

Attention Is All You Need

How Transformers and GPT work

Not only ChatGPT, Gemini, Grok, Deepseek, etc

Transformer Architecture introduced in 2017 paper named “Attention is All You Need” and initially was intended for Language Translation

Found also usages in

- [AlphaFold](#), [AlphaGenome](#)
- Robotics (Google RT-1 (Robotics Transformer 1))
- Weather Forecast ([TENT](#), etc)
- [Anomaly Detection](#), [Trajectory Prediction](#)
- Audio, Video
- etc

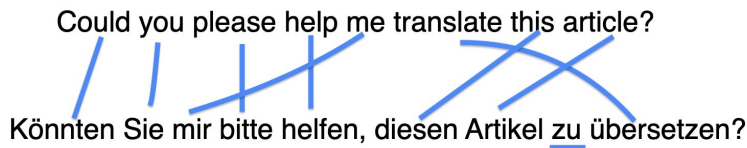


Translation problems

- Different input and output length



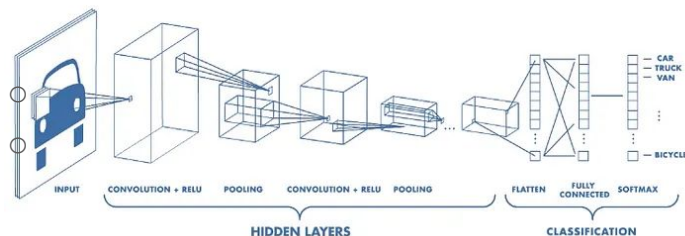
- Changing ordering of the words



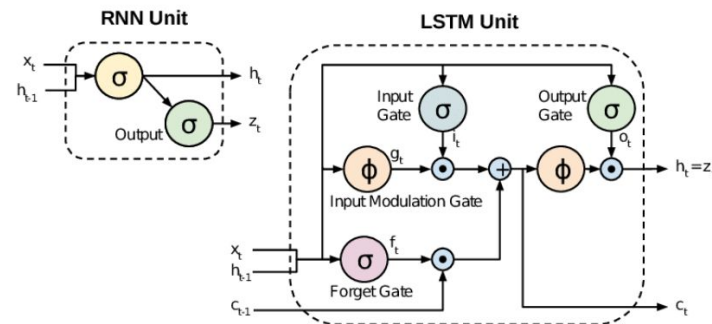
- Morphologically rich languages
- Multiple meanings
- etc

Common Neural Networks for Sequential Data

- RNN - Recurrent Neural Network (1982)
 - LSTM - Long Short-Term Memory
 - GRU - Gated Recurrent Unit
- CNN - Convolutional Neural Network (1989)
 - TCN - Temporal Convolutional Network



- Transformers (2017)



Yann LeCun, CNN inventor

Images from Deepseek, Youtube

<https://ashutoshtrpathi.com/2021/07/02/what-is-the-main-difference-between-rnn-and-lstm-nlp-rnn-vs-lstm/>

Generating Text using RNN

Generating Text with Recurrent Neural Networks

Ilya Sutskever
James Martens
Geoffrey Hinton

University of Toronto, 6 King's College Rd., Toronto, ON M5S 3G4 CANADA

ILYA@CS.UTORONTO.CA
JMARTENS@CS.TORONTO.EDU
HINTON@CS.TORONTO.EDU

6.1.1.1. SAMPLES FROM THE WIKIPEDIA MODEL

We now present a sample from the Wikipedia model. We use ? to indicate the “unknown” character. The sample below was obtained by running the MRNN less than 10 times and selecting the most intriguing sample. The beginning of the paragraph and the parentheses near the end are particularly interesting. The MRNN was initialized with the phrase “The meaning of life is”:

The meaning of life is the tradition of the ancient human reproduction: it is less favorable to the good boy for when to remove her bigger. In the show's agreement unanimously resurfaced. The wild pastured with consistent street forests were incorporated by the 15th century BE. In 1996 the primary rapford undergoes an effort that the reserve conditioning, written into Jewish cities, sleepers to incorporate the .St Eurasia that activates the population. Mar??a Nationale, Kelli, Zedlat-Dukastoe, Florendon, Ptu's thought is. To adapt in most parts of North America, the dynamic fairy Dan please believes, the free speech are much related to the

June 12, 2017 - Transformers
June 11, 2018 - GPT-1

https://icml.cc/2011/papers/524_icmlpaper.pdf



[Geoffrey Hinton](#)

