Autonomous Systems: Deep Learning Transformers

HOCHSCHULE HEILBRONN

Prof. Dr.-Ing. Nicolaj Stache

Heilbronn University of Applied Sciences

Motivation

So far: RNN based Seq2Seq

- Processes data sequentially
- Captures timely dependencies in sequences

Problem:

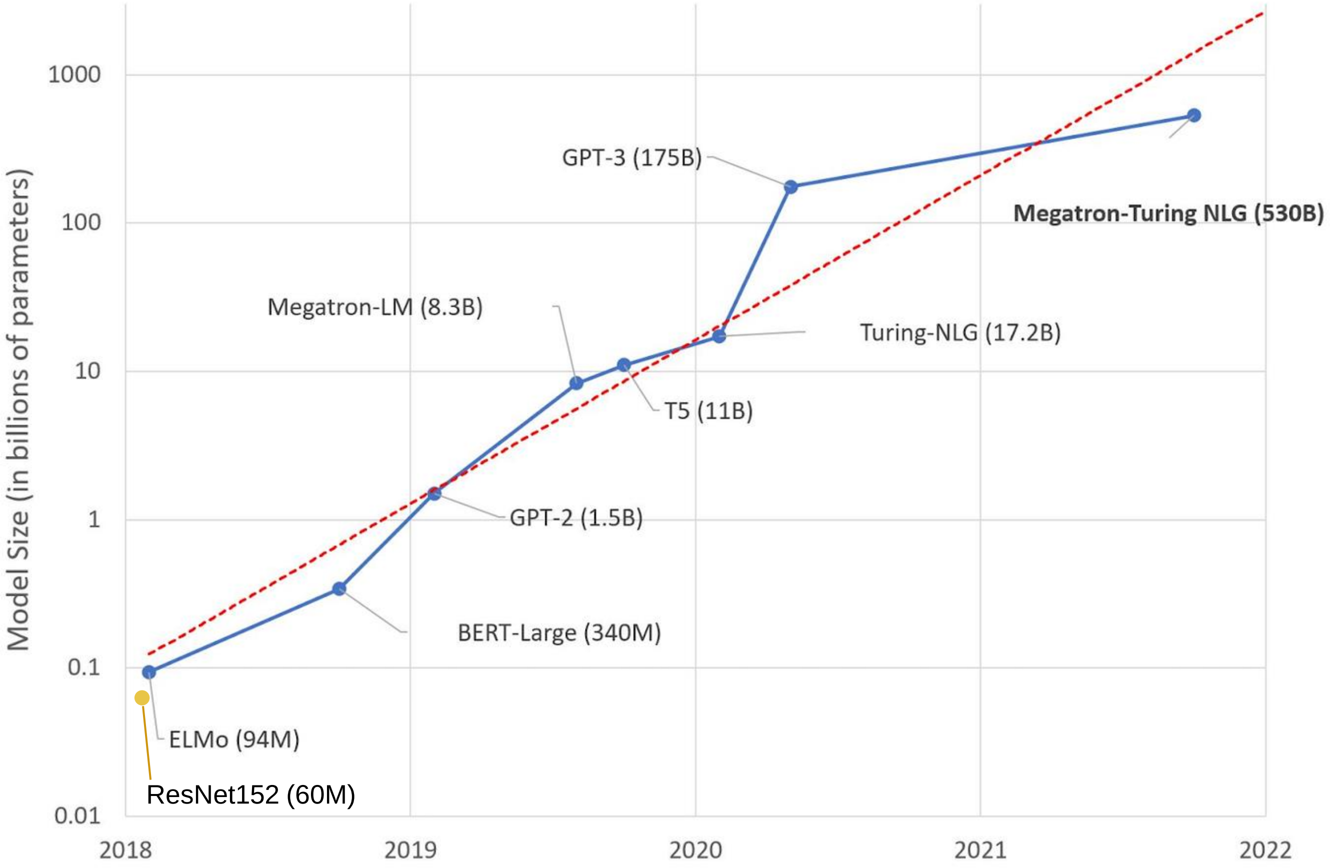
- The sequential nature prevents parallelization
- Struggles with long-range dependencies

Solution:

- Transformers

Transformers

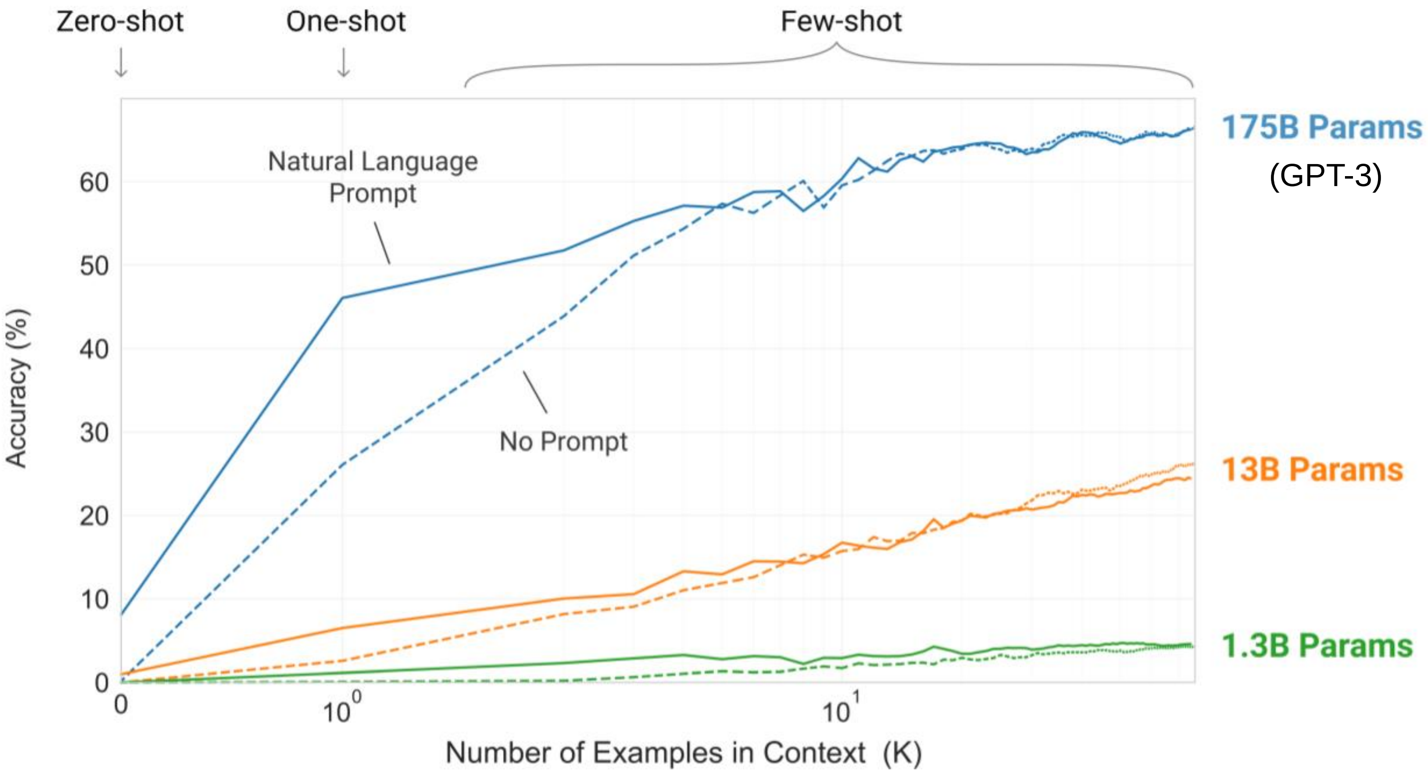
HEILBRONN UNIVERSITY
OF APPLIED SCIENCES



https://huggingface.co/blog/large-language-models 01.02.2022

- Size of transformers grows at an exponential rate

Transformers



(GPT-3)

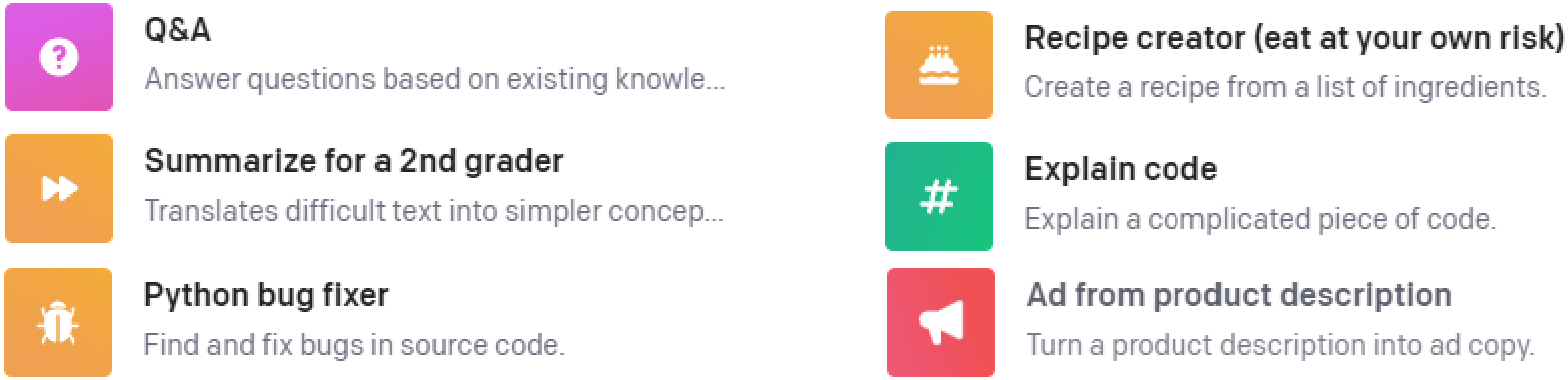
https://arxiv.org/pdf/2005.14165.pdf

- More parameters result in better accuracy

GPT-3

- 175 billion trainable parameters
- Trained on about 45 TB of text data
- Training would cost $4.5 million and would take 355 years on a V100 GPU server (28 TFLOPS capacity)

Example Tasks:



Q&A
Answer questions based on existing knowle...

