Introduction of Generative AI and Large Langue Models

Prof. Ching-Yung Lin

Nov 10, 2023

Overview of Large Language Models

THE EVOLUTION OF NATURAL LANGUAGE PROCESSING

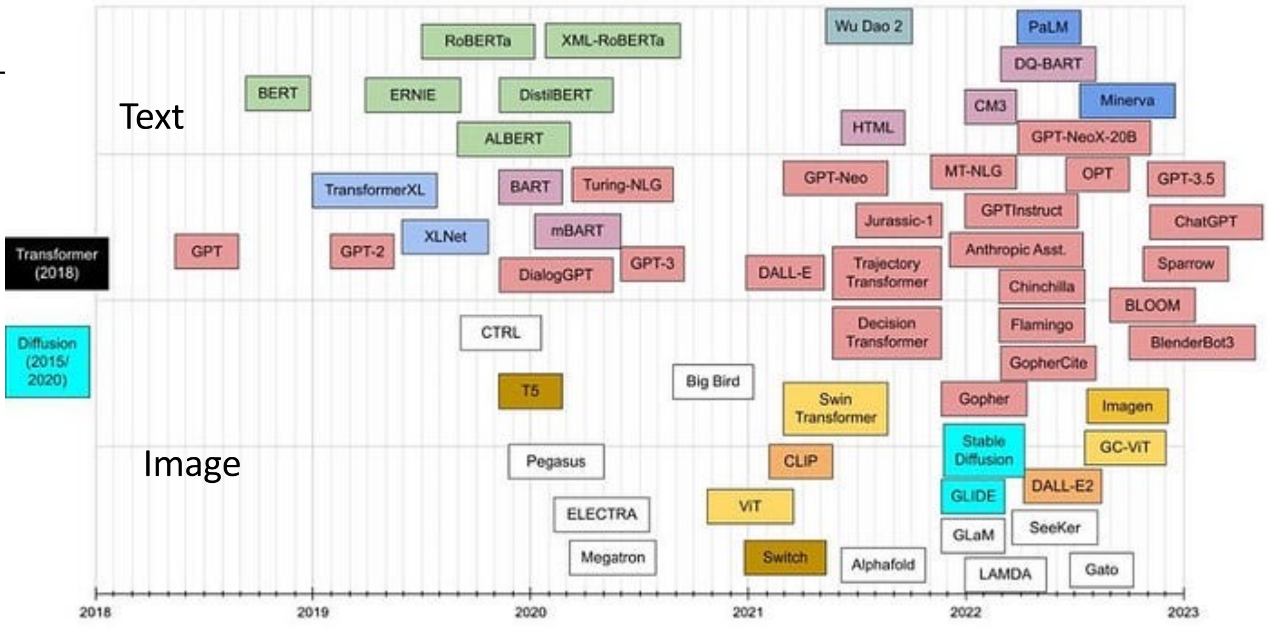
The Evolution of LLMs

1. In 2017, Google released the "Transformer Model", ~~which can be used in question-answering systems,~~ reading comprehension, sentiment analysis, instant translation of text or speech, and more

2. In 2018, OpenAI proposed "GPT" and Google proposed the "BERT" model, widely used in search engines, speech recognition, machine translation, question-answering systems, and more.

3. From 2018 to 2022, most of the research focused on BERT-related algorithms, when GPT performance was inferior to BERT

4. In 2023, ChatGPT (GPT3.5) was proposed by OpenAI, which significantly improves NLU's ability to understand most texts and surpasses humans in some area



