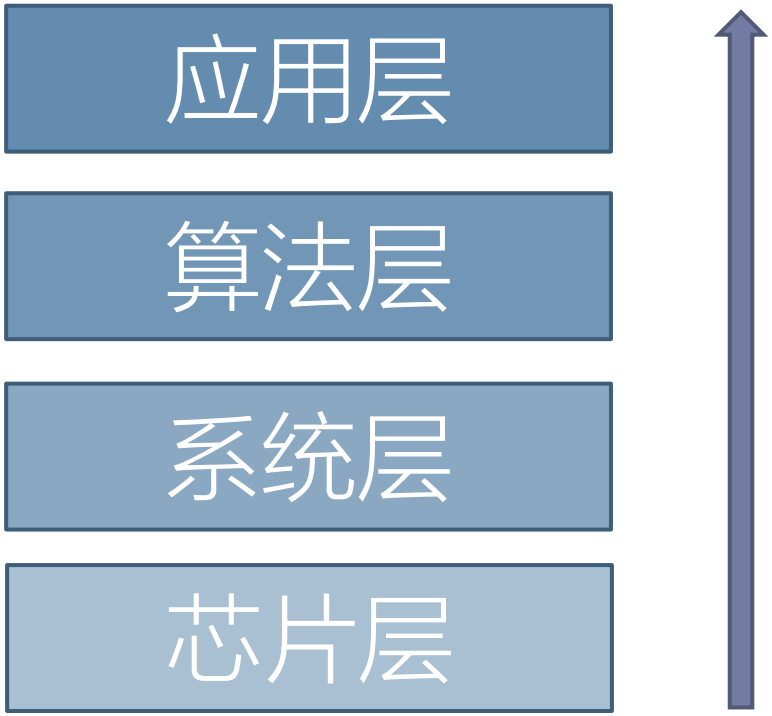
Ch11- 智能计算系统概论

王 超

提纲

- 为什么要引入智能计算系统
- 为什么要学习智能计算系统
- 人工智能软件与算法
- 智能计算系统
- 驱动范例

人工智能技术分层



人工智能底层科技的缺失可能使得我国智能产业成为空中楼阁

人工智能方向应该培养什么样的人才?

两个参考问题

人工智能方向应该培养人工智能（子）系统的设计者和研究者

力学、汽车试验学等方面工作的基本能力

- 计算机专业应该培养什么样的人才?
  - 计算机专业当培养计算机整机或子系统的设计者和研究者

对课程体系的发展建议

- 只包含各类机器学习算法、视听觉应用这条软件线，只能算是“人工智能应用专业”或者“人工智能算法专业”
- 谷歌有世界上最大的AI算法研究团队，然而
  - 谷歌董事长John Hennessy是计算机体系结构科学家，图灵奖得主
  - 谷歌AI的总领导者Jeff Dean是计算机系统研究者
  - 谷歌AI最令人瞩目的三个进展都是系统（Tensorflow、AlphaGo、TPU），而不仅仅是某个特定算法，算法只是系统的一个环节
- OpenAI ChatGPT的成功很大程度上来源于系统的发展
  - 10000个A100芯片组成的复杂系统
  - 每次训练花费超过1000万美元

应当包含系统线的课程，帮助学生理解系统到底是怎样执行的

对课程体系的建议

在高年级本科生（或者硕士研究生）阶段，应当设置一门系统类课程，能帮助同学实现对当前主流智能软硬件体系的融会贯通，具备自己动手完成一个完整智能系统的能力。这门课程就是智能计算系统

智能计算系统课程对学生的价值

- 全面的实践能力
  - 没有系统知识、只会调参，对整个系统的耗时、耗电毫无感觉，不具备把一个算法在实际系统上部署起来的能力的学生做不出真东西
  - 会用Tensorflow赚20万人民币，会设计Tensorflow赚20万美元
- 更强的研究能力
  - 能够从更广阔的视野和维度开展研究，不只是盯着准确率
  - 形成系统思维，拥有科研道路更广阔的舞台

什么是智能计算系统?

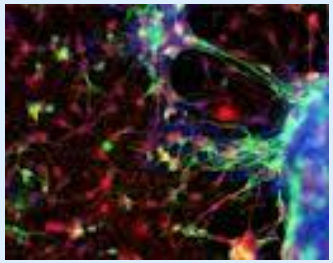
智能计算系统是智能的物质载体

现阶段的智能计算系统通常是集成CPU和智能芯片的异构系统，软件上通常包括一套面向开发者的智能计算编程环境（包括编程框架和编程语言）

智能计算系统的形态

超级计算机

商业分析



药物研制

数据中心

Baidu 百度

YouTube

广告推荐

自动翻译

智能手机

语音识别

图像分析

嵌入式设备

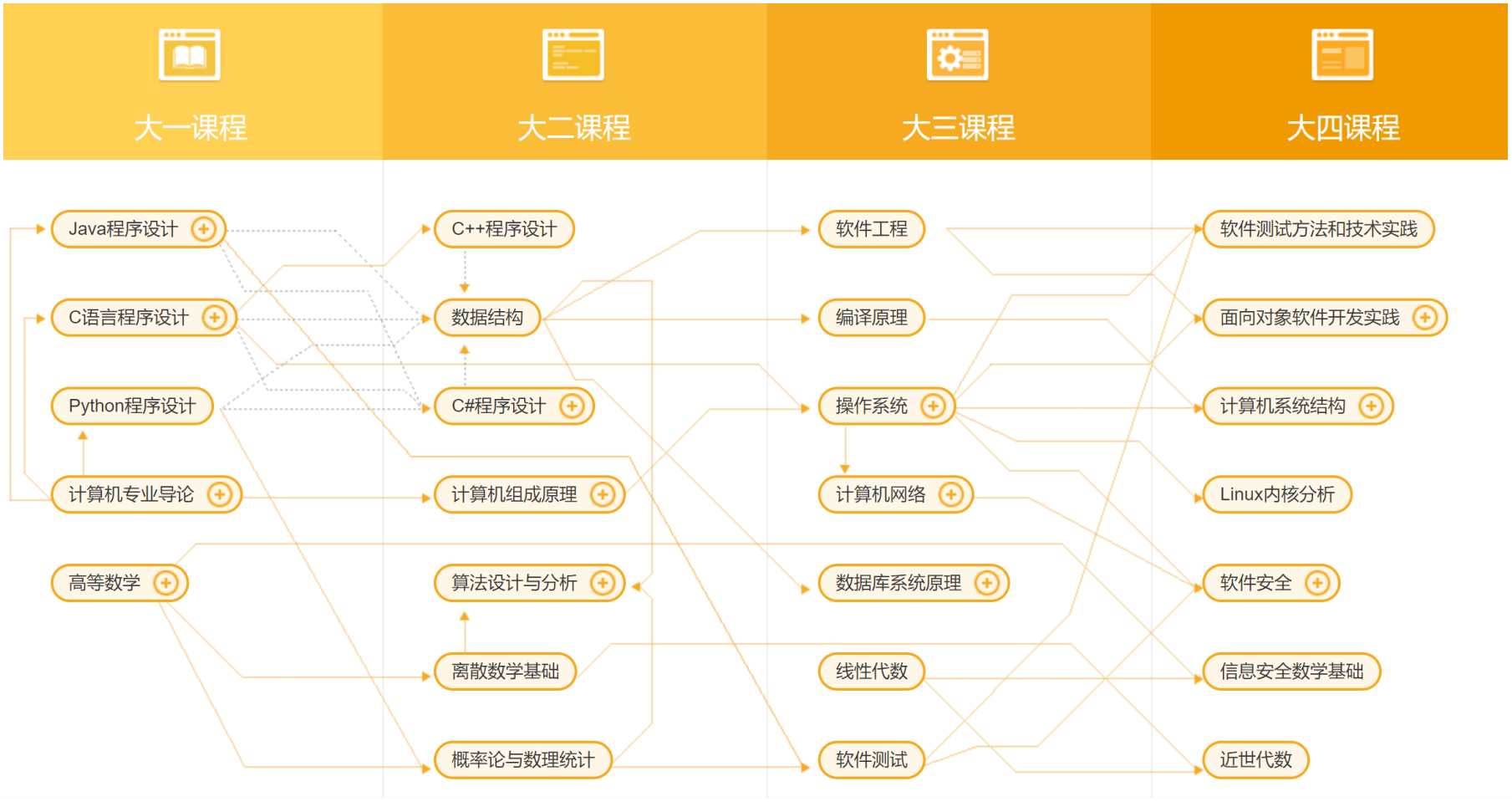
机器人

消费类电子

提纲

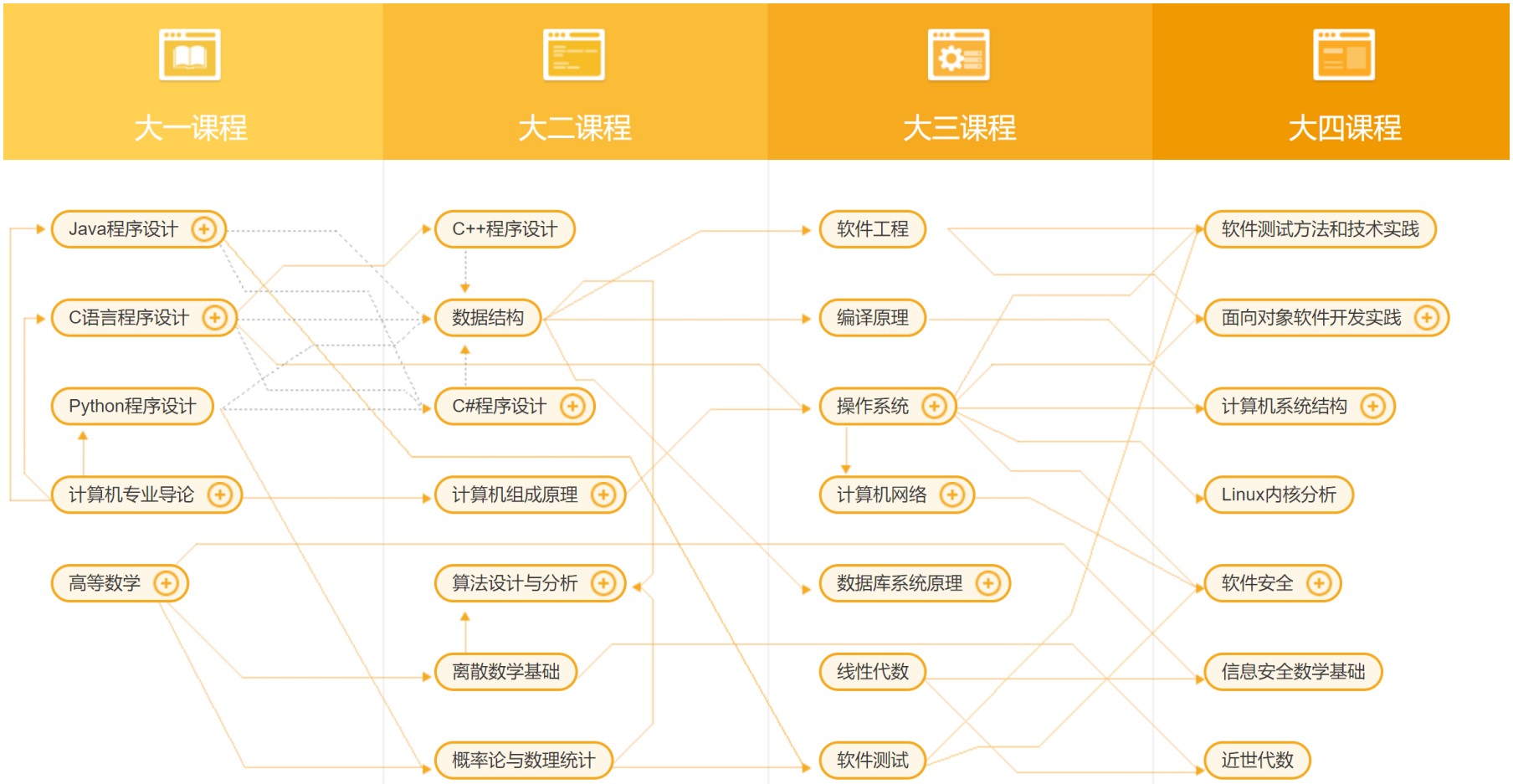
- 为什么要开这门课
- 为什么来上这门课
- 人工智能
- 智能计算系统
- 驱动范例