Summarize for a 2nd grader
Translates difficult text into simpler concep...

Python bug fixer
Find and fix bugs in source code.

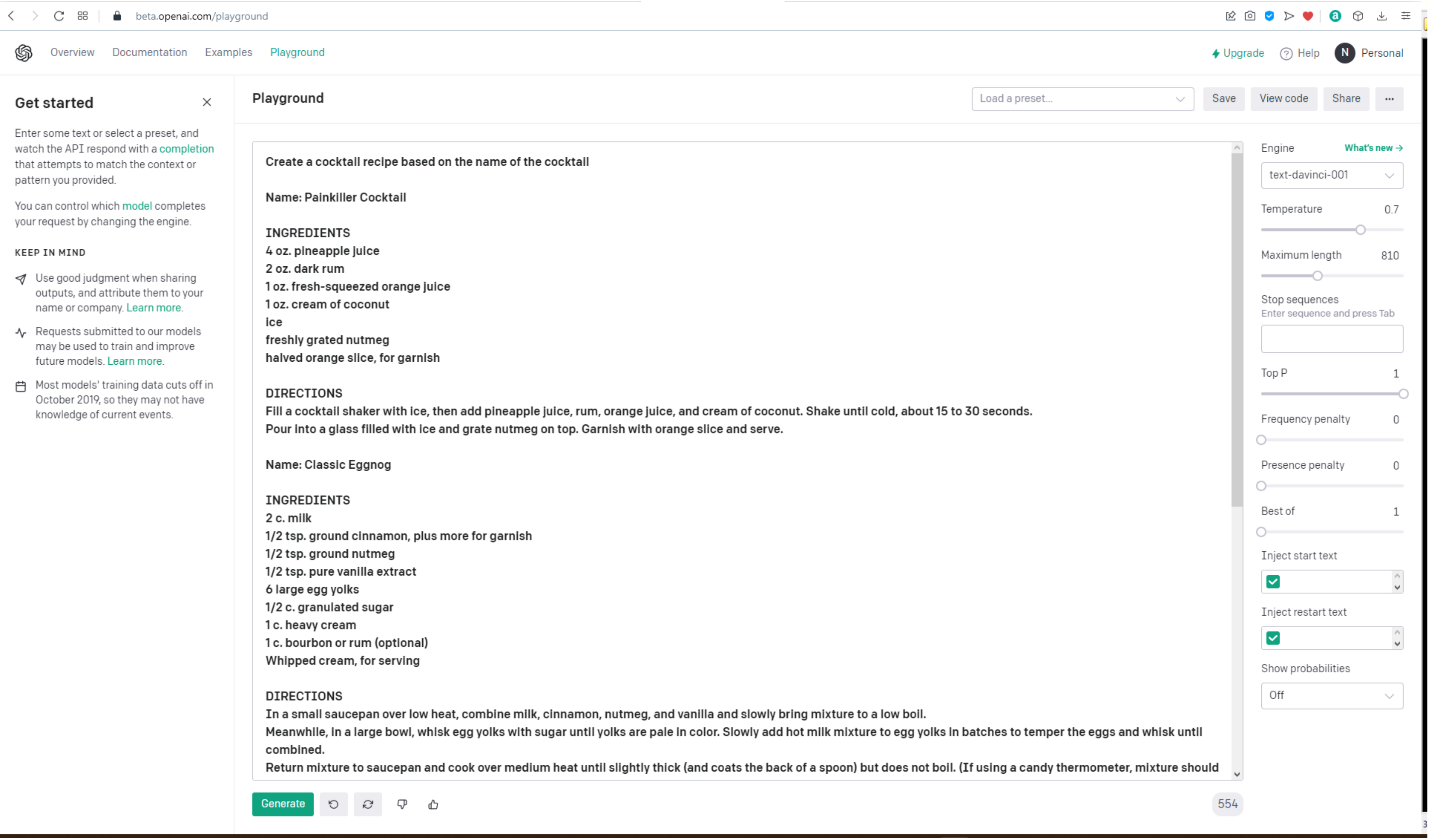
Recipe creator (eat at your own risk)
Create a recipe from a list of ingredients.

Explain code
Explain a complicated piece of code.

Ad from product description
Turn a product description into ad copy.

→ Try it out on: https://beta.openai.com/

Example



Get started

Enter some text or select a preset, and watch the API respond with a completion that attempts to match the context or pattern you provided.

You can control which model completes your request by changing the engine.

KEEP IN MIND

- Use good judgment when sharing outputs, and attribute them to your name or company. Learn more.
- Requests submitted to our models may be used to train and improve future models. Learn more.
- Most models' training data cuts off in October 2019, so they may not have knowledge of current events.

Playground

Create a cocktail recipe based on the name of the cocktail

Name: Painkiller Cocktail

INGREDIENTS
4 oz. pineapple juice
2 oz. dark rum
1 oz. fresh-squeezed orange juice
1 oz. cream of coconut
ice
freshly grated nutmeg
halved orange slice, for garnish

DIRECTIONS
Fill a cocktail shaker with ice, then add pineapple juice, rum, orange juice, and cream of coconut. Shake until cold, about 15 to 30 seconds.
Pour into a glass filled with ice and grate nutmeg on top. Garnish with orange slice and serve.

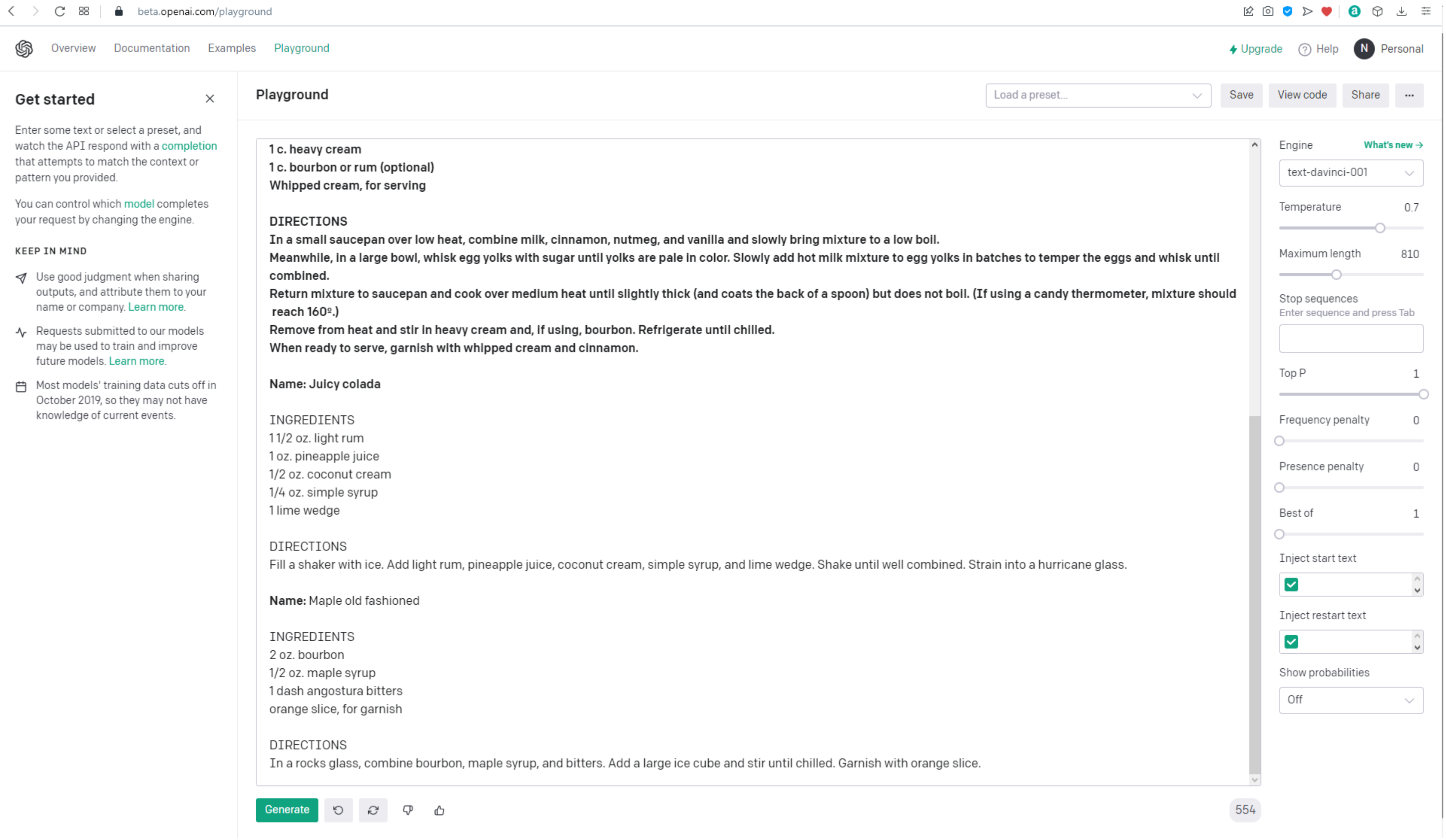
Name: Classic Eggnog

INGREDIENTS
2 c. milk
1/2 tsp. ground cinnamon, plus more for garnish
1/2 tsp. ground nutmeg
1/2 tsp. pure vanilla extract
6 large egg yolks
1/2 c. granulated sugar
1 c. heavy cream
1 c. bourbon or rum (optional)
Whipped cream, for serving