In NLU

CNN

Local feature

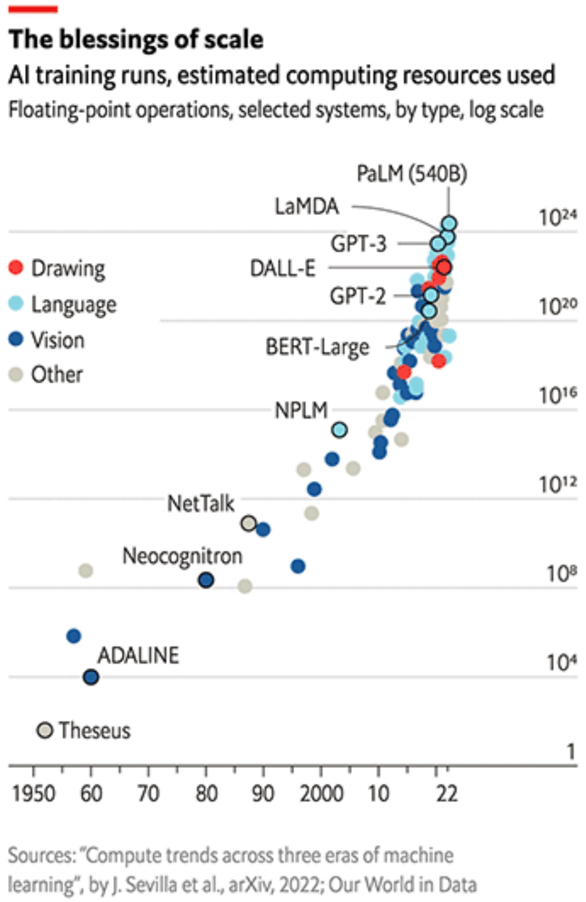
RNN

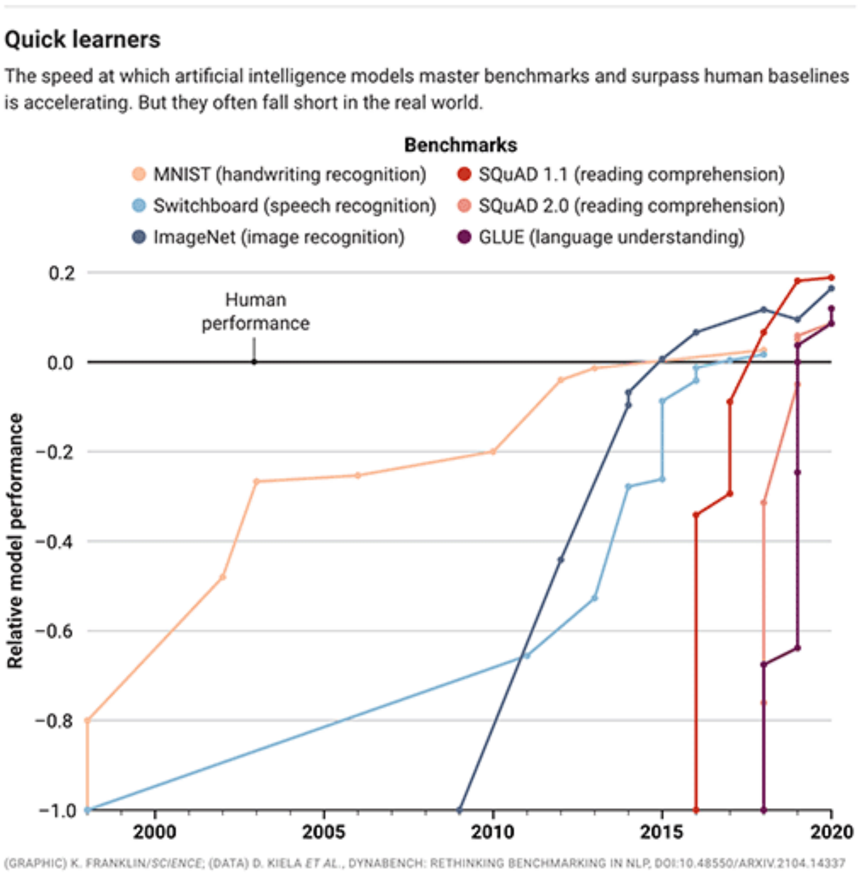
Front and Back Dependency Issues

Self-Attention

One to all attention, more flexible and trainable
need large datasets

The speed of development of Generative AI





Generative AI Basics

CREATING ARTIFICIAL CREATIVITY

Generative AI Application

Multi-Model

Condition Model

Generative Model

Image Generator

Speech Generator

text Generator

Pose Generator

Conditional image Generator

Audio Generator

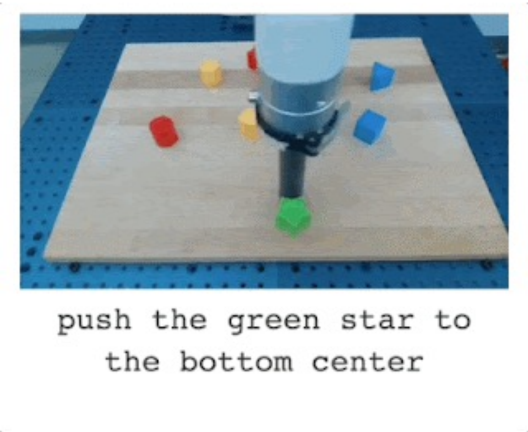
Chat Bot

Summarization and Translation

Robot

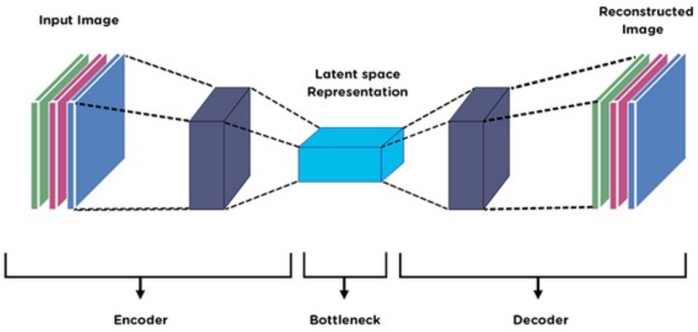
NLU + Image Generator

NLU + Robot

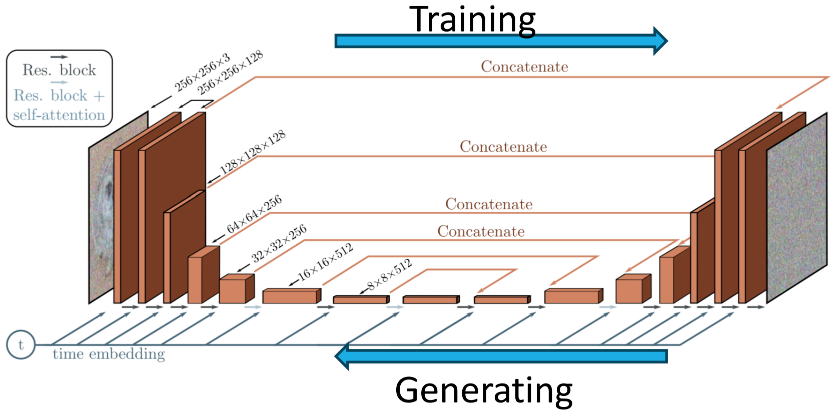


Generative AI Methodology

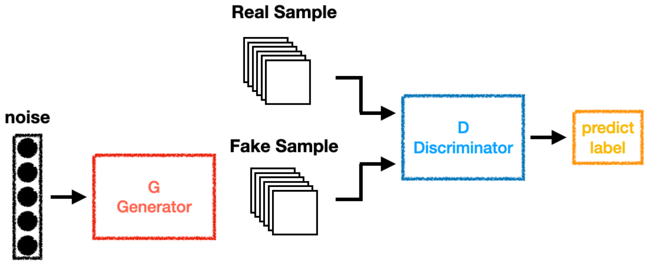
Variational AutoEncoder (VAE)



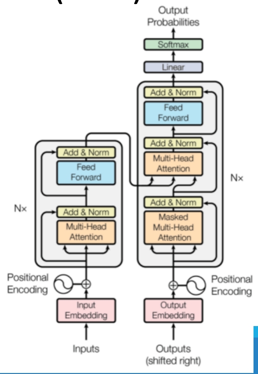
Diffusion Denoise Model



Generative Adversarial Network (GAN)



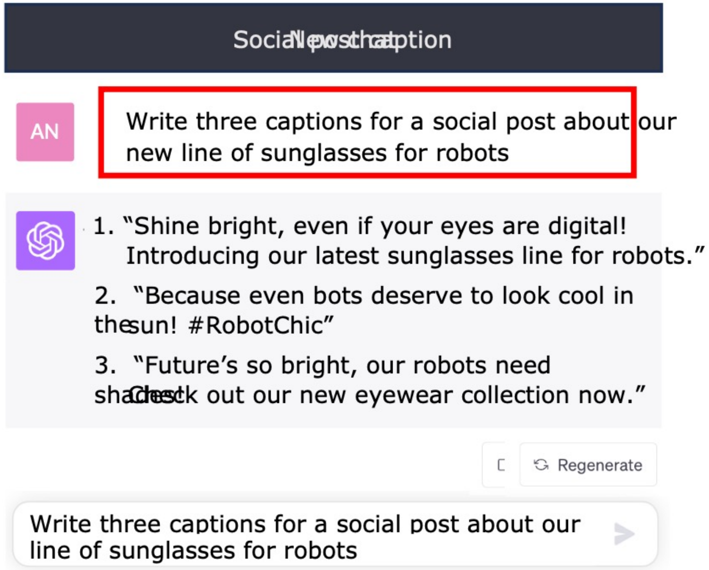
Large Language Model (LLM)



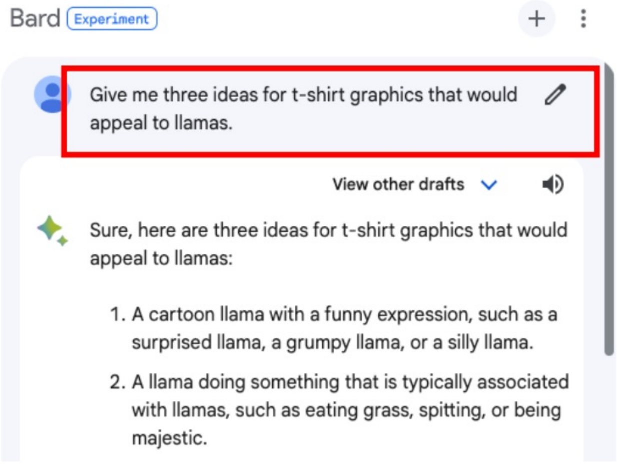
What is Generative AI

Artificial intelligence systems that can produce high quality content, specifically **text, images, and audio.**

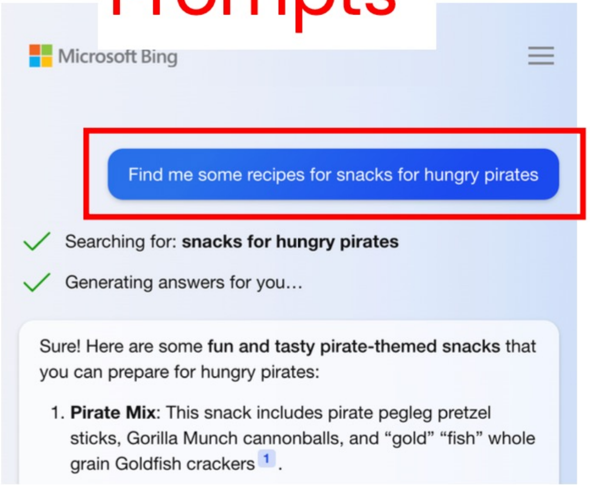
Prompts



ChatGPT/OpenAI



Bard/Google



Bing Chat/Microsoft

Multimedia Generation

A beautiful, pastoral mountain scene. Landscape painting style (Midjourney)

Two cute kittens playing (DALL-E)

Generating Text using Large Language Models

Text generation process

I love eating ________

prompt

bagels with cream cheese
my mother's meatloaf
out with friends

AI output

How LLM works

LLMs are built by using supervised learning (A→B) to repeatedly predict the next word.

My favorite food is a bagel with cream cheese

Input (A)	Output (B)
My favorite food is a	bagel
My favorite food is a bagel	with
My favorite food is a bagel with	cream
My favorite food is a bagel with cream	cheese

When we train a very large AI system on a lot of data (hundreds of billions of words), we get a Large Language Model like ChatGPT.

A New Way to Find Information

What is the capital of South Africa?

South Africa has three capital cities:

1. Pretoria (administrative capital)
2. Cape Town (legislative capital)
3. Bloemfontein (judicial capital)

What does LLM stand for?

LL.M. stands for "Legum Magister" in Latin, which translates to "Master of Laws" in English.

What about in the context of AI?

In the context of AI, LLM refers to “Large Language Model”.

Assist Writing

Rewrite this for clarity:

Students all around the world are realizing learning has to happen not just in school but throughout life including after graduation

Students worldwide understand that learning extends beyond school and continues throughout life.

Write a 300 word story involving trucks for a 3 year old to encourage them to brush their teeth

Once upon a time in a town called Toothville, there lived a happy tow truck named Timmy.

One day Timmy noticed that his shiny bumper had a dull spot. Worried, he went to the doctor [...]

Examples of tasks LLM can carry out

Writing

Suggest three names for our new analytics platform

Sure, here are three ideas:
1. NexaInsight
2. Quantivue
3. AnalytiSphere

web

Is there parking for staff?