计算机专业培养计划的正面启示



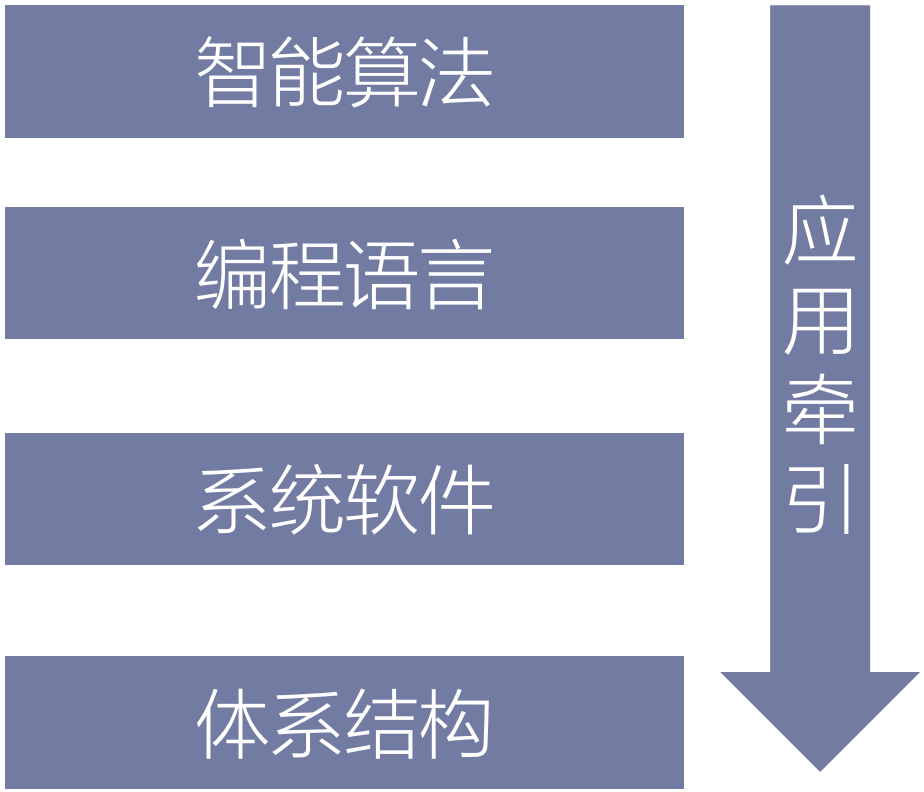
我国各高校计算机专业培养计划都包含计算机组成原理、操作系统、编译原理、计算机体系结构等硬件系统类课程

计算机专业培养计划的负面启示



课程条块分割，学生不能融会贯通做出一个完整系统，导致我国信息产业全栈式人才缺乏，核心硬科技竞争力缺失

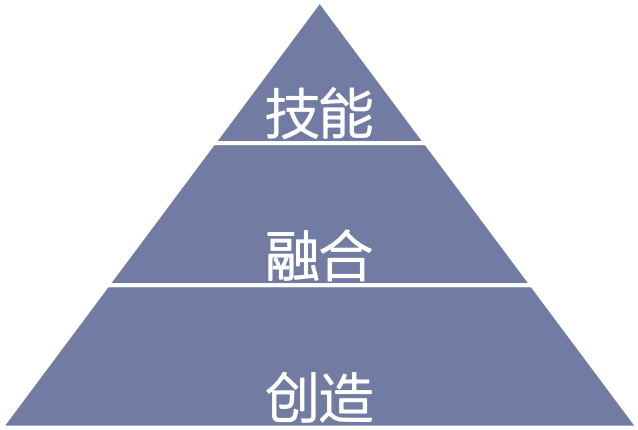
应用驱动、全栈贯通的课程体系



一门帮助学生学以致用、形成全局系统观的工科课程

课程目标和目的

- 中国需要一大批智能基础设施的开发者和设计者
- 专业普及课程
  - 应用驱动，全栈贯通
- 智能计算系统
  - 建立智能计算系统设计及应用的知识体系
  - 掌握智能应用开发的基本技能
  - 培养开展智能计算系统基础研究的兴趣和能力



能用智能计算系统，能开发智能应用

实现对智能计算软硬件技术栈的贯通理解

创造未来智能计算系统的科研基础

课程要求

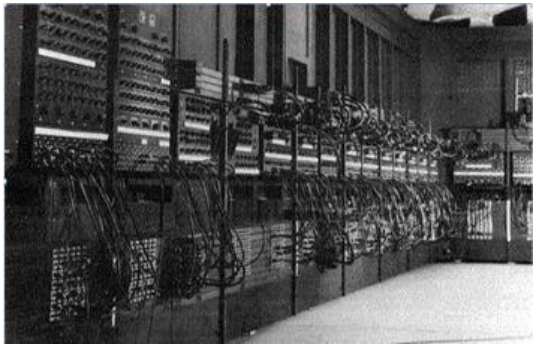
- C/C++
- 计算机组成原理/计算机体系结构
- 机器学习/算法导论

提纲

- 为什么要开这门课
- 为什么来上这门课
- 人工智能
- 智能计算系统
- 驱动范例

智能时代

蒸汽机



集成电路

智能计算系统

上世纪人类从工业时代过渡到信息时代
现在已经发展到向智能时代进化的拐点

中国需要一大批智能计算系统的开发者和设计者

国家战略

"AI holds the potential to be a major driver of economic growth and social progress" [White House report, 2016]

Released domestic strategic plan to become world leader in AI by 2030 [2017]

"Whoever becomes the leader in this sphere [AI] will become the ruler of the world" [Putin, 2017]

国家战略

Yearly Release of AI National Strategies by Country

Source: AI Index, 2022 | Table: 2023 AI Index Report

Year	Country
2017	Canada, China, Finland
2018	Australia, France, Germany, India, Mauritius, Mexico, Sweden
2019	Argentina, Austria, Bangladesh, Botswana, Chile, Colombia, Cyprus, Czech Republic, Denmark, Egypt, Estonia, Japan, Kenya, Lithuania, Luxembourg, Malta, Netherlands, Portugal, Qatar, Romania, Russia, Sierra Leone, Singapore, United Arab Emirates, United States of America, Uruguay
2020	Algeria, Bulgaria, Croatia, Greece, Hungary, Indonesia, Latvia, Norway, Poland, Saudi Arabia, Serbia, South Korea, Spain, Switzerland
2021	Brazil, Ireland, Peru, Philippines, Slovenia, Tunisia, Turkey, Ukraine, United Kingdom, Vietnam
2022	Italy, Thailand

加拿大于2017年3月正式推出了第一个国家人工智能战略；迄今为止，总共发布了62个国家人工智能战略。发布战略的数量在2019年达到峰值。

企业投入

Google "An important shift from a mobile first world to an AI first world" [CEO Sundar Pichai @ Google I/O 2017]

Microsoft Created AI and Research group as 4th engineering division, now 8K people [2016]

Meta Created Facebook AI Research, Mark Zuckerberg very optimistic and invested

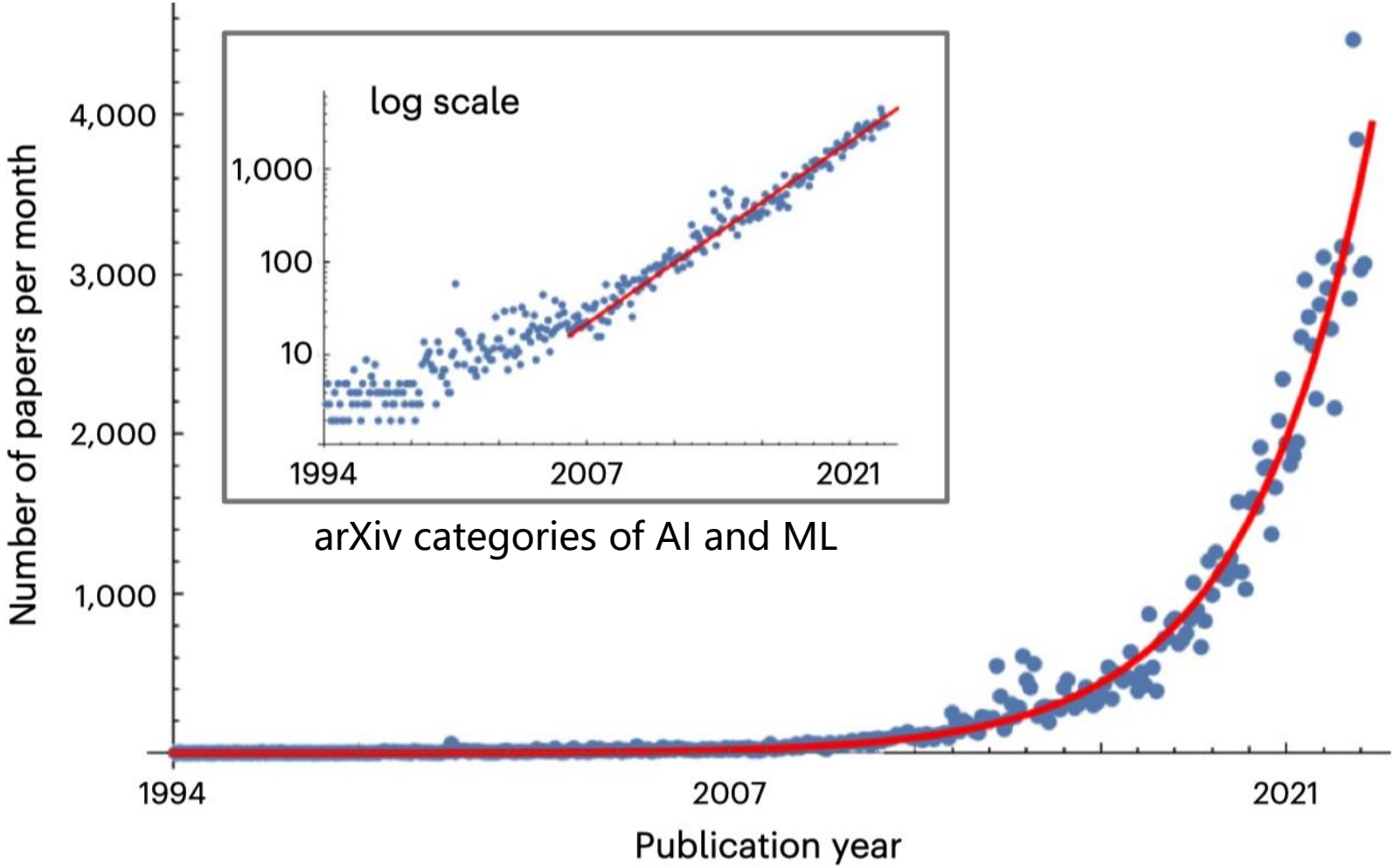
OpenAI Pioneering research on the path to AGI [2015]

Baidu 百度 百度将All in AI，我们在AI时代的核心战略就是开放赋能，我们的将来必须建立在与每个开发者共赢的基础上。[前COO陆奇@百度开发者大会2017]

Tencent 腾讯 AI in All，AI技术...能够真正和各行各业实际应用结合在一起，从而让AI新技术能够得到实际价值的发挥。[COO任宇昕@腾讯全球合作伙伴大会2017]

Alibaba Group 阿里巴巴集团 未来3年阿里巴巴在技术研发上的投入将超过1000亿人民币.[马云@2017杭州·云栖大会]

