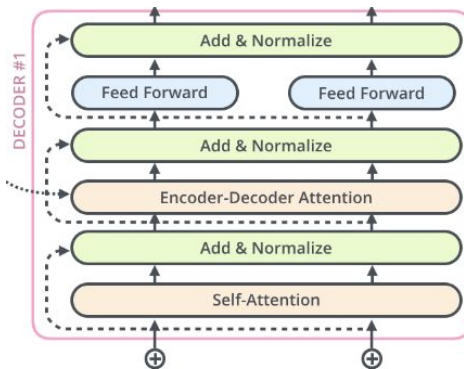# Transformer block in Encoder

1. Input: A sequence of vectors
2. Output: A sequence of vectors
3. Key Components:
   a. **Self-attention Layer**
   b. **Positional Embeddings**
   c. Residual and Normalization Layer
   d. Fully-connected Layer

Transformer block in Decoder

z^1 z^2 z^3 z^4

Transformer Block

x^1 x^2 x^3 x^4

Encoder outputs

DECODER #1

Add & Normalize

Feed Forward   Feed Forward

Add & Normalize

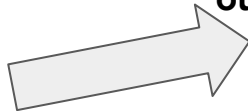Encoder-Decoder Attention

Add & Normalize

Self-Attention

1. Input: A sequence of vectors
2. Output: A sequence of vectors
3. Key Components:
   a. **Masked Self-attention Layer**
   b. Positional Embeddings
   c. **Encoder-Decoder Attention**
   d. Residual and Normalization Layer
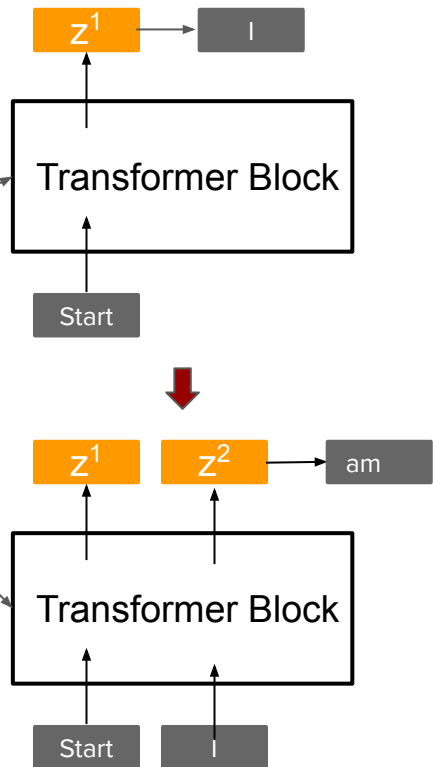   e. Fully-connected Layer

Decoding Process

- The decoder is **autoregressive**
  - Begins with a start token
  - Before the stop token is generated, repeat
    - Take the list of previous outputs with the encoder outputs that contain the attention information from the input
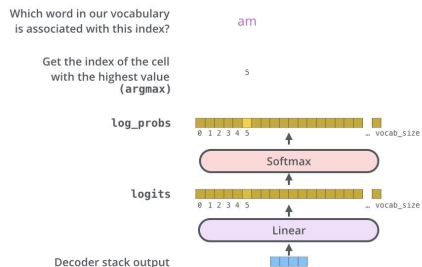   - Generate the current output

Decoding Process

19

Decoding Process

- Linear Classifier with Final softmax for output probabilities
  - The output of the classifier would be the size of vocabulary
  - After softmax, probability scores between 0 and 1 will be generated
  - Decoding Strategies:
    - Greedy search: the index of the highest probability score would be taken to predict the current word
    - Beam search: takes into account the N most likely tokens
    - Other advanced sampling: https://deci.ai/blog/from-top-k-to-beam-search-llm-decoding-strategies/

Which word in our vocabulary
is associated with this index?   am

Get the index of the cell
with the highest value
(argmax)   5

log_probs   ▭▭▭▭▭▭▭▭▭▭▭ ▭ … vocab_size
            0 1 2 3 4 5
               ┌─────────────┐
               │   Softmax   │
               └─────────────┘
logits   ▭▭▭▭▭▭▭▭▭▭▭ ▭ … vocab_size
            0 1 2 3 4 5
               ┌─────────────┐
               │   Linear   │
               └─────────────┘

Decoder stack output   ▭▭▭▭

This figure starts from the bottom with the vector produced as the output of the decoder stack. It is then turned into an output word.

Temperature Sampling

$$p_i = \text{Softmax}(z_i) = \frac{\exp(z_i)}{\sum_{j=1}^{n} \exp(z_j)}$$
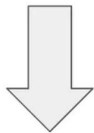
$$p_i(T) = \text{Softmax}(z_i/T) = \frac{\exp(z_i/T)}{\sum_{j=1}^{n} \exp(z_j/T)}$$

Temperature is a scaling function applied to softmax inputs

Mode

Chat Beta

Model

gpt-4

Controls randomness: Lowering results in less random completions. As the temperature approaches zero, the model will become deterministic and repetitive.

Temperature   2

Maximum length   2048

Top P   1

Frequency penalty   0

Presence penalty   0

Temperature Sampling
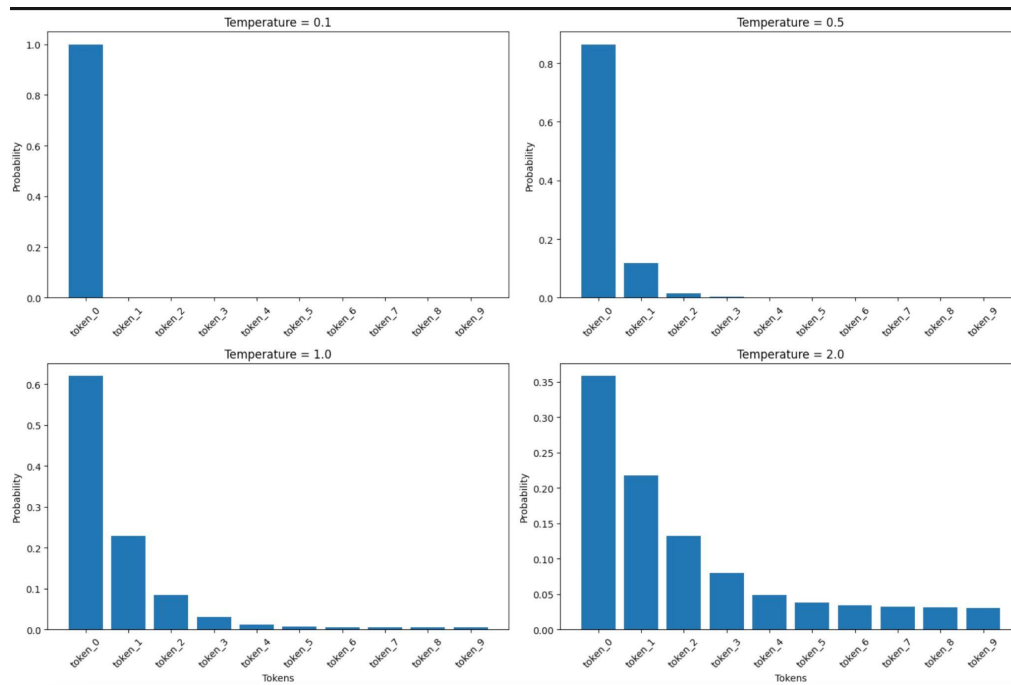
Encoder-Decoder
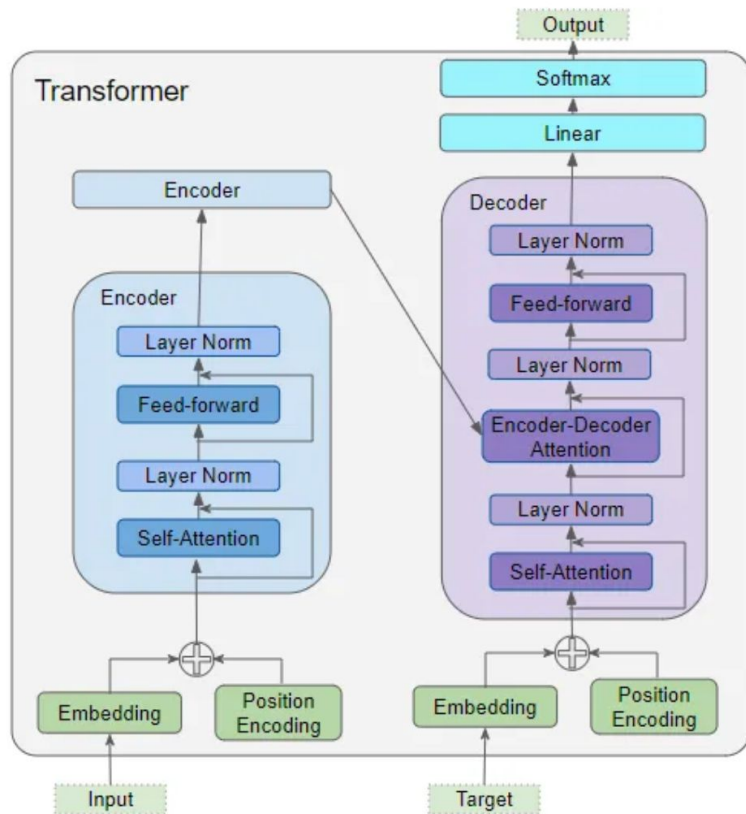
Three kinds of attention in transformers:
- Self-attention
  - Input sequence <> Input sequence
- Masked self-attention
  - Previous steps in output sequence <> current steps in output sequence
- Encoder-Decoder attention
  - Input sequence <> Output sequence