Image from <https://www.nobelprize.org/prizes/physics/2024/hinton/facts/>

Bahdanau Attention

NEURAL MACHINE TRANSLATION
BY JOINTLY LEARNING TO ALIGN AND TRANSLATE

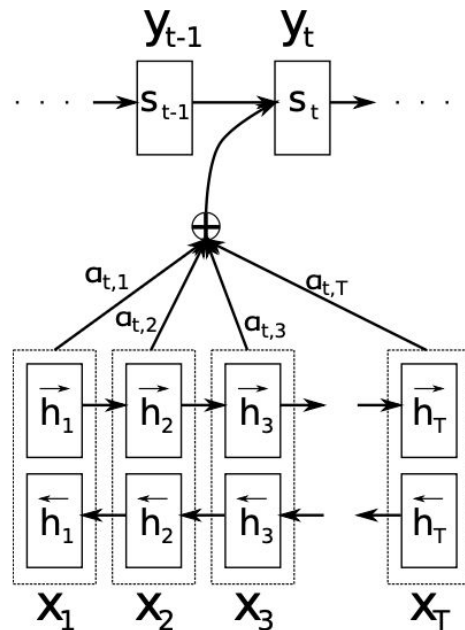
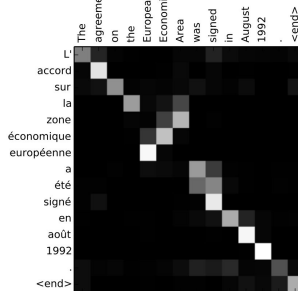
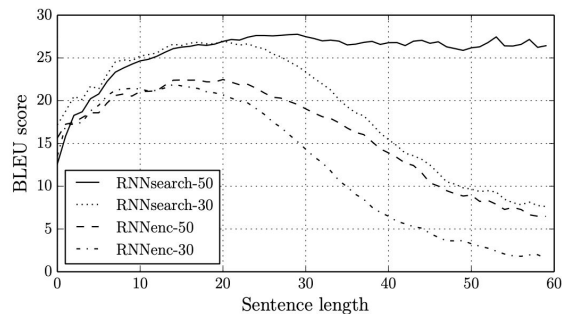
Dzmitry Bahdanau
Jacobs University Bremen, Germany

KyungHyun Cho Yoshua Bengio*
Université de Montréal

<https://arxiv.org/abs/1409.0473>

by D Bahdanau · 2014 · Cited by 39758

Introduced attention mechanism (*additive*)



Images from PDF

Transformer architecture

Introduced in 2017, paper “Attention Is All You Need”

- For text translation
- Consists from **Encoder** and **Decoder**
 - GPT is Decoder only architecture
- Introduced Dot-Product Attention
- Introduced Multi-Head Attention
- RNN -> FNN (MLP)
- Autoregressive Decoder
- Scalable

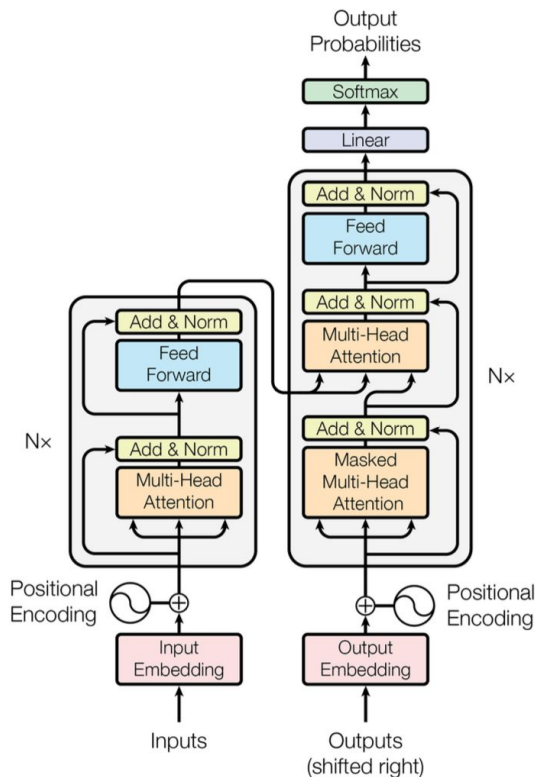


Figure 1: The Transformer - model architecture.

Tokenization

- Algorithm: BPE (byte-pair encoding)
 - originally designed for data compression
- Online: <https://platform.openai.com/tokenizer>

Many words map to one token, but some don't: indivisible.

- Spaces are included !
- <bos> <eos> and other special tokens.

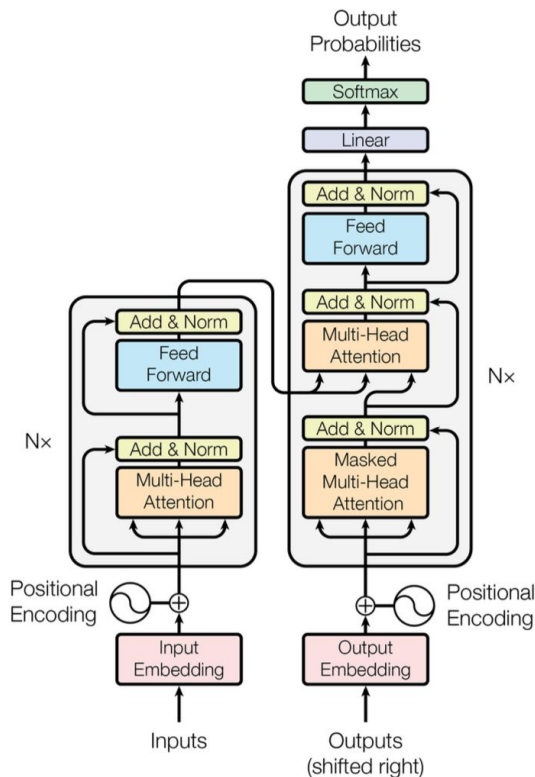


Figure 1: The Transformer - model architecture.