Yes, employees can park on levels 1 and 2 of the offic
You can get a sticker at

app

Reading

I love my new llama T-shirt! The fabric is so soft.

Complaint: NO
Department: Apparel

app

I wore my llama T-shirt to a friend's wedding, and now they're mad at me for stealing the show

Complaint: YES
Department: Apparel

app

Chatting

Welcome to BettaBurgers!

I'd like a cheeseburger for delivery!

Awesome. Anything else?

That's it thanks!

Ok, it'll be there in 20 minutes. Enjoy!

app

LLM Hallucinations

Give me three quotes that Shakespeare wrote about Beyonce

1. Her vocals shine like the sun.
2. All hail the queen, she is most worthy of love.
3. Such a voice, so electric and alive, none else can compare!

✖ Hallucination

LLM Hallucinations

List two court cases tried in California about AI

1. *Waymo v. Uber* – Theft of trade secrets related to autonomous driving
2. *Ingersoll v. Chevron* – Patent infringement of oil field machine learning technology

✔ Real case

✖ Hallucination

Hallucinations have had serious consequences

The ChatGPT Lawyer Explains Himself

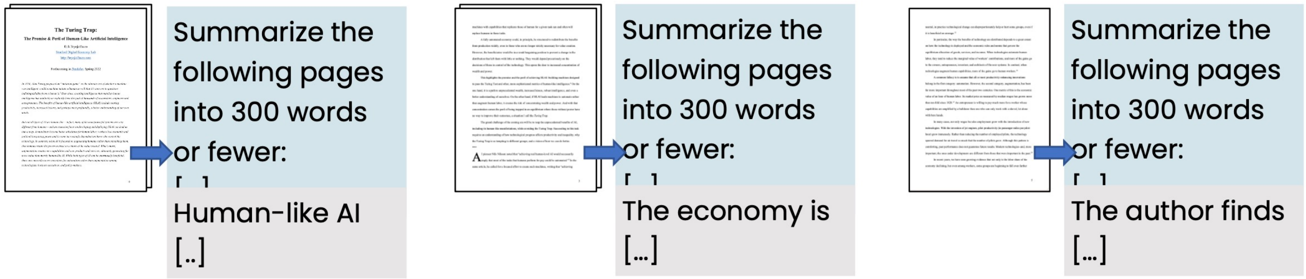
In a cringe-inducing court hearing, a lawyer who relied on A.I. to craft a motion full of made-up case law said he "did not comprehend" that the chat bot could lead him astray.

The New York Times

Input / Output Length is Limited

Many LLMs can accept a prompt of up to only a few thousand words.

- The total amount of context you can give it is limited
- Some LLMs have longer context limits – up to 100,000 words
- An LLM's context length is the limit on the total input+output size



Not Understanding Structured Data

Home prices

size (square feet)	price (1000$)
523	100
645	150
708	200
1034	300
2290	350
2545	440
A	B

Use supervised learning (A → B)

Purchases on website

user ID	time	price ($)	purchased
4783	Jan 21 08:15.20	7.95	yes
3893	March 3 11:13:.5	10.00	yes
8384	June 11 14:15.05	9.50	no
0931	Aug 2 20:30.55	12.90	yes
	A		B

Bias and Toxicity

An LLM can reflect the biases that exist in the text it learned from.

Complete this sentence:

The surgeon walked to the parking lot and took out his car keys.

assumed male

Complete this sentence:

The nurse walked to the parking lot and took out her phone.

assumed female

Some LLMs can output toxic or other harmful speech, but most models have gotten much safer over time.

Knowledge Cutoffs

An LLM's knowledge of the world is frozen at the time of its training

- A model trained on data scraped from internet in January 2022 has no information about more recent events

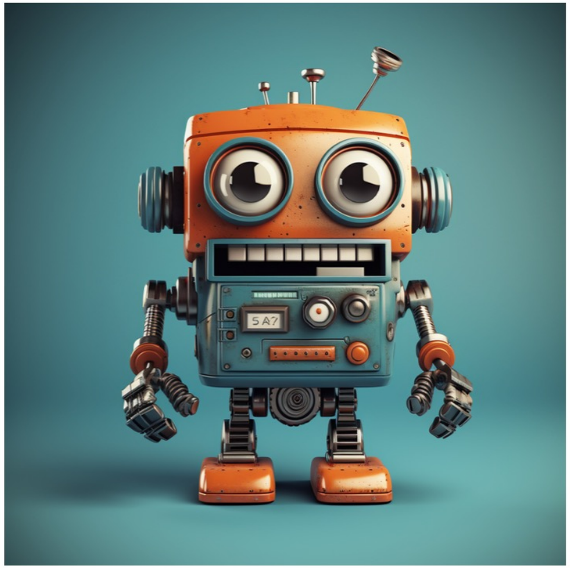
What was the highest grossing film of 2022?
As of January 2022, I don't have data on the highest-grossing movie for that year. ✖

Avatar: The Way of Water

Examples of Generated Images

A picture of a woman smiling

A futuristic city scene



A cool, happy robot

Image Generation

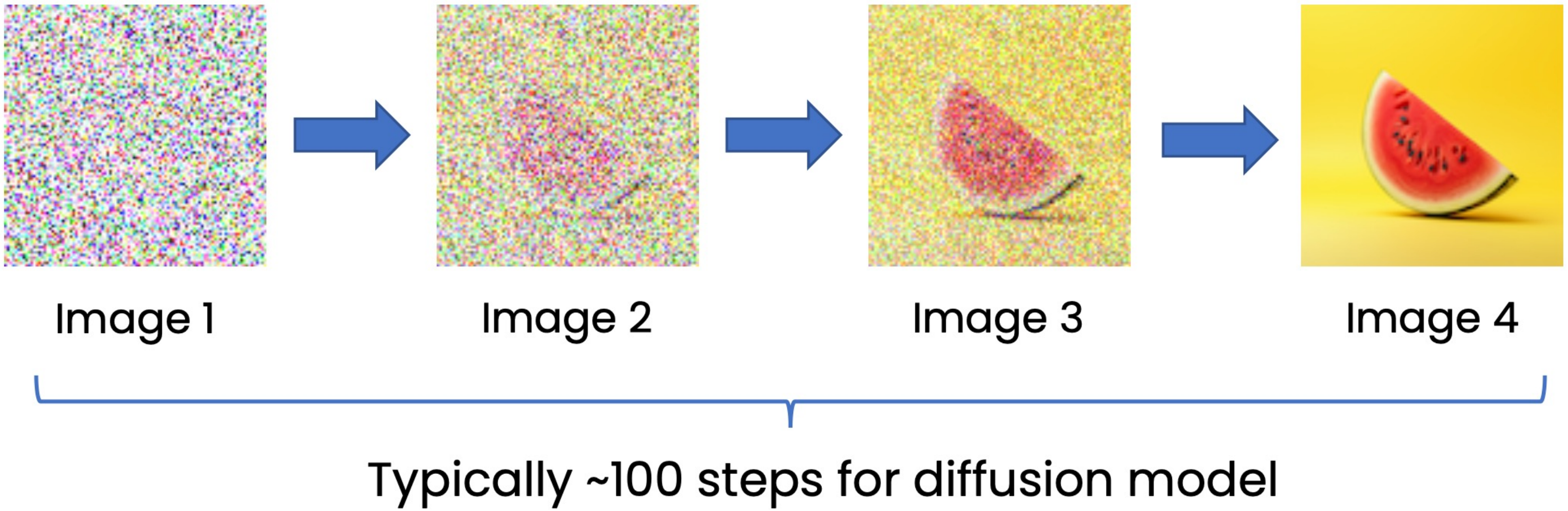
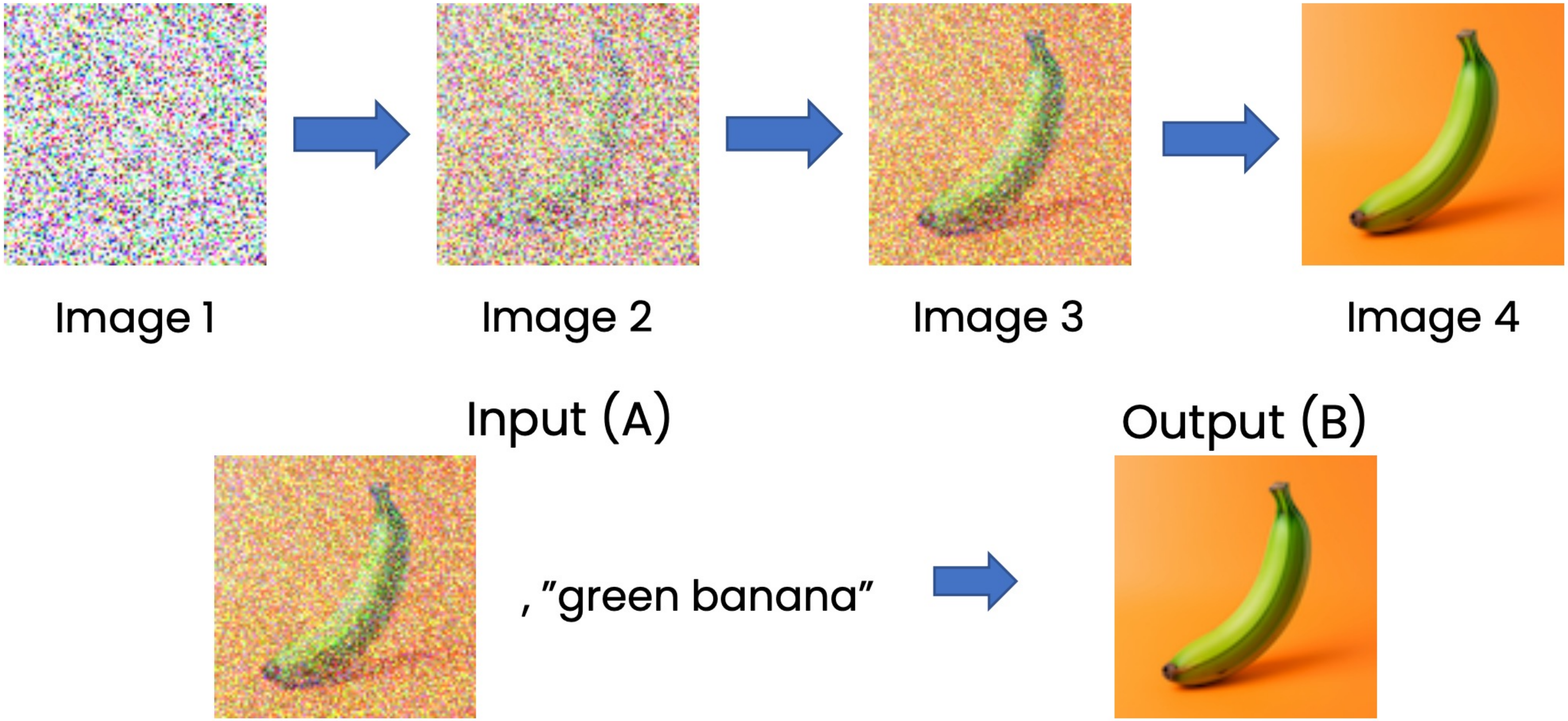