DIRECTIONS
In a small saucepan over low heat, combine milk, cinnamon, nutmeg, and vanilla and slowly bring mixture to a low boil.
Meanwhile, in a large bowl, whisk egg yolks with sugar until yolks are pale in color. Slowly add hot milk mixture to egg yolks in batches to temper the eggs and whisk until combined.
Return mixture to saucepan and cook over medium heat until slightly thick (and coats the back of a spoon) but does not boil. (If using a candy thermometer, mixture should

Engine: text-davinci-001 (What's new →)
Temperature: 0.7
Maximum length: 810
Stop sequences: Enter sequence and press Tab
Top P: 1
Frequency penalty: 0
Presence penalty: 0
Best of: 1
Inject start text: ☑
Inject restart text: ☑
Show probabilities: Off

Example



Get started

Enter some text or select a preset, and watch the API respond with a completion that attempts to match the context or pattern you provided.

You can control which model completes your request by changing the engine.

KEEP IN MIND

- Use good judgment when sharing outputs, and attribute them to your name or company. Learn more.
- Requests submitted to our models may be used to train and improve future models. Learn more.
- Most models' training data cuts off in October 2019, so they may not have knowledge of current events.

Playground

1 c. heavy cream
1 c. bourbon or rum (optional)
Whipped cream, for serving

DIRECTIONS
In a small saucepan over low heat, combine milk, cinnamon, nutmeg, and vanilla and slowly bring mixture to a low boil.
Meanwhile, in a large bowl, whisk egg yolks with sugar until yolks are pale in color. Slowly add hot milk mixture to egg yolks in batches to temper the eggs and whisk until combined.
Return mixture to saucepan and cook over medium heat until slightly thick (and coats the back of a spoon) but does not boil. (If using a candy thermometer, mixture should reach 160º.)
Remove from heat and stir in heavy cream and, if using, bourbon. Refrigerate until chilled.
When ready to serve, garnish with whipped cream and cinnamon.

Name: Juicy colada

INGREDIENTS
1 1/2 oz. light rum
1 oz. pineapple juice
1/2 oz. coconut cream
1/4 oz. simple syrup
1 lime wedge

DIRECTIONS
Fill a shaker with ice. Add light rum, pineapple juice, coconut cream, simple syrup, and lime wedge. Shake until well combined. Strain into a hurricane glass.

Name: Maple old fashioned

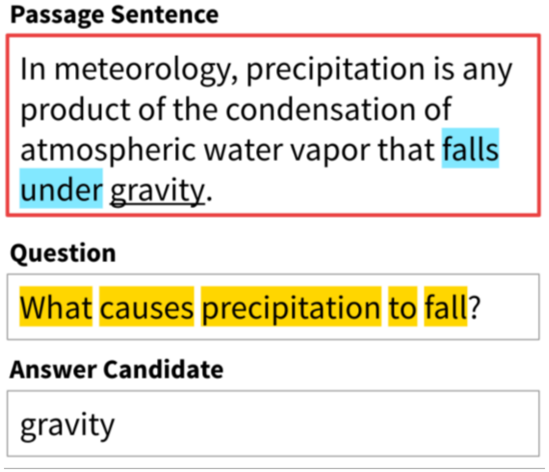
INGREDIENTS
2 oz. bourbon
1/2 oz. maple syrup
1 dash angostura bitters
orange slice, for garnish

DIRECTIONS
In a rocks glass, combine bourbon, maple syrup, and bitters. Add a large ice cube and stir until chilled. Garnish with orange slice.

Engine: text-davinci-001 (What's new →)
Temperature: 0.7
Maximum length: 810
Stop sequences: Enter sequence and press Tab
Top P: 1
Frequency penalty: 0
Presence penalty: 0
Best of: 1
Inject start text: ☑
Inject restart text: ☑
Show probabilities: Off

Applications

Question Answering



Passage Sentence

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity.

Question

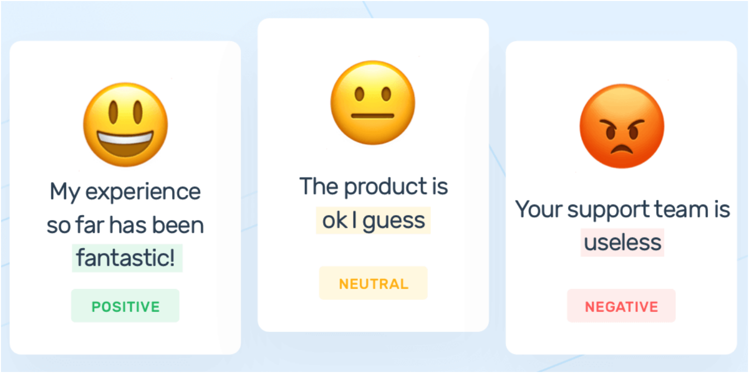
What causes precipitation to fall?

Answer Candidate

gravity

https://rajpurkar.github.io/mlx/qa-and-squad/ 01.02.2022

Sentiment Analysis



My experience so far has been fantastic!

POSITIVE

The product is ok I guess

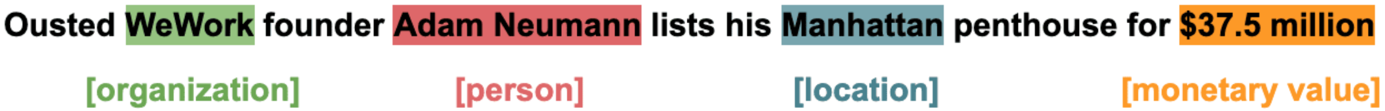
NEUTRAL

Your support team is useless

NEGATIVE

https://monkeylearn.com/sentiment-analysis/ 01.02.2022

Named Entity Recognition



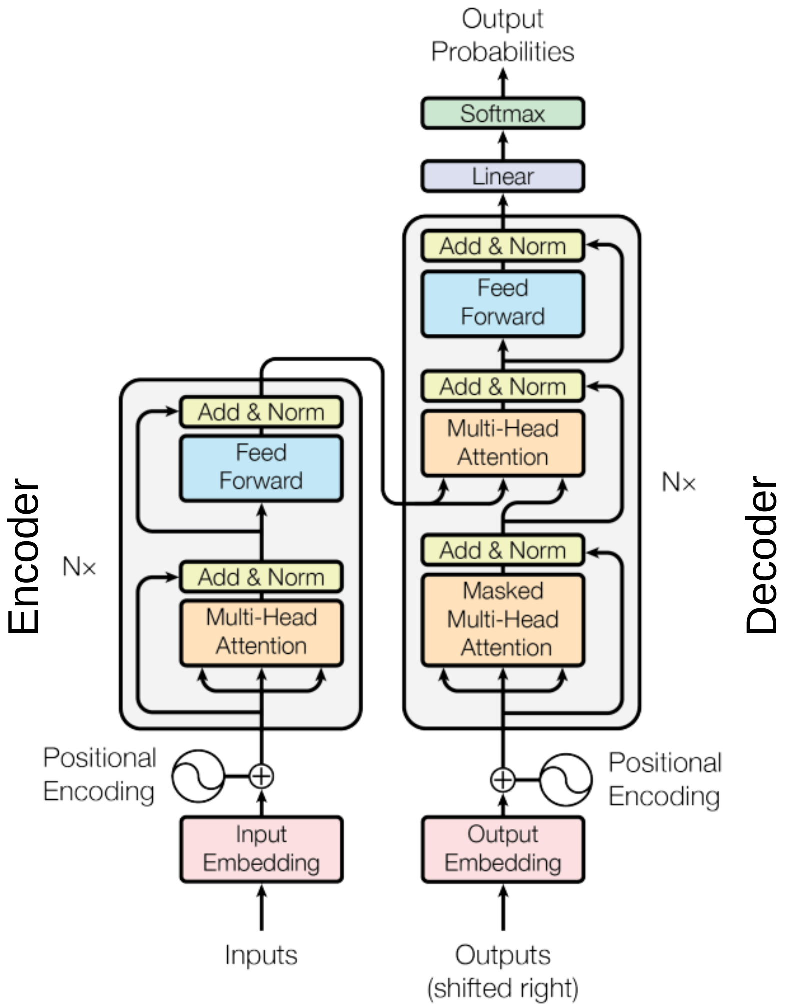
Ousted WeWork founder Adam Neumann lists his Manhattan penthouse for $37.5 million

