

# Was leisten grosse Sprachmodelle und wie funktionieren sie ?

Prof. Christian Bauckhage



**Fraunhofer**  
IAIS

**Lamarr Institute for Machine Learning and Artificial Intelligence**

# AI in the 2010s

$$\begin{array}{r} \text{big data} \\ + \text{affordable HPC} \\ + \text{open source software} \\ + \text{deep learning systems} \\ \hline = \text{progress in AI} \end{array}$$

## AI in the 2020s

+ very big data  
+ affordable HPC  
+ open source software  
+ transformer networks

---

= dramatic progress in AI

fun with ChatGPT

## quantum computing knowledge (December 2022)

Can you draw a quantum circuit that computes the logical AND of two qubits?

The following diagram shows a quantum circuit that computes the logical AND of two qubits:

The two qubits are first initialized in the 0 state. Then, a controlled-NOT gate is applied to the first qubit, which flips it to the 1 state if the second qubit is in the 1 state. The output of the circuit is then determined by the result of the AND gate. If both inputs are 1, the output is 1. If either input is 0, the output is 0.

Can you produce corresponding qiskit code?

The corresponding Qiskit code for this circuit is as follows:

```
from qiskit import QuantumCircuit, QuantumRegister, ClassicalRegister

# Create a Quantum Register with 2 qubits
q = QuantumRegister(2)

# Create a Classical Register with 1 bit
c = ClassicalRegister(1)

# Create the Quantum Circuit
circuit = QuantumCircuit(q, c)

# Initialize both qubits to 0
circuit.x(q[0])
circuit.x(q[1])

# Apply a controlled-NOT gate to the first qubit
circuit.cx(q[1], q[0])

# Measure the output
circuit.measure(q[0], c[0])
```



on careful inspection, the answer and the code are *wrong*

we all know that ChatGPT tends to give incorrect answers

technically, this is not a problem at all

experts know how to iron out this kind of stupidity  
(see e.g. Wolfram's blog post from January 2023)

by now, ChatGPT and Wolfram Alpha are coupled

sociologically, trustworthiness is a whole different story . . .

# quantum computing homework (April 2023)

the following is a matrix representation of a quantum operator written in LaTeX

```
R = \begin{bmatrix} \cos \frac{\theta}{2} & -\sin \frac{\theta}{2} \\ \sin \frac{\theta}{2} & \cos \frac{\theta}{2} \end{bmatrix}
```

can you show that this operator is unitary?

To show that the operator  $R$  is unitary, we need to demonstrate that it satisfies the condition  $R^\dagger R = R R^\dagger = I$ , where  $R^\dagger$  is the Hermitian conjugate of  $R$ , and  $I$  is the identity matrix.

First, let's find the Hermitian conjugate of  $R$ . Since  $R$  is a real matrix, its Hermitian conjugate is just its transpose:

```
R^\dagger = \begin{bmatrix} \cos \frac{\theta}{2} & \sin \frac{\theta}{2} \\ -\sin \frac{\theta}{2} & \cos \frac{\theta}{2} \end{bmatrix}
```



Next, let's calculate  $R^{\dagger}R$ :

$$R^{\dagger}R = \begin{bmatrix} \cos \frac{\theta}{2} & \sin \frac{\theta}{2} \\ -\sin \frac{\theta}{2} & \cos \frac{\theta}{2} \end{bmatrix} \begin{bmatrix} \cos \frac{\theta}{2} & -\sin \frac{\theta}{2} \\ \sin \frac{\theta}{2} & \cos \frac{\theta}{2} \end{bmatrix} = \begin{bmatrix} \cos^2 \frac{\theta}{2} + \sin^2 \frac{\theta}{2} & 0 \\ 0 & \cos^2 \frac{\theta}{2} + \sin^2 \frac{\theta}{2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I.$$

Similarly, let's calculate  $RR^{\dagger}$ :

$$RR^{\dagger} = \begin{bmatrix} \cos \frac{\theta}{2} & -\sin \frac{\theta}{2} \\ \sin \frac{\theta}{2} & \cos \frac{\theta}{2} \end{bmatrix} \begin{bmatrix} \cos \frac{\theta}{2} & \sin \frac{\theta}{2} \\ -\sin \frac{\theta}{2} & \cos \frac{\theta}{2} \end{bmatrix} = \begin{bmatrix} \cos^2 \frac{\theta}{2} + \sin^2 \frac{\theta}{2} & 0 \\ 0 & \cos^2 \frac{\theta}{2} + \sin^2 \frac{\theta}{2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I.$$