Source:
https://towardsdatascience.com/transformers-explained-visually-part-2-how-it-works-step-by-step-b49fa4a64f34

23

2. Attention

Word Embeddings

- **Apple** in two sentences:
  - Sentence 1: My favorite fruit is **apple**
  - Sentence 2: Solution: My favorite brand is **apple**

apple

One embedding has multiple senses

Word embeddings' space

Contextualized Word Embeddings

- Telling context in words
  - Sentence 1: My favorite fruit is **apple1**
  - Sentence 2: Solution: My favorite brand is **apple2**

Word embeddings' space

From nearby words, we can guess two different meanings of this word (i.e., food and brand)
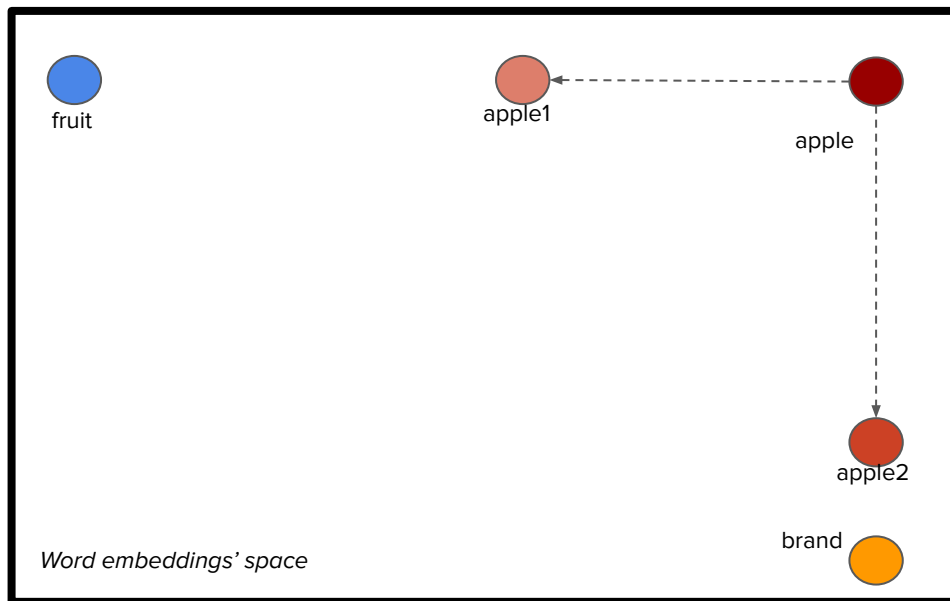
Contextualized Word Embeddings

- Telling context in words
  - Sentence 1: My favorite fruit is **apple1**
  - Sentence 2: Solution: My favorite brand is **apple2**

Word embeddings' space

1. In sentence 1, move the word embedding of apple towards the word "fruit"
2. In sentence 2, move the word embedding of apple toward the word "brand"

   This is how attention will work

How to move one word closer to another one

- Average two words
  - apple1 = 0.8*apple + 0.2*fruit   = 0.8*[5,7] + 0.2*[0,8]= [4.0, 7.2]
  - apple2 = 0.9*apple + 0.1*brand = 0.9*[5,7] + 0.1*[6,0]= [5.1, 6.3]

Similarity/Attention

Embeddings
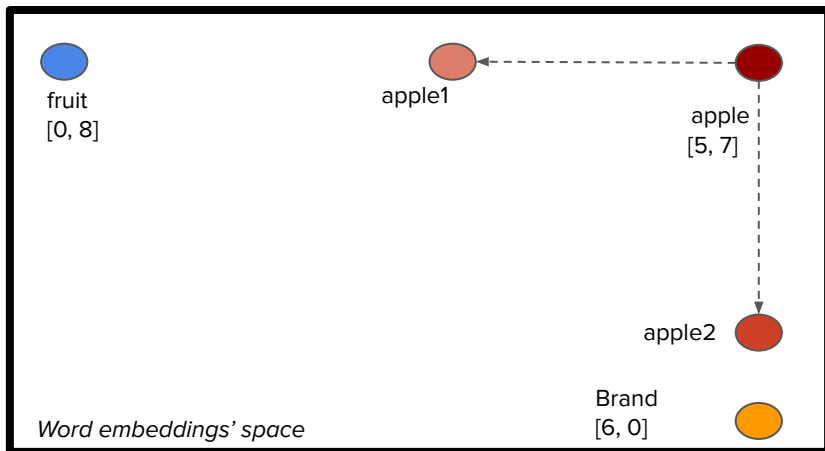
Attention mechanism is able to learn multiple embeddings for the same word in multiple sentences

fruit
[0, 8]

apple1

apple
[5, 7]

apple2

Brand
[6, 0]

Word embeddings' space

How to derive similarity

- **Apple** in two sentences:
  - Sentence 1: My favorite fruit is **apple**
  - Sentence 2: My favorite brand is **apple**
- Why we move apple to fruit in sentence 1? Instead of other words as "my" and "is"
- It is based on the similarity!
- Assume every word has its own base vector (as word2vec), the contextualized word embedding of **apple** in the sentence: my favorite fruit is apple

= Attention(apple, my)*base_vec(my) + Attention(apple, favorite)*base_vec(favorite) + Attention(apple, fruit)*base_vec(fruit) + Attention(apple, is)*base_vec(is) + Attention(apple, apple)*base_vec(apple)

Target word   **Context word**
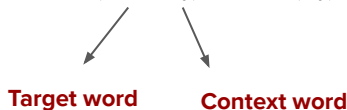
How to derive similarity

- Via embeddings, the similarity between two irrelevant words would be zero, while the similarity between the related pair would be high

	my	favourite	fruit	is	apple
my	1	0	0	0	0
favourite	0	1	0	0	0
fruit	0	0	1	0	0.25
is	0	0	0	1	0
apple	0	0	0.25	0	1

	my	favourite	brand	is	apple
my	1	0	0	0	0
favourite	0	1	0	0	0
brand	0	0	1	0	0.11
is	0	0	0	1	0
apple	0	0	0.11	0	1

