



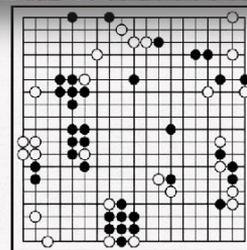
大规模 预训练 AI模型 发展,现状和未来

华南师范大学人工智能学院

范晨悠, 马宇函, 江金刚

1. AI发展里程碑 (2016年)

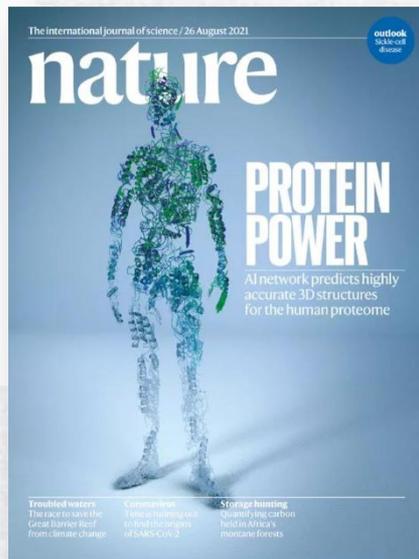
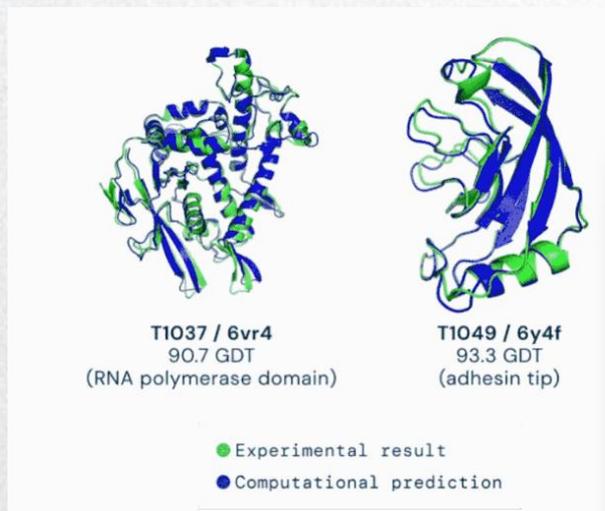
围棋AI模型AlphaGo/Zero, 击败人类世界冠军并登上 N&S 杂志封面。



AI发展里程碑 (2021年)

蛋白质结构预测算法AlphaFold
登上 Nature 杂志封面。

-- “可准确预测98.5%的人类蛋白结构”

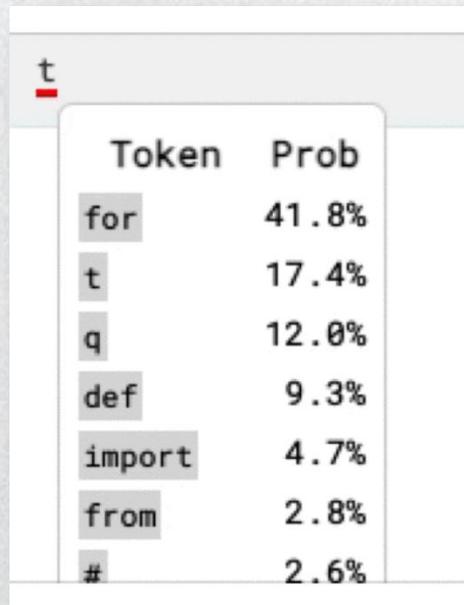


Highly accurate protein structure prediction with AlphaFold, Nature, 2021

AI发展里程碑 (2022年)

AI编程模型AlphaCode登上Science封面。

该模型在全球顶级计算机编程评测系统Codeforces中击败了半数的人类程序员。



A screenshot of a token probability table. The table has two columns: 'Token' and 'Prob'. The tokens listed are 'for', 't', 'q', 'def', 'import', 'from', and '#'. The probabilities are 41.8%, 17.4%, 12.0%, 9.3%, 4.7%, 2.8%, and 2.6% respectively. The token 't' is highlighted with a red underline above it.

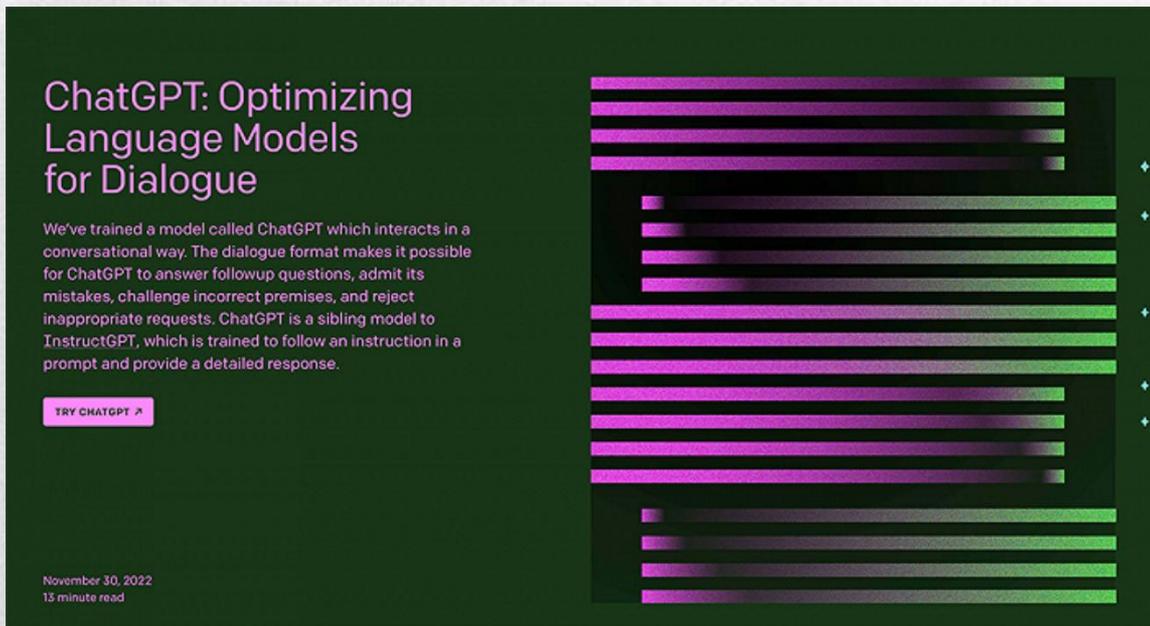
Token	Prob
for	41.8%
t	17.4%
q	12.0%
def	9.3%
import	4.7%
from	2.8%
#	2.6%



AI learns to write computer code in 'stunning' advance, Science, 2022

AI发展里程碑 (2022年)

人工智能对话模型
ChatGPT 发布并突破
1亿用户。



深度学习三巨头获图灵奖 (2018年)

“They led significant breakthroughs in AI technologies ...”