Embedding

- Token -> Vector

Vector size: Transformer - 512, GPT3 - 12288

- Learnable parameter (Initially random)
- Word2Vec, GLoVe are not used!

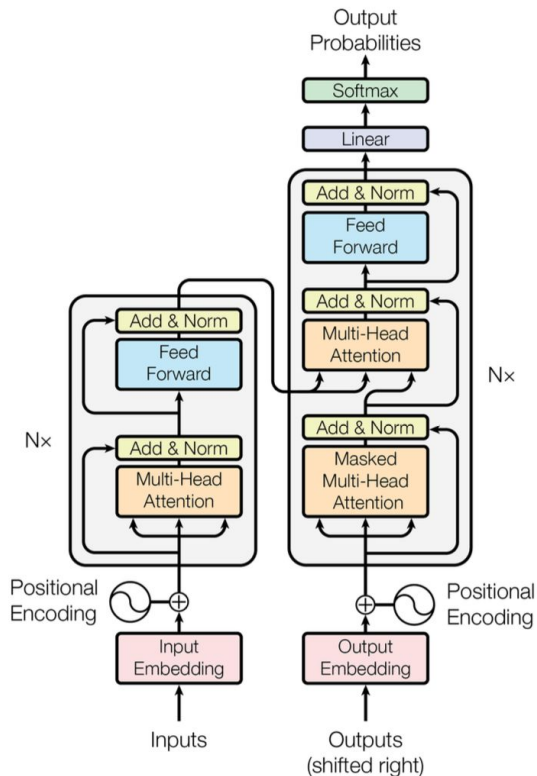
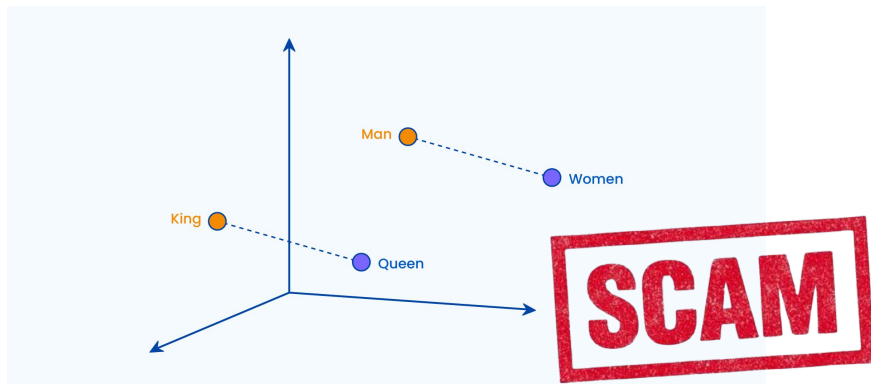


Figure 1: The Transformer - model architecture.

Image from <https://www.searchunify.com/su/sudo-technical-blogs/demystifying-contextual-query-embedding/>

Positional Encoding

- The cat likes to chase the mouse
not equal to
The mouse likes to chase the cat
- $X = X + PE$

$$PE_{(pos, 2i)} = \sin(pos/10000^{2i/d_{\text{model}}})$$

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

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

Better: RoPE (LLaMa, Deepseek), ALiBi

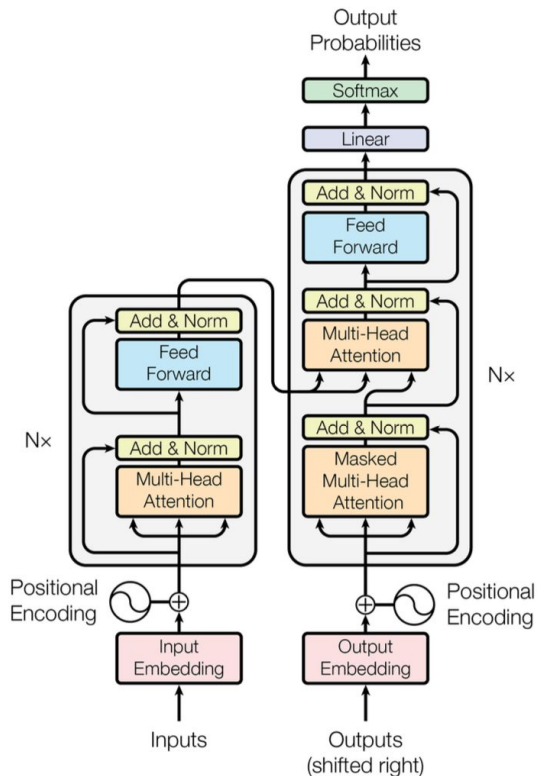
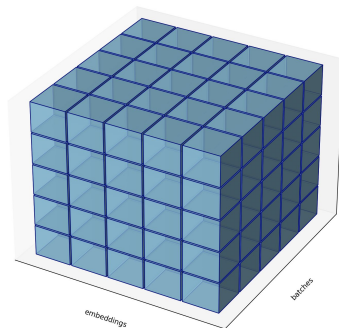


Figure 1: The Transformer - model architecture.