[organization] [person] [location] [monetary value]

https://monkeylearn.com/blog/named-entity-recognition/ 01.02.2022

And many more…

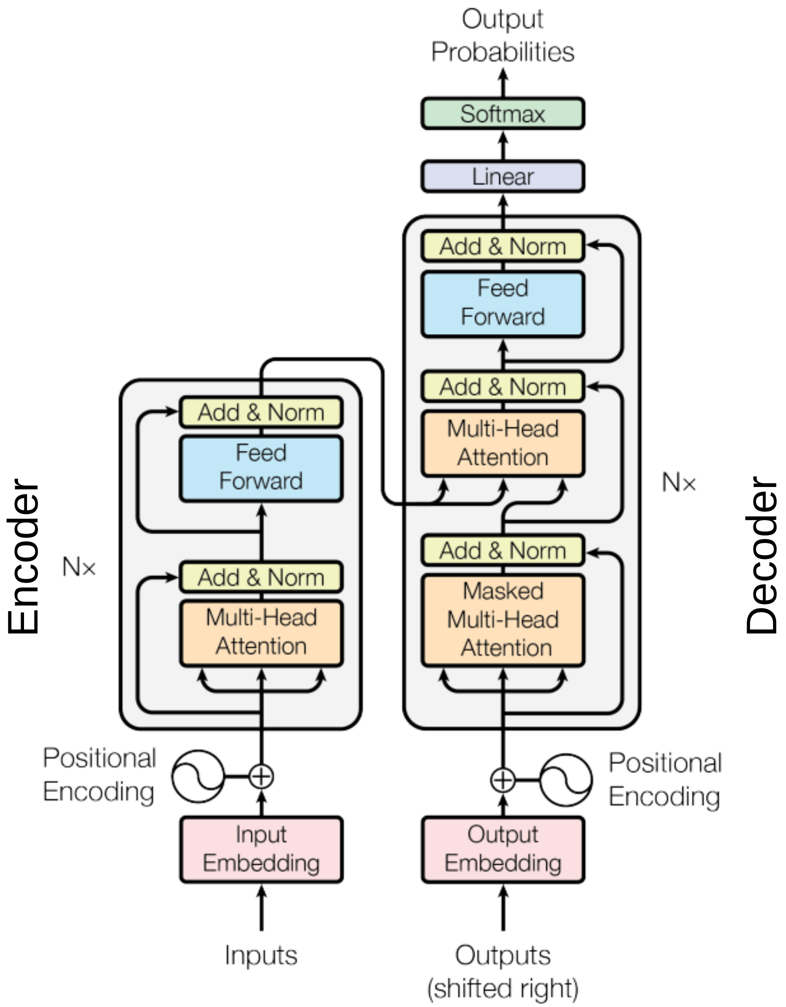
“Attention is all you need” 2017



- Encoder-Decoder Structure
- No Recurrence
- Outperformed all SOTA Algorithms in NLP Tasks

https://arxiv.org/pdf/1706.03762.pdf

Transformers

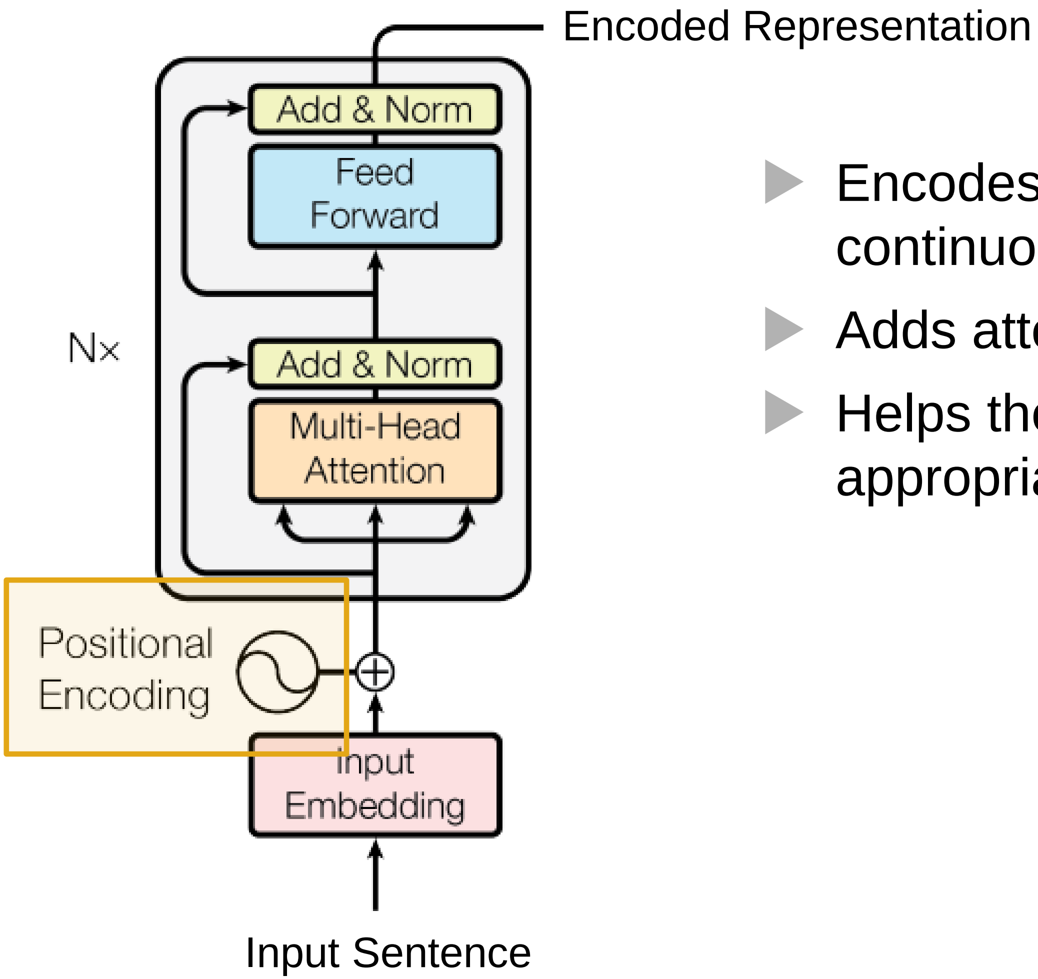


https://arxiv.org/pdf/1706.03762.pdf

Whats new?

- Positional Encodings
- Self-Attention
- Multi-Head Attention
- Masked Multi-Head Attention

Encoder



- Encodes the input into a continuous representation
- Adds attention information
- Helps the decoder to focus on appropriate words

https://arxiv.org/pdf/1706.03762.pdf

Positional Encodings

- It should output a unique encoding for each time-step
- Distance between any two time-steps should be consistent across sentences with different lengths.
- It must be deterministic.

$$\vec{p_t} = \begin{bmatrix} \sin(\omega_1 . t) \\ \cos(\omega_1 . t) \\ \\ \sin(\omega_2 . t) \\ \cos(\omega_2 . t) \\ \\ \vdots \\ \\ \sin(\omega_{d/2} . t) \\ \cos(\omega_{d/2} . t) \end{bmatrix}_{d\times 1}$$

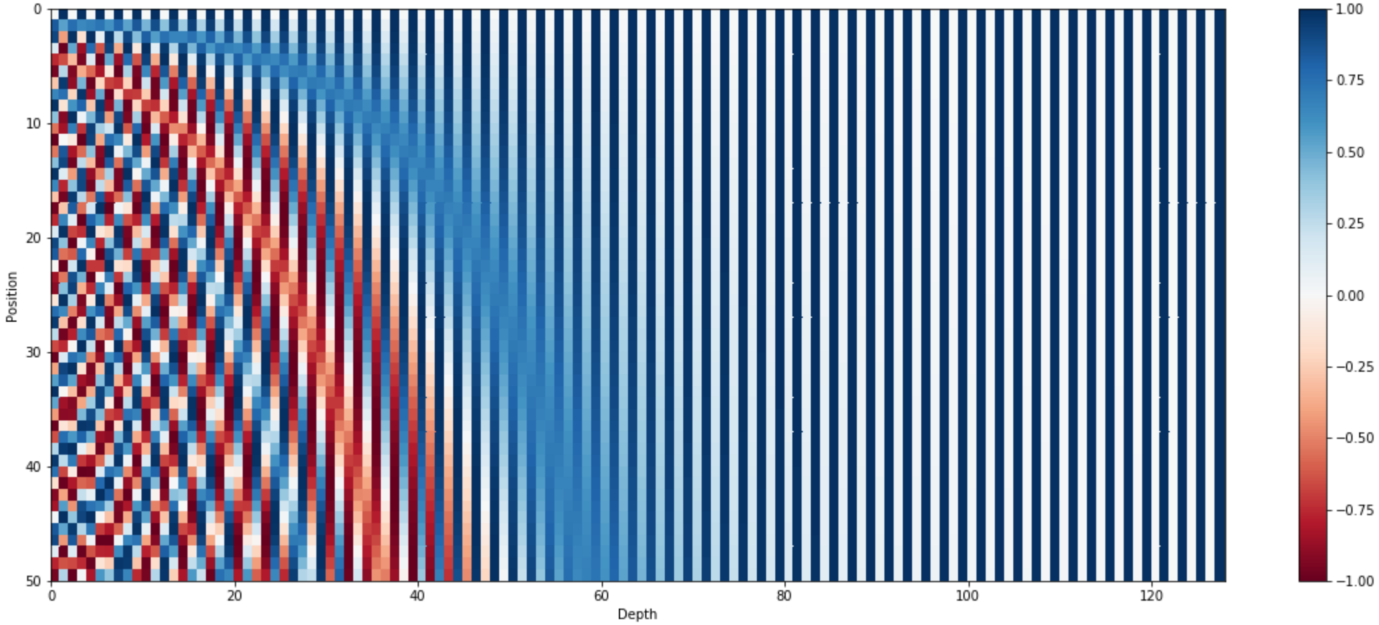
$$\vec{p_t}^{(i)} = f(t)^{(i)} := \begin{cases} \sin(\omega_k . t), & \text{if } i = 2k \\ \cos(\omega_k . t), & \text{if } i = 2k+1 \end{cases}$$

$$\omega_k = \frac{1}{10000^{2k/d}}$$

https://kazemnejad.com/blog/transformer_architecture_positional_encoding/ 01.02.2022