- 约书亚·本吉奥(Yoshua Bengio)
 - 加拿大蒙特利尔大学
- 杰弗里·辛顿(Geoffrey Hinton)
 - 多伦多大学, Google Brain
- 杨乐昆(Yann LeCun)
 - 纽约大学, Meta首席科学家



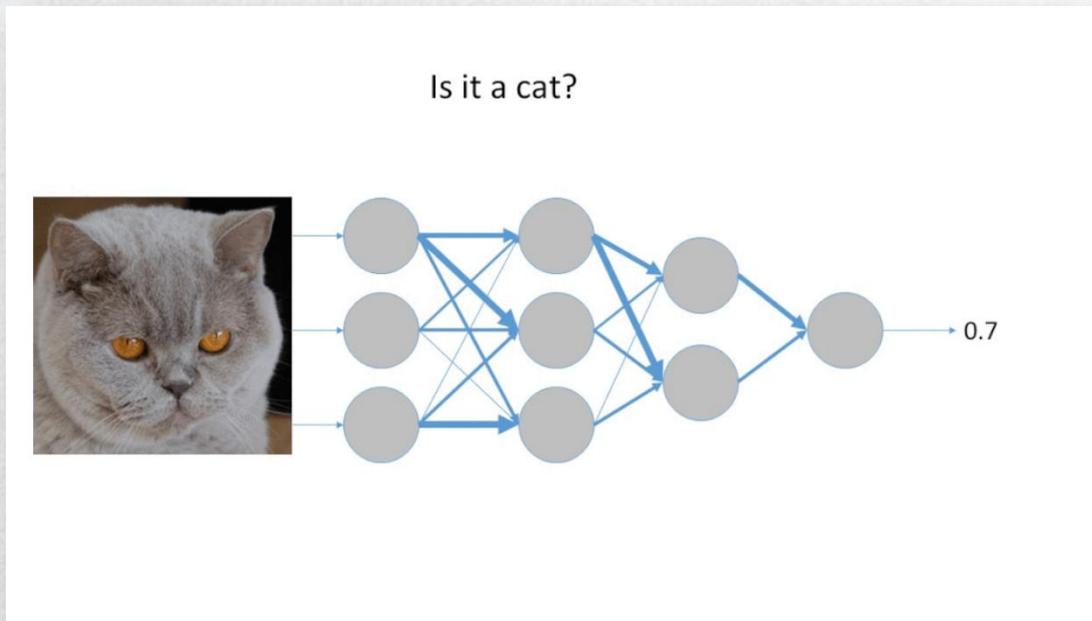
2.人工智能模型发展 (技术篇)

- 2012年，以AlexNet为代表的卷积神经网络(百万参数)
 - AI进入**深度学习**时代
- 2017年，以Transformer为代表的自然语言模型(>1亿参数)
 - AI进入**大模型**时代
- 2020~至今，以GPT-3为代表的预训练大模型(>1000亿参数)
 - 进入**超大模型**时代

现代深度神经网络(DNN) - 2012年

DNN是一种用于模拟人类神经系统的计算模型。

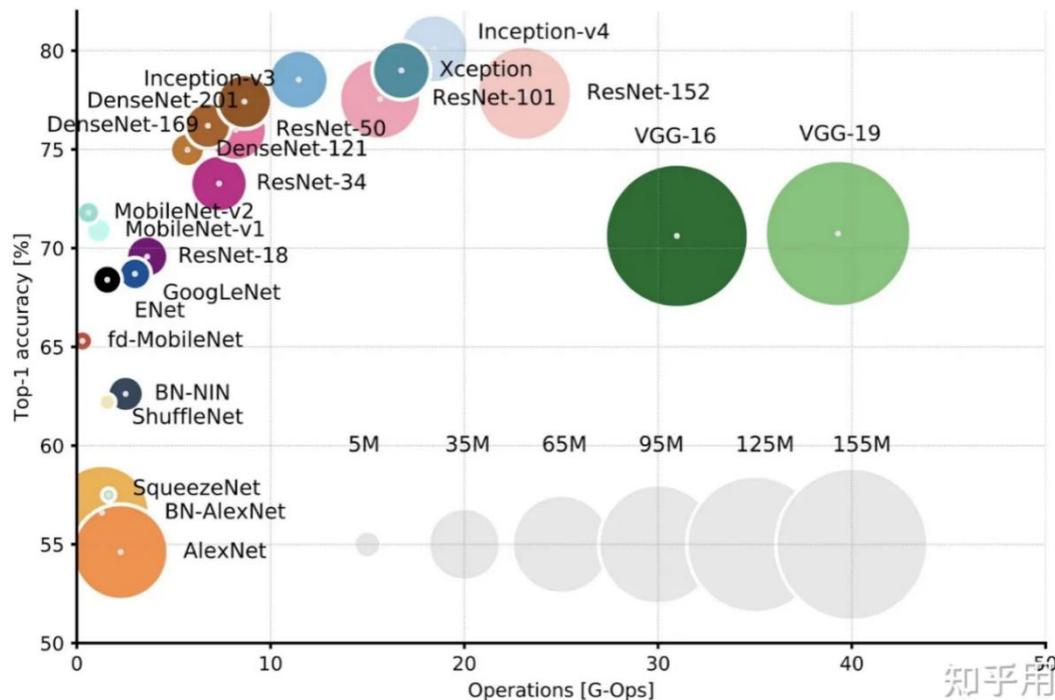
- 使用多层神经元组合来处理复杂的输入数据。
- 可以学习到高维输入特征的复杂关系，如图像视频、语音、自然语言等。
- 可扩展, 依赖算力和数据。



ImageNet Classification with Deep Convolutional Neural Networks, 2012.
作者: Alex Krizhevsky, [Ilya Sutskever \(OpenAI\)](#), Geoffrey E. Hinton

2012-2017 DNN模型规模稳步增长

深度学习初期模型越来越大



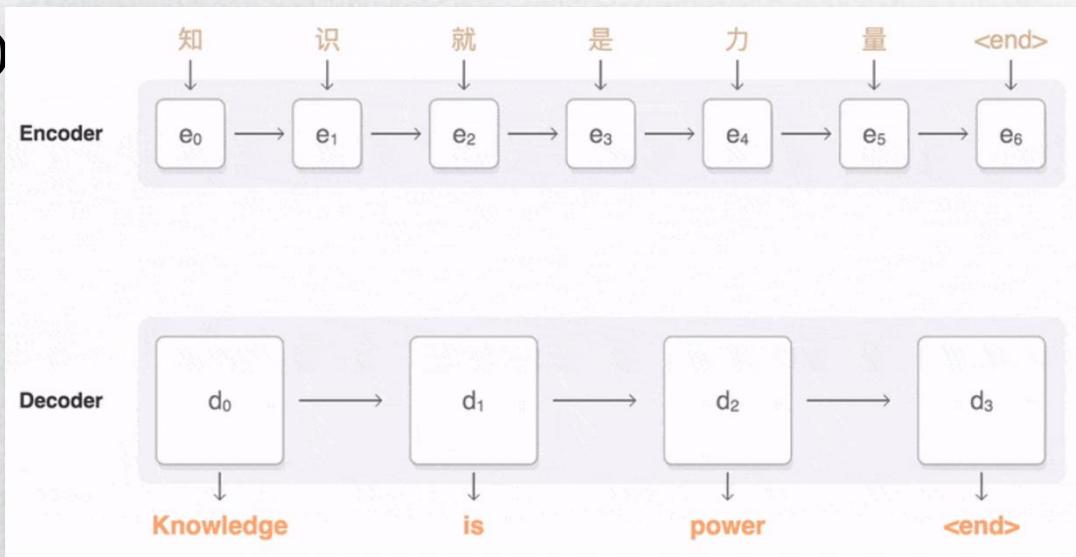
知乎用户

图片来源于知乎

语言模型 Transformer - 2017年

新一代序列数据模型 (参数过亿)