Therefore, since  $R^{\dagger}R = RR^{\dagger} = I$ , we conclude that the operator  $R$  is unitary.



apart from minute syntax errors in the  $\text{\LaTeX}$  code, the answer is correct

if we compile the  $\text{\LaTeX}$  expressions, we find that ChatGPT proposes to prove that  $R^\dagger R = RR^\dagger = I$  and then proceeds as follows

$$R^\dagger = \begin{bmatrix} \cos \frac{\theta}{2} & \sin \frac{\theta}{2} \\ -\sin \frac{\theta}{2} & \cos \frac{\theta}{2} \end{bmatrix}$$

$$R^\dagger R = \begin{bmatrix} \cos \frac{\theta}{2} & \sin \frac{\theta}{2} \\ -\sin \frac{\theta}{2} & \cos \frac{\theta}{2} \end{bmatrix} \begin{bmatrix} \cos \frac{\theta}{2} & -\sin \frac{\theta}{2} \\ \sin \frac{\theta}{2} & \cos \frac{\theta}{2} \end{bmatrix} = \begin{bmatrix} \cos^2 \frac{\theta}{2} + \sin^2 \frac{\theta}{2} & 0 \\ 0 & \cos^2 \frac{\theta}{2} + \sin^2 \frac{\theta}{2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I$$

$$RR^\dagger = \begin{bmatrix} \cos \frac{\theta}{2} & -\sin \frac{\theta}{2} \\ \sin \frac{\theta}{2} & \cos \frac{\theta}{2} \end{bmatrix} \begin{bmatrix} \cos \frac{\theta}{2} & \sin \frac{\theta}{2} \\ -\sin \frac{\theta}{2} & \cos \frac{\theta}{2} \end{bmatrix} = \begin{bmatrix} \cos^2 \frac{\theta}{2} + \sin^2 \frac{\theta}{2} & 0 \\ 0 & \cos^2 \frac{\theta}{2} + \sin^2 \frac{\theta}{2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I$$

# take home message

we need to rethink how to teach, how to prepare exercises, exams, . . .

we (probably) need to revise learning goals, the extent of curricula, . . .

THE STRAITS TIMES

SINGAPORE

▲ LOG IN

Students, teachers will learn to properly  
use tools like ChatGPT: Chan Chun  
Sing



ChatGPT can be a useful tool for learning only when students have mastered basic concepts and thinking skills. PHOTO: ST FILE

# real work with GPT-4

but first an introductory remark . . .

# our group works on financial document analysis

Deußer, Pielka, Pucknat, Jacob, Dilmaghani, Nourimand, Kliem, Loitz, Bauckhage, Sifa: *Contradiction Detection in Financial Reports*, Proc. Northern Lights Deep Learning Workshop, 2023

Biesner, Ramamurthy, Stenzel, Lübbering, Hillebrand, Ladi, Pielka, Loitz, Bauckhage, Sifa: *Anonymization of German Financial documents using neural network-based language models with contextual word representations*, Int. J. of Data Science and Analytics, 2022

Hillebrand, Deußer, Dilmaghani, Kliem, Loitz, Bauckhage, Sifa: *KPI-BERT: A Joint Named Entity Recognition and Relation Extraction Model for Financial Reports*, Proc. ICPR, 2022

Chapman, Hillebrand, Stenzel, Deußer, Biesner, Bauckhage, Sifa: *Towards Generating Financial Reports from Tabular Data Using Transformers*, Proc. ICPR, 2022

Ramamurthy, Pielka, Stenzel, Bauckhage, Sifa, Dilmaghani, Warning, Kliem, Loitz: *ALIBERT: Improved Automated List Inspection (ALI) with BERT*, Proc. ACM Symp. Document Engineering, 2021

Pielka, Sifa, Hillebrand, Biesner, Ramamurthy, Ladi, Bauckhage: *Tackling Contradiction Detection in German Using Machine Translation and End-to-End Recurrent Neural Networks*, Proc. ICPR, 2020

⋮

# KPI extraction

What are the relevant KPIs in the following text?

Example:

Input:

As of September 28, 2019 and September 29, 2018, the Company had total deferred revenue of \$8.1 billion and \$8.8 billion, respectively.

Output:

1. deferred revenue (2018): \$8.8 billion
2. deferred revenue (2019): \$8.1 billion

Input:

The Company's investment policy and strategy are focused on the preservation of capital and supporting the Company's liquidity

[...]

performed a sensitivity analysis to determine the impact a change in interest rates would have on the value of the investment portfolio assuming a 100 basis point parallel shift in the yield curve. Based on investment positions as of September 28, 2019 and September 29, 2018, a hypothetical 100 basis point increase in interest rates across all maturities would result in a \$2.8 billion and \$4.9 billion decline

[...]