How to derive similarity

- The diagonal entries are all 1
- The similarity between any irrelevant words is 0 (for simplicity)
- The similarity between apple and fruit is 0.25 while the one between apple and brand is 0.11 considering apple is used more often in the same context as fruit

	my	favourite	fruit	is	apple
my	1	0	0	0	0
favourite	0	1	0	0	0
fruit	0	0	1	0	0.25
is	0	0	0	1	0
apple	0	0	0.25	0	1

	my	favourite	brand	is	apple
my	1	0	0	0	0
favourite	0	1	0	0	0
brand	0	0	1	0	0.11
is	0	0	0	1	0
apple	0	0	0.11	0	1

How to derive similarity

- Contextualized Target Word = The sum of a product between the similarity between target word and context word * context word embeddings
- We should also normalize the similarity along the sentence (**softmax**)
- Therefore
  - my (in the sentence 1) = my
  - apple (in the sentence 1) = 0.2 * fruit + 0.8 * apple
  - apple (in the sentence 2) = ?

	my	favourite	fruit	is	apple
my	1	0	0	0	0
favourite	0	1	0	0	0
fruit	0	0	0.8	0	0.2
is	0	0	0	1	0
apple	0	0	0.2	0	0.8

Normalized

	my	favourite	brand	is	apple
my	1	0	0	0	0
favourite	0	1	0	0	0
brand	0	0	0.9	0	0.9
is	0	0	0	1	0
apple	0	0	0.1	0	0.1

Contextualized Word Embeddings

- Transformers is proposed to learn better feature for NLP data
- The core layer is self-attention layer which can map a sequence of word embeddings to another sequence of word embeddings which is contextualized

Encoded Embeddings

Contextualized Word Embedding

money   in   the   bank

Contextualized Word Embedding

along   the   river   bank

Contextualized Word Embedding

blood   bank   collects   blood

**Different Contexts,
Different Encoded Embeddings for bank.**

33
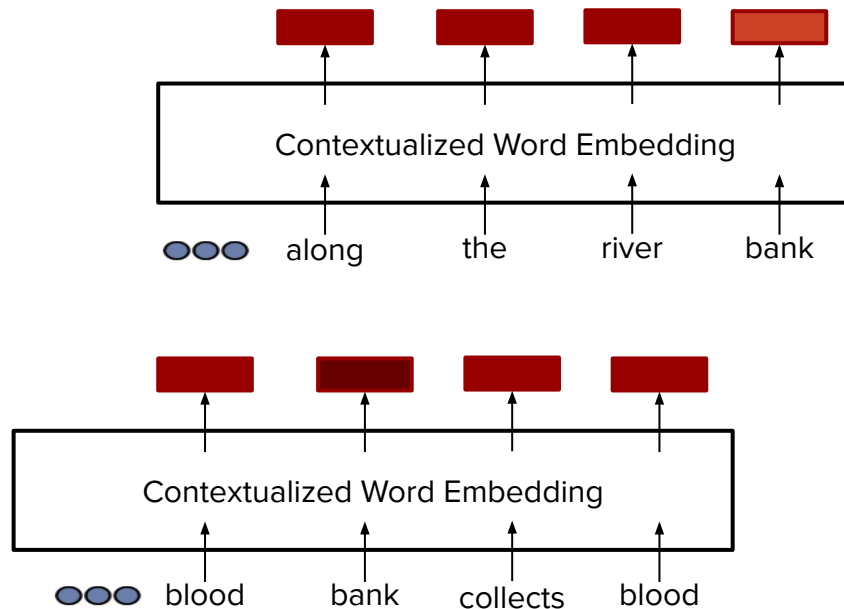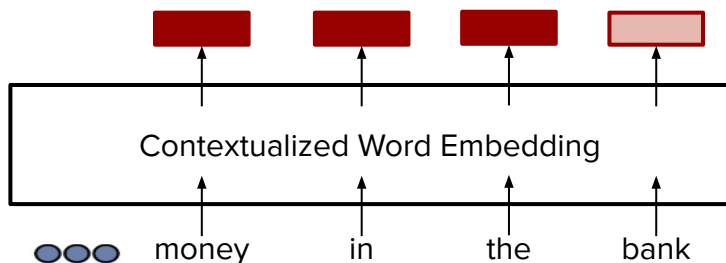
What is BERT

- **B**idirectional **E**ncoder **R**epresentations from **T**ransformers (**BERT**)
- BERT: Encoder of Transformer,

BERT

along   the   river   bank
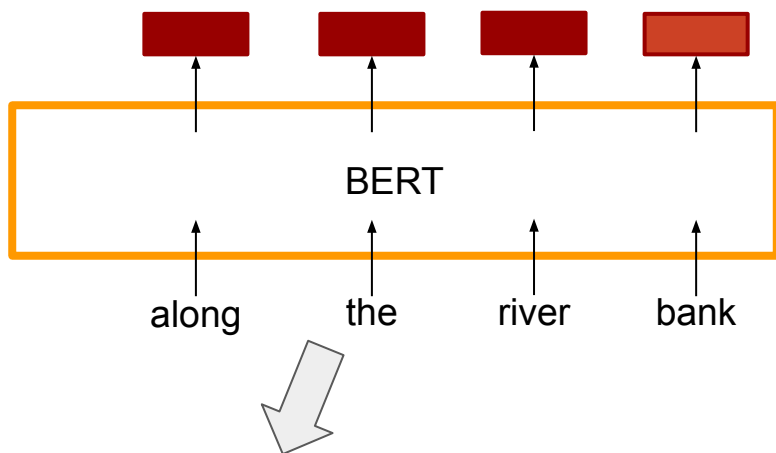
Given a sequence of words, generate a sequence of vectors and then can be used for various NLP tasks