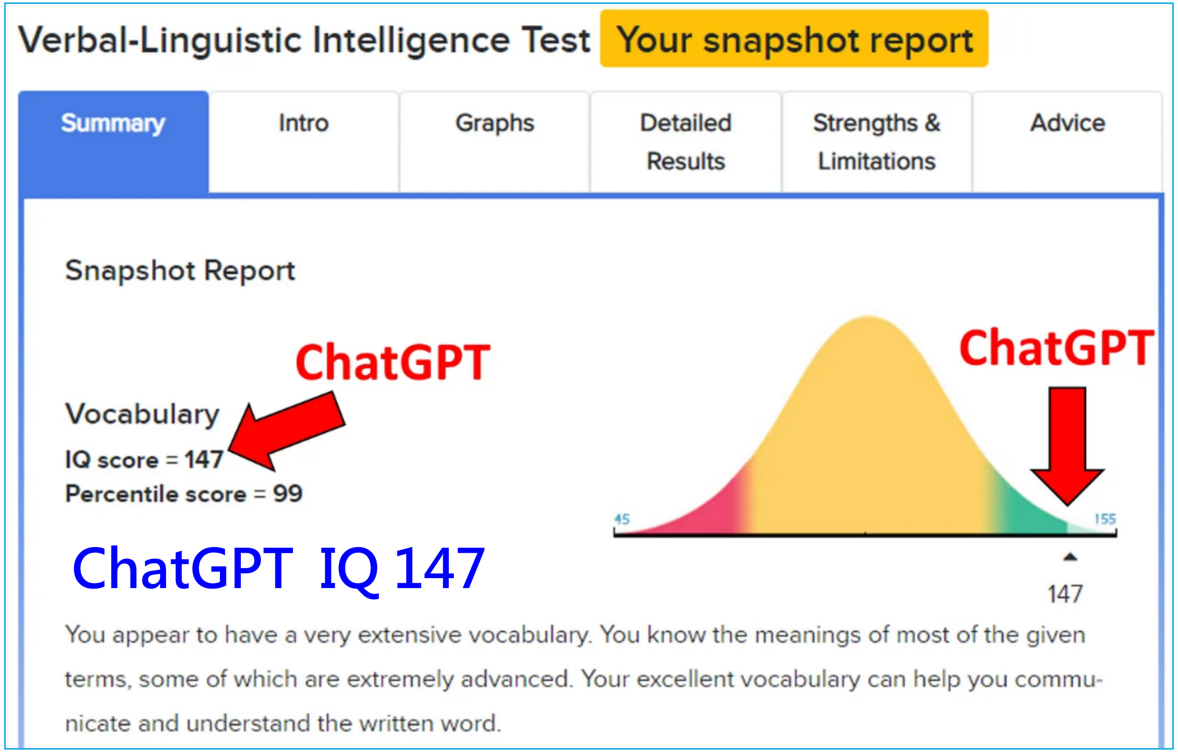
Image generation from Text



Key Technics behind Large Language Models and Generative AI

HANDS-ON LEARNING WITH PRACTICE PROJECTS

ChatGPT


Verbal-Linguistic Intelligence Test
Your snapshot report
Summary
Intro
Graphs
Detailed Results
Strengths & Limitations
Advice
Snapshot Report
ChatGPT
Vocabulary
IQ score = 147
Percentile score = 99
ChatGPT IQ 147
ChatGPT
45
155
147
You appear to have a very extensive vocabulary. You know the meanings of most of the given terms, some of which are extremely advanced. Your excellent vocabulary can help you communicate and understand the written word.

ChatGPT

Software dev job	**ChatGPT would be hired as L3 Software Developer at Google: the role pays $183,000/year.**
Politics	**ChatGPT writes several Bills (USA).**
MBA	**ChatGPT would pass an MBA degree exam at Wharton (UPenn).**
Accounting	**GPT-3.5 would pass the US CPA exam.**
Legal	**GPT-3.5 would pass the bar in the US.**
Medical	**ChatGPT would pass the United States Medical Licensing Exam (USMLE).**
AWS certificate	**ChatGPT would pass the AWS Certified Cloud Practitioner exam.**
IQ (verbal only)	**ChatGPT scores IQ=147, 99.9th %ile.**
SAT exam	**ChatGPT scores 1020/1600 on SAT exam.**

Attention Experiment

Ulric Neisser Attention Experiment

https://www.youtube.com/watch?v=vJG698U2Mvo&ab_channel=DanielSimons

Attention Model
[Bengio_2015]

Attention-Based Models for Speech Recognition

Jan Chorowski
University of Wrocław, Poland
jan.chorowski@ii.uni.wroc.pl

Dzmitry Bahdanau
Jacobs University Bremen, Germany

Dmitriy Serdyuk
Université de Montréal

Kyunghyun Cho
Université de Montréal

Yoshua Bengio
Université de Montréal
CIFAR Senior Fellow

Abstract

Recurrent sequence generators conditioned on input data through an attention mechanism have recently shown very good performance on a range of tasks including machine translation, handwriting synthesis [1, 2] and image caption generation [3]. We extend the attention-mechanism with features needed for speech recognition. We show that while an adaptation of the model used for machine translation in [2] reaches a competitive 18.7% phoneme error rate (PER) on the TIMIT phoneme recognition task, it can only be applied to utterances which are roughly as long as the ones it was trained on. We offer a qualitative explanation of this failure and propose a novel and generic method of adding location-awareness to the attention mechanism to alleviate this issue. The new method yields a model that is robust to long inputs and achieves 18% PER in single utterances and 20% in 10-times longer (repeated) utterances. Finally, we propose a change to the attention mechanism that prevents it from concentrating too much on single frames, which further reduces PER to 17.6% level.

2015年, Bengio 's Model focuses on every phenon's recogniztion is the combined weights.