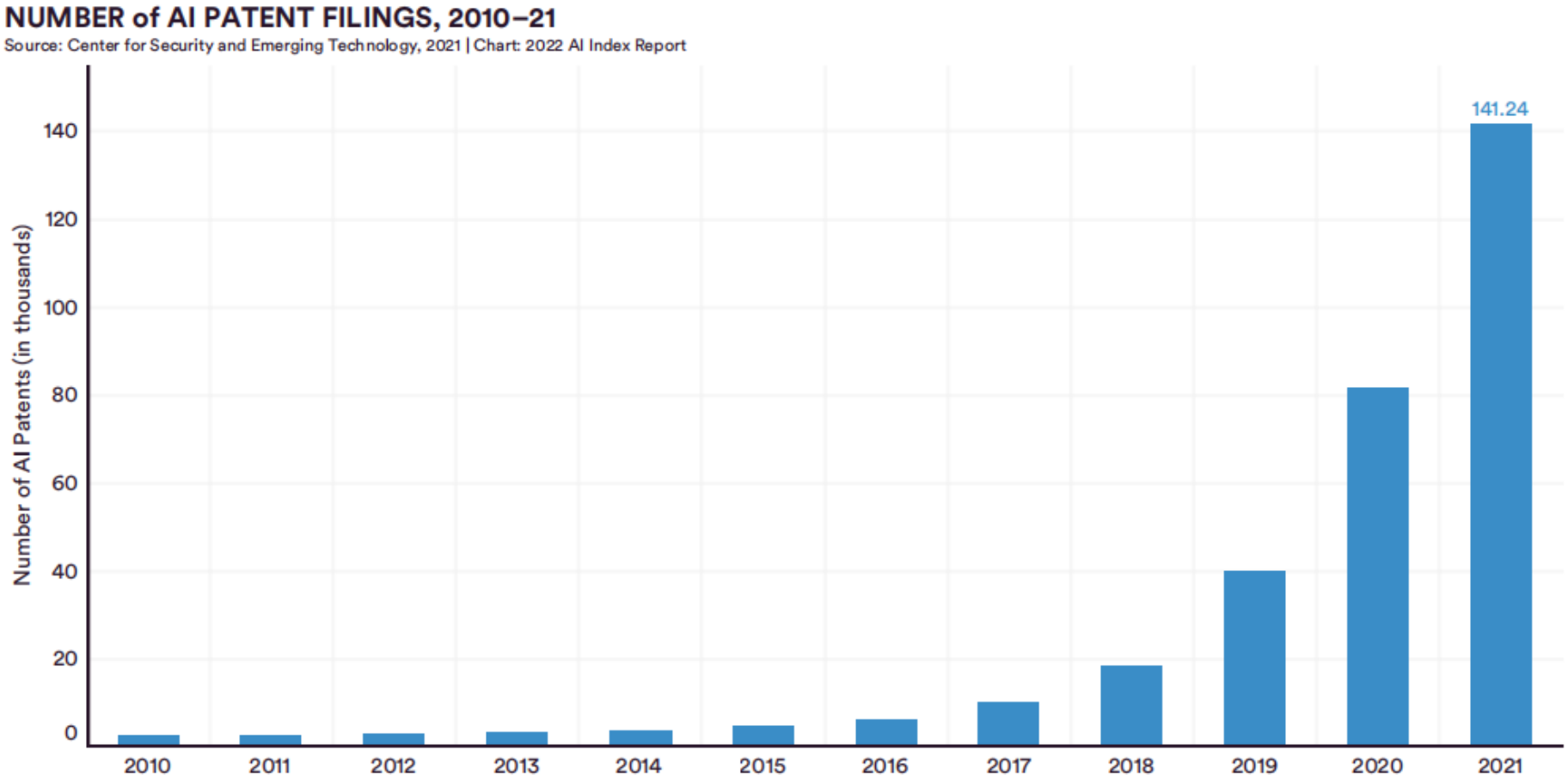
研究趋势



AI相关论文指数增长

From: Forecasting the future of artificial intelligence with machine learning-based link prediction in an exponentially growing knowledge network, nature, 2023