Transformer
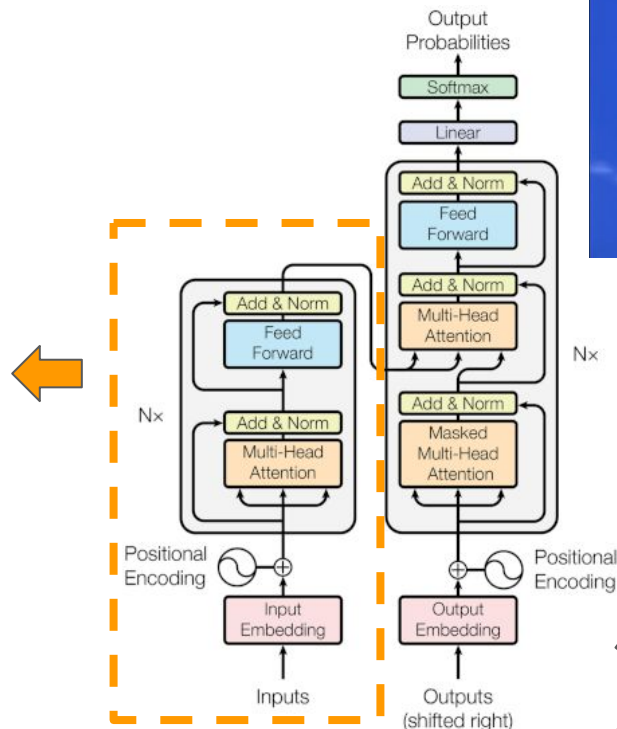
Solve Seq2Seq Task

Contextualized Word Embeddings

Encoded Embeddings

Contextualized Word Embedding

money   in   the   bank

Contextualized Word Embedding

along   the   river   bank

Contextualized Word Embedding

blood   bank   collects   blood
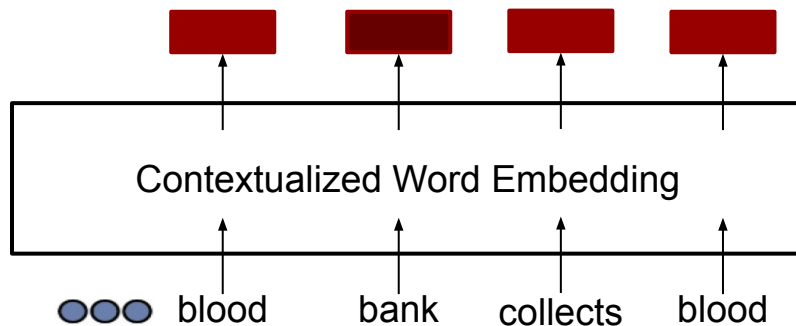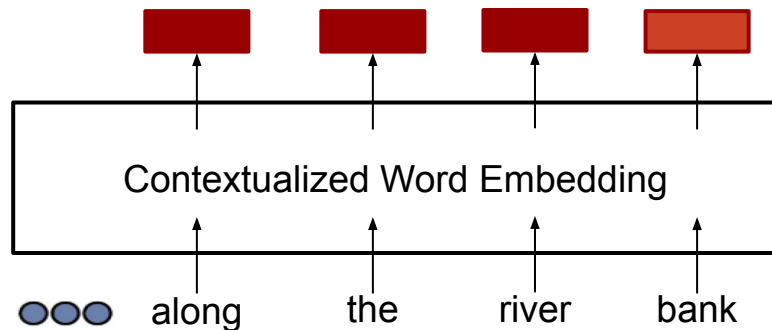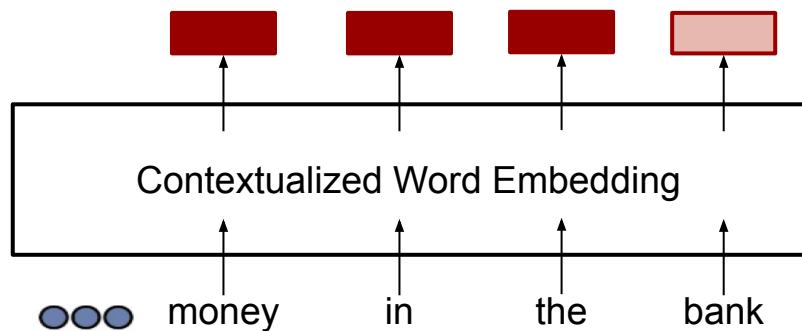
**Different Contexts,
Different Encoded Embeddings for bank.**

Embeddings generated from BERT

Cos-similarities among vectors of "apple" in different context

	An apple is sweet in taste.	We can make juice and jam from apple.	An apple a day keeps the doctor away.	Investors in Apple have a love-hate relationship with the iPhone.	Both Apple and Huawei have wireless charging features.	Steve Jobs returned to Apple in 1196.
An apple is sweet in taste.	1	0.74	0.82	0.53	0.48	0.5
We can make juice and jam from apple.	0.74	1	0.71	0.51	0.48	0.47
An apple a day keeps the doctor away.	0.82	0.71	1	0.46	0.42	0.44
Investors in Apple have a love-hate relationship with the iPhone.	0.53	0.51	0.46	1	0.82	0.77
Both Apple and Huawei have wireless charging features.	0.48	0.48	0.42	0.82	1	0.72
Steve Jobs returned to Apple in 1196.	0.5	0.47	0.44	0.77	0.72	1

How does BERT compute

dim: 768

Bert

| 101 | 4274 | 9861 | 2003 | 1037 | 2828 | 1997 | 16941 | 26775 | 14428 | 9861 | 102 |

3. Replace tokens with their indexes in vocab

BertTokenizer

| [CLS] | internet | fraud | is | a | type | of | cyber | ##cr | ##ime | fraud | [SEP] |

1. Break words into tokens
2. Add [CLS] and [SEP] tokens

"Internet fraud is a type of cybercrime fraud"

In the case of ChatGPT the generated numbers are probabilities. ChatGPT has a limited vocabulary, and the probabilities indicate how likely each vocabulary word is based on the input word sequence. ChatGPT has a limited reading range, and the input sequence has a maximum length of about 3000 words, broken into 4000 sub-word tokens. Once ChatGPT generates a word, it adds that word to the input sequence, and generates a new word. This process continues until it produces a special word called a "stop" token, or it hits a preset word limit.

Why is the reading range limited?

2.1 Self-Attention

Basic Self-Attention

- A sequence-to-sequence operation taking a sequence of vectors in and generate a sequence of vectors out
  - [x1, x2, x3] -> [z1, z2, z3]
- Relating different positions of the input sequence in order to compute the representation
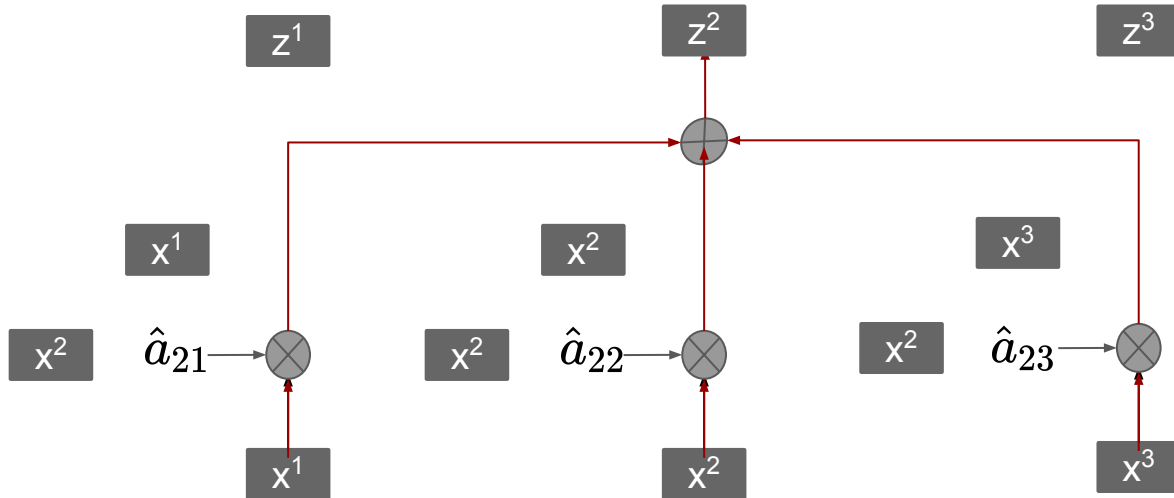
$$z^i = \sum_j \hat{a}_{ij} x^j$$

$$a_{ij} = (x^i)^T (x^j)$$

$$\hat{a}_{ij} = \frac{e^{a_{ij}}}{\sum_j e^{a_{ij}}}$$

Basic Self-Attention

Sentence i: My favourite fruit is apple

	my	favourite	fruit	is	apple
my	1	0	0	0	0
favourite	0	1	0	0	0
fruit	0	0	1	0	0.25
is	0	0	0	1	0
apple	0	0	0.25	0	1

New Word Index

my_i

favourite_i

fruit_i

is_i

apple_i

Attention Step

my

favourite

0.8*fruit+0.2*apple

is

0.2*fruit+0.8*apple

Basic Self-Attention

- There are no model parameters. It is totally determined by the embedding layer
    - Solution: introduce model parameters -> using three sets of embeddings to get contextualized embeddings
- Self attention is permutation equivariant. It ignores the order information.
    - Solution: add positional embeddings

Self-Attention layer

Step 1: Generate **query**, **key**, and **value** vector for the **input** vector at each time step.

q — Query (to match others): $q^i = W^q x^i$

k — Key (to be matched): $k^i = W^k x^i$

Model parameters are introduced here.

v — Value (representation): $v^i = W^v x^i$

In practice, bias vectors may be added to the product of matrix multiplication

q^1 k^1 v^1 q^2 k^2 v^2 q^3 k^3 v^3

x^1 x^2 x^3
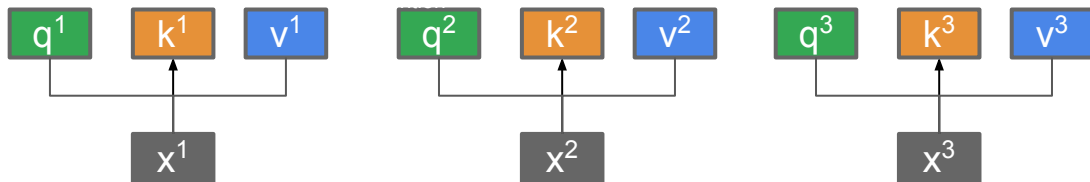
Word embeddings

The key/value/query formulation of attention is from the paper Attention Is All You Need.