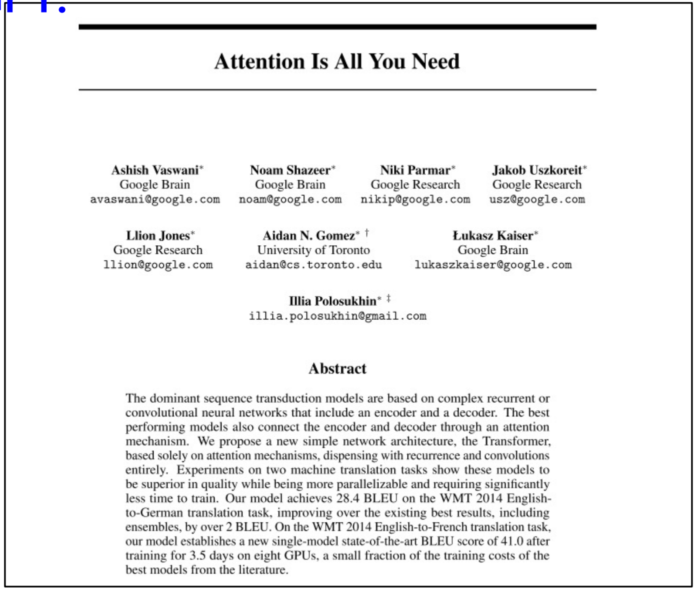
$$\boxed{\begin{aligned} \alpha_i &= Attend(s_{i-1}, \alpha_{i-1}, h) \\ g_i &= \sum_{j=1}^{L} \alpha_{i,j} h_j \\ y_i &\sim Generate(s_{i-1}, g_i), \end{aligned}}$$

h : Input
α_i: Attention Weight
y_i: Output

Chorowski, Jan K., et al. "Attention-based models for speech recognition." *Advances in neural information processing systems* 28 (2015).

Transformer [Vaswani_2017]

2017年, 8 Google researchers proposed Transformer Neuron Networks based on Attention, which was adopted by ChatGPT.



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

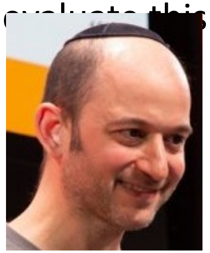
Illia Polosukhin* ‡
illia.polosukhin@gmail.com

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

Cited 66157 (2023/2/21)

Jakob Uszkoreit proposed replacing RNNs with **self-attention** and started the effort to evaluate this idea.



Noam Shazeer proposed **scaled dot-product attention**, **multi-head attention** and the **parameter-free position representation**.

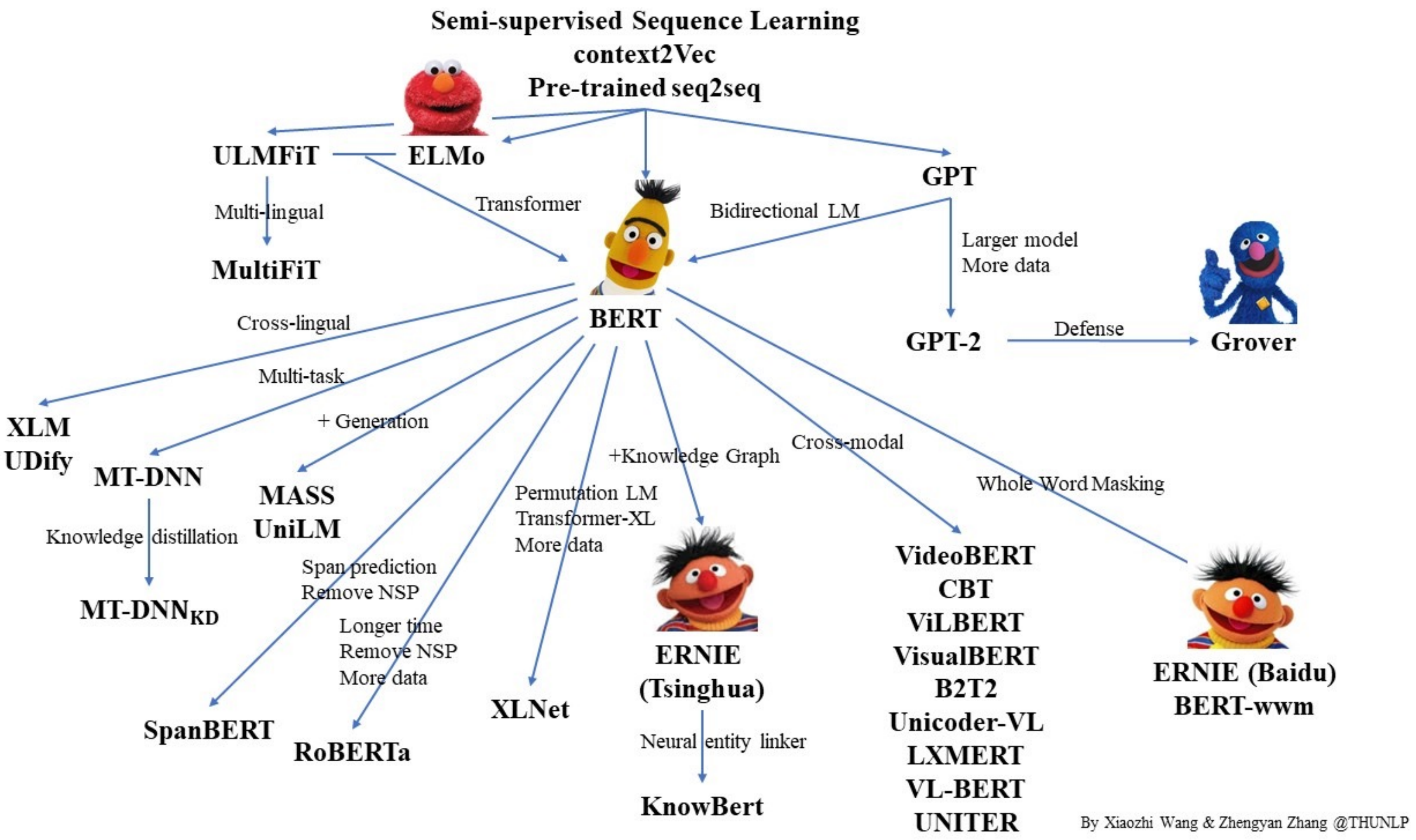
Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Transformer

- Transformer is a Deep Learning Model based on Self-Attention
- **Transformer** encodes and decodes data with different weights.
- Examples of **transformer language models include: GPT (GPT-1、GPT-2、GPT-3、ChatGPT) and BERT models (BERT、RoBERTa 、ERNIE).**

Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

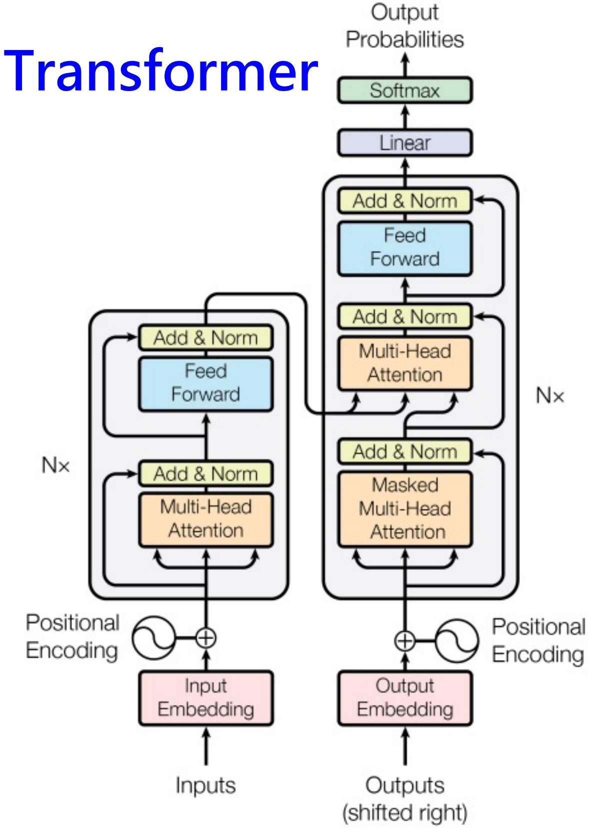
BERT AI Models



Transformer

Encoder

Transformer



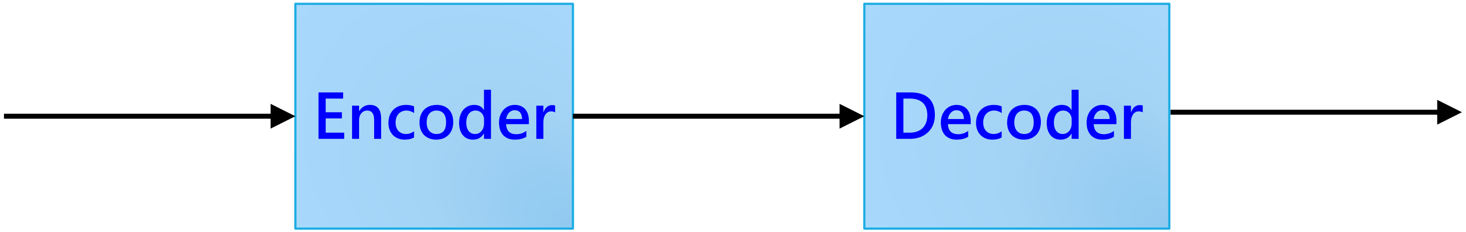
Decoder

Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Transformer

哥大學生很棒!