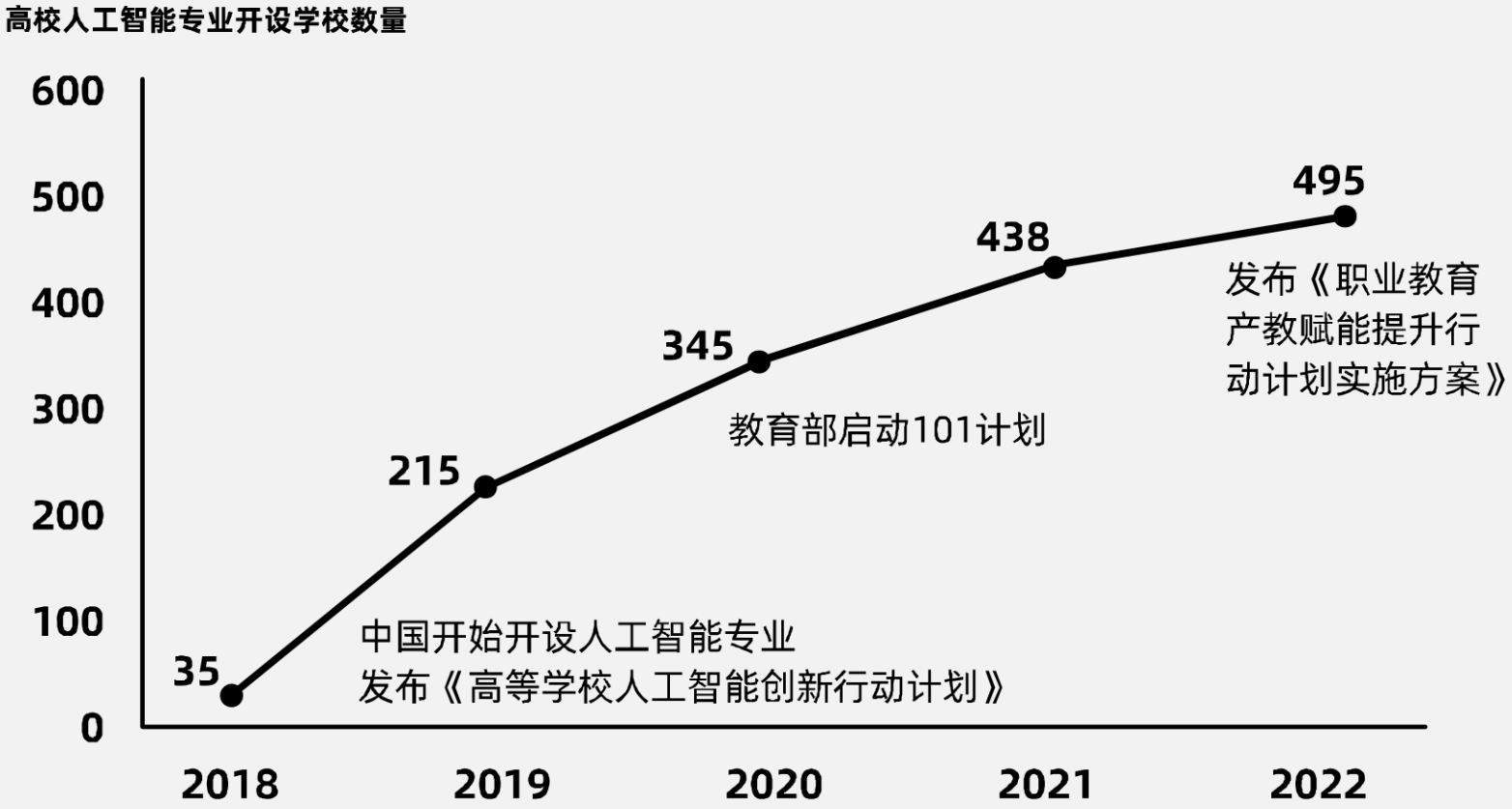
研究趋势



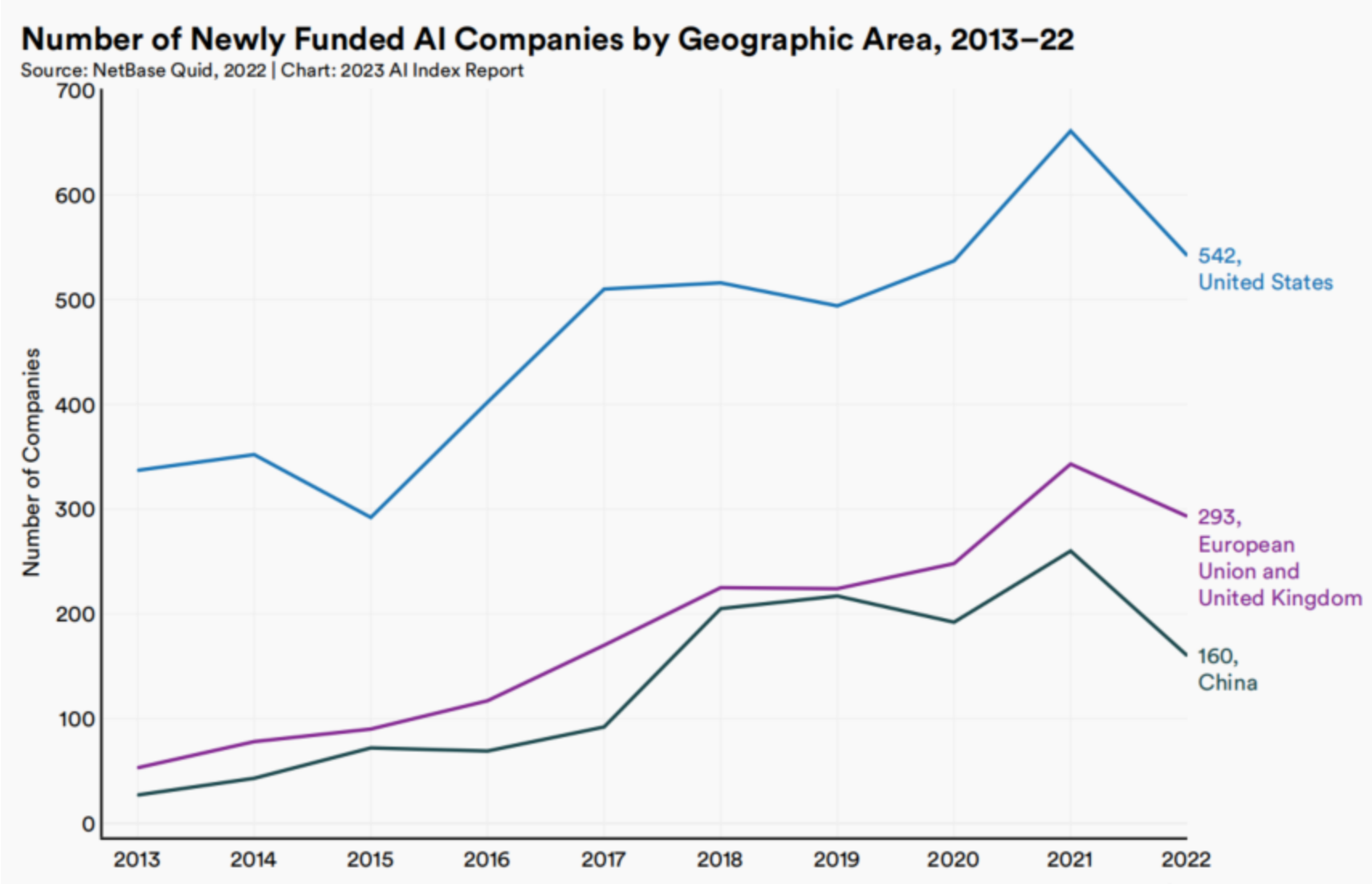
2015至2021年，AI相关专利数量增长30多倍，复合年增长率76.9%

高等教育

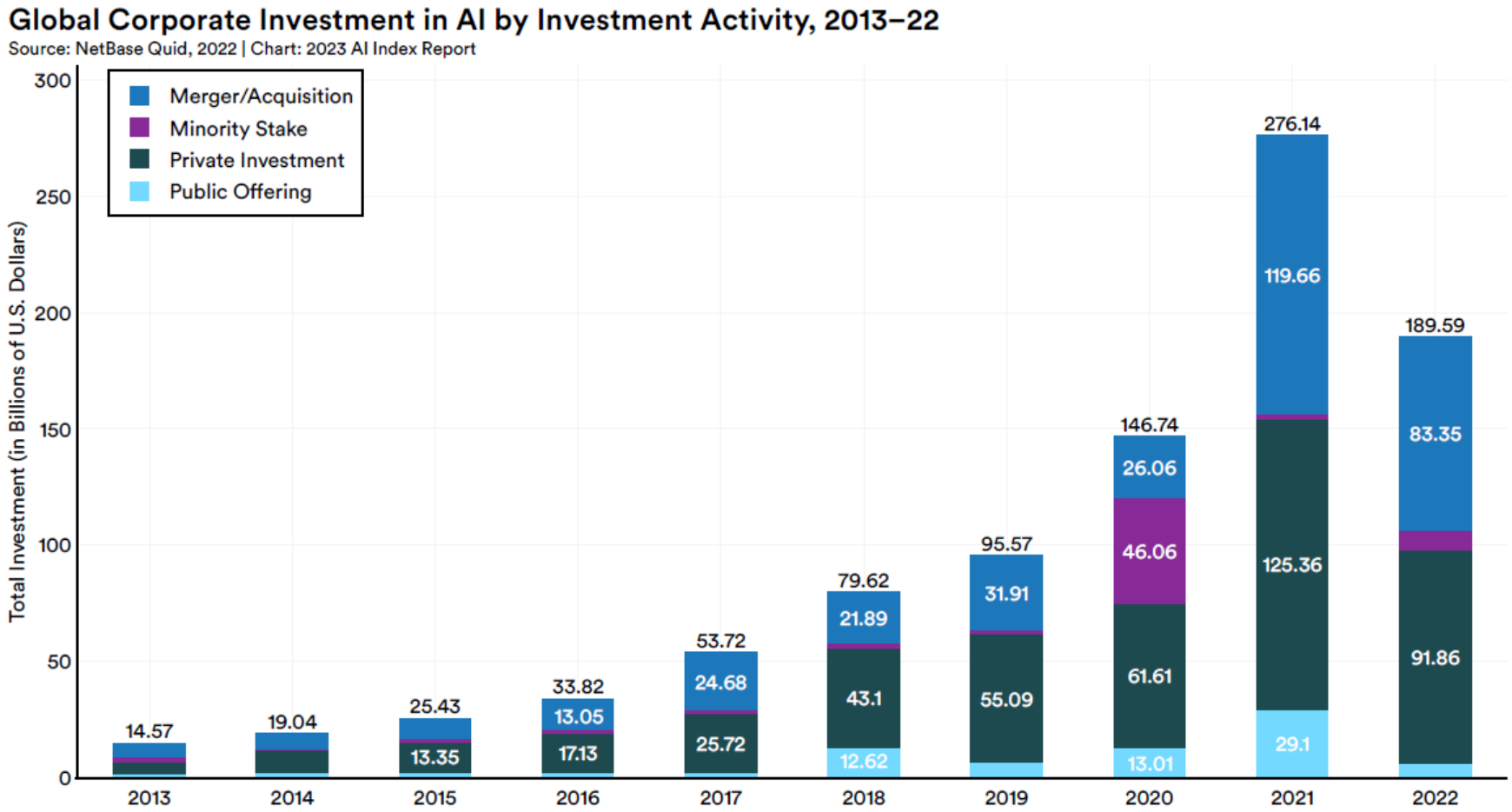
中国人工智能专业建设情况



AI新创公司

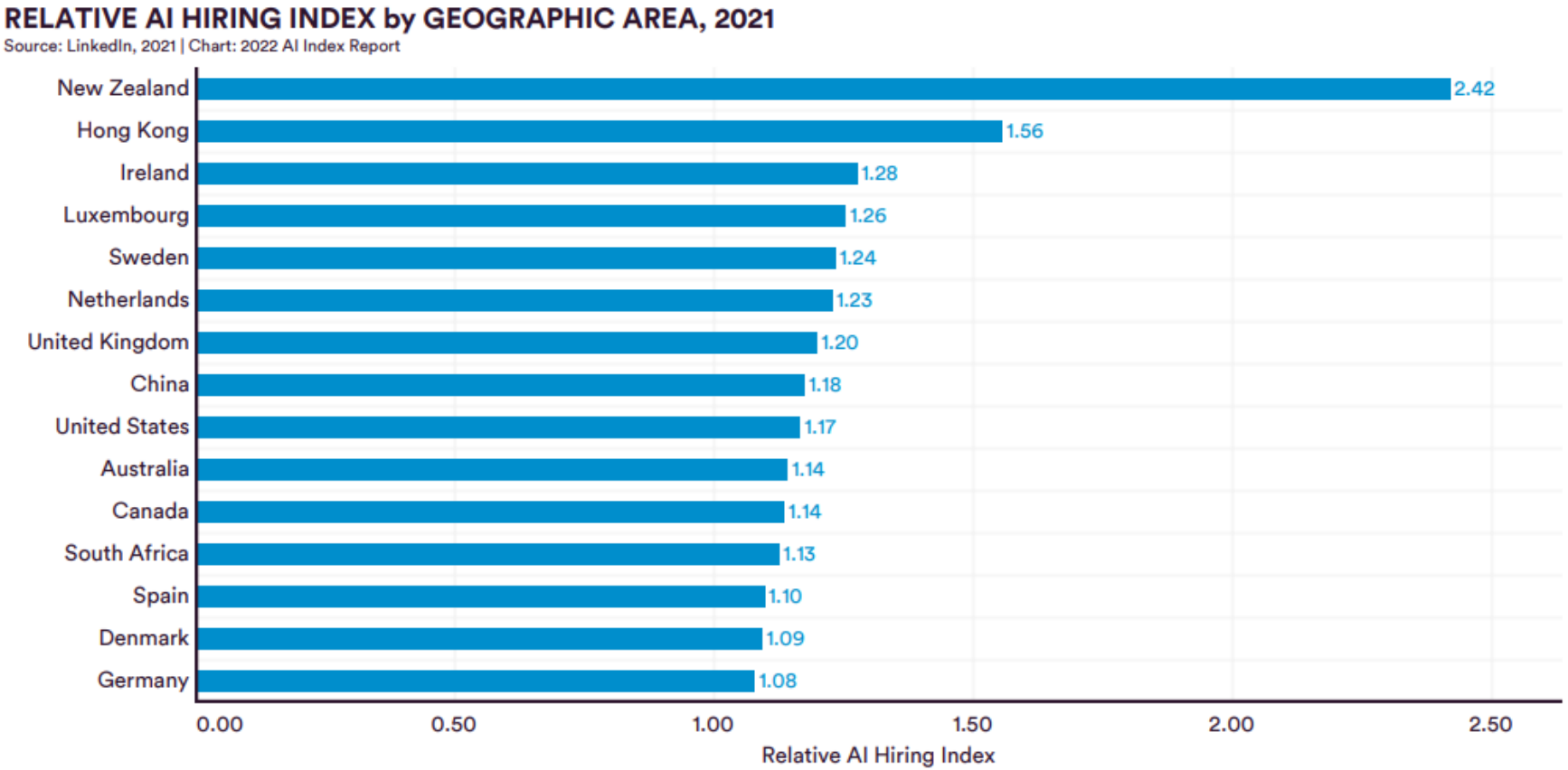

Number of Newly Funded AI Companies by Geographic Area, 2013–22
Source: NetBase Quid, 2022 | Chart: 2023 AI Index Report
Number of Companies
700
600
500
400
300
200
100
0
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
542, United States
293, European Union and United Kingdom
160, China

投资规模

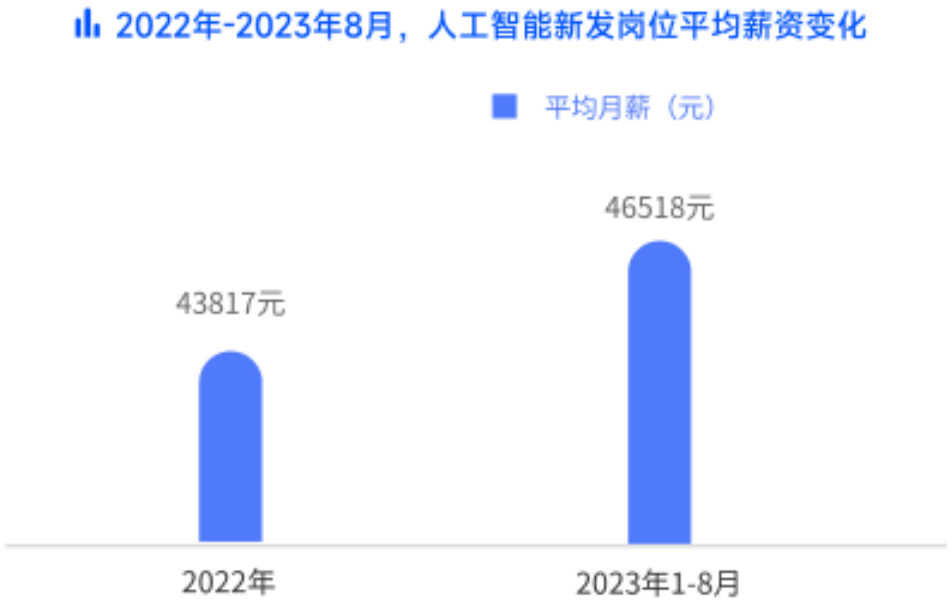


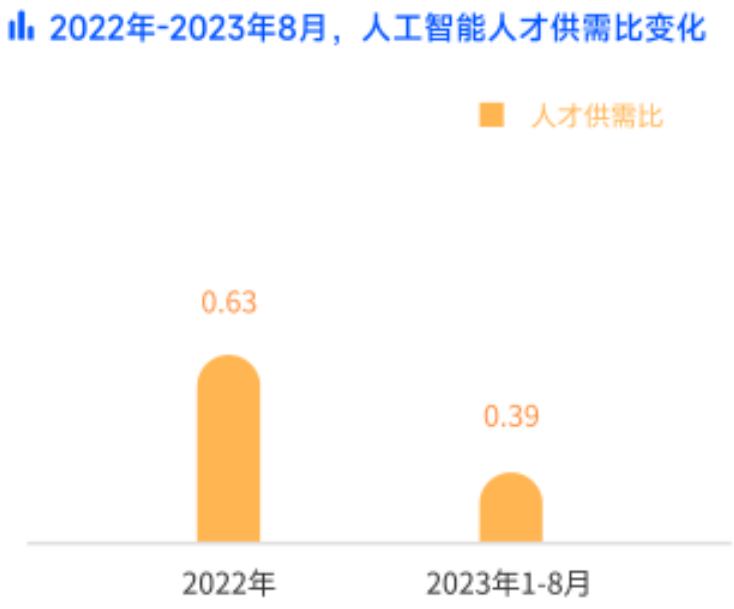
在过去的十年里，与人工智能相关的投资增长了十三倍

就业机会


RELATIVE AI HIRING INDEX by GEOGRAPHIC AREA, 2021
Source: LinkedIn, 2021 | Chart: 2022 AI Index Report
New Zealand 2.42
Hong Kong 1.56
Ireland 1.28
Luxembourg 1.26
Sweden 1.24
Netherlands 1.23
United Kingdom 1.20
China 1.18
United States 1.17
Australia 1.14
Canada 1.14
South Africa 1.13
Spain 1.10
Denmark 1.09
Germany 1.08
0.00
0.50
1.00
1.50
2.00
2.50
Relative AI Hiring Index