264

How should one understand the queries, keys, and values

The key/value/query concept is analogous to retrieval systems. For example, when you search for videos on Youtube, the search engine will map your **query** (text in the search bar) against a set of **keys** (video title, description, etc.) associated with candidate videos in their database, then present you the best matched videos (**values**).

+50

The attention operation can be thought of as a retrieval process as well.

As mentioned in the paper you referenced (Neural Machine Translation by Jointly Learning to Align and Translate), attention by definition is just a weighted average of values,

$$c = \sum_j \alpha_j h_j$$

where $\sum \alpha_j = 1$.

44

Self-Attention layer

Step 2: Compute attention scores using query vectors and key vectors

To encode the i-th word in the sequence, we need to compute the attention scores between this i-th word and all the words in the sequence.

1. Pick the query vector from the i-th word: q^i
2. Attention score computation between q^i and all key vectors of the nearby words (including the target word itself)

$$a_{i,j} = \frac{q^i \cdot k^j}{\sqrt{d_k}}$$

Dim of key vectors
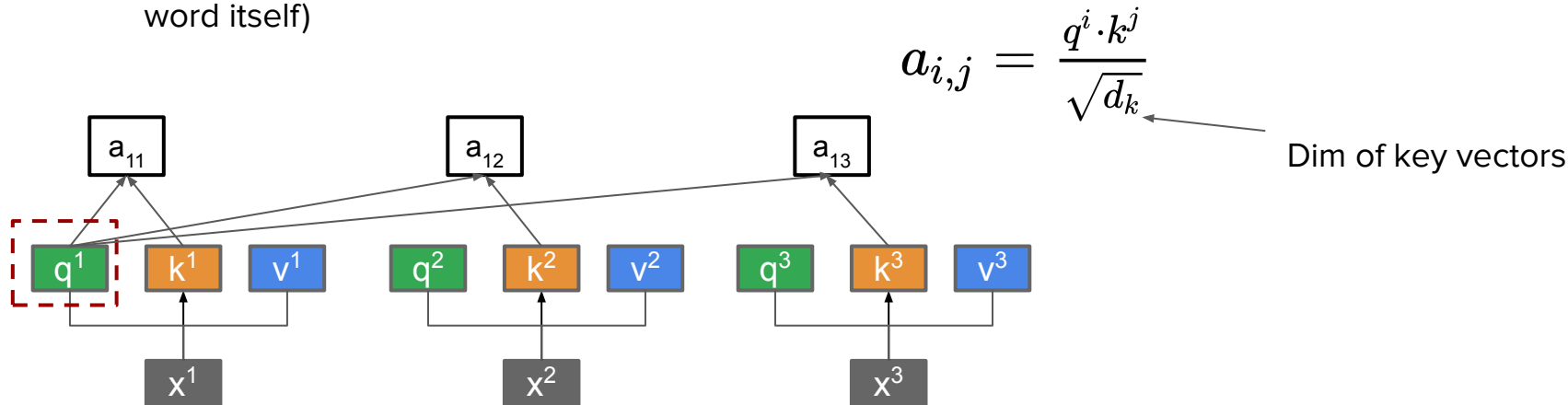
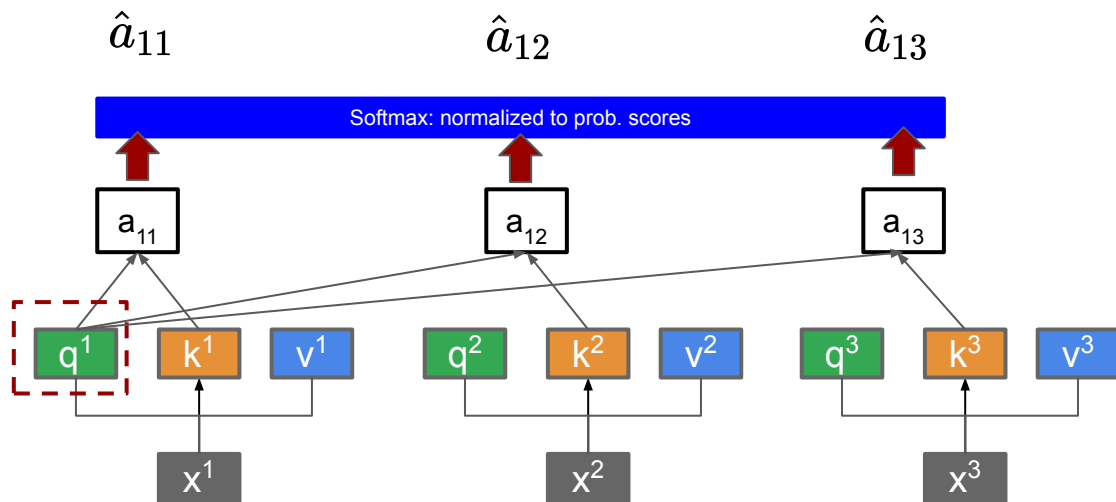Word vectors

Self-Attention layer

Step 3: Fed unscaled attention scores into softmax layers $\hat{a}_{1i} = \dfrac{e^{a_{1i}}}{\sum_j e^{a_{1j}}}$

$\hat{a}_{11}$ $\hat{a}_{12}$ $\hat{a}_{13}$

Softmax: normalized to prob. scores

a_{11} a_{12} a_{13}

q^1 k^1 v^1 q^2 k^2 v^2 q^3 k^3 v^3

x^1 x^2 x^3

Word vectors
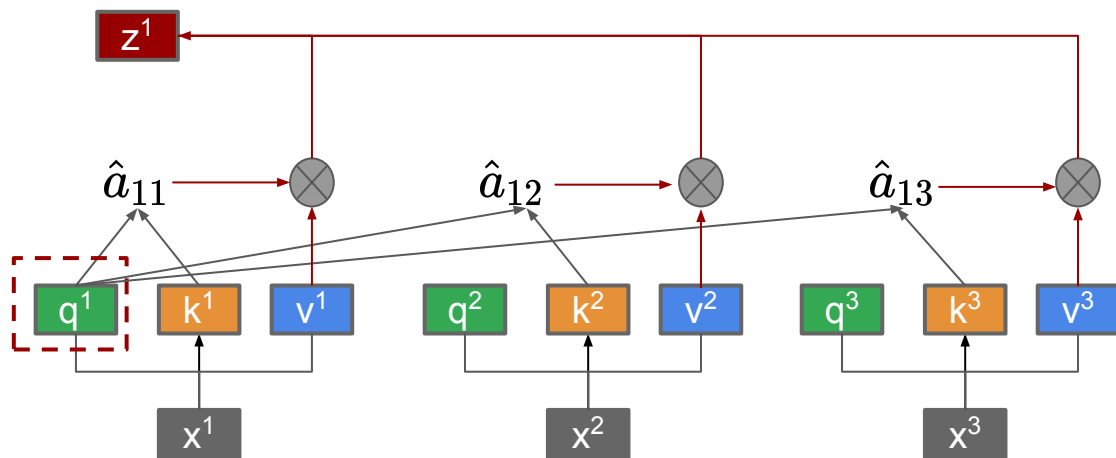
Self-Attention layer

Step 4: Take the sum of all the value vectors weighted by the attention scores.