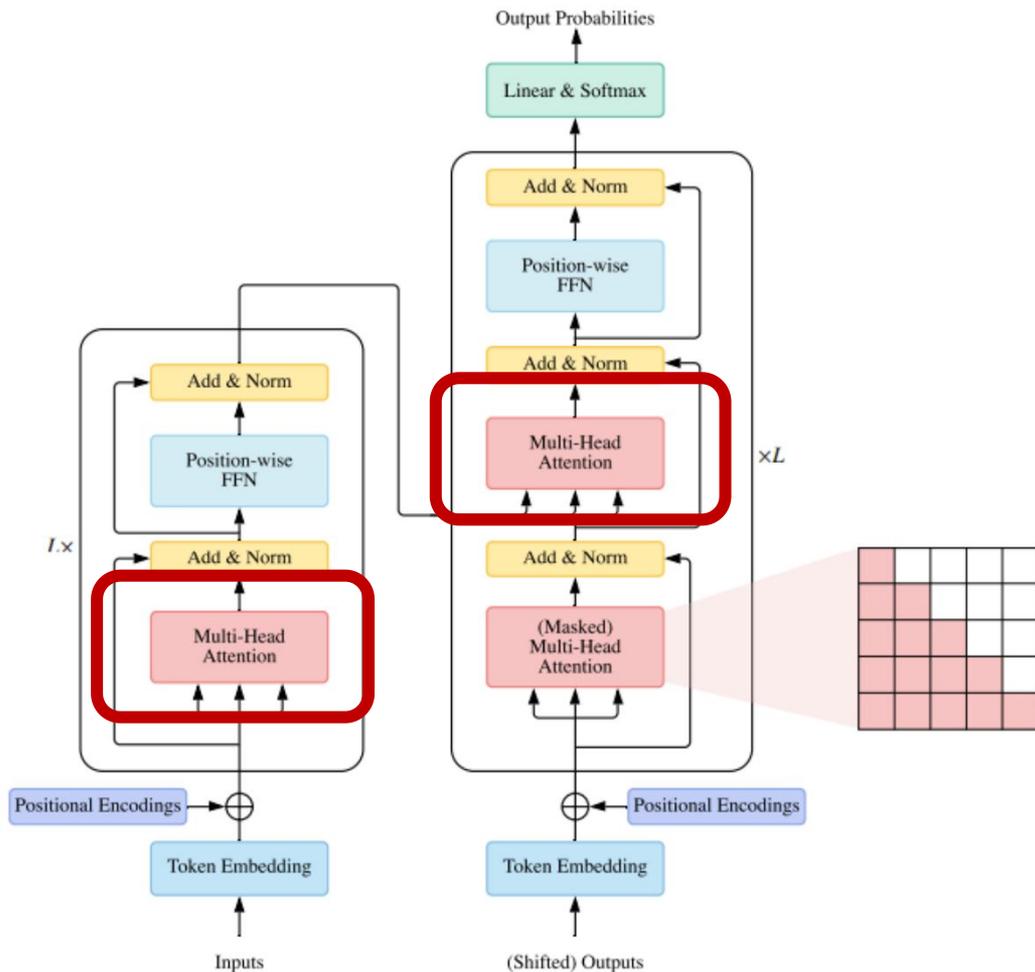
- 基于编码-解码模型
- 计算所有输入单词之间的相关程度(编码)
- 计算所有输入-输出单词之间的相关程度(解码)
- 模型计算量大大增加



基本单元: 注意力模块

使用基于键值对的注意力机制, 以实现:

- 自注意力(self-attention) 对输入进行编码
- 交叉注意力(cross-attention) 对输出进行解码
- 更好理解上下文
- 可以输出更长的文字, 理解历史对话

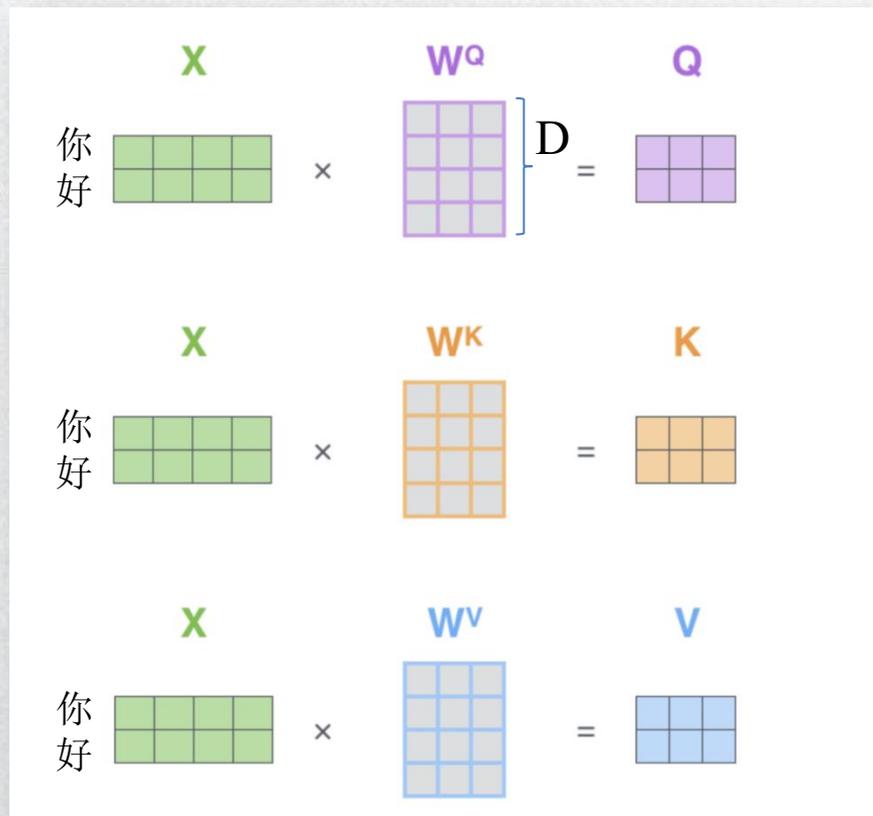


基于键值对的注意力机制(主要的计算消耗单元)

对一句话每一个单词

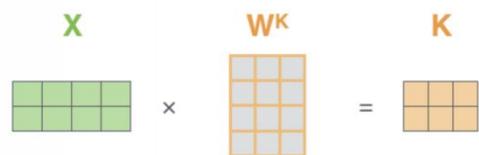
- 计算Query (查询)
- 计算Key (键)
- 计算Value (值)

参数量为 $O(D^2)$



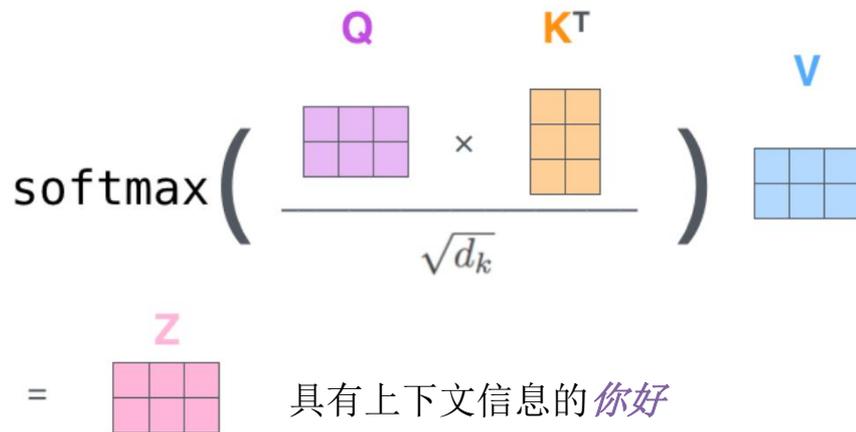
基于键值对的注意力机制(主要的计算消耗单元)