国内就业机会

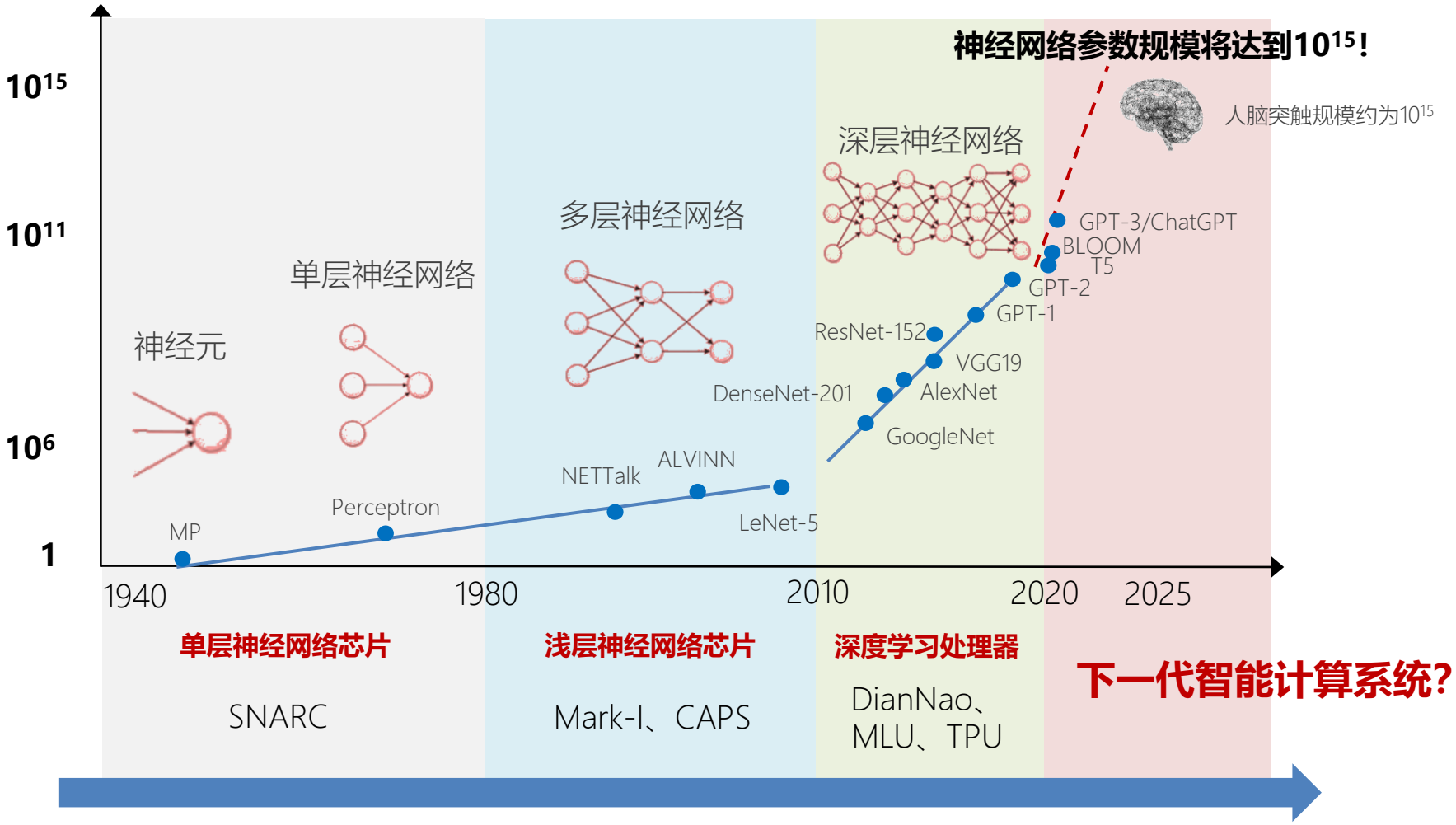




统计时间：2022.1.1 — 2023.8.31 数据来源：脉脉高聘人才智库
人才供需比=岗位供给/人才需求。人才供需比 >1，说明供大于求；<1，说明供给小于需求

AI领域人才紧缺度增加，研发岗位平均薪资涨幅明显

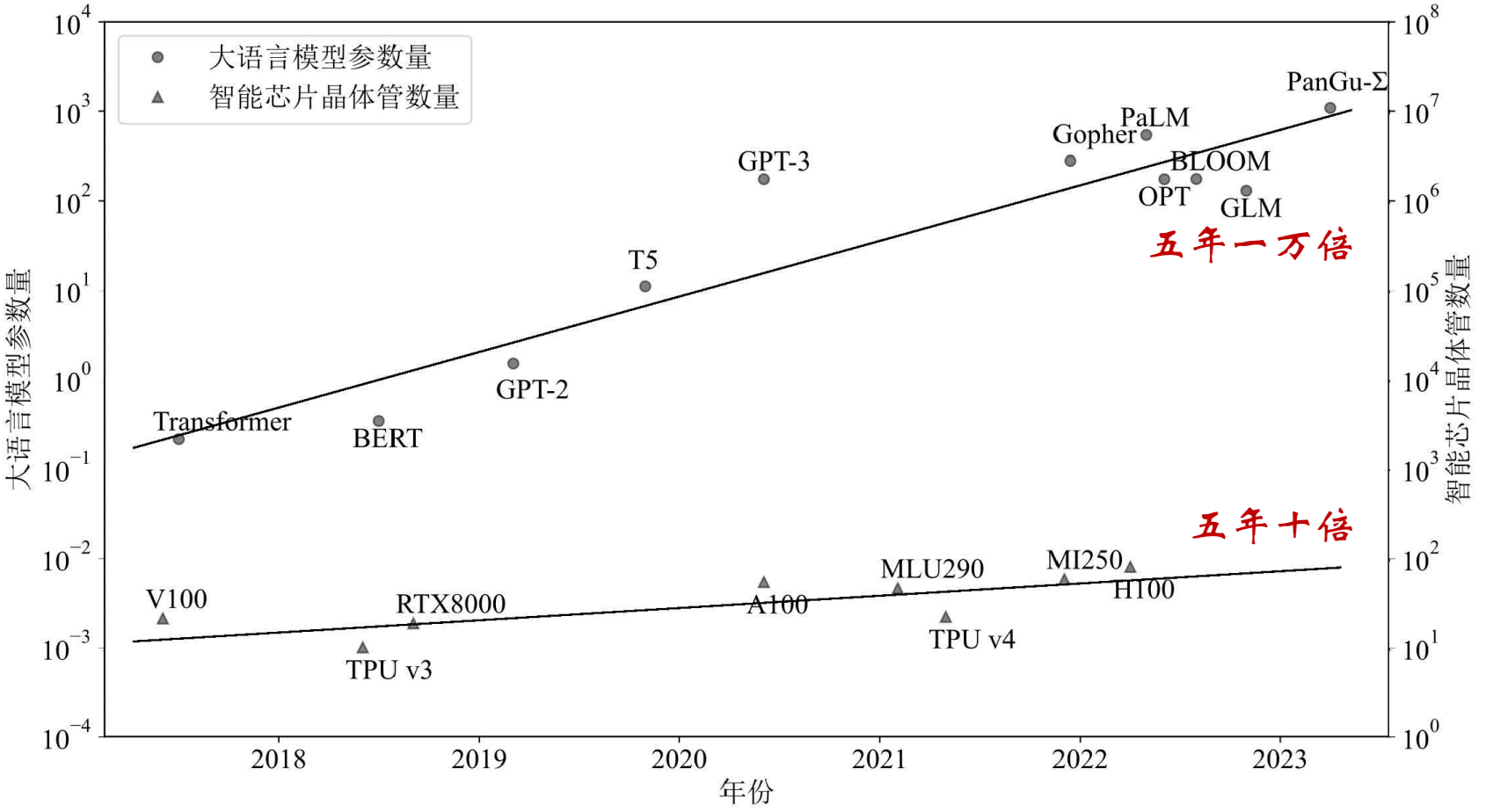
参数规模



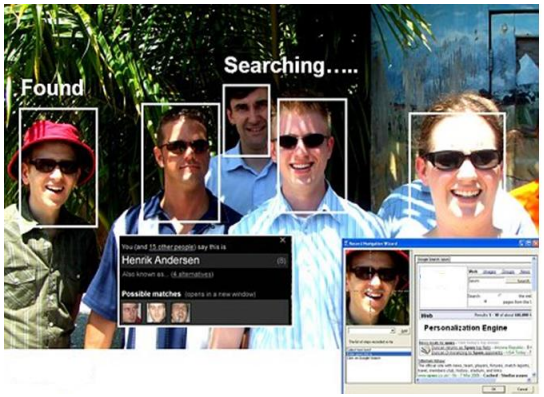
大模型参数规模已达万亿级!

28

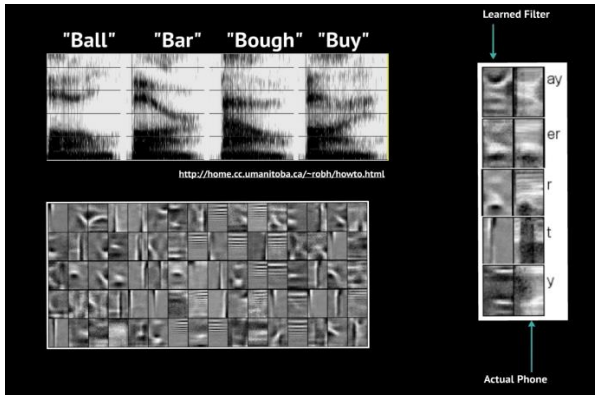
算力需求


大语言模型参数量
智能芯片晶体管数量
PanGu-Σ
PaLM
Gopher
BLOOM
OPT
GLM
GPT-3
T5
GPT-2
Transformer
BERT
五年一万倍
五年十倍
MI250
MLU290
H100
V100
RTX8000
A100
TPU v4
TPU v3
大语言模型参数量
智能芯片晶体管数量
年份
2018
2019
2020
2021
2022
2023

人工智能不断飞速发展



lfw人脸测试准确度99%
(成人仅97%)