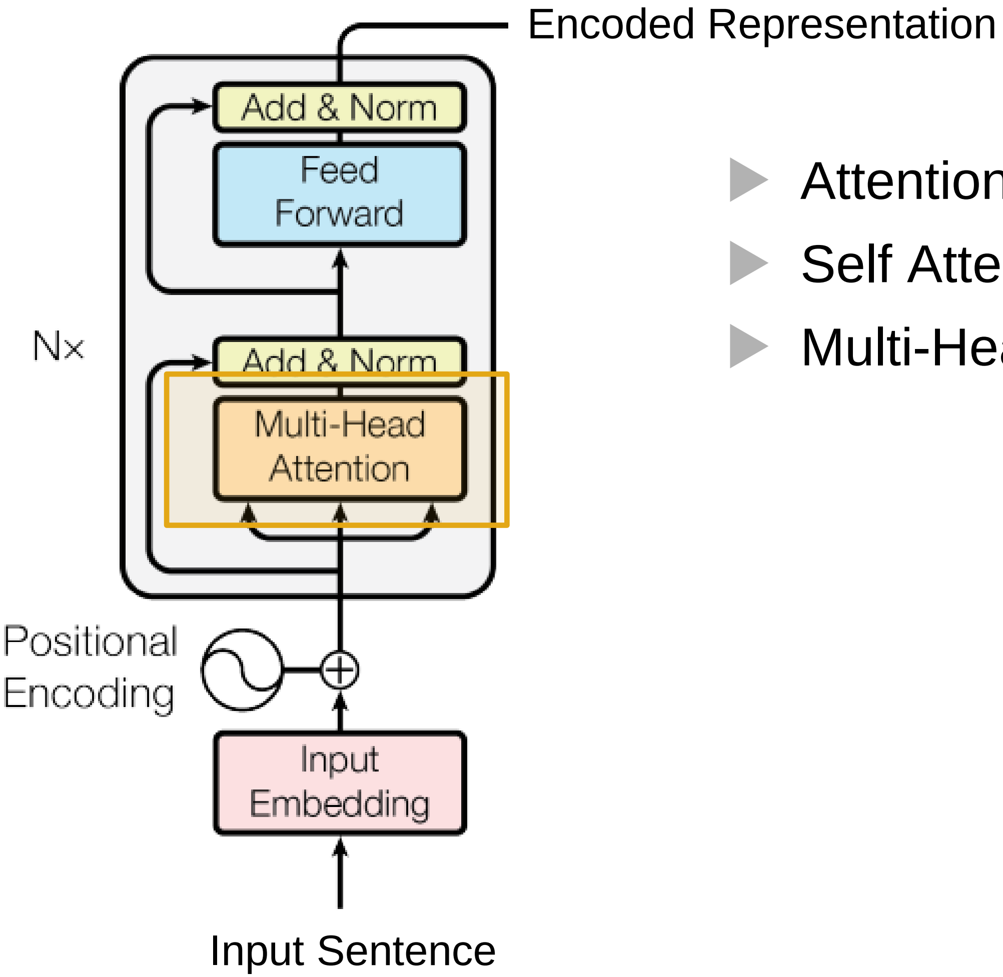
Positional Encodings

0 :	0 0 0 0	8 :	1 0 0 0
1 :	0 0 0 1	9 :	1 0 0 1
2 :	0 0 1 0	10 :	1 0 1 0
3 :	0 0 1 1	11 :	1 0 1 1
4 :	0 1 0 0	12 :	1 1 0 0
5 :	0 1 0 1	13 :	1 1 0 1
6 :	0 1 1 0	14 :	1 1 1 0
7 :	0 1 1 1	15 :	1 1 1 1



https://kazemnejad.com/blog/transformer_architecture_positional_encoding/ 01.02.2022

Encoder

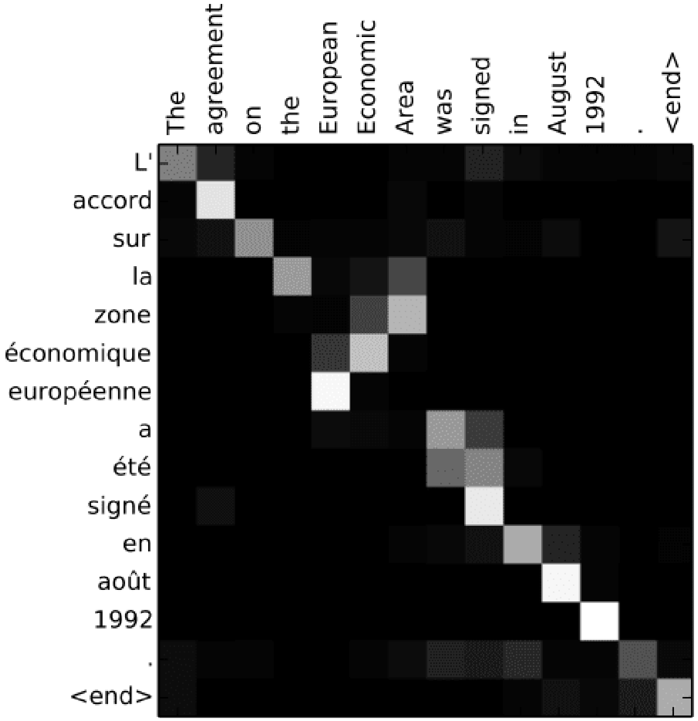


- Attention
- Self Attention
- Multi-Head Attention

https://arxiv.org/pdf/1706.03762.pdf

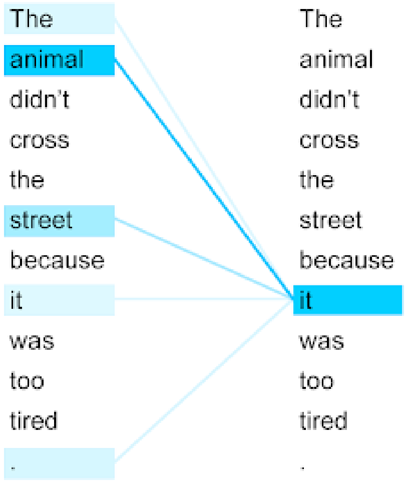
Attention

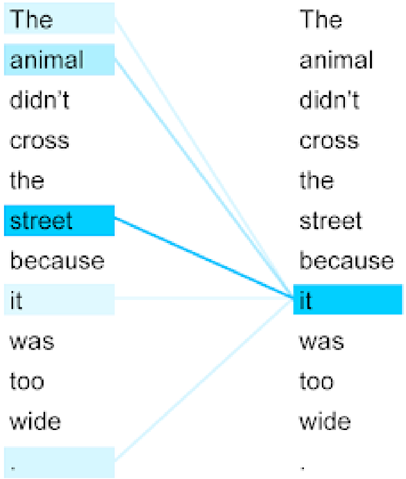
- Mimics the human attention
- Focusing on a few relevant parts
- Relations between Input und Output Sequence
- Enables better processing of very long sequences