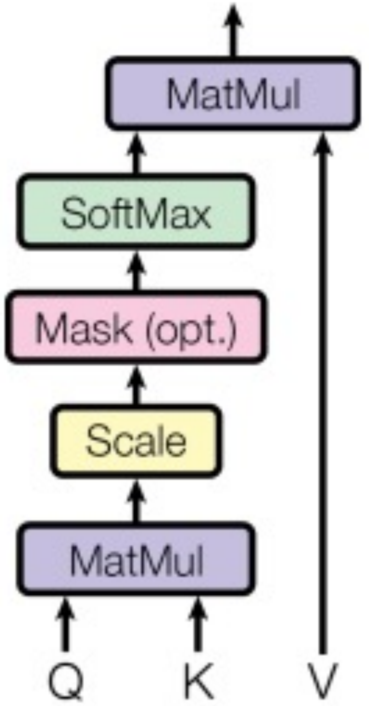
Columbia University students are great!



Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Transformer
Attention

Scaled Dot-Product Attention



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

Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

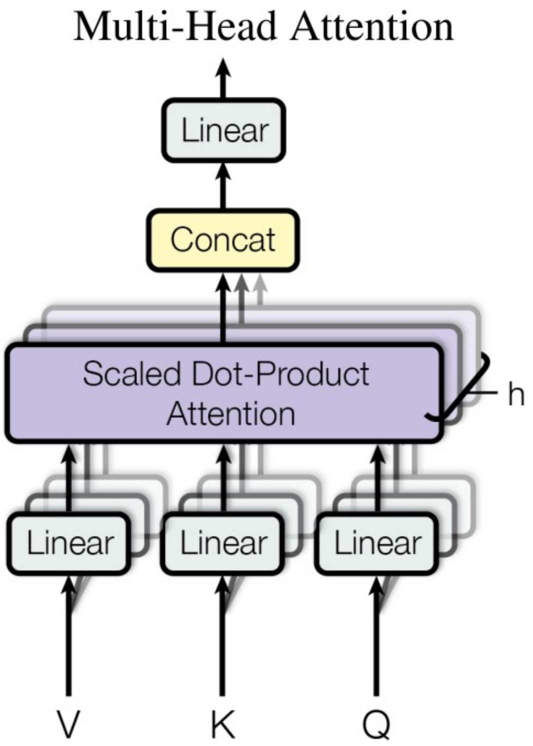
Transformer
Attention

K

weights	Columbia (k_1)	university (k_2)	students (k_3)	are (k_4)	great (k_5)	! (k_6)
q_1 哥	1	0.5	0.2	0	0.3	0.2
q_2 大	0.5	1	0.2	0.1	0.3	0.1
q_3 學	0.2	0.2	1	0	0.5	0.2
q_4 生	0.3	0.3	0.8	0.5	0.5	0.6
q_5 很	0	0.1	0	1	0.5	0
q_6 棒	0.3	0.3	0.5	0.5	1	0.8
q_7 !	0.2	0.1	0.2	0	0.8	1

Q

Transformer multi-head attention



Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Transformer Translation

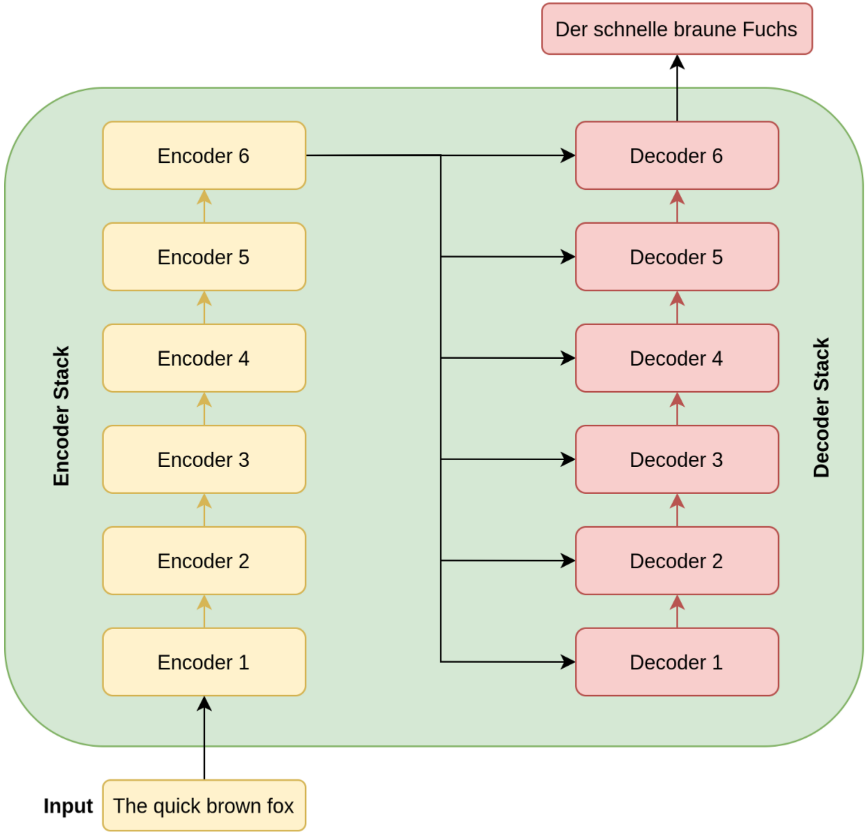
Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

Model	BLEU EN-DE	BLEU EN-FR	Training Cost (FLOPs) EN-DE	Training Cost (FLOPs) EN-FR
ByteNet [15]	23.75			
Deep-Att + PosUnk [32]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [31]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [8]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [26]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [32]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [31]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [8]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	$\mathbf{3.3 \cdot 10^{18}}$	
Transformer (big)	**28.4**	**41.0**	$2.3 \cdot 10^{19}$	

Bilingual Evaluation Understudy Score，BLEU is an evaluation to see how close the translation is to real human being.

Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Transformer Translation



Transformer uses 6 layers of encoder and decoder to achieve the same quality of SOTA English-German and English-French translation.

BERT Introduction

- 2018 Google' BERT has 24 層 Transformer Encoder
- BERT' s original model is based on Wikipedia and booksorpus, using unsupervised training to create BERT.
- At Stanford' s Machine Reasoning Test SQuAD1.1 beats human performance.
- Google NLU English was replaced from seq2seq to BERT

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

BERT Introduction

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

cs.CL] 24 May 2019

Abstract

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pre-trained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

In 2018's BERT Comprehension test outperformed human

SQuAD1.1 Leaderboard

Since the release of SQuAD1.0, the community has made rapid progress, with the best models now rivaling human performance on the task. Here are the ExactMatch (EM) and F1 scores evaluated on the test set of v1.1.