语音速记战胜
人类专业速记员

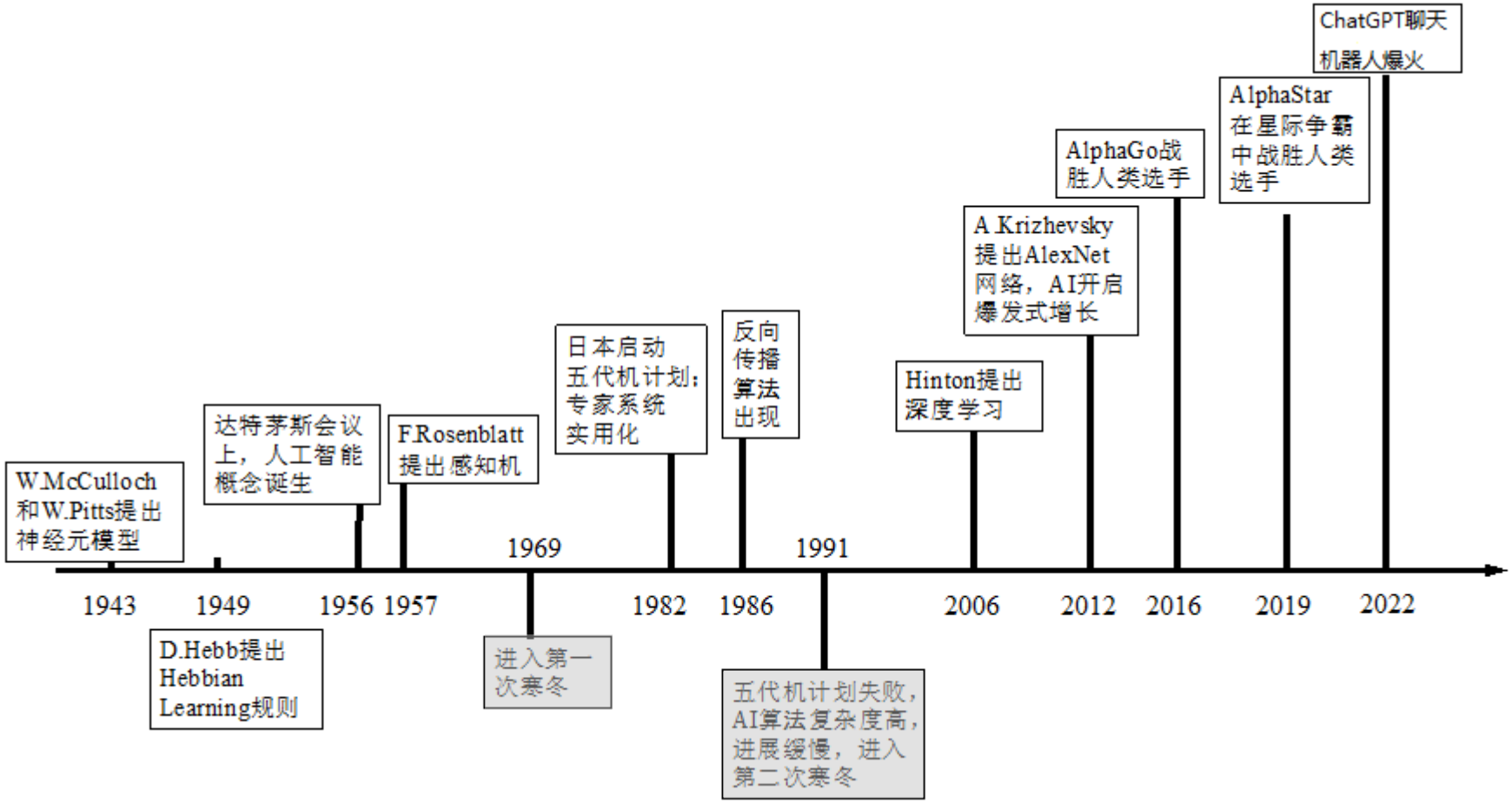
AlphaGo战胜李世石

人工智能算法在多种应用上接近或超过了人类水平

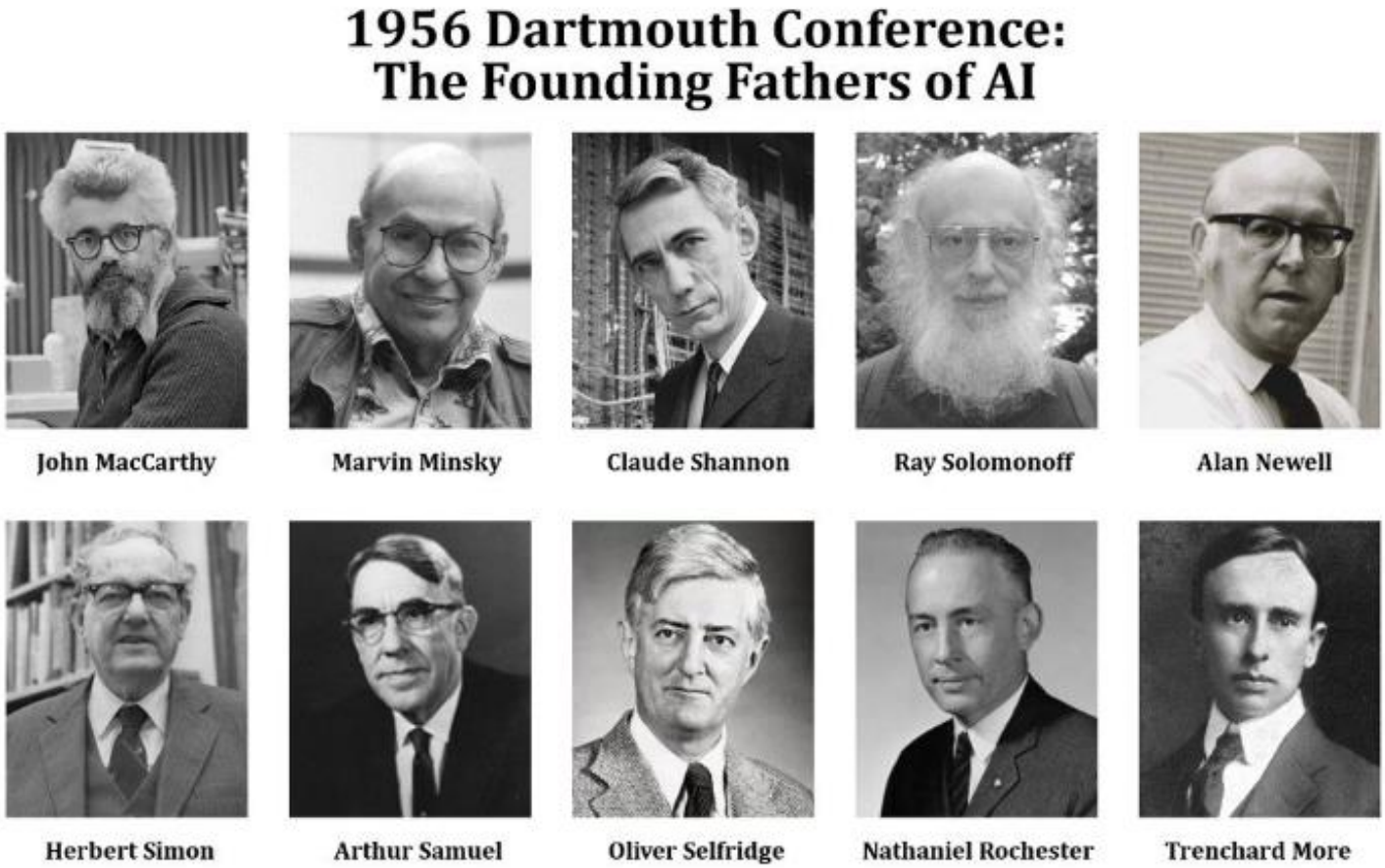
什么是人工智能

- 人工智能：人制造出来的机器所表现出来的智能
- 强人工智能或通用人工智能：具备与人类同等智慧、或超越人类的人工智能，能表现正常人类所具有的所有智能行为
- 弱人工智能：能完成某种特定具体任务的人工智能，计算机科学的非平凡应用

人工智能的三次热潮

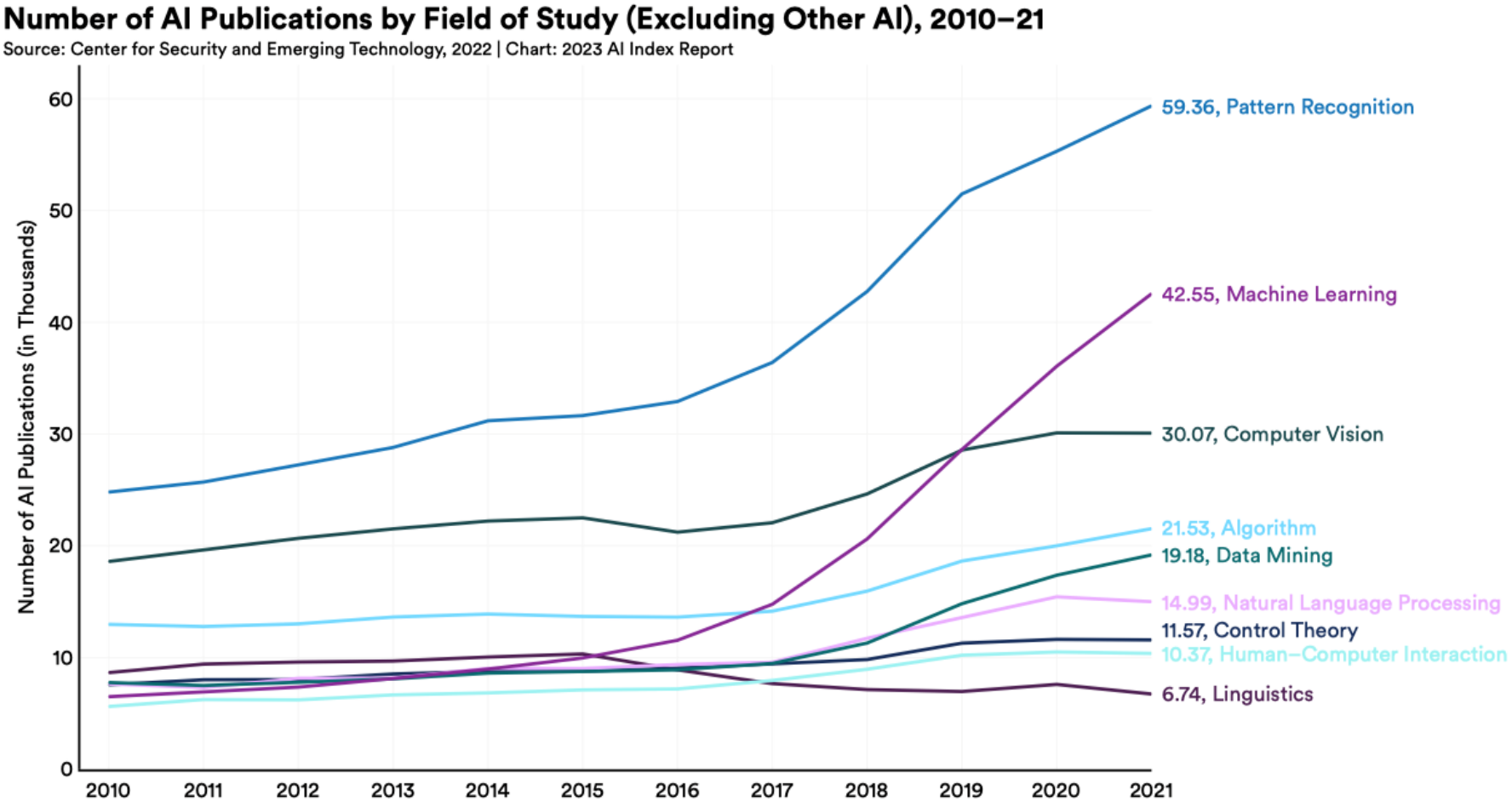

ChatGPT聊天机器人爆火
AlphaStar在星际争霸中战胜人类选手
AlphaGo战胜人类选手
A.Krizhevsky提出AlexNet网络，AI开启爆发式增长
Hinton提出深度学习
反向传播算法出现
日本启动五代机计划：专家系统实用化
F.Rosenblatt提出感知机
达特茅斯会议上，人工智能概念诞生
W.McCulloch和W.Pitts提出神经元模型
1943
1949
1956
1957
1969
1982
1986
1991
2006
2012
2016
2019
2022
D.Hebb提出Hebbian Learning规则
进入第一次寒冬
五代机计划失败，AI算法复杂度高，进展缓慢，进入第二次寒冬

1956年达特茅斯人工智能研讨会



Founding fathers of AI. Courtesy of scienceabc.com

人工智能都在研究什么?