As of September 28, 2019 and September 29, 2018, the Company had outstanding floating- and fixed-rate notes with varying maturities for an aggregate carrying amount of \$102.1 billion and \$102.5 billion

[...]

Impact of a 100 basis point increase in interest rates on investment portfolio value (2018):  
\$4.9 billion decline

Impact of a 100 basis point increase in interest rates on investment portfolio value (2019):  
\$2.8 billion decline

Aggregate carrying amount of outstanding floating- and fixed-rate notes (2018):  
\$102.5 billion

Aggregate carrying amount of outstanding floating- and fixed-rate notes (2019):  
\$102.1 billion

[...]



GPT-4 correctly understands and generalizes complex, example-based instructions

it can solve higher cognitive tasks such KPI extraction, contradiction detection, . . .

extensive tests suggest that it performs better than special purpose models based on small transformers which were specifically trained on financial documents



# take home message

the market for high-skill white collar jobs is headed for disruption



Home > News > Research > Goldman: Artificial Intelligence Will Boost Global GDP by 7%

Research April 12, 2023

## Goldman: Artificial Intelligence Will Boost Global GDP by 7%

*There will also be a job-loss toll, but, as in previous technology leaps, many displaced workers will fit into newly created positions, the firm's study says.*

By Larry Light



HOME > TECH

## AI systems like ChatGPT could impact 300 million full-time jobs worldwide, with administrative and legal roles some of the most at risk, Goldman Sachs report says

Business Insider May 26, 2023 1:00 PM GMT+0



Illustration of a trading floor with "Goldman Sachs" branding. © 2023

there is a price to pay ...

consider OpenAI's progress ...

model	release	model size	training data	training cost
GPT-1	2018	$1.17 \times 10^8$	4.5 GB	$1.7 \times 10^{19}$ FLOP
GPT-2	2019	$1.50 \times 10^9$	40.0 GB	$1.5 \times 10^{21}$ FLOP
GPT-3	2020	$1.75 \times 10^{11}$	570.0 GB	$3.1 \times 10^{23}$ FLOP
GPT-3.5	2022	$1.75 \times 10^{11}$	unknown	unknown
GPT-4	2023	$\sim 1.00 \times 10^{12}$	unknown	$\sim 2.1 \times 10^{25}$ FLOP

source: Wikipedia

## take home messages

**AI systems based on large language models should not (yet) be seen as lexicons or improved search engines**

**however, they show astounding *emergent* problem solving skills**

⇒ **we may need to rethink how we do everything**

**alas, this might be much easier said than done**

a considerable challenge is that most current leading AI systems are proprietary and *not open*

recent ongoing regulation debates suggest that the silicon valley aims at securing monopolies

transformers and self-attention

preliminaries

# an artificial neuron

synaptic summation

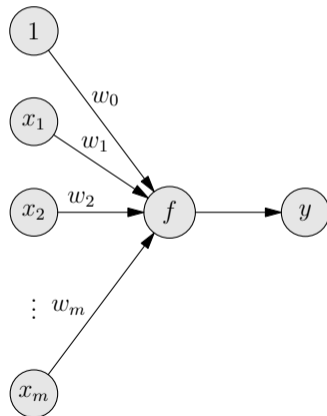
$$s = \mathbf{w}^T \mathbf{x}$$

activation

$$y = f(s)$$

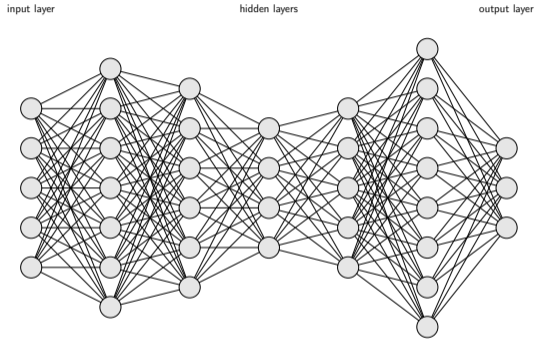
$$f \in \{\tanh, \text{ReLU}, \text{LReLU}, \dots\}$$

input      activation      output



# a deep artificial neural network

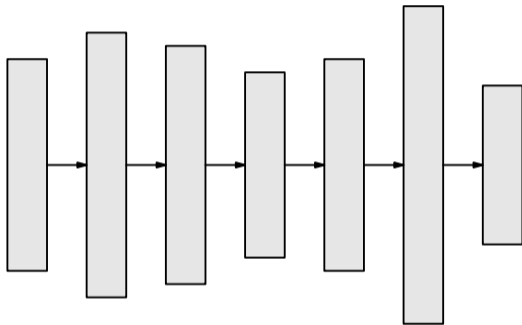
nowadays, people work with  
 $O(10^9)$  neurons and  $O(10^{12})$   
connections