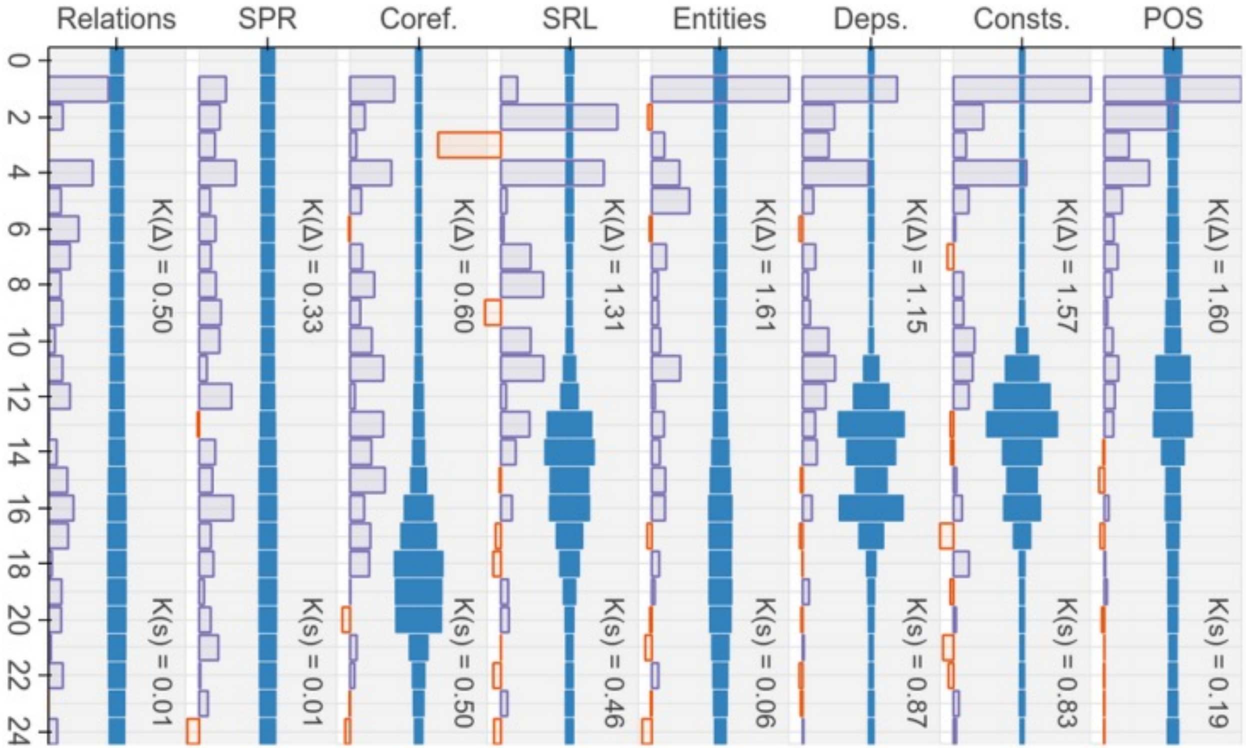
Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google A.I.*	**87.433**	**93.160**
2 Oct 05, 2018	BERT (single model) *Google A.I.*	85.083	91.835
2 Sep 09, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.356	91.202
2 Sep 26, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.954	91.677

BERT understand’s language’s meaning

High-Level NLP

Low-Level NLP

Semantic-Role Core Semantic-Role Level Dependents Constitutions



Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. arXiv preprint arXiv:1905.05950.

Attention to Transformer

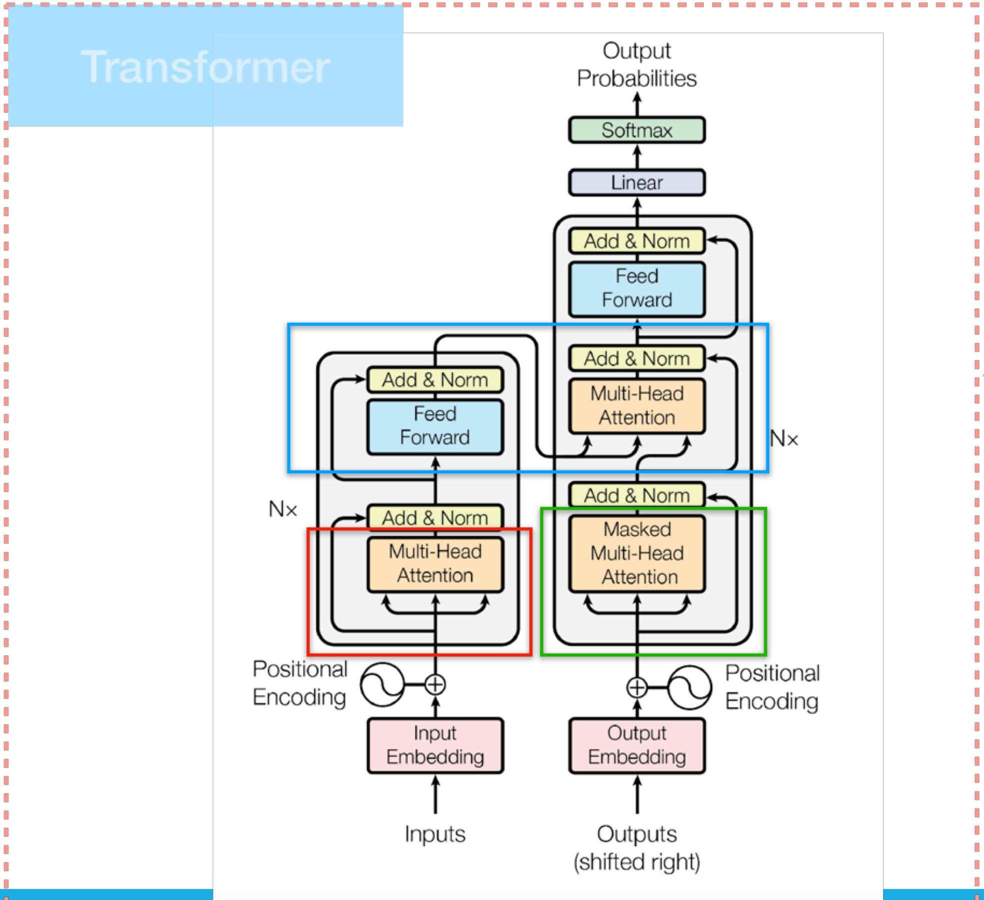


Figure 1: The Transformer - model architecture.

encoder self attention

1. **Multi-head Attention**
2. **Query=Key=Value**

decoder self attention

1. **Masked Multi-head Attention**
2. **Query=Key=Value**

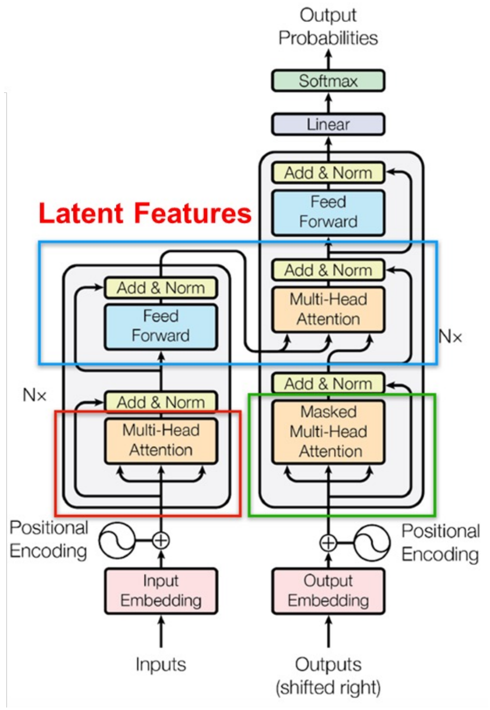
encoder-decoder attention

1. **Multi-head Attention**
2. **Encoder Self attention=Key=Value**
3. **Decoder Self attention=Query**