$$X \times W^Q = Q$$


$$X \times W^K = K$$


$$X \times W^V = V$$

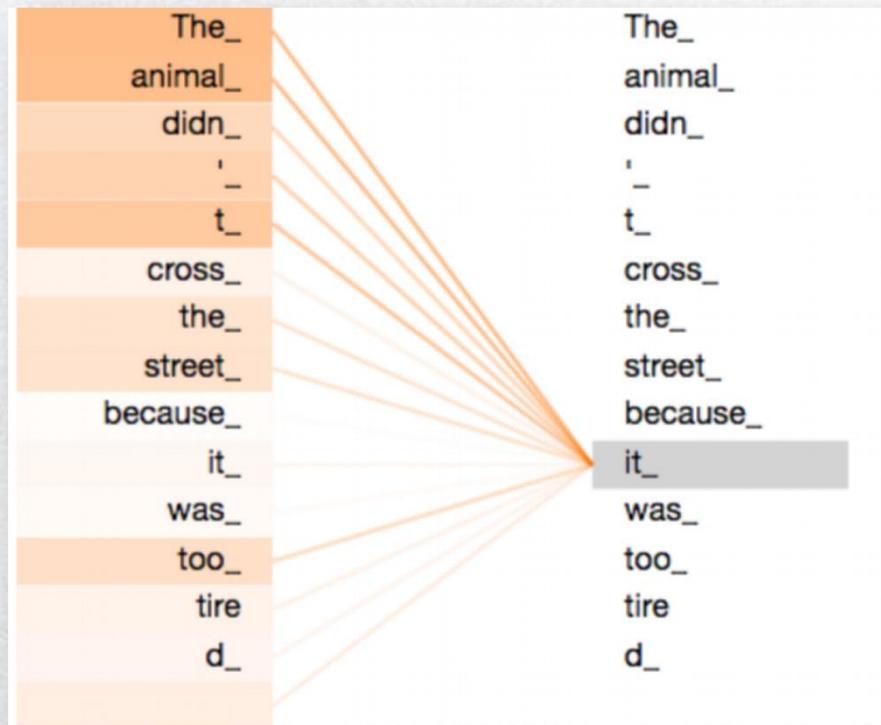


$$\text{softmax} \left(\frac{Q \times K^T}{\sqrt{d_k}} \right) V$$
$$= Z \quad \text{具有上下文信息的} \text{你好}$$


计算量为 $O(N \cdot D^2)$, N 为单词个数

注意力机制的目的 – 更好的编码和解码!

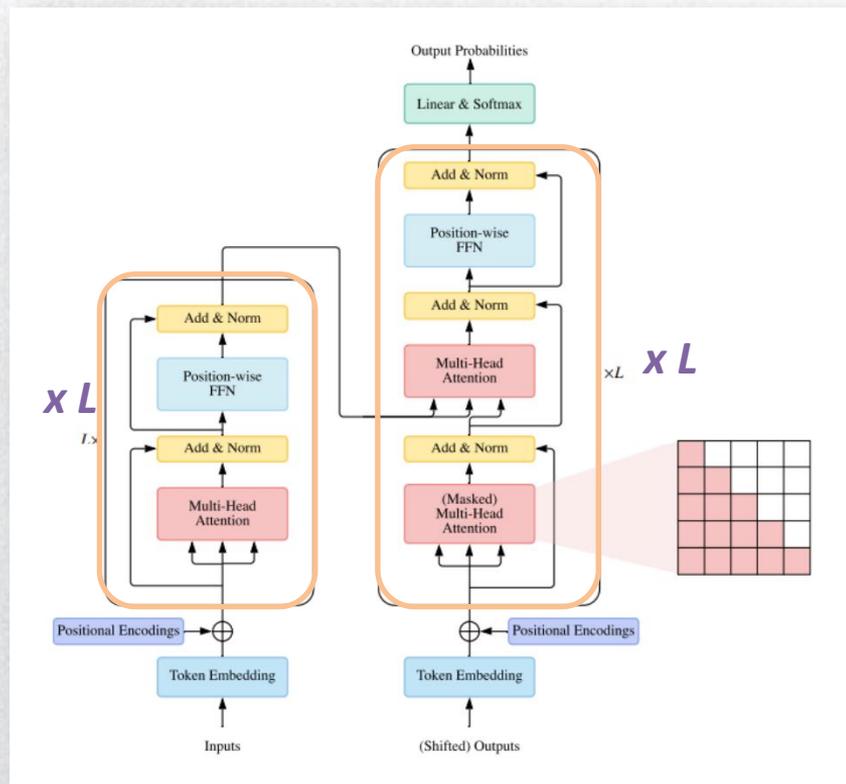
- 通过注意力机制编码
- 每一个单词均能找到合适的上下文信息
- 例如, it 关联到
 - animal,
 - cross-the-street,
 - too tired



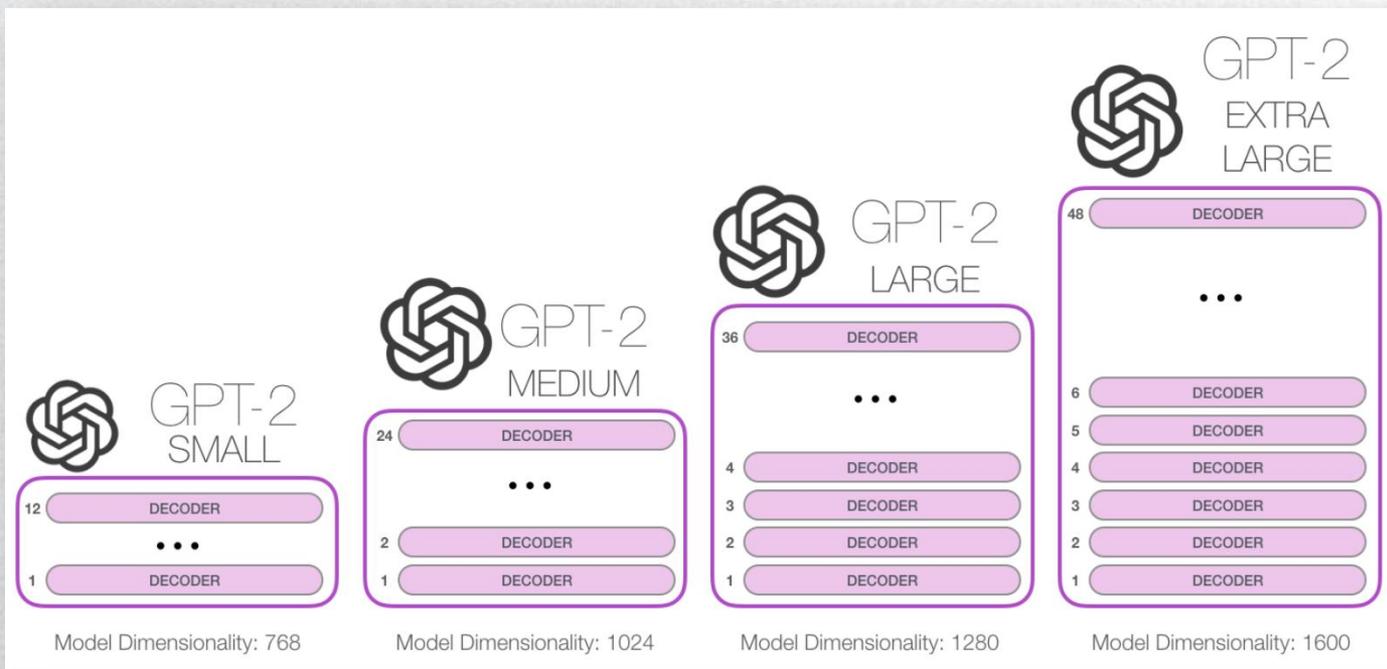
The animal didn't cross the street because it was too tired.