# a deep artificial neural network

nowadays, people work with  $O(10^9)$  neurons and  $O(10^{12})$  connections

in reports, they therefore do not discuss details





# a deep artificial neural network

nowadays, people work with  
 $O(10^9)$  neurons and  $O(10^{12})$   
connections

in reports, they therefore do  
not discuss details, seriously



down the rabbit hole

# transformers

developed by a (mainly)  
Google team in 2017/18

---

## 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  
lukaszkaizer@google.com

**Illia Polosukhin\* ‡**  
illia.polosukhin@gmail.com

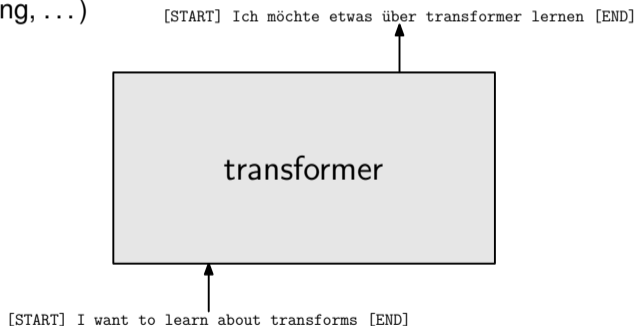
### Abstract

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 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

# purpose

originally conceived for text translation

however, the idea generalizes to other tasks and data (image captioning, ...)



# favorable characteristics

transformers excel at modeling sequential data but don't make assumptions about temporal / spatial relationships across data points

⇒ transformers are versatile and widely applicable

unlike RNNs and CNNs, transformers can capture distant / long-range contexts or dependencies among (input / output) data

*attention* allows each location in each layer to access the entire input while RNNs or CNNs build context via a hierarchy of processing steps

⇒ transformers learn fast

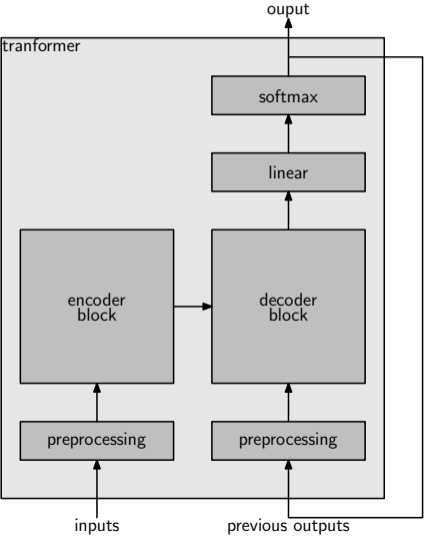
unlike RNNs, transformers involve computations that easily allow for parallelization

⇒ transformers run fast

# encoder / decoder architecture

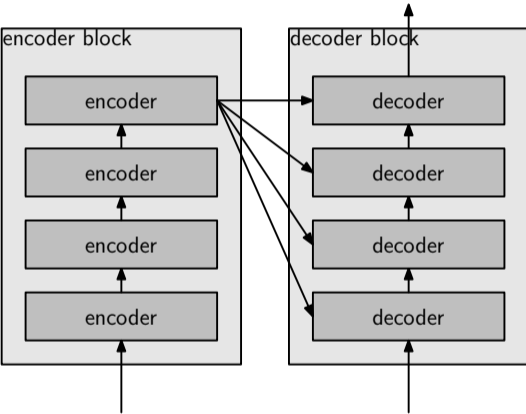
preprocessing details depend on the task- and data at hand

for text translation, one considers  
word embedding +  
positional encoding



# zooming in

the encoder- and decoder block consist of stacks of  $N$  encoders and decoders, respectively