Encoded vector for the first element

$$z^1 = \sum_i \hat{a}_{1i} v^i$$

Word vectors

Self-Attention layer

Step 5: All elements in input sequence x^i will be encoded into new vectors z^i

Encoded vector for the second element

$$z^2 = \sum_i \hat{a}_{2i} v^i$$
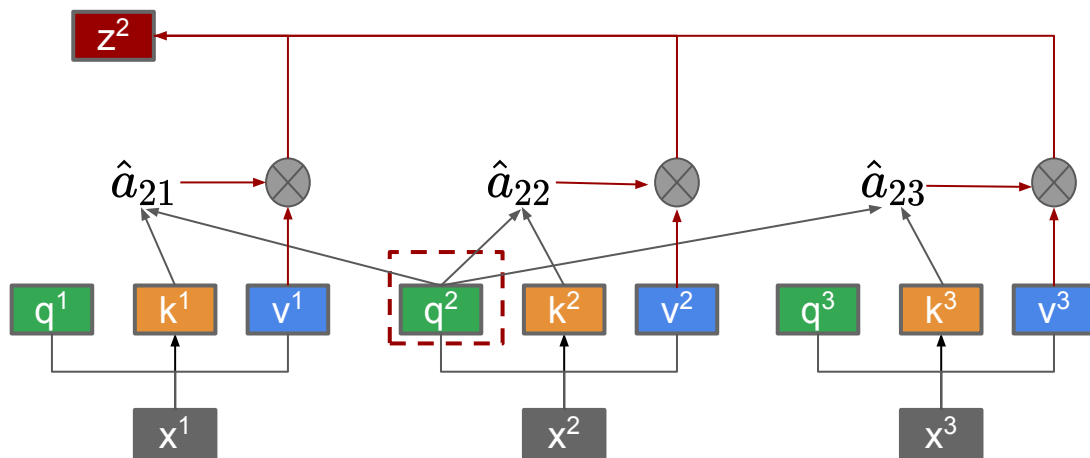
Word vectors

Matrix formulation

$$Z = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

Self-Attention Layer

z^i can be computed **parallelly** based on **the whole input sequence.**

X

$Q = W^q X$

$K = W^k X$

$V = W^v X$

Matrix Multiplication
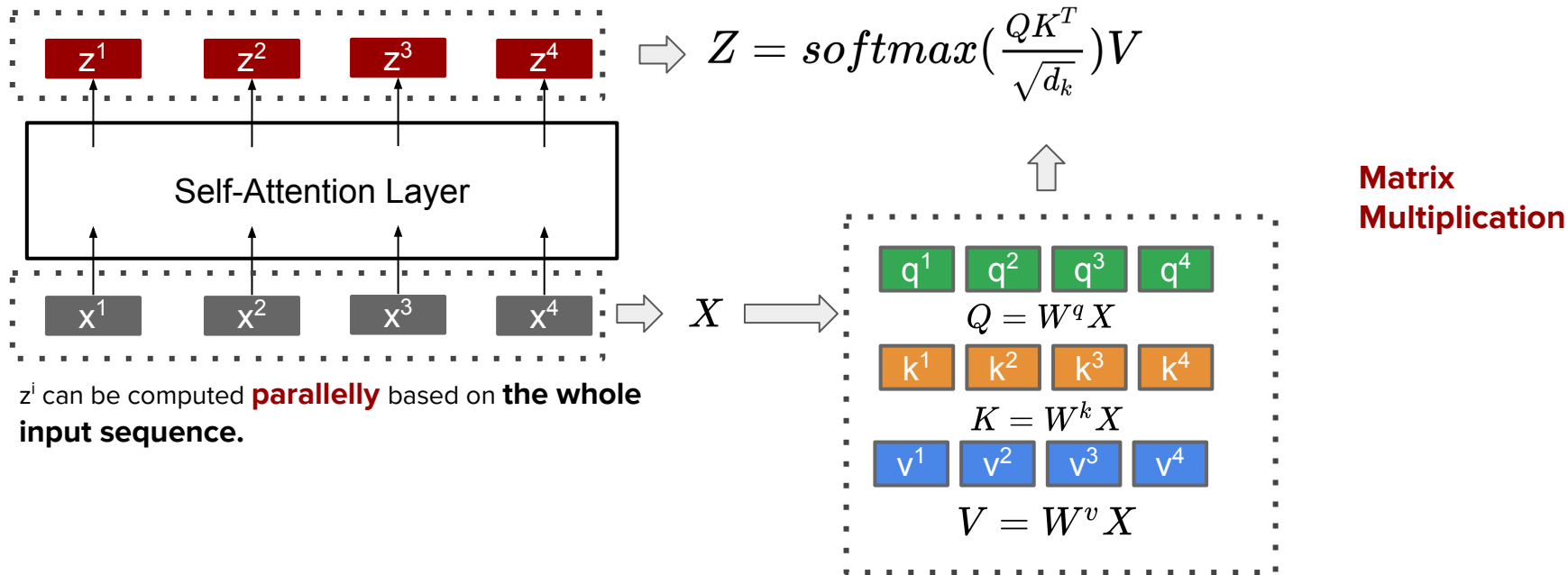
Multi-head Self-Attention

- Model parameters: W^k, W^q, W^v specific one kind of attention
- We can have multiple set of W^k, W^q, W^v
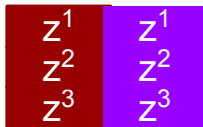
Concatenating all heads
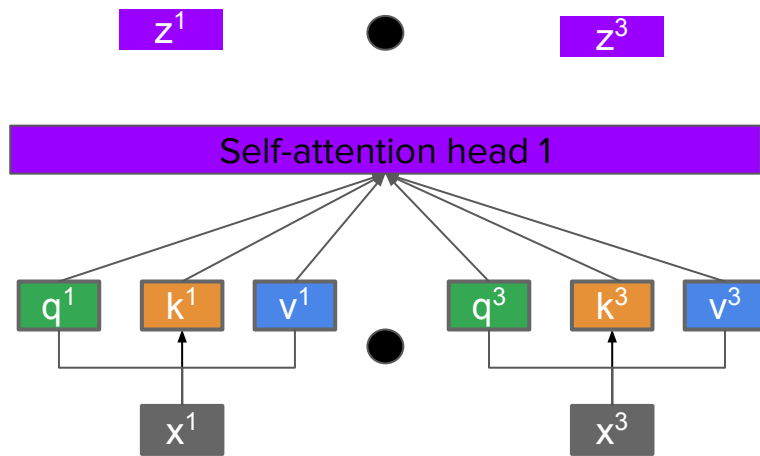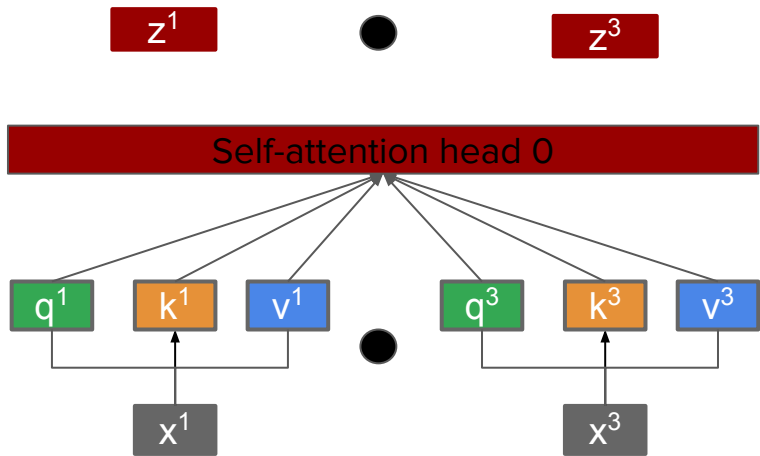
Illustration of Self-Attention

- Multi-head means separate W^k, W^q, W^v matrices
    - Expands the model's ability to focus on different positions
    - Gives the attention layer multiple "representation subspaces"
    - For example, two-head self-attention

High attention from one head

High attention from another head
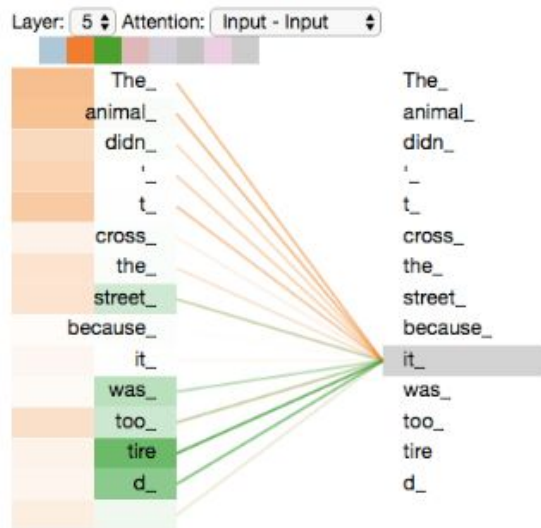
Head here is similar to
1. **the filter in convolutional layer**
2. **the neuron in fully-connected layer**