Data format. Scalability

Transformer accepts data as 3d matrix:

- Embeddings (fixed size)
- Sequences
- Batches (max load to GPU)



Uses matrix tricks inside.

<https://www.calculator.net/matrix-calculator.html>

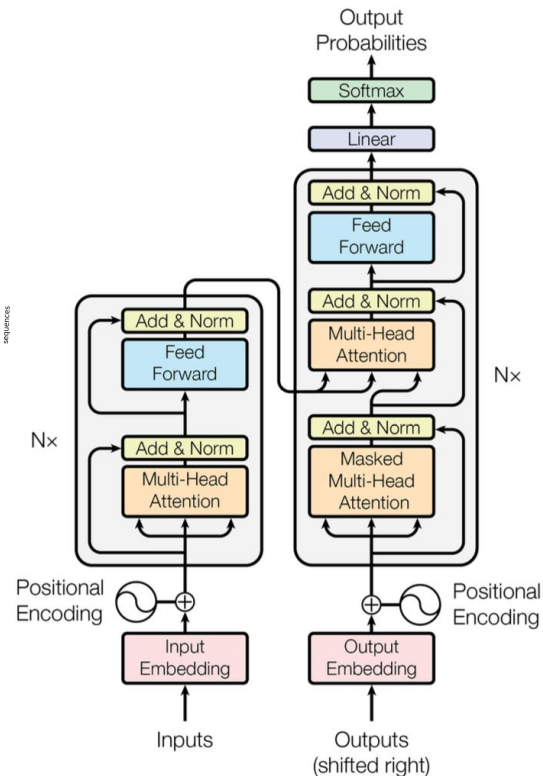


Figure 1: The Transformer - model architecture.

Add (= Residual Connection)

$$X = X + \text{attention}(X)$$

attention: embeddings see each other

and

$$X = X + \text{ffn}(X)$$

ffn: embeddings do not see each other

Residual function refers to the idea that instead of learning a direct mapping from input to output, the network learns the difference (or residual) between the input and the desired output.

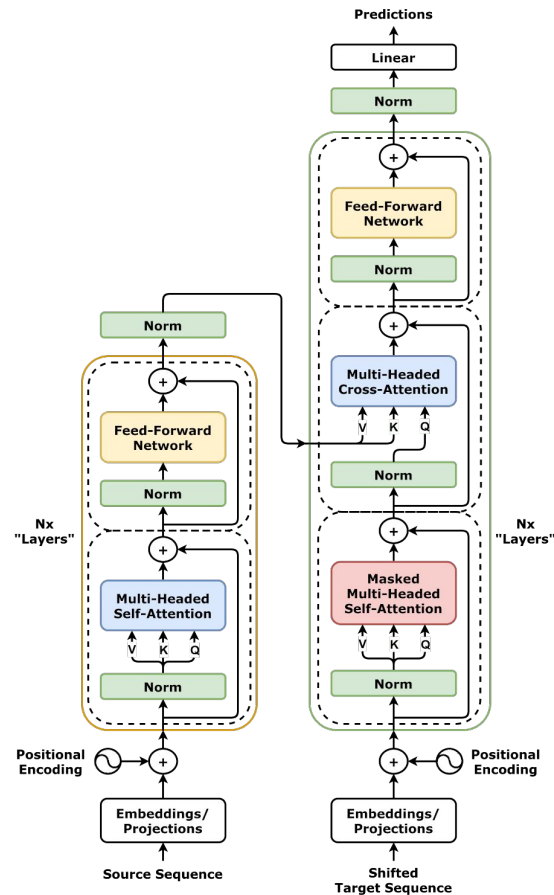
Introduced in ResNet (2015, <https://arxiv.org/pdf/1512.03385>)

Deep Residual Learning for Image Recognition

Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun

Microsoft Research

{kahe, v-xiangz, v-shren, jiansun}@microsoft.com



Scaled Dot-Product Attention (original name)