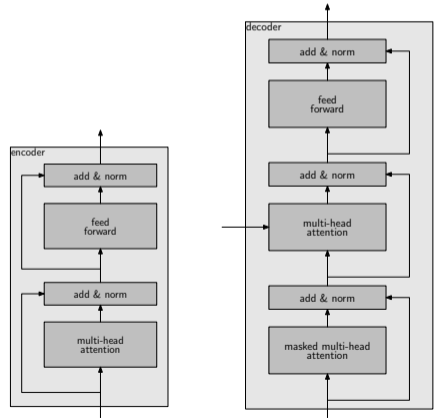


# zooming in

encoders and decoders involve *attention heads* and feed forward layers for feature extraction

normalization layers ensure numerical stability

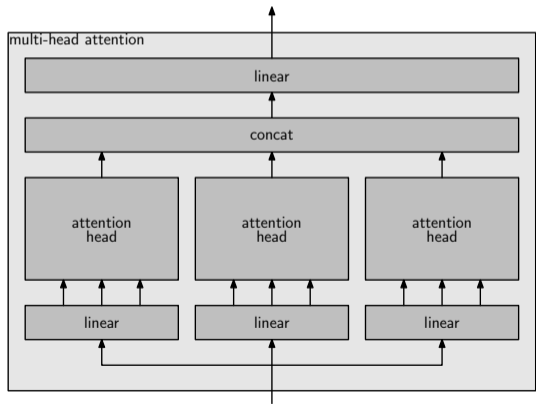
residual connections ensure training efficiency





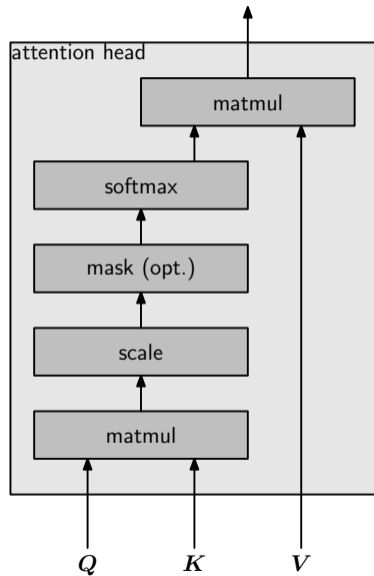
# zooming in

multi-head attention modules run *several attention heads* in parallel, concatenate, and process their results



# zooming in

(self) *attention heads* are the key components of a transformer



## (self) attention

the input to each attention head is a matrix of feature vectors

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]$$

during training, matrices  $\mathbf{W}_Q$ ,  $\mathbf{W}_K$ , and  $\mathbf{W}_V$  are learned which allow for computing *queries*, *keys*, and *values*

$$\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_n] = \mathbf{W}_Q \mathbf{X}$$

$$\mathbf{K} = [\mathbf{k}_1, \mathbf{k}_2, \dots, \mathbf{k}_n] = \mathbf{W}_K \mathbf{X}$$

$$\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n] = \mathbf{W}_V \mathbf{X}$$

**note:** we are working with column vectors  
Web tutorials usually work with row vectors

## queries, keys, values

the terms *queries*, *keys*, and *values* allude to database technology where we have a query, look for a matching key, and retrieve the respective value

in a transformer, each attention head uses (normalized) inner products to determine how much input item  $i$  (query  $\mathbf{q}_i$ ) should attend to item  $j$  (key  $\mathbf{k}_j$ )

$$a_{ij} = \operatorname{softmax}_j \left( \frac{\mathbf{q}_i^\top \mathbf{k}_j}{\sqrt{d}} \right) \quad (\text{divisor } \sqrt{d} = \sqrt{\dim(\mathbf{q}_i)} = \sqrt{\dim(\mathbf{k}_j)} \text{ is for numerics})$$

attention scores are then used to weigh the contribution of item  $j$  (value  $\mathbf{v}_j$ )

$$\mathbf{y}_i = \sum_j a_{ij} \mathbf{v}_j$$

⇒ attention from each  $x_i$  to each  $x_j$  can be efficiently computed as

$$Y = \text{softmax} \left( \frac{Q^T K}{\sqrt{d}} \right) V = \text{softmax} \left( \frac{X^T W_Q^T W_K X}{\sqrt{d}} \right) W_V X$$

the  $Y_1, \dots, Y_l$  from multiple attention heads are concatenated, multiplied by yet another weight matrix  $W$ , and passed on

in a *masked* attention head of a decoder, input  $x_i$  is only allowed to attend to those  $x_j$  with  $j < i$  (i.e. is only allowed backward attention)

# remarks

the (self) attention mechanism is plausible / reasonable

but the whole combined architecture is a scientific mess

why all the add & norm layers ?

why all the feed-forward layers ?

why stacking encoders / decoders ?

what about all the craftiness (dropout, ReLU activation, ADAM training, ...) ?

⇒ we cannot argue with practical results and performance but should keep in mind ...

 **take home message**

nature



technology



**Prof. Christian Bauckhage**

**Lamarr Institute for ML and AI**

**University of Bonn**

**Fraunhofer IAIS**