Multi-head Self-Attention

- If the layer has k heads, the output would be k sets of embeddings
- We need to reduce the dimensionality by concatenating and projection into the low dimensional
  - For example, as below: 4->32->4

dim-4

X

Thinking Machines

Calculating attention separately in eight different attention heads

ATTENTION HEAD #0   ATTENTION HEAD #1   ...   ATTENTION HEAD #7

Z_0 Z_1 Z_7

1) Concatenate all the attention heads   dim-32

Z_0 Z_1 Z_2 Z_3 Z_4 Z_5 Z_6 Z_7

2) Multiply with a weight matrix W^O that was trained jointly with the model

X

dim-4

W^O

3) The result would be the Z matrix that captures information from all the attention heads. We can send this forward to the FFNN

Z

=

Self-attention layer

That's pretty much all there is to multi-headed self-attention. It's quite a handful of matrices, I realize. Let me try to put them all in one visual so we can look at them in one place

1) This is our input sentence*

2) We embed each word*

3) Split into 8 heads. We multiply X or R with weight matrices

4) Calculate attention using the resulting $Q/K/V$ matrices

5) Concatenate the resulting Z matrices, then multiply with weight matrix W^O to produce the output of the layer

Thinking Machines

X

W_0^Q
W_0^K
W_0^V

Q_0
K_0
V_0

Z_0

W^O

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

W_1^Q
W_1^K
W_1^V

Q_1
K_1
V_1

Z_1

Z

R

...

W_7^Q
W_7^K
W_7^V

...

Q_7
K_7
V_7

...

Z_7
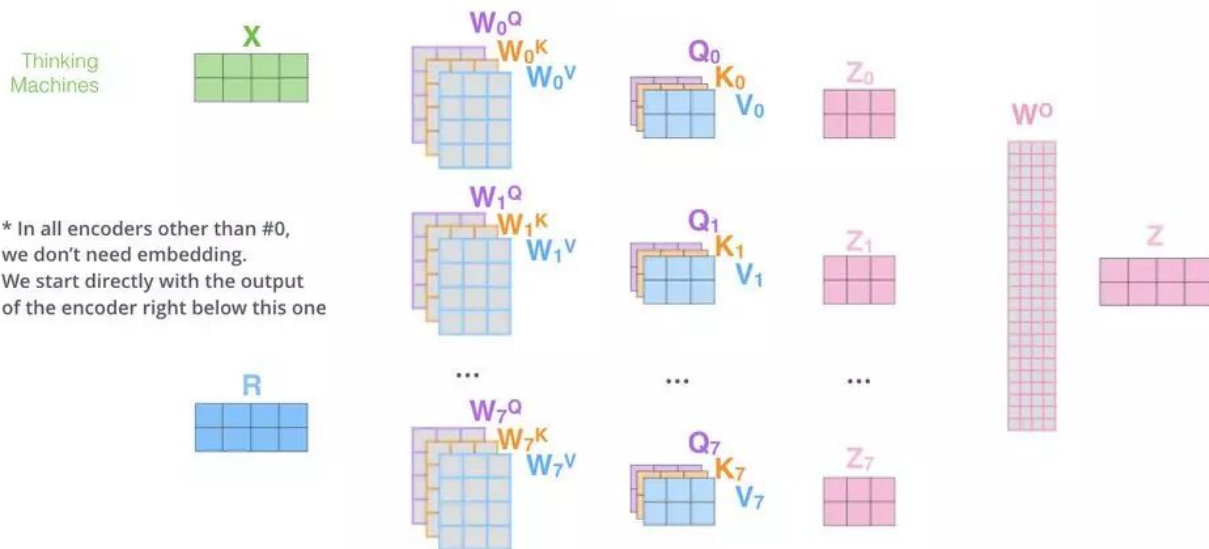
MultiHeadAttention in Keras

```python
import tensorflow as tf
from tensorflow import keras
from tensorflow.keras.layers import MultiHeadAttention
```

```python
target = tf.keras.Input(shape=[6, 16])

# assume it is a sentence of 6 words. Then, each word has a
```

```python
layer = MultiHeadAttention(num_heads=1, key_dim=2)
output_tensor, attention_scores = layer(target, target, return_attention_scores=True)
print(output_tensor.shape)
print(attention_scores.shape)
```

```
(None, 6, 16)
(None, 1, 6, 6)
```

```python
for matrix in layer.weights:
    print(matrix.shape)
```

```
(16, 1, 2)
(1, 2)
(16, 1, 2)
(1, 2)
(16, 1, 2)
(1, 2)
(1, 2, 16)
(16,)
```

MultiHeadAttention in Keras

```
layer = MultiHeadAttention(num_heads=3, key_dim=2)
target = tf.keras.Input(shape=[6, 16])
output_tensor, attention_scores = layer(target, target, return_attention_scores=True)
print(output_tensor.shape)
print(attention_scores.shape)
```

```
(None, 6, 16)
(None, 3, 6, 6)
```

```
for matrix in layer.weights:
    print(matrix.shape)
```

```
(16, 3, 2)
(3, 2)
(16, 3, 2)
(3, 2)
(16, 3, 2)
(3, 2)
(3, 2, 16)
(16,)
```

MultiHeadAttention in Keras

If we change the key_dim from 2 to 5?

```
layer = MultiHeadAttention(num_heads=3, key_dim=2)
target = tf.keras.Input(shape=[6, 16])
output_tensor, attention_scores = layer(target, target, return_attention_scores=True)
print(output_tensor.shape)
print(attention_scores.shape)
```

```
(None, 6, 16)
(None, 3, 6, 6)
```

```
for matrix in layer.weights:
    print(matrix.shape)
```

```
(16, 3, 2)
(3, 2)
(16, 3, 2)
(3, 2)
(16, 3, 2)
(3, 2)
(3, 2, 16)
(16,)
```

https://github.com/rz0718/BT5153_2024
/blob/main/codes/lab_lecture08/Attention_Layer_in_Keras.ipynb

2.2 Position Embedding

Positional embeddings

- No position information in self-attention
- Positional Embeddings: each position has a unique positional vector PE(pos)
  - Add this vector to each input embeddings
  - Expands the model's ability to focus on different positions.

Dim.

Vocab size

Dim.

Max. Length of Input Sequence

Transformer Block #1

Transformer Block #0

x^1 x^2 x^3 x^4

x^1 x^2 x^3 x^4

$\oplus$ $\oplus$ $\oplus$ $\oplus$

PE^1 PE^2 PE^3 PE^4

Input Embedding Matrix (Learned from the data)