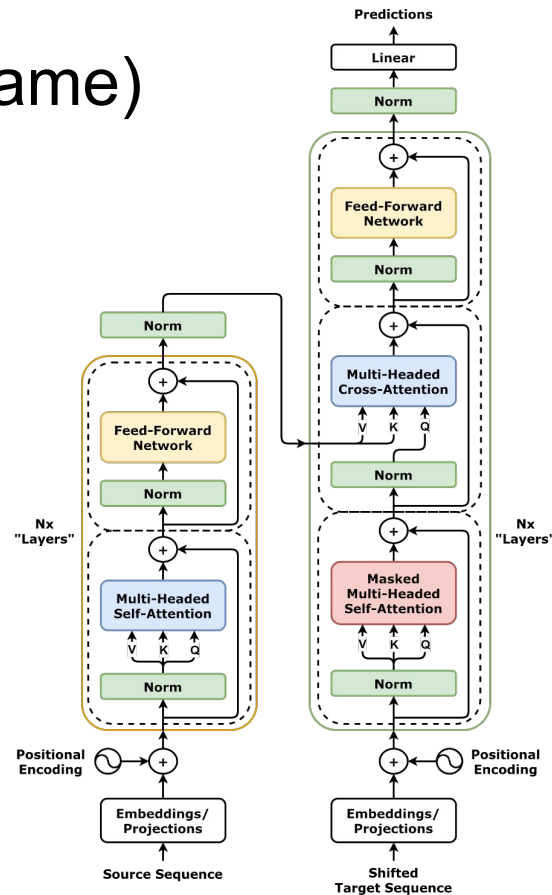
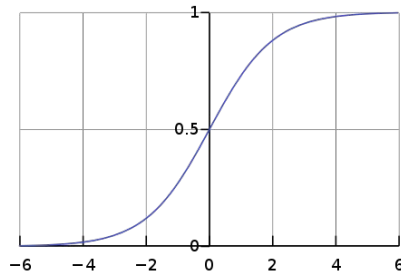
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where $Q = X * W_q$, $K = X * W_k$, $V = X * W_v$

W_q , W_k , W_v - Learnable parameters

softmax:

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}}$$

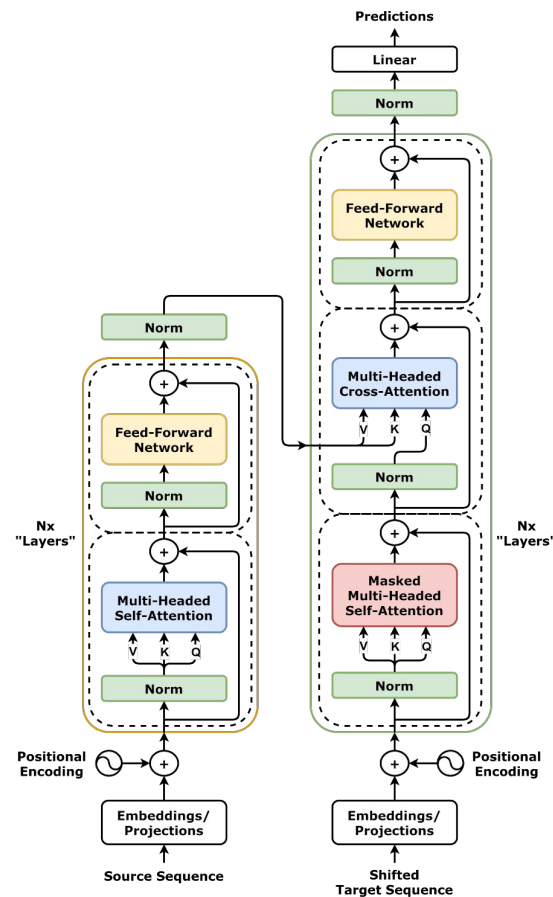


Multi-Head Attention

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O$$

where $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$

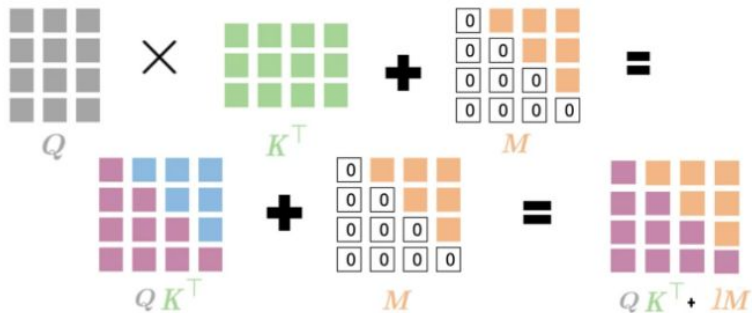
W^O returns matrix back to embeddings dimension



Masked (Casual) Attention

Causal attention math

→ Minus infinity -in practice, a huge negative number



For progressive learning:

The
The cat
The cat likes
The cat likes to

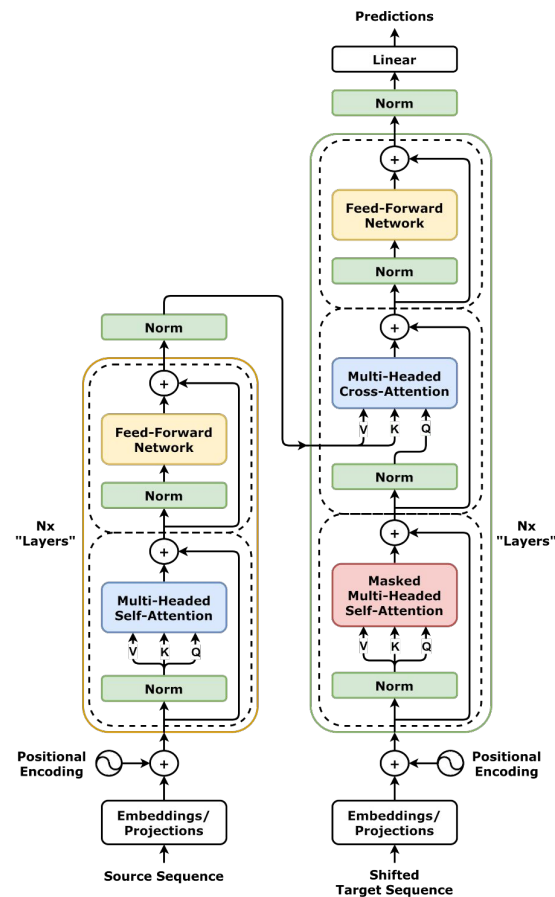


Image from <https://blog.sailor.plus/deep-learning/attention/>

Cross-Attention and Self-Attention

In Cross Attention decoder is calculating K and V from encoder sequence.

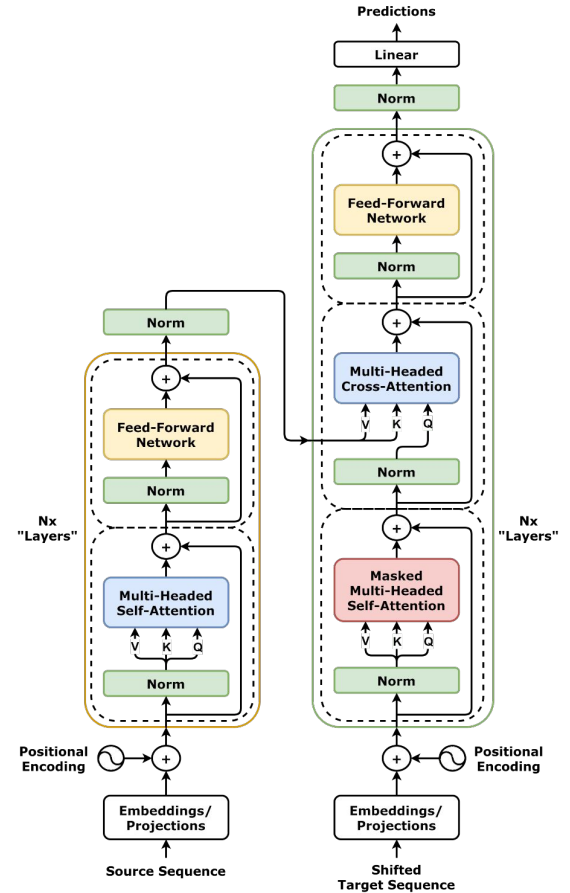
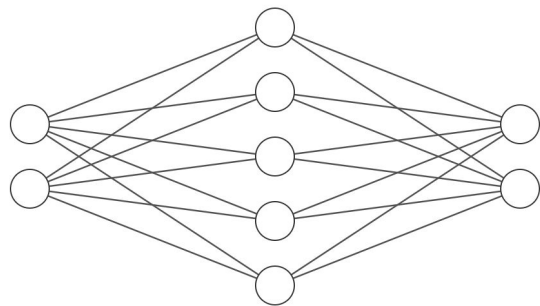


Image from Wikipedia

Feed-Forward Neural Network Theory

- Forward Propagation
- Weights and Biases (W and b)
- Activation Functions
 - ReLU, GeLU, tanh, sigmoid, ...
- Loss Functions
 - measure how well predictions match the actual data
 - Mean Squared Error (MSE), Cross-Entropy, ...
- Backpropagation
- Learning Rate
- Dropout (2016, [Hinton paper](https://arxiv.org/abs/1207.0593))
 - prevents overfitting, neuron dying



Input Layer $\in \mathbb{R}^2$

Hidden Layer $\in \mathbb{R}^5$

Output Layer $\in \mathbb{R}^2$

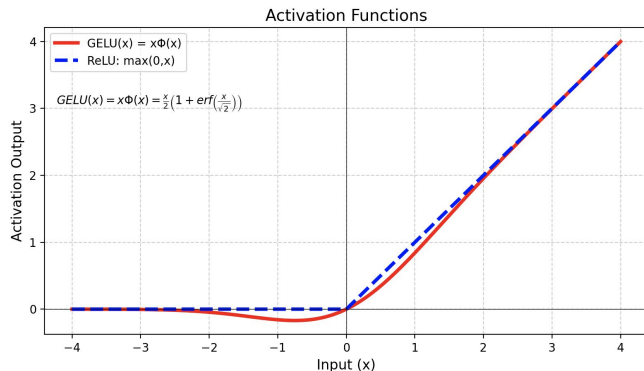
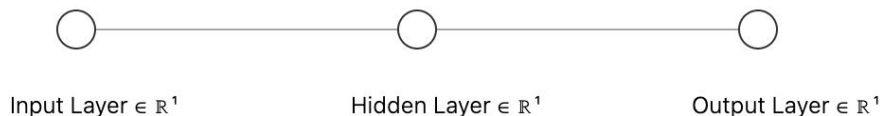