https://arxiv.org/pdf/1409.0473.pdf 01.02.2022

Self-Attention

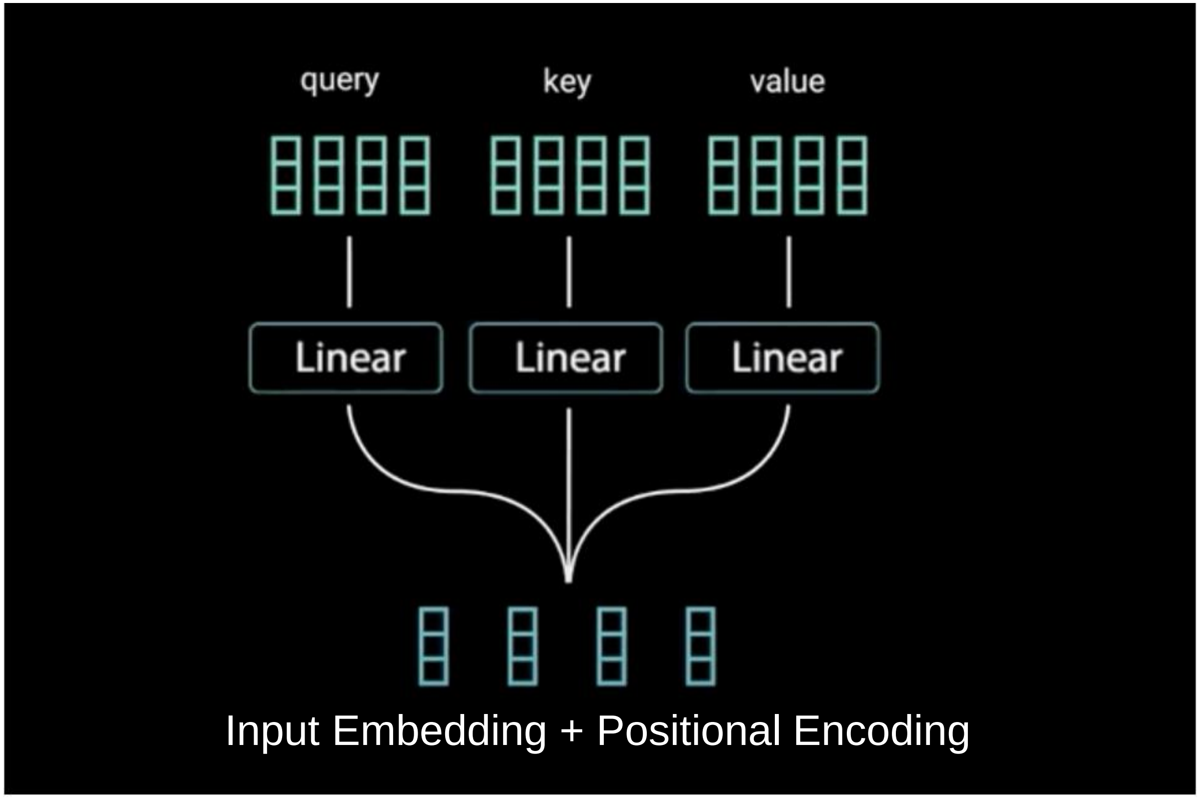




- Relations inside the same sequence
- Better understanding of context

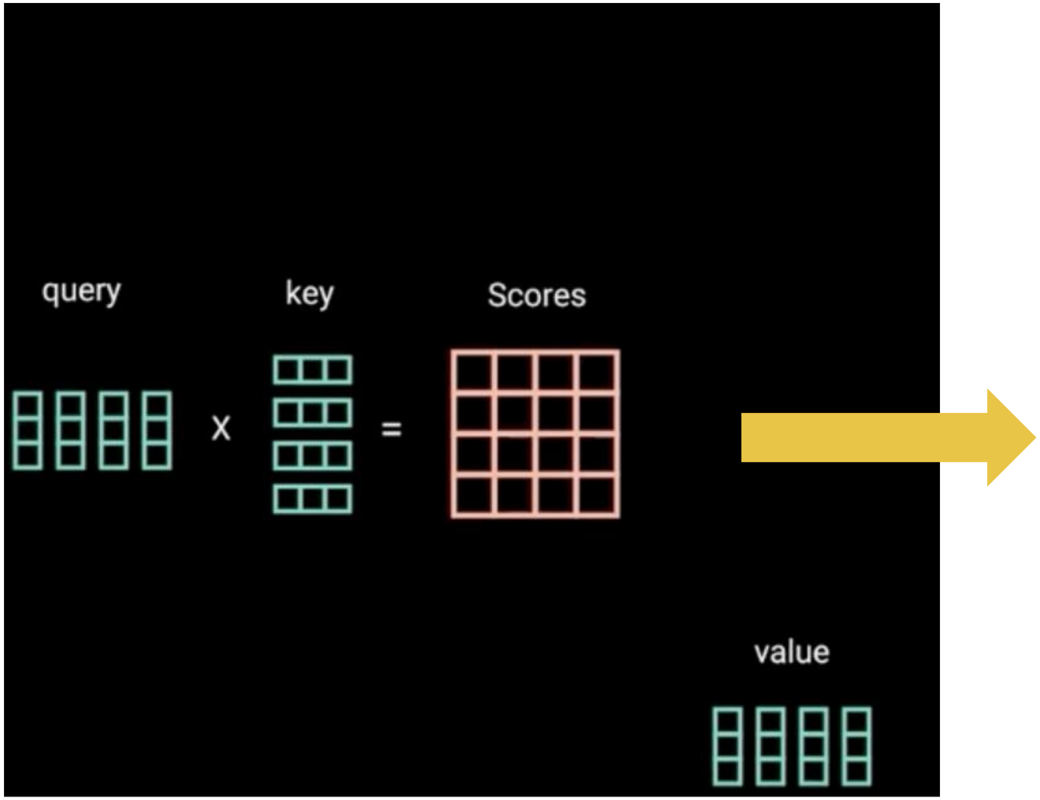
https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html 01.02.2022

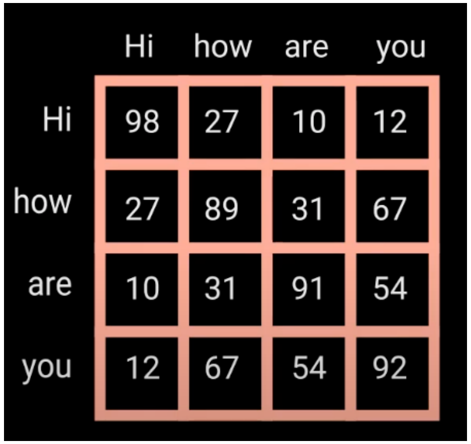
Self-Attention



https://towardsdatascience.com/illustrated-guide-to-transformers-step-by-step-explanation-f74876522bc0

Self-Attention

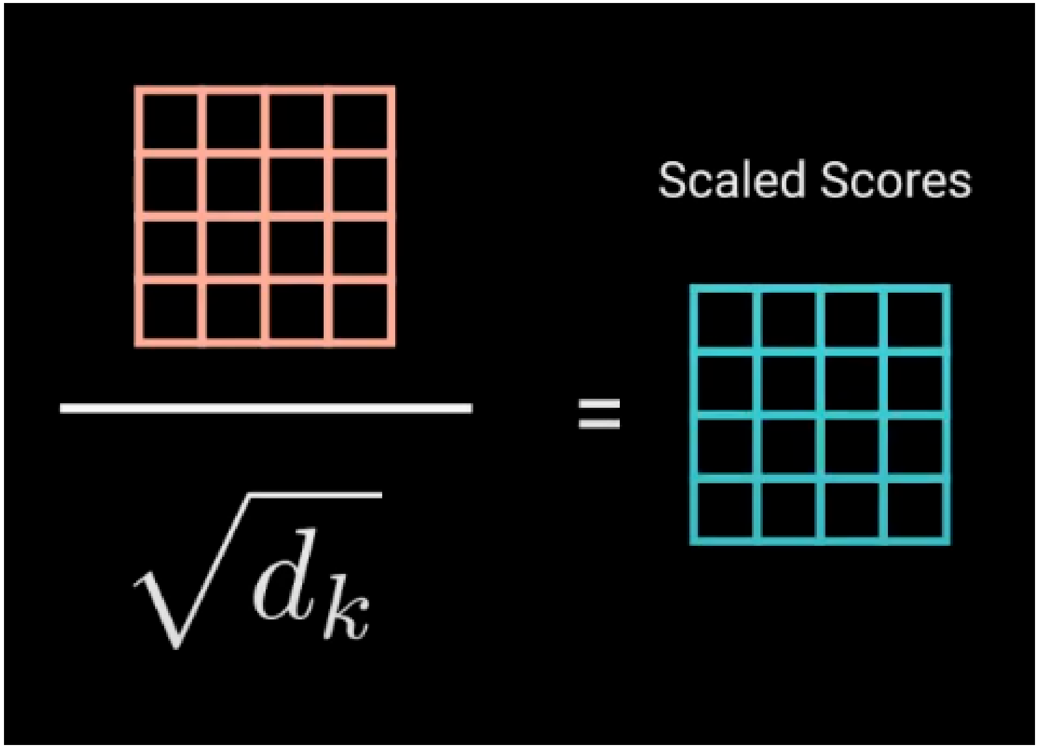




	Hi	how	are	you
Hi	98	27	10	12
how	27	89	31	67
are	10	31	91	54
you	12	67	54	92

https://towardsdatascience.com/illustrated-guide-to-transformers-step-by-step-explanation-f74876522bc0