Number of AI Publications by Field of Study (Excluding Other AI), 2010–21
Source: Center for Security and Emerging Technology, 2022 | Chart: 2023 AI Index Report
Number of AI Publications (in Thousands)
0
10
20
30
40
50
60
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
59.36, Pattern Recognition
42.55, Machine Learning
30.07, Computer Vision
21.53, Algorithm
19.18, Data Mining
14.99, Natural Language Processing
11.57, Control Theory
10.37, Human–Computer Interaction
6.74, Linguistics

人工智能三个流派

- 行为主义：基于控制论，构建感知-动作型控制系统
- 符号主义：基于符号逻辑的方法，用逻辑表示知识和求解问题
- 连接主义：基于大脑中神经元细胞连接的计算模型，用人工神经网络来拟合智能行为

符号逻辑的一个例子

$$\neg(\forall x(B(x) \rightarrow P(x)))$$

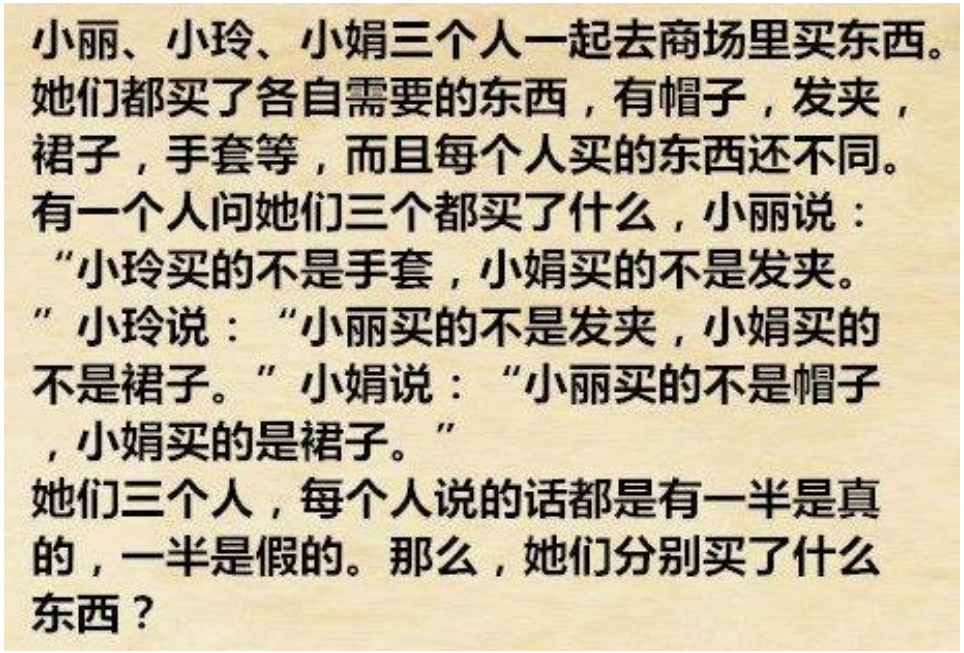
一阶谓词逻辑

符号主义的困难：逻辑，常识，求解器

- 逻辑：未找到能表述世间所有知识的简洁逻辑体系
- 常识：无穷无尽的常识
- 求解器：命题逻辑判定NP完全，一阶谓词逻辑不可判定

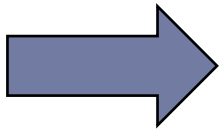
更本质的问题

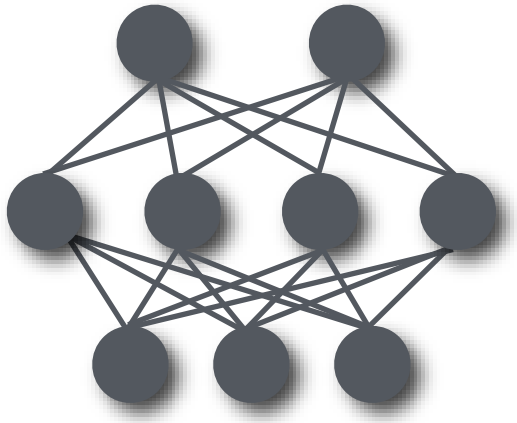
- 人的智能主要是符号智能吗?



- 符号主义最本质的问题是只考虑了理性认识的智能。人类的智能包括感性认识（感知）和理性认识（认知）两个方面
  - 人类语言的例子：词汇，时态，格，数字

连接主义：人工神经网络



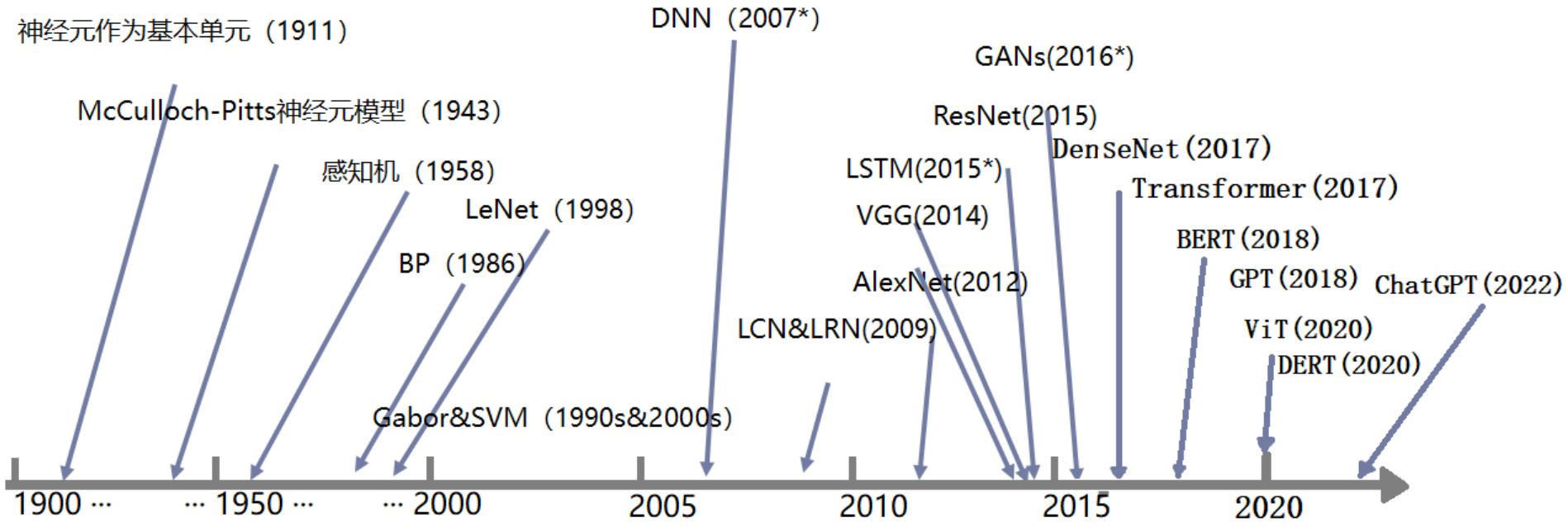


- 1943， 神经元模型， McCulloch 和 Pitts，第一波神经网络
- 1949，《The Organization of Behaviour》，Hebb学习
- 1958，感知机模型（preceptron），Rosenblatt
- 1986，BP反向传播训练方法， Rumelhart、Hinton 和 Williams，第二波神经网络
- 1998，卷积神经网络， Lecun
- 2000，自然语言模型， Bengio
- 2006，深度置信网络（DBN），Hinton, 第三波神经网络
- 2012，AlexNet （Dropout），Hinton团队赢得ImageNet比赛ILSVRC的冠军
- 2015，Deep Residual Network，AlphaGo
- 2016，生成对抗网络，GAN
- 2017, Transformer自然语言处理
- 2018，BERT, GPT自然语言处理
- 2020，ViT, GPT图像处理
- 2022，ChatGPT聊天机器人

2024，Sora根据提示词生成的视频

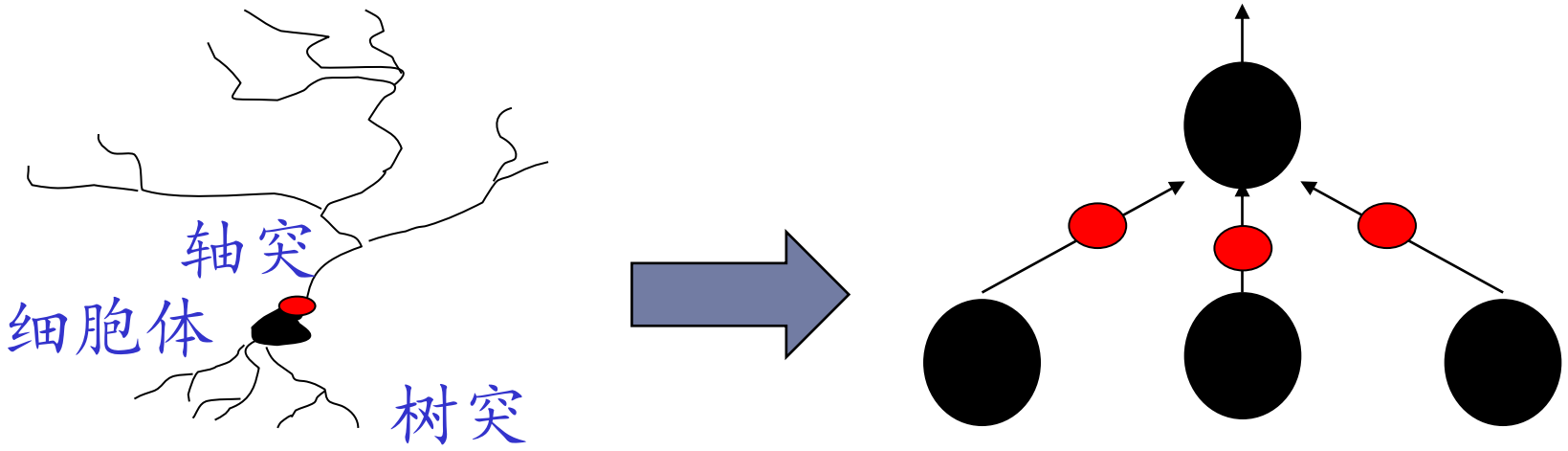
Prompt: Photorealistic closeup video of two pirate ships battling each other as they sail inside a cup of coffee.