模型参数膨胀的秘密

- 大量堆叠注意力模块层数 L
e.g., $L=96$ (ChatGPT)
- 隐藏层维度 D 增长为 2048
- 注意力特征参数 $O(D^2 \cdot L)$
- 其他参数, 如词嵌入, 输出层
 - 达到上亿级别



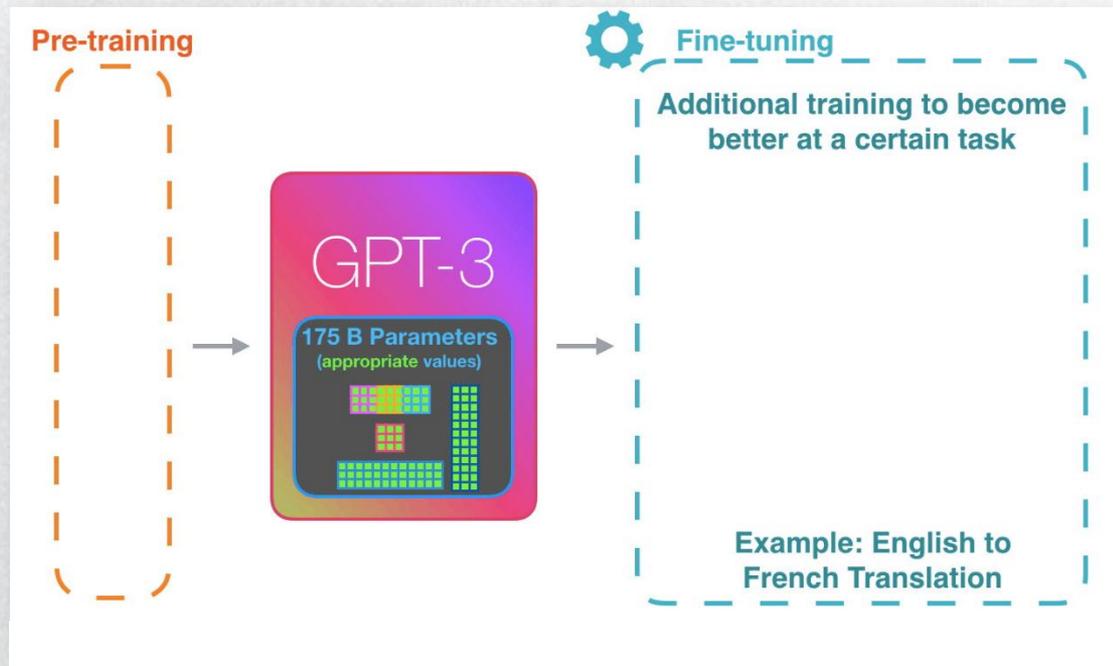
层数(纵向)叠加



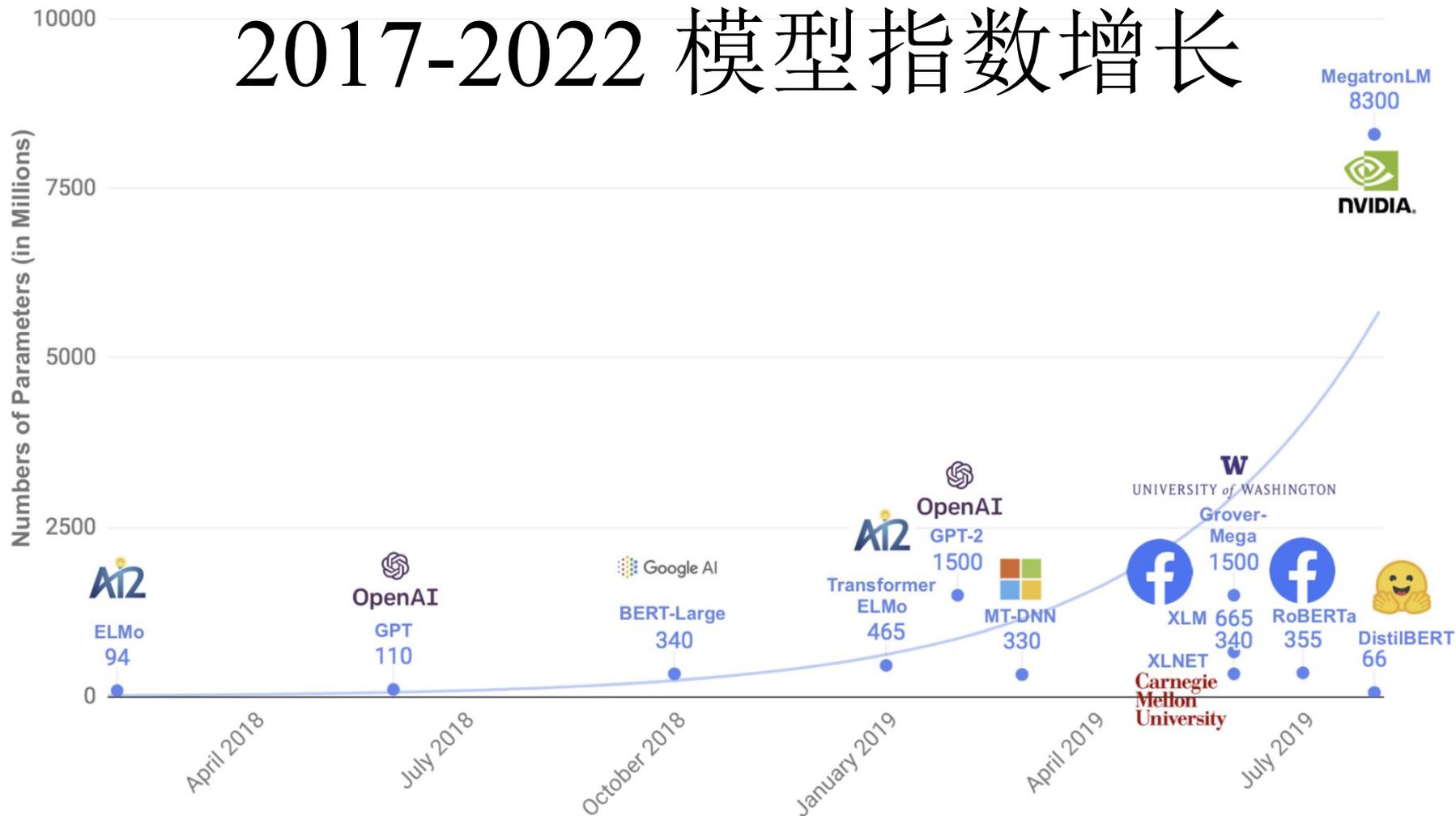
维度(横向)增加

大语言模型 (Large Language Model)