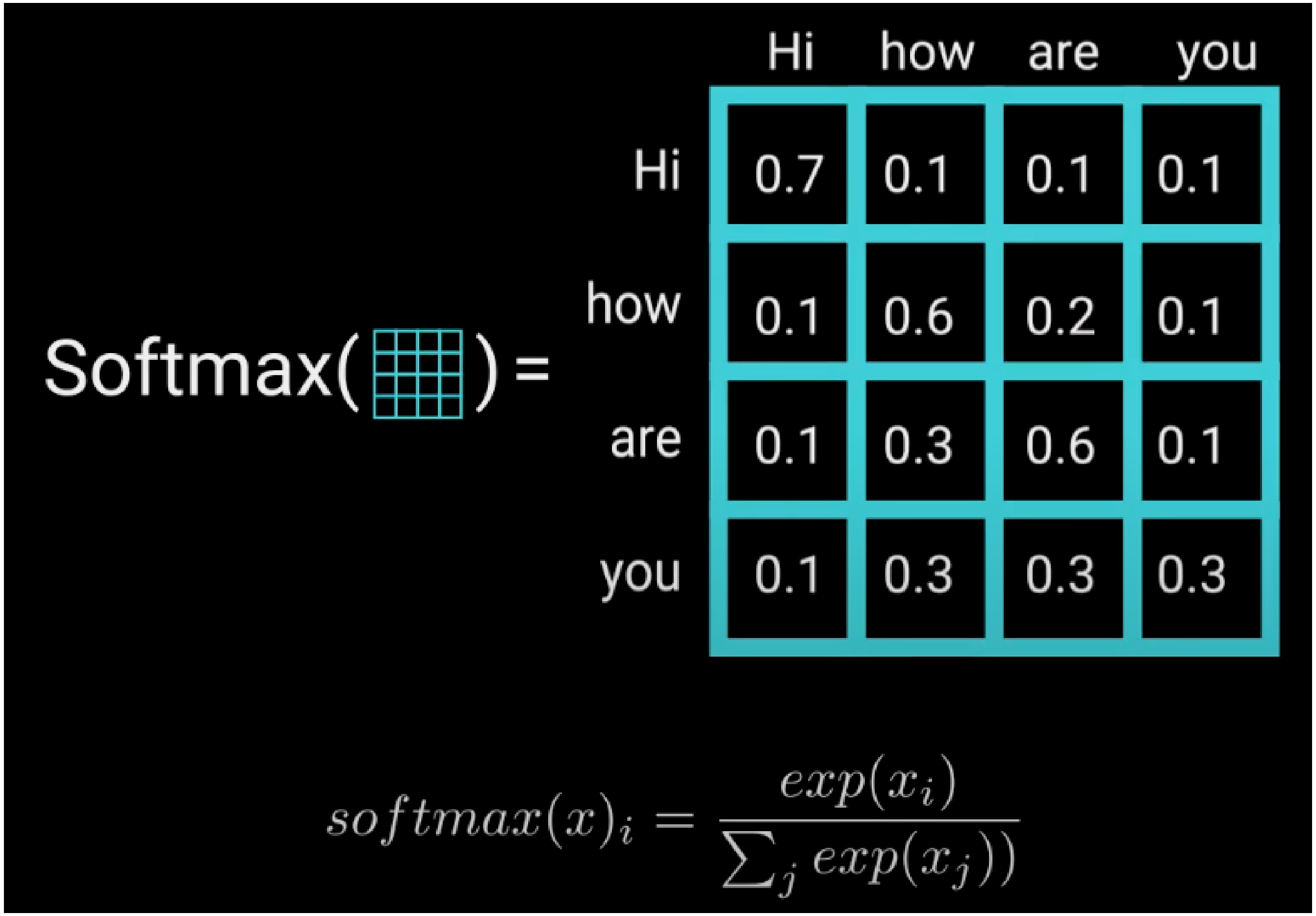
Self-Attention



https://towardsdatascience.com/illustrated-guide-to-transformers-step-by-step-explanation-f74876522bc0

- More stable gradients
- Prevents exploding effects

Self-Attention



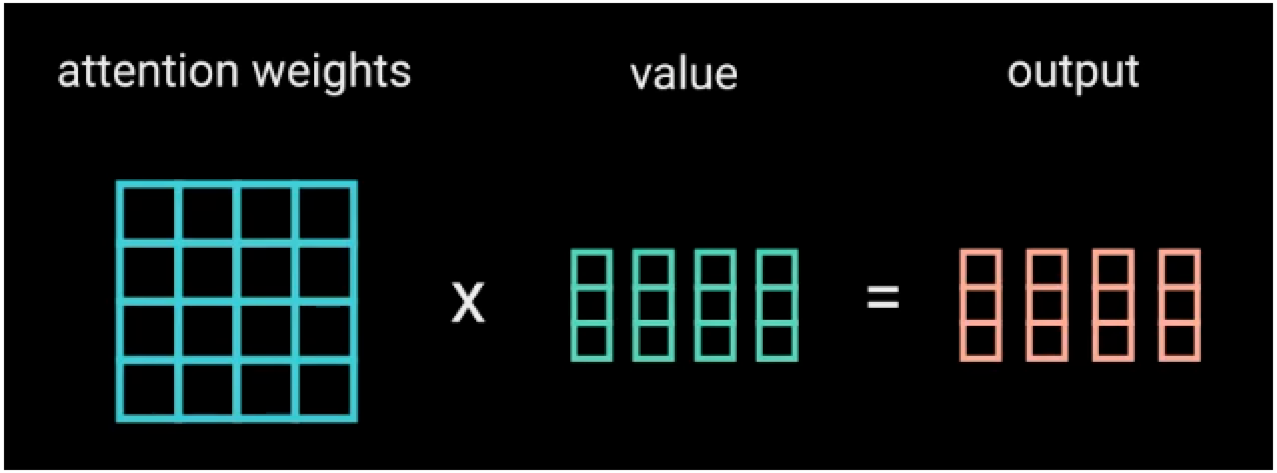
Softmax(▦) =

	Hi	how	are	you
Hi	0.7	0.1	0.1	0.1
how	0.1	0.6	0.2	0.1
are	0.1	0.3	0.6	0.1
you	0.1	0.3	0.3	0.3

$$softmax(x)_i = \frac{exp(x_i)}{\sum_j exp(x_j))}$$

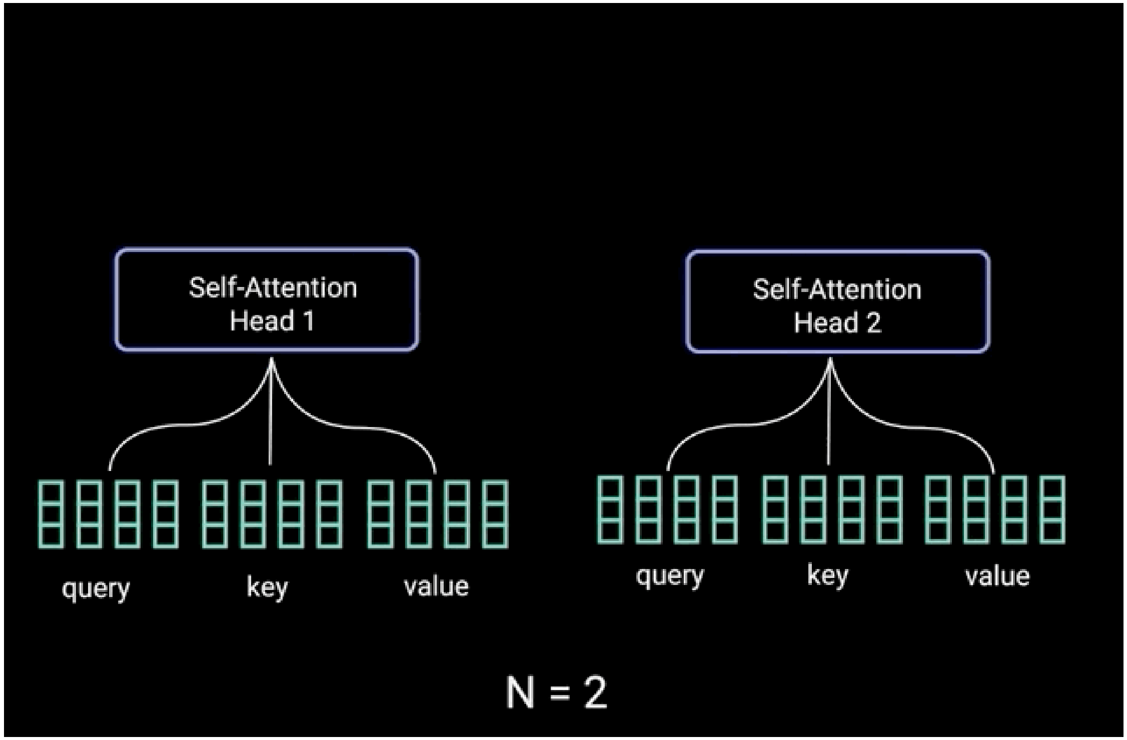
https://towardsdatascience.com/illustrated-guide-to-transformers-step-by-step-explanation-f74876522bc0

Self-Attention



https://towardsdatascience.com/illustrated-guide-to-transformers-step-by-step-explanation-f74876522bc0

Multi-Head Attention



https://towardsdatascience.com/illustrated-guide-to-transformers-step-by-step-explanation-f74876522bc0

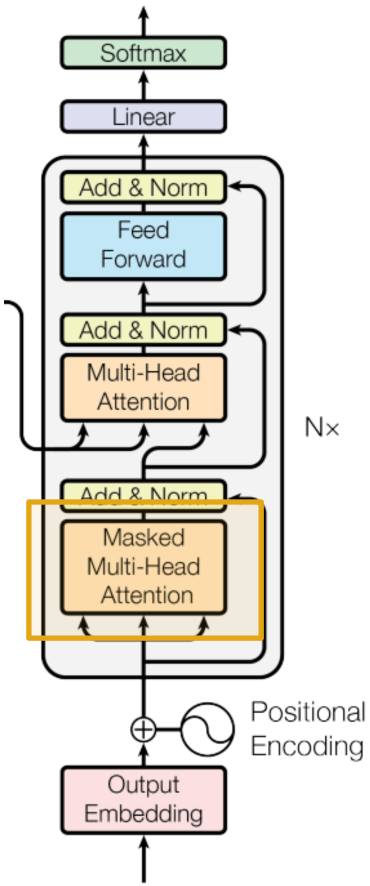
- Calculate Self-Attention multiple times
- Each head learns something different