Image from <https://alexlenail.me/NN-SVG/index.html>

FNN in Transformer

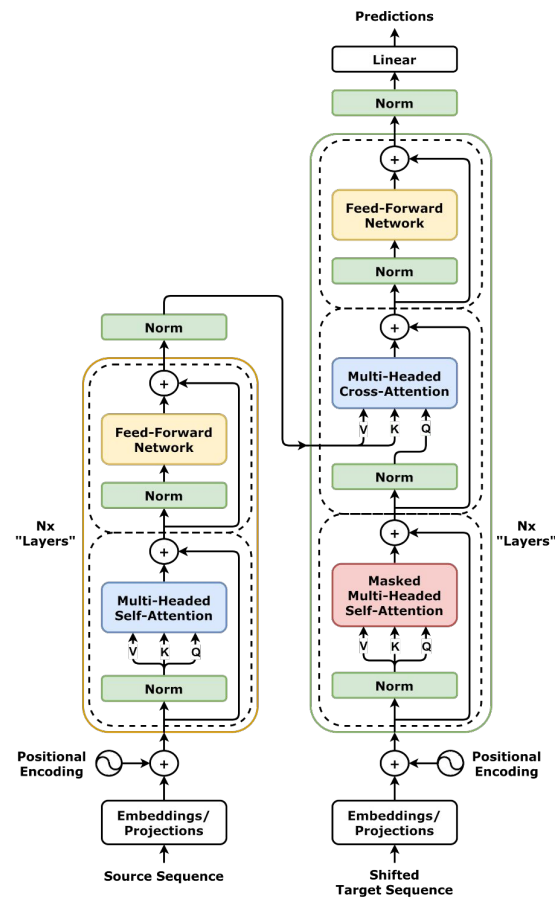
Minimalistic



$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

Uses matrix tricks inside.

<https://www.calculator.net/matrix-calculator.html>



Norm

$$X = \text{layer_norm}(X)$$

The key of layer norm is to normalize the input to the layer using the mean and standard deviation.

Layer norm plays two roles in neural networks:

- Projects the key vectors onto a hyperplane.
- Scales the key vectors to have the same length.

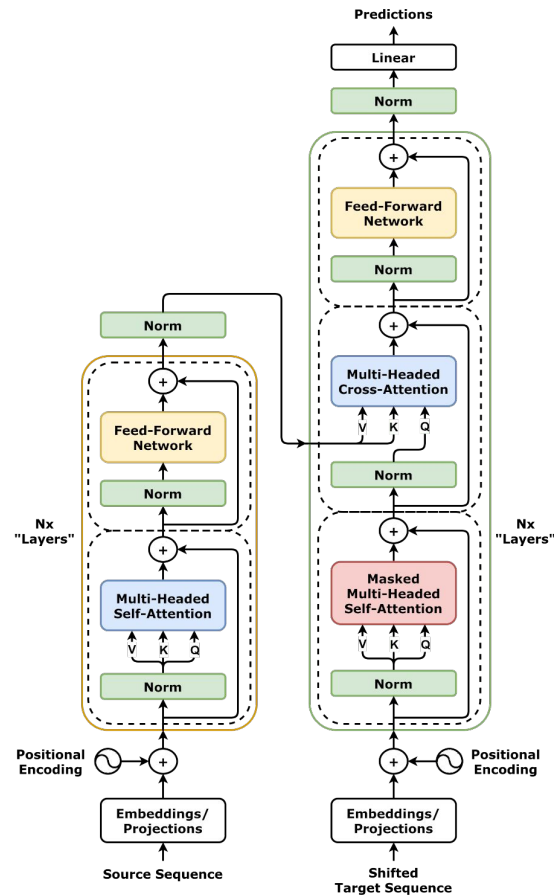
Paper: <https://arxiv.org/abs/1607.06450> (2016)

Layer Normalization

Jimmy Lei Ba
University of Toronto
jimmy@psi.toronto.edu

Jamie Ryan Kiros
University of Toronto
rkiros@cs.toronto.edu

Geoffrey E. Hinton
University of Toronto
and Google Inc.
hinton@cs.toronto.edu



Inference in the Transformer Model

Or back to the next token (classification task)

Actions:

- **logits** = $X * \text{Head}$
Head has the same dimension like vocabulary!
- **softmax**
Convert to probabilities
- **argmax / multinomial** (if inference)
to find next index for embedding (forward) or
- **cross entropy** (if learning)
and backpropagation if learning