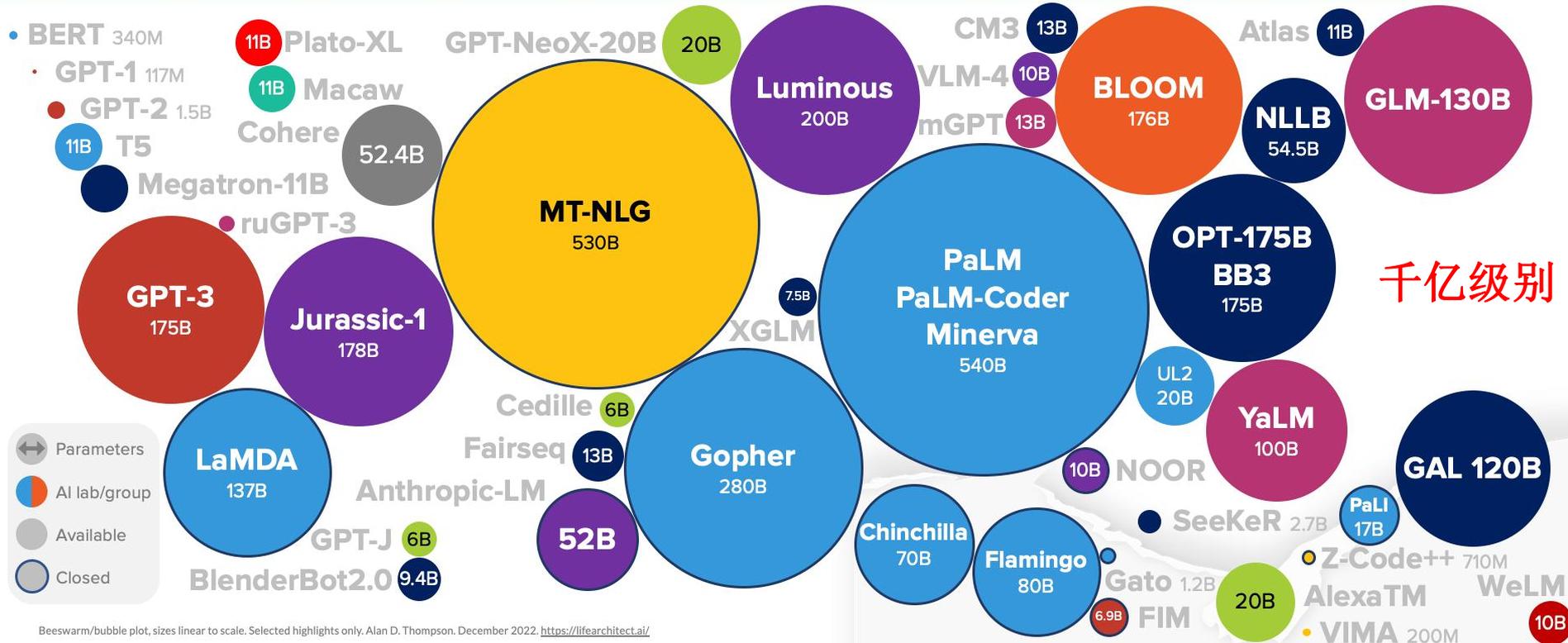
- 参数个数1千亿-2万亿
- 采用预训练加任务微调
- 超级数据中心进行训练



2017-2022 模型指数增长



LANGUAGE MODEL SIZES TO DEC/2022



Beeswarm/bubble plot, sizes linear to scale. Selected highlights only. Alan D. Thompson, December 2022. <https://lifearchitct.ai/>



3. 大模型算力成本(载荷)

2021年 (公布数据):

1. GPT-3 在三月份有1百万注册用户
2. GPT-3 输出 3百万 单词/分钟 (WPM)

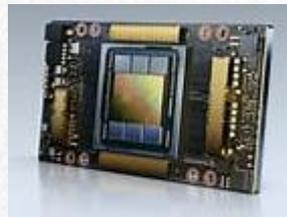
2022/23 估算:

1. ChatGPT 有 1亿月活用户
2. 估算ChatGPT 输出量为3亿单词/分钟
3. 5000台服务器可满足全世界的访问要求

大模型算力成本(软硬件)

模型训练

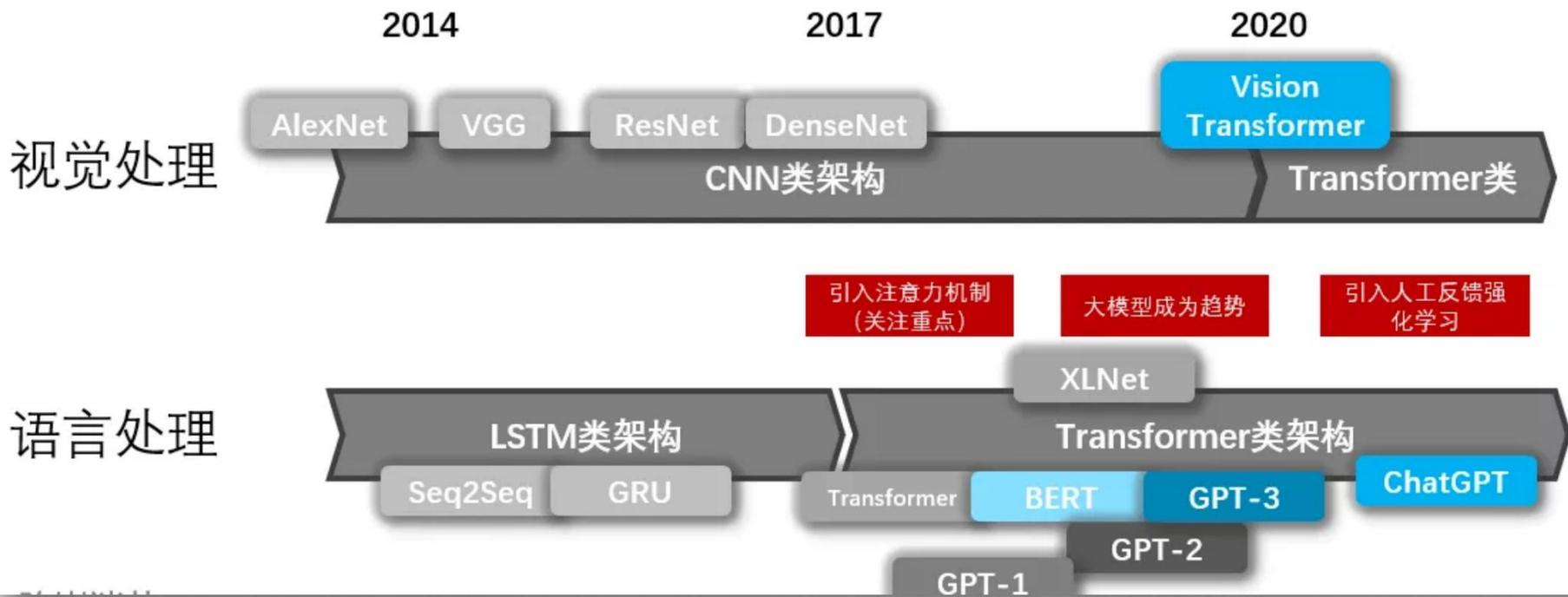
- 硬件: 384块A100, 单张卡显存80GB, 满足1750亿参数批量训练
- 成本: 960万美元 (48台DGX-A100服务器, 20万美元/台)
- 电费: 单次训练300万美元, 每天5万 × 2个月



部署推理

- 硬件: 8张A100单次运算, 每秒钟能产生大约15-20个单词
- 部署: 3亿单词/分钟=> 5000台服务器满足访问=> ~ 10亿美元
- API 收费(最新公开): 100万单词 / \$2.7