Positional Embedding Matrix (Pre-defined and Fixed)
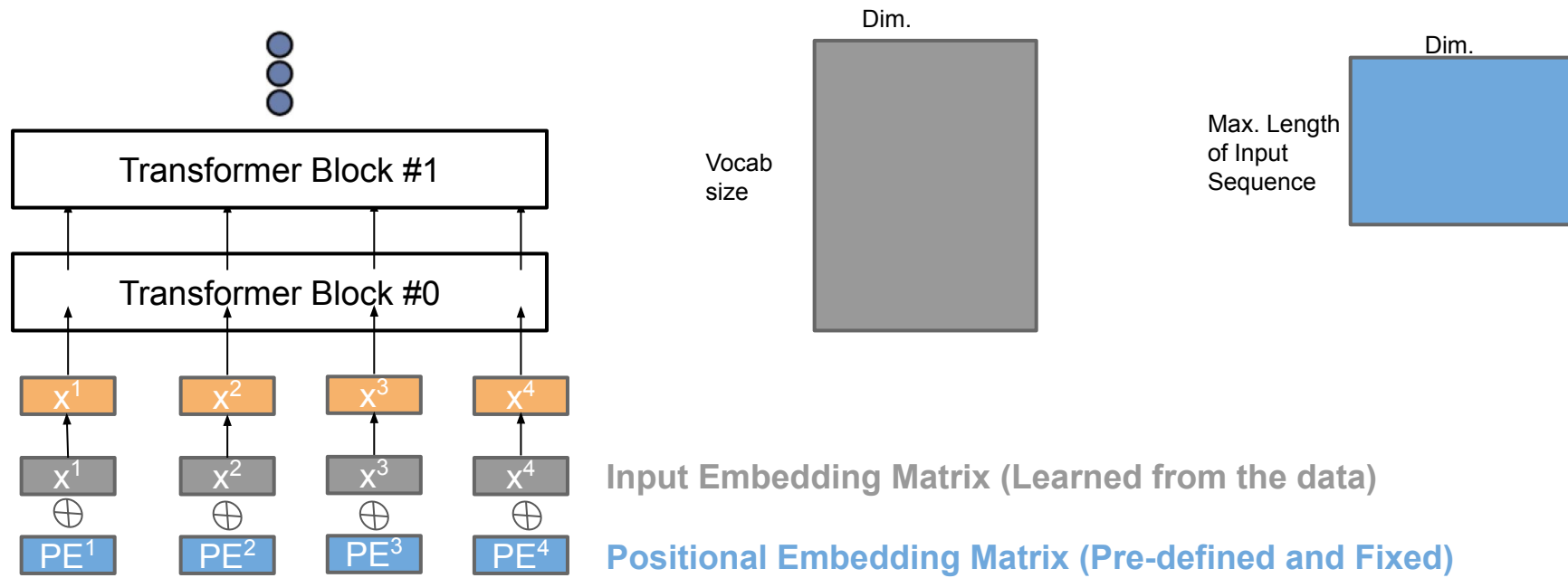
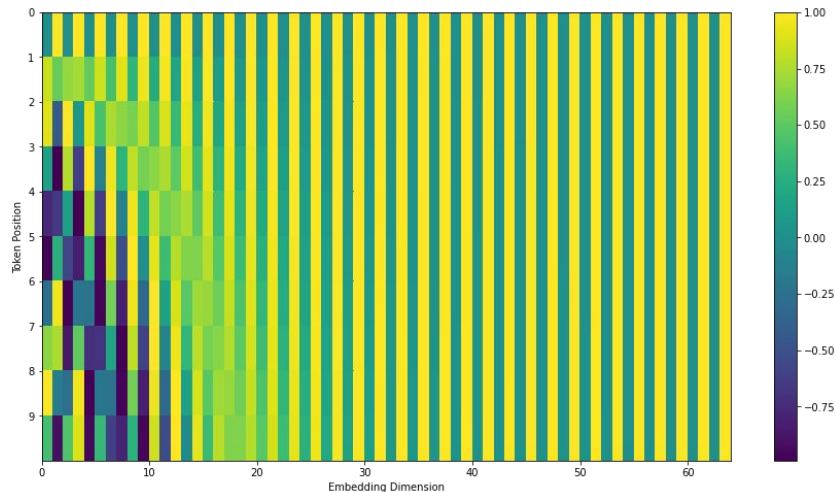# Positional embeddings

- The equation in the original paper:

$$PE_{(pos, 2i)} = sin(pos/10000^{2i/d_{model}})$$
$$PE_{(pos, 2i+1)} = cos(pos/10000^{2i/d_{model}})$$

Core idea: using fixed weights which encode information related to a specific position of a token in a sentence

More details: https://kazemnejad.com/blog/transformer_architecture_positional_encoding/

3. Summary

A "functional" viewpoint on Transformer

Output
Probabilities

Softmax

Linear

Loss function (predict next token given previous context)

Sequence to Sequence mapping with multiple matrix multiplication

Input: [batch, d_embedding, length]
Output: [batch, d_embedding, length]

Function parameters are updated via backpropagation.

Positional Encoding

Input Embedding

Inputs

Positional Encoding

Output Embedding

Outputs (shifted right)

Transformers is replacing RNN and CNN

- Compared to Transformers, RNN
  - can not be trained in parallel
  - suffers from long dependency issues
- Compared to Transformers, CNN
  - is unable to capture all possible combinations of words (filter size is predefined)
- Compared to the previous NN, Transformers
  - **Non sequential**: the input sequence are processed as a whole
  - **Self Attention**: contextualized word embeddings
  - **Positional embeddings**: a better way than recurrence to capture order information
- Podcast about transformers
  - https://www.youtube.com/watch?v=9uw3F6rndnA

62

Must Read !!

[The illustrated transformer](#) (the source of the awesome visualizations)

Implementations of Transformers

- [Build Transformer from Scratch](#)
- We can import it from huggingface

Next Class: LLM I

Appendix

Masked Self-Attention

Masked Self-Attention

- This is the attention layer used to compute the dependency among the target words
- Since the sequence is generated word by word, we need to prevent it from conditioning to the future tokens
- For example:
  - to generate "a", we should not have access to "student"

Target Self Attention Score

	I	am	a	student
I	0.7	0.1 ✖	0.1 ✖	0.1 ✖
am	0.1	0.6	0.2 ✖	0.1 ✖
a	0.1	0.3	0.6	0.1 ✖
student	0.1	0.3	0.3	0.3

＋

Look-ahead Bias

	I	am	a	student
I	0	-inf	-inf	-inf
am	0	0	-inf	-inf
a	0	0	0	-inf
student	0	0	0	0

＝

Masked Self Attention Score

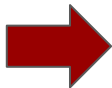|	I	am	a	student
I	0.7	-inf	-inf	-inf
am	0.1	0.6	-inf	-inf
a	0.1	0.3	0.6	-inf
student	0.1	0.3	0.3	0.3

Masked Self-Attention

- Add look-ahead mask matrix
- Apply softmax to get the probabilistic scores
  - The negative infinities would become zero after softmax
  - For example, the attention score for "a"
    - has values for itself and all words before it
    - Zero for the word "student"

Masked Self Attention Score

	I	am	a	student
I	0.7	-inf	-inf	-inf
am	0.1	0.6	-inf	-inf
a	0.1	0.3	0.6	-inf
student	0.1	0.3	0.3	0.3

softmax

Normalized Masked Self-Attention Scores

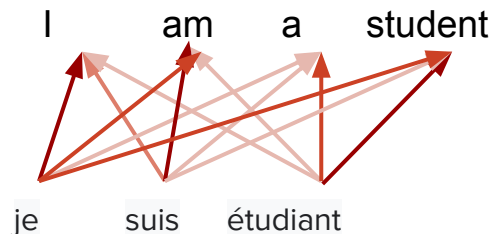|	I	am	a	student
I	1	0	0	0
am	0.37	0.62	0	0
a	0.26	0.3	0.43	0
student	0.21	0.26	0.26	0.26

69

Encoder-decoder Attention

Encoder-decoder attention

Attention in decoder layer:

1. Attention vectors: a vector of importance weights (measure the interaction between each target word with each input word)

Output   I   am   a   student

Input   je   suis   étudiant

High Attention
Low Attention

2. The target is approximated by the sum of their input values weighted by the attention scores.

$$Vec_{student}=0.15*Vec_{je} + 0.05*Vec_{suis} + 0.8*Vec_{etudiant}$$

Encoder-Decoder attention layer

Different from Self attention layer
1. Generate **query** vector for the generated output sequence (from itself: Decoder)
2. Generate **key** and **value** vector for the **input** sequence at each time step (from Encoder)

Self Attention Score n

	je	suis	eludient
je			
suis			
eludient			

Encoder-decoder Attention Score

	I	am	a	student
je				
suis				
eludient				

Matrix formulation

Encoder-Output

z^1 z^2 z^3

Value & Key

Encoder-decoder Attention Layer

h^1 h^2

y^1 y^2

Query

$$Z = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

X

q^1 q^2

$Q = W^q X$

k^1 k^2 k^3 k^4

$K = W^k X$

v^1 v^2 v^3 v^4

$V = W^v X$