→ More representation power

Decoder

Output Probabilities

Encoded Representation

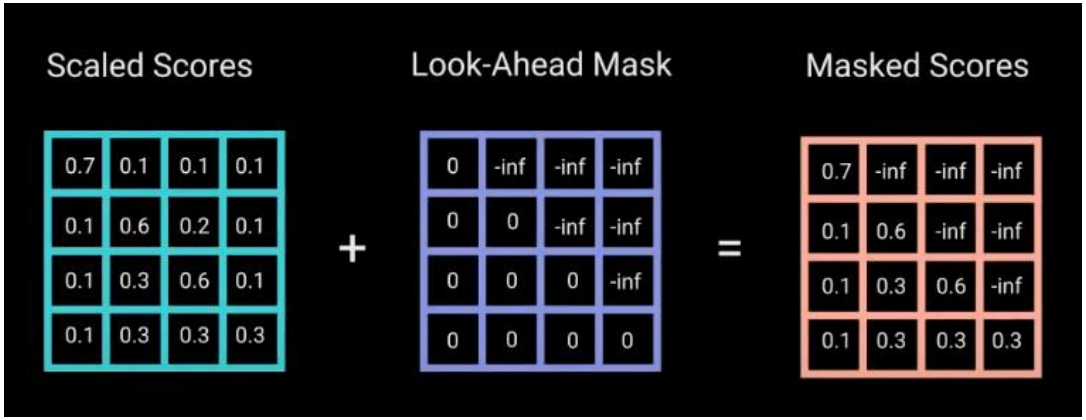


Old Outputs

- Decodes continuous representation
- Generates text sequences
- Works autoregeressiv

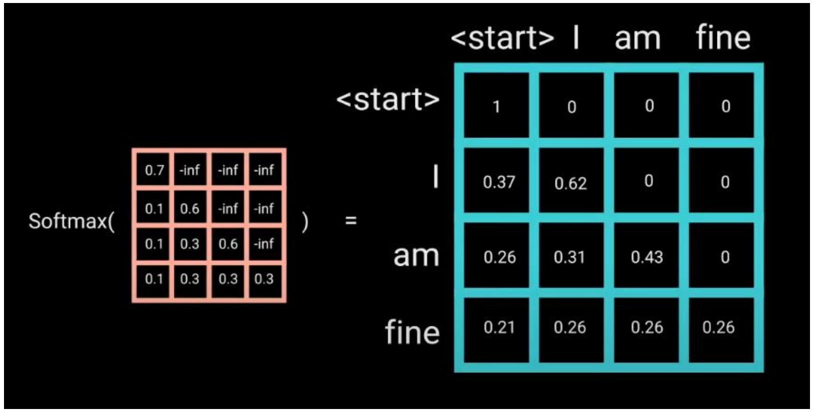
https://arxiv.org/pdf/1706.03762.pdf

Masked-Multi-Head Attention



$$\begin{bmatrix} 0.7 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.6 & 0.2 & 0.1 \\ 0.1 & 0.3 & 0.6 & 0.1 \\ 0.1 & 0.3 & 0.3 & 0.3 \end{bmatrix} + \begin{bmatrix} 0 & -inf & -inf & -inf \\ 0 & 0 & -inf & -inf \\ 0 & 0 & 0 & -inf \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0.7 & -inf & -inf & -inf \\ 0.1 & 0.6 & -inf & -inf \\ 0.1 & 0.3 & 0.6 & -inf \\ 0.1 & 0.3 & 0.3 & 0.3 \end{bmatrix}$$

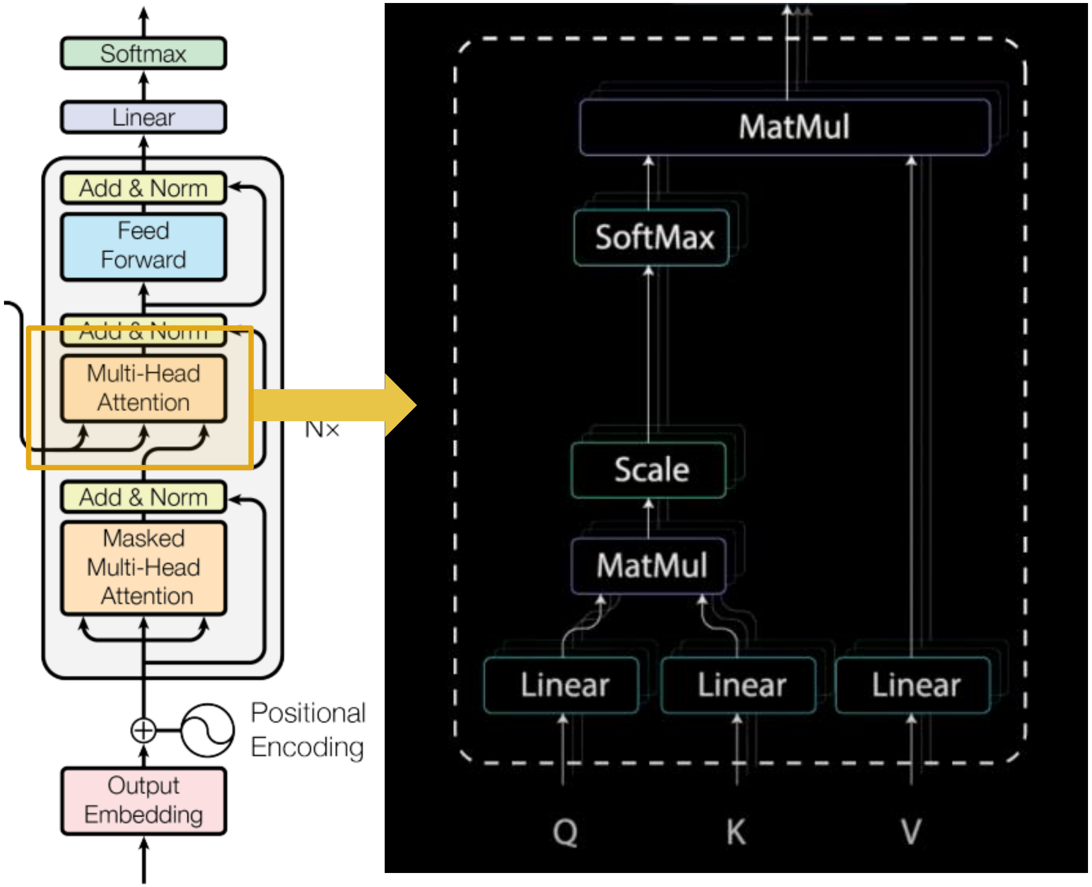
- Only used for training
- Each word gets only the attention score for itself, and the words generated before.



$$\text{Softmax}\left(\begin{bmatrix} 0.7 & -inf & -inf & -inf \\ 0.1 & 0.6 & -inf & -inf \\ 0.1 & 0.3 & 0.6 & -inf \\ 0.1 & 0.3 & 0.3 & 0.3 \end{bmatrix}\right) = \begin{array}{c|cccc} & \text{<start>} & \text{I} & \text{am} & \text{fine} \\ \hline \text{<start>} & 1 & 0 & 0 & 0 \\ \text{I} & 0.37 & 0.62 & 0 & 0 \\ \text{am} & 0.26 & 0.31 & 0.43 & 0 \\ \text{fine} & 0.21 & 0.26 & 0.26 & 0.26 \end{array}$$

https://towardsdatascience.com/illustrated-guide-to-transformers-step-by-step-explanation-f74876522bc0

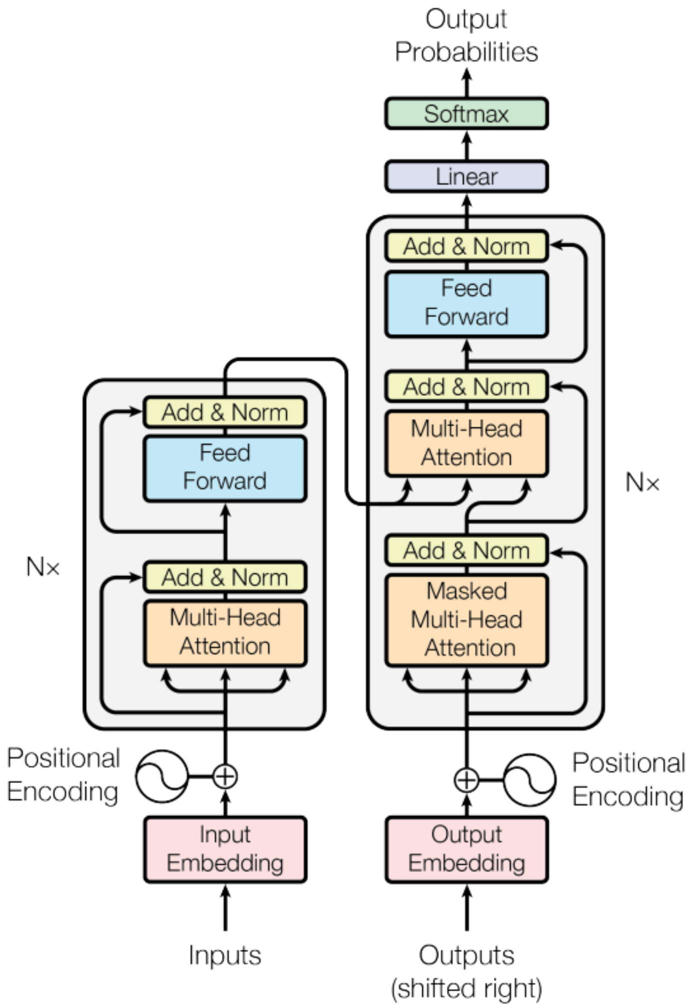
Multi-Head Attention Decoder



- Q and K from Encoder
- V from Masked-Multihead Attention of Decoder
- Helps to focus on the right parts of the encoders output

https://towardsdatascience.com/illustrated-guide-to-transformers-step-by-step-explanation-f74876522bc0

Transformers



https://arxiv.org/pdf/1706.03762.pdf

What you know now!

- Positional Encodings
- Self-Attention
- Multi-Head Attention
- Masked Multi-Head Attention

Hands-on part: Translation with transformers

- Please work through the tutorial 12_Transformers_Translation_Example.ipynb