4. 视觉及多模态大模型



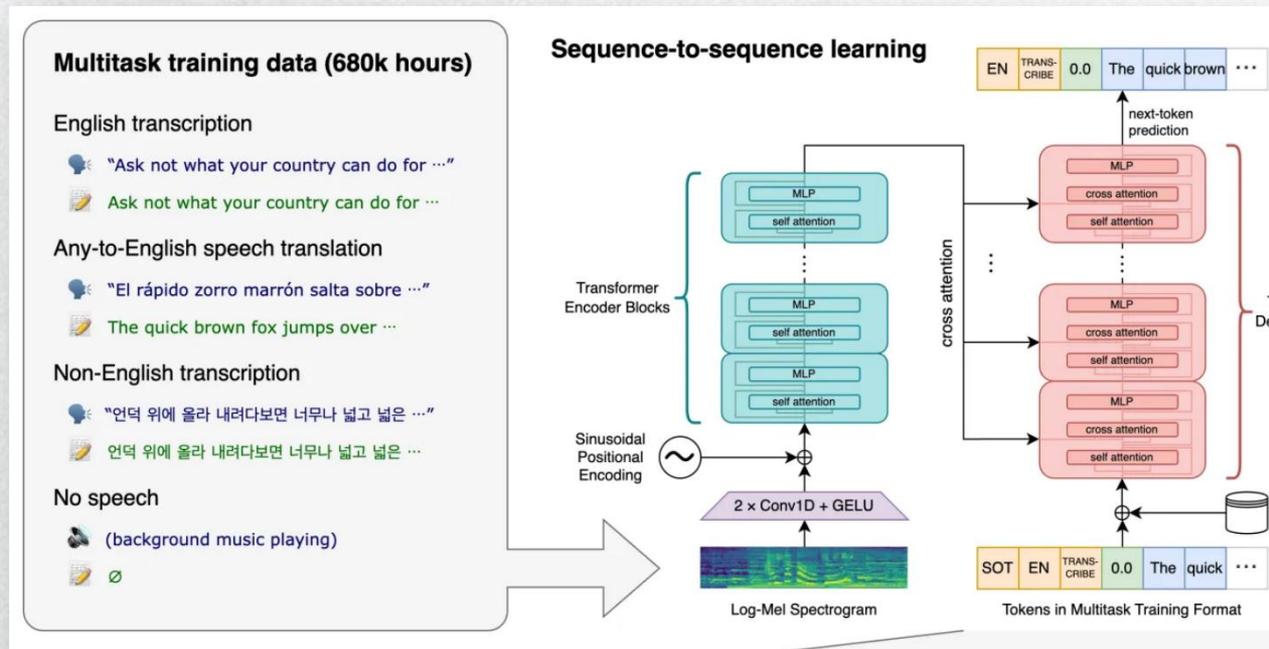
多模态 预训练大模型 (PLM)

- 图像识别
- 文本-图像生成
- 语音识别和翻译
- 情感分析



Whisper 语音识别模型

- 15 亿参数
- 96种语言
- 68万小时音频/字幕数据
- 接近人类水平



Whisper 语音模型

- 英文语音转录成英文文本;
- 其他语言语音翻译成英文;
- 其他语言转录成该语言;
- 识别环境是否有人说话。

Multitask training data (680k hours)

English transcription

 "Ask not what your country can do for ..."

 Ask not what your country can do for ...

Any-to-English speech translation

 "El rápido zorro marrón salta sobre ..."

 The quick brown fox jumps over ...

Non-English transcription

 "언덕 위에 올라 내려다보면 너무나 넓고 넓은 ..."

 언덕 위에 올라 내려다보면 너무나 넓고 넓은 ...

No speech

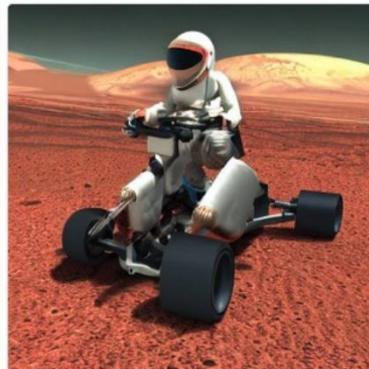
 (background music playing)

 ∅

图像生成 - Diffusion扩散模型

student drive a KARTING on mars

Generate image



astronaut rides a horse

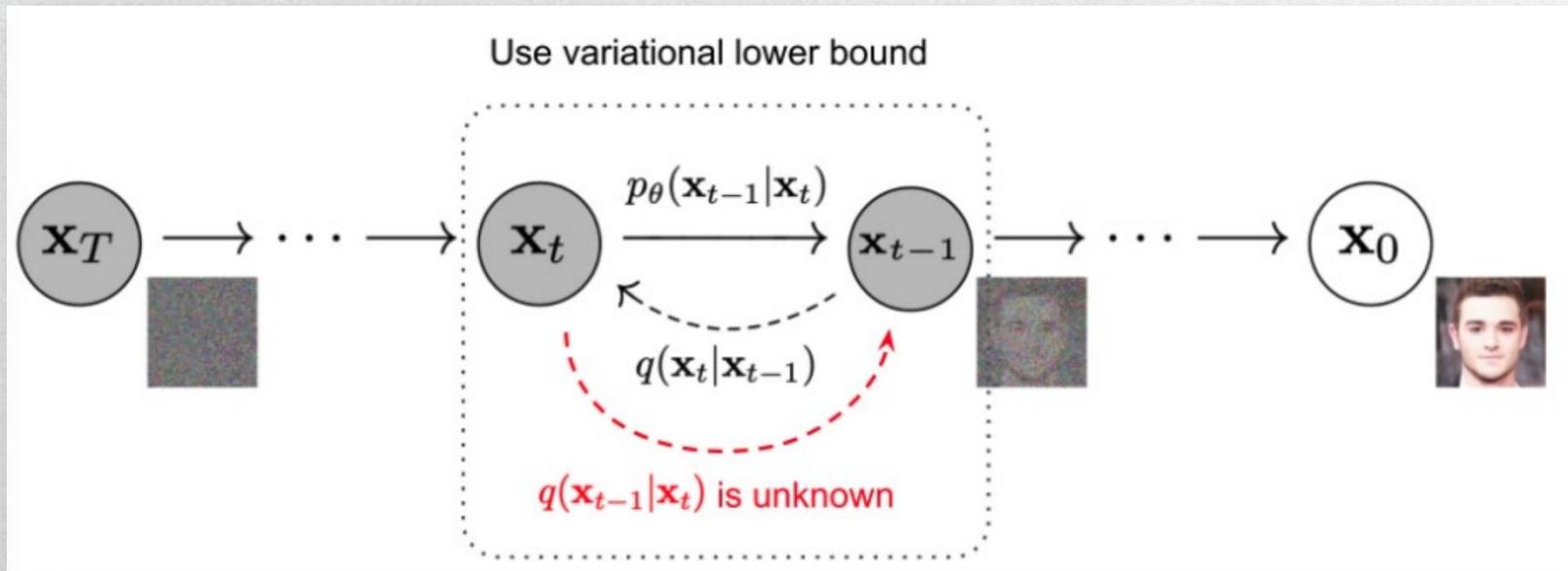
Generate image



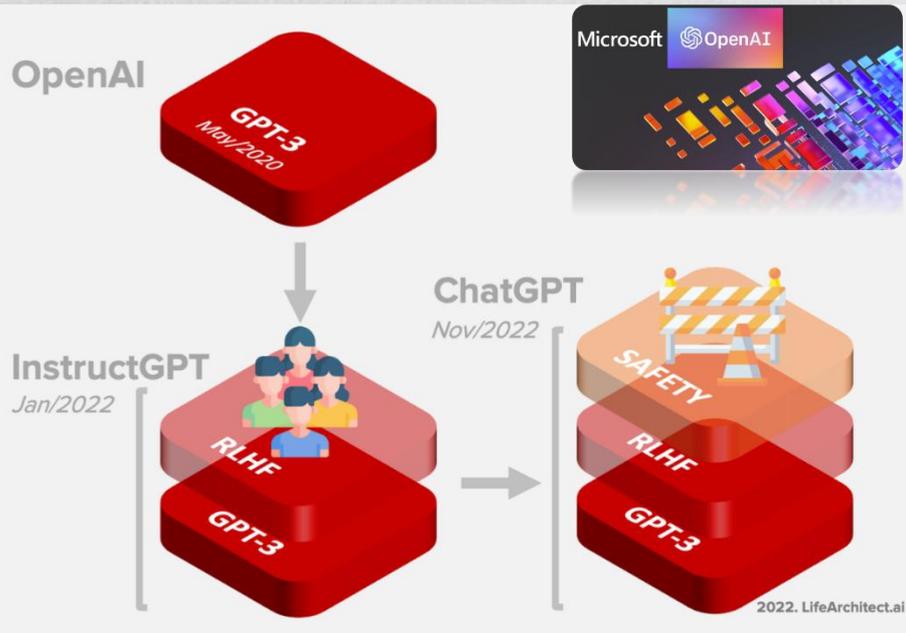
<https://stablediffusionweb.com/#demo>

扩散模型 (~50亿参数)