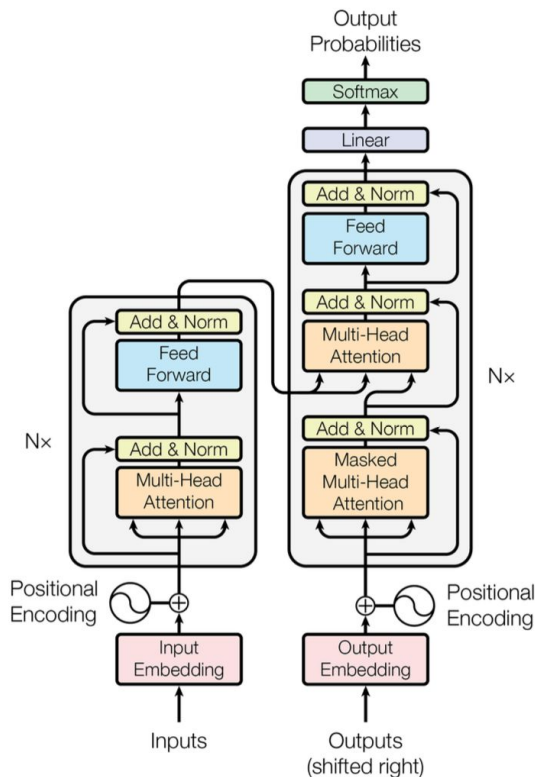


Figure 1: The Transformer - model architecture.

90 gigawatts needed

Schmidt Warns Congress: U.S. Must Win Race to Superintelligent AI

“The typical nuclear power plant is 1 gigawatt. We have roughly 90 of them. We’re talking about [an additional] 90 gigawatts needed in the next 3-5 years in order to maintain this leadership,” Schmidt said, highlighting the massive burden AI is putting on the energy grid. He said that supporting the energy needs of AI will require using every form of energy.

Sam Altman Proposes 5-Gigawatt AI Data Center



Image from leonardo.ai

Deep Learning. Backprop.

Autograd (automatic differentiation) is a system that automatically computes gradients (derivatives) of tensors in machine learning frameworks like PyTorch, TensorFlow.

- TensorFlow (2015 by Google)
 - define-then-run
- PyTorch (2016 by Facebook)
 - define-by-run

Letter | Published: 09 October 1986

Learning representations by back-propagating errors

[David E. Rumelhart](#), [Geoffrey E. Hinton](#) & [Ronald J. Williams](#)

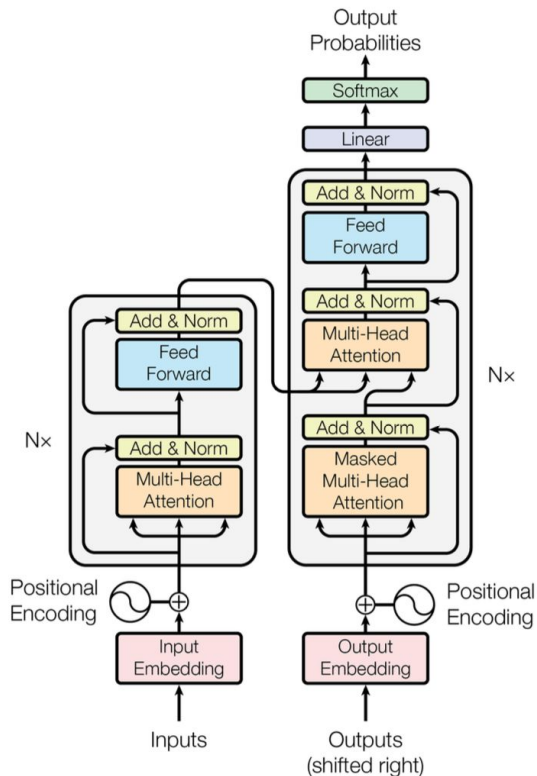


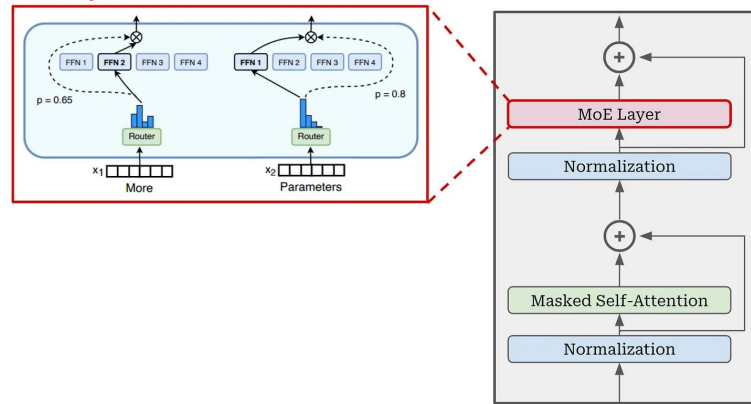
Figure 1: The Transformer - model architecture.

MoE - Mixture of Experts (GPT)

The concept of Mixture of Experts (MoE) was first introduced in 1991 by Robert Jacobs and *Geoffrey Hinton* in the paper [Adaptive Mixtures of Local Experts](#)

Used in DeepSeek, Mixtral, etc

MoE Layer



Demo. Q&A

Code and presentaion available at

<https://github.com/hza/askGPT>