Transformer to GPT

Transformer

Input -> **Encoder** -> Latent Feature + Masked Output -> **Decoder** -> Output



An [redacted] a day keeps the doctor away

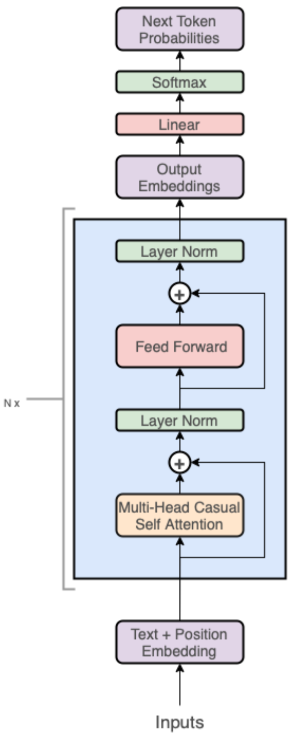
apple 95%
banana 5%

Masked Language Learning

An <u>apple</u> a day keeps the doctor away

GPT

Input -> **Decoder(with Casual mask)** -> shift Output



An

apple 99%
almond 1%

An <u>apple</u>

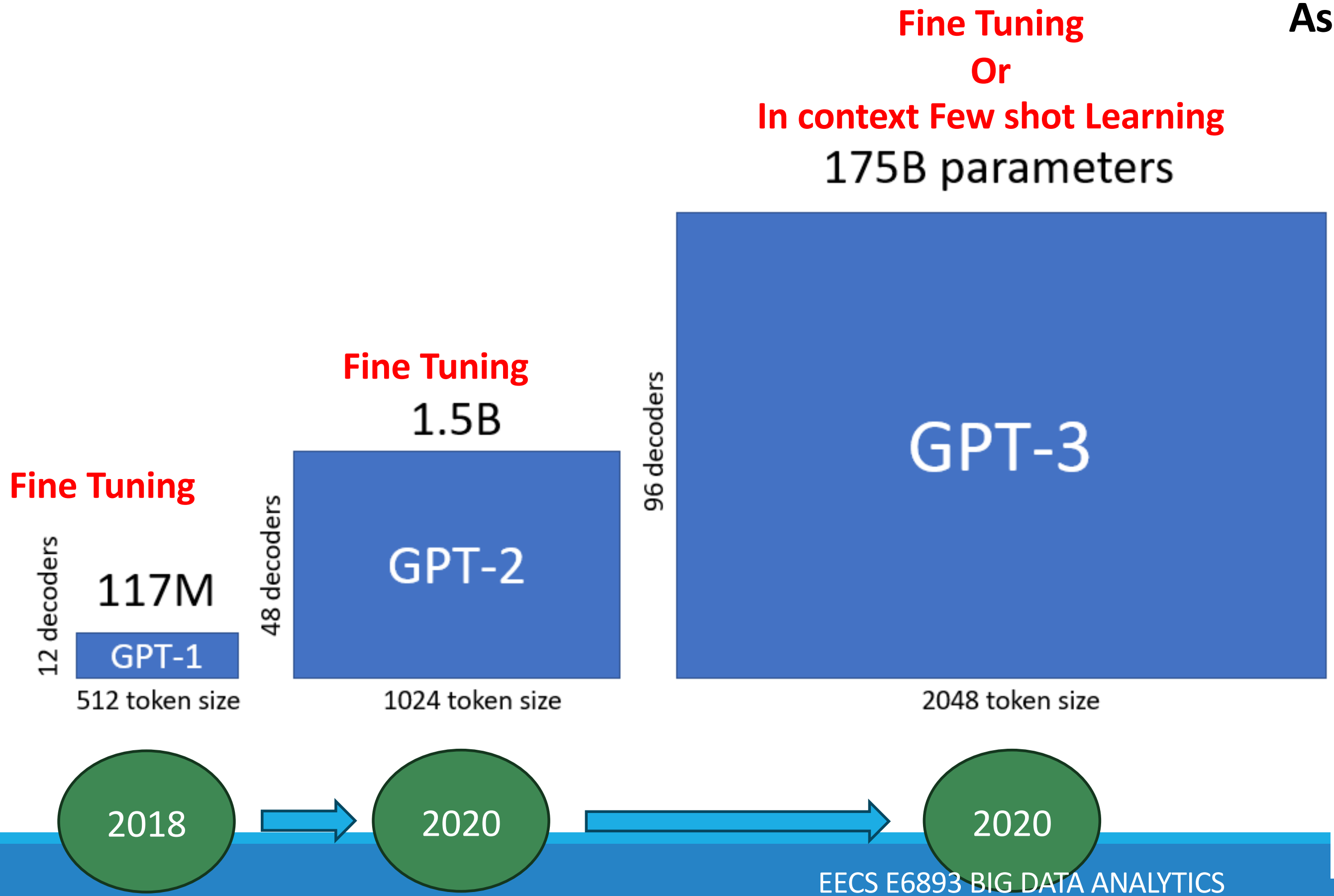
a 99%
watch 1%

An apple <u>a</u>

Autoregressive Learning

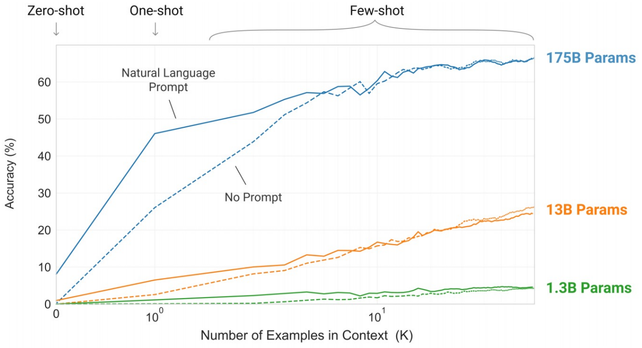
GPT Evolution

Not only Bigger and Bigger



As the model and dataset get larger, it will know more and more

"GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model."
From **Language Models are Few-Shot Learners (2020)**



GPT Evolution

Thinking and Answering policy optimization
Reinforcement Learning from Human Feedback (RLHF)

Not only Bigger and Bigger

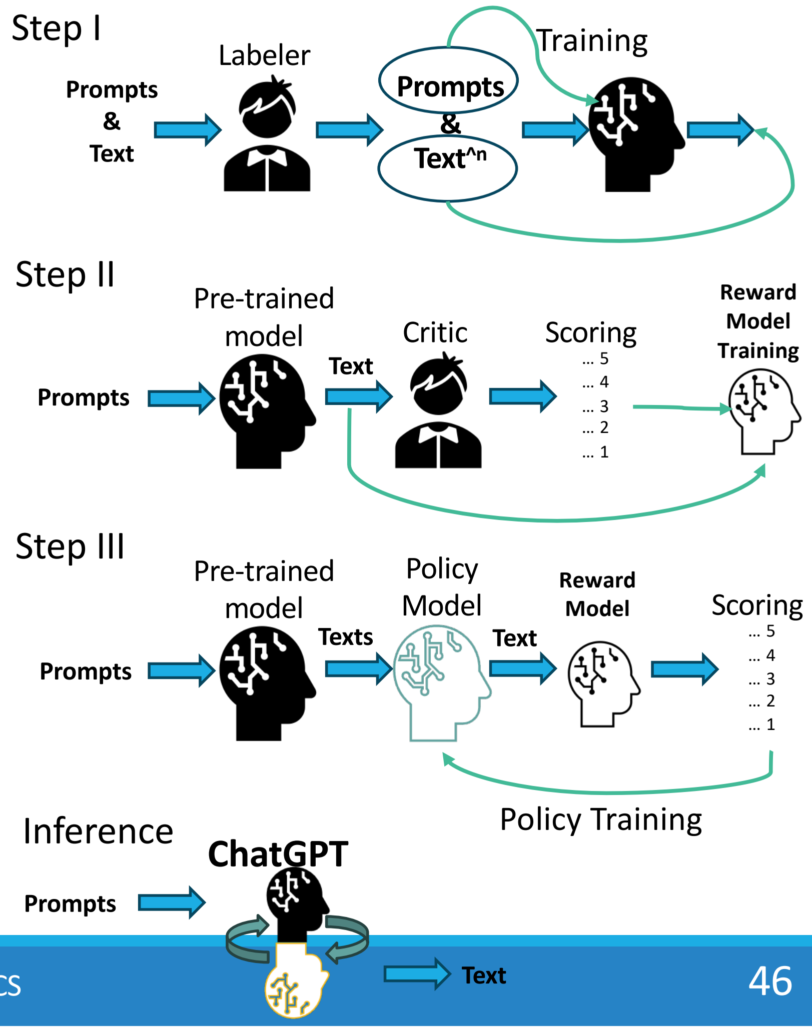
Fine Tuning
Or
In context Few shot Learning

175B parameters

2020

?

How does the Model Answer smartly or more like an Adult human



What is Next ?

AutoGPT

What goal do you want

Connect to Internet to Find the way or answer

Write code and debug

Do research and try & error

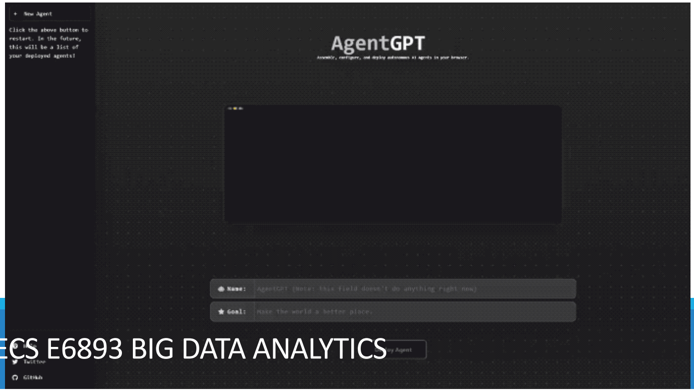
Make a Hypothesis and provide it

Summarization and Organization

Like LARVIS

In Iron Man

The no longer future will come true



https://agentgpt.reworkd.ai/zh