学习到人类图片的所有纹理和语义特征



GPT (Generative Pre-Training) 发展历史



模型	年份	参数量	数据量
GPT-1	2018	1亿	5GB
GPT-2	2019	15亿	40GB
GPT-3	2020	1750亿	45TB
GPT-3.5	2022	1750亿	-

Chat-GPT：丰富多彩的生成能力

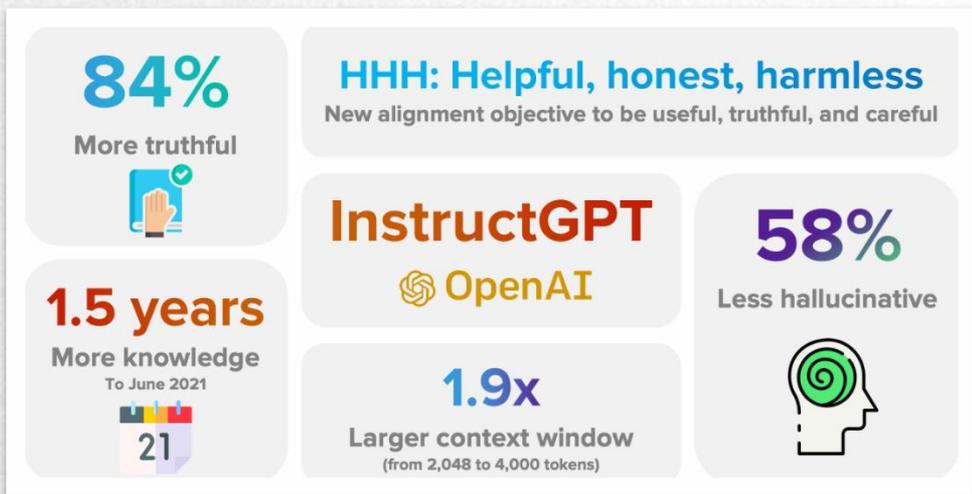
- 多种体裁：
 - 小说,诗歌,邮件,学术,代码
- 角色: 模仿各种人物写作风格
 - 李白,丘吉尔,夏洛克
- 指定语气和情感进行
- 语言: 超过90种语言支持

用幽默的口气向老师建议考试简单一些.

老师,你让我们太难了,让我们来一场轻松的考试吧!也许可以考考我们的“把一只猫放进冰箱”的答案!

ChatGPT 数据 - 人类反馈

- 工程师设计提示文本，由人类回答进行有监督地微调训练
 - 1万条对话
- 人工对模型输出打分排序，训练模型分辨不当内容
 - 5万条对话
- 基于打分，使用强化学习算法进行微调，输出无害的、人类可理解的答案



应用举例

- 作为学生, 可以使用大模型做哪些工作 ?
- 作为AI/CS专业的学生, 如何提高自己的代码和写作水平?
- 作为日常使用工具, 如何使用大模型提高工作效率?

示例代码生成

Q:用python写一个快速排序

```
# 快速排序
def quick_sort(array):
    if len(array) < 2:
        return array
    else:
        pivot = array[0]
        less = [i for i in array[1:] if i <= pivot]
        greater = [i for i in array[1:] if i > pivot]
        return quick_sort(less) + [pivot] + quick_sort(greater)
```

代码 debug

Q:how to correct the code `a[1,2,3,4,5] print(a[5])`

A:

The code should be changed to `print(a[4])` since the index of the last element in the list is 4.

修改英文语法

Q:修改英文语法Please revise. He go to work early everyday.

A:

He goes to work early every day.

写一封请假信

Q:写一封中文请假信，请假10天，说明我要陪父母去海南玩

A:

尊敬的老师:

你好!

我是你的学生XXX，因为家里有事，我需要请假10天。具体原因是，我的父母要去海南旅游，我想陪着他们一起去，所以我需要请假10天。

文献检索



提供一些多模态和chatgpt结合的论文链接



以下是一些关于多模态和ChatGPT结合的论文链接，供您参考：



1. "M-MIMIC: Multimodal Medical Information Generation with Iterative Refinement and Context Fusion" (2022) by Yifan Zhang et al. <https://arxiv.org/abs/2201.01553>
2. "Multimodal Language Generation with Recurrent Variational Autoencoders" (2021) by Xinyu Zhang et al. <https://arxiv.org/abs/2106.07228>
3. "Multi-Modal Conversational AI with Dense Generative Transformer Networks" (2021) by Sumanth Dathathra et al. <https://arxiv.org/abs/2103.03229>
4. "MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations" (2019) by Soujanya Poria et al. <https://arxiv.org/abs/1810.02508>

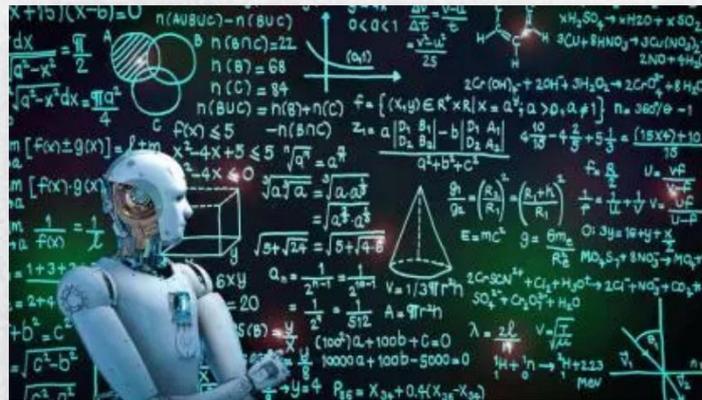
这些论文探讨了多模态和ChatGPT结合的应用，从不同的角度研究了该领域的问题和挑战，并提出了一些创新性的方法和解决方案。希望这些链接对您有所帮助！

令人深思的话题

- 我们能够信任AI给的答案的准确性吗? AI 医生的建议可靠吗?
- AI 真的能够理解一段完整的对话吗? 它会遗忘我们的历史对话吗? 它有逻辑推理能力吗?
- 人类还有必要思考吗, 如果我们的专业知识都能够被AI轻易的记住并回答. 我们会被AI取代吗?
 - **No. 小学数学题平均得分55分*!**

模型滥用的担忧

- 输出虚假、不实的信息和知识
 - 论文引用和链接不存在
 - 专业领域的事实错误
- 数据泄露问题
 - 隐私泄露被AI爬取
 - 研究资料通过对话被套取
- 恶意内容生成
 - 病毒代码,“毁灭人类计划书”





THANKS

QA