人工神经网络

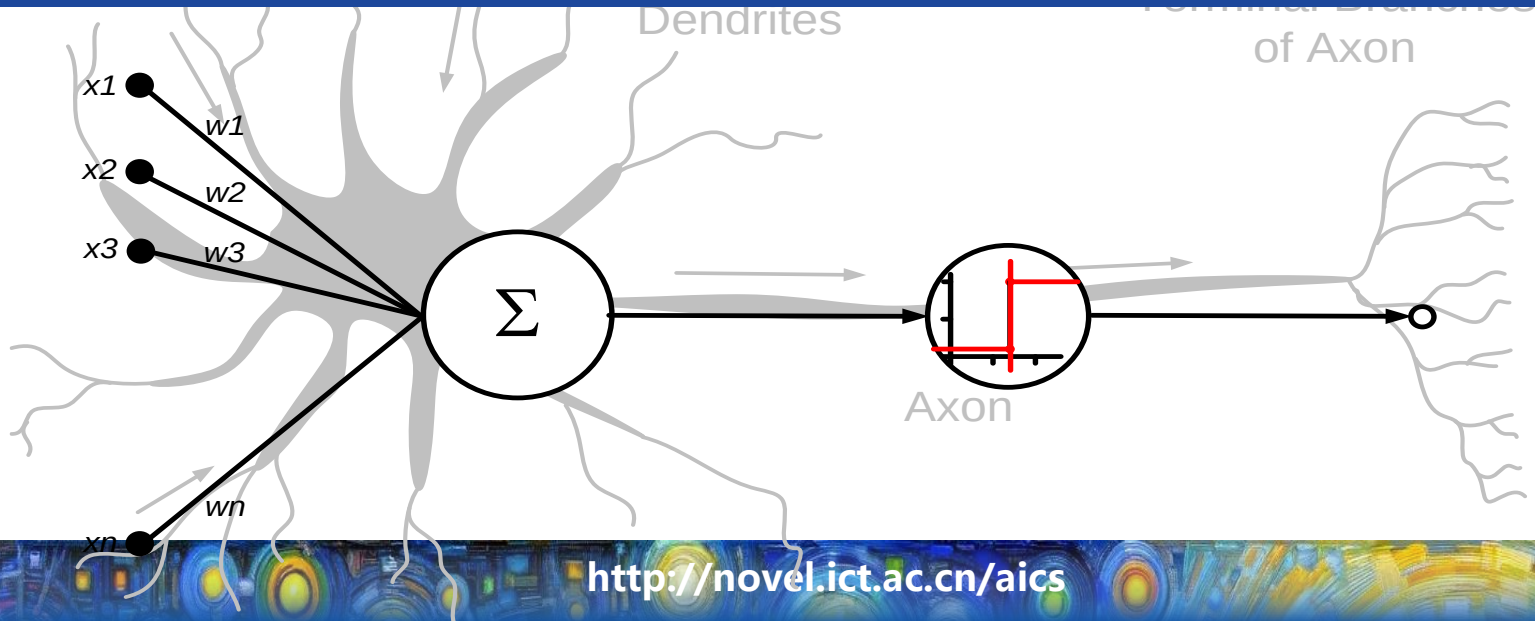


*开始流行时间

人工神经元

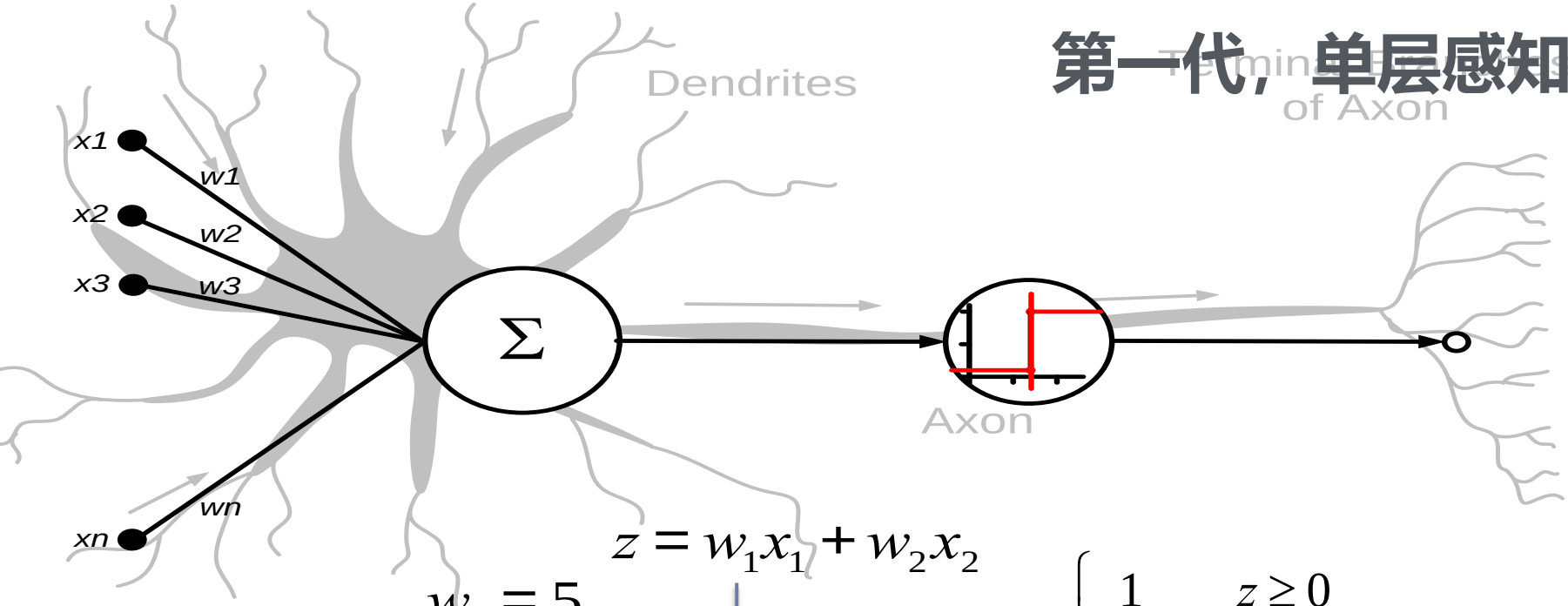


生物神经元：人工神经元=老鼠：米老鼠

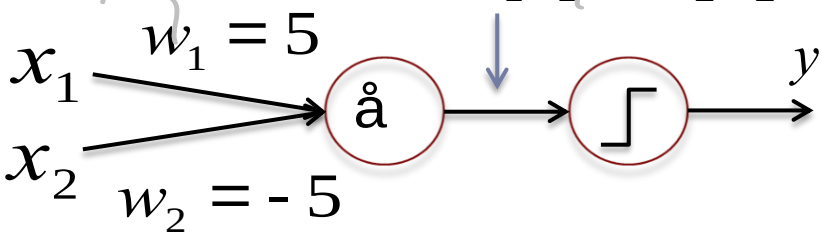


一个神经元的单层感知机

第一代，单层感知机



$$z = w_1 x_1 + w_2 x_2$$



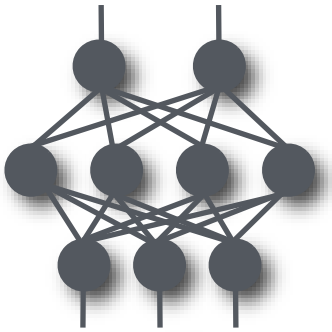
$$y = \begin{cases} 1 & z \geq 0 \\ 0 & otherwise \end{cases}$$

x_1	x_2	z	y
1	-1	10	1
-1	1	-10	0

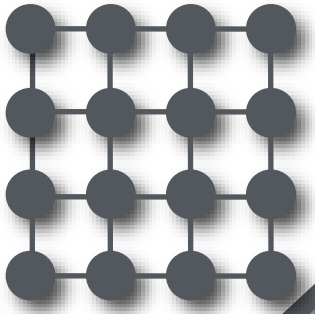
多层+多个神经元

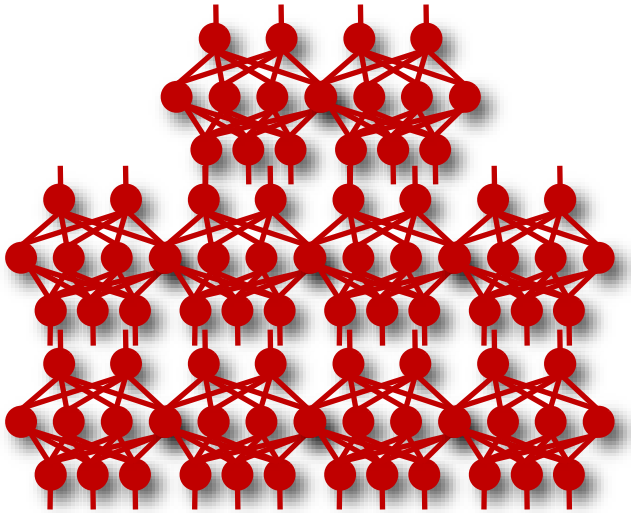
第三代，深度神经网络

第二代，MLP



SVMs

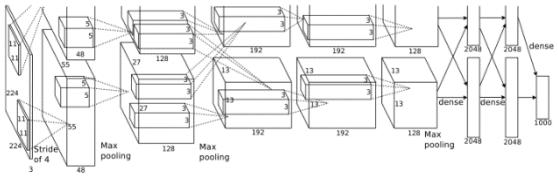




1990s

如今

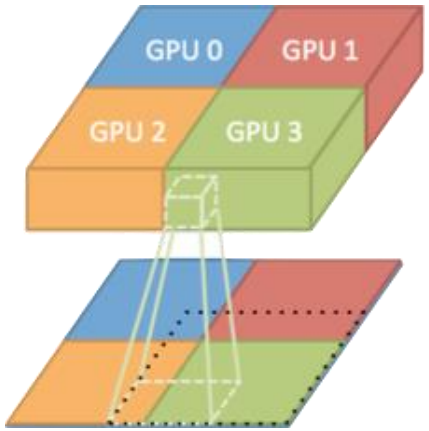
深而大的深度神经网络



6千万参数

十亿参数

ResNet152 层



110亿参数

1750亿

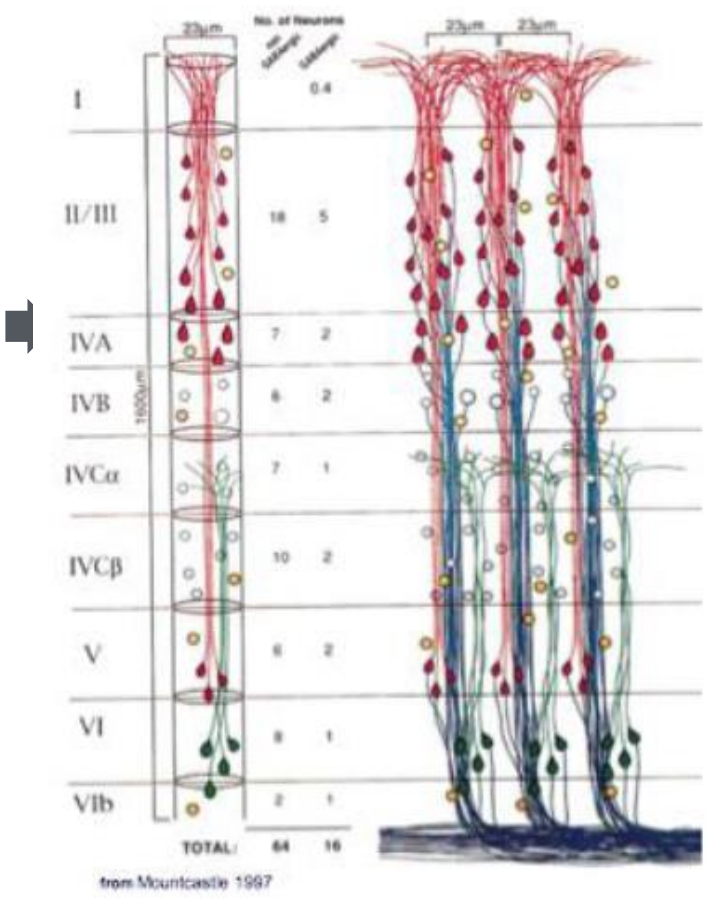
Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems (pp. 1–9).

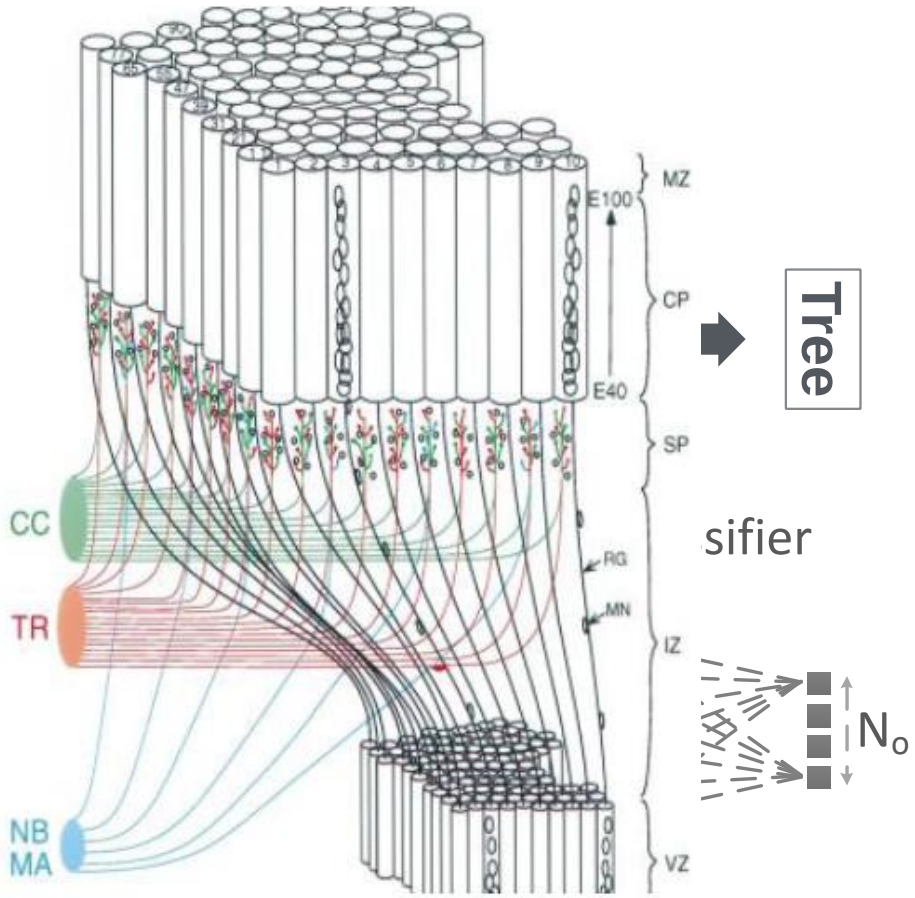
Le, Q. V., Ranzato, M. A., Monga, R., Devin, M., Chen, K., Corrado, G. S., … Ng, A. Y. (2012). Building High-level Features Using Large Scale Unsupervised Learning. In International Conference on Machine Learning.

Coates, A., Huval, B., Wang, T., Wu, D. J., & Ng, A. Y. (2013). Deep learning with cots hpc systems. In International Conference on Machine Learning.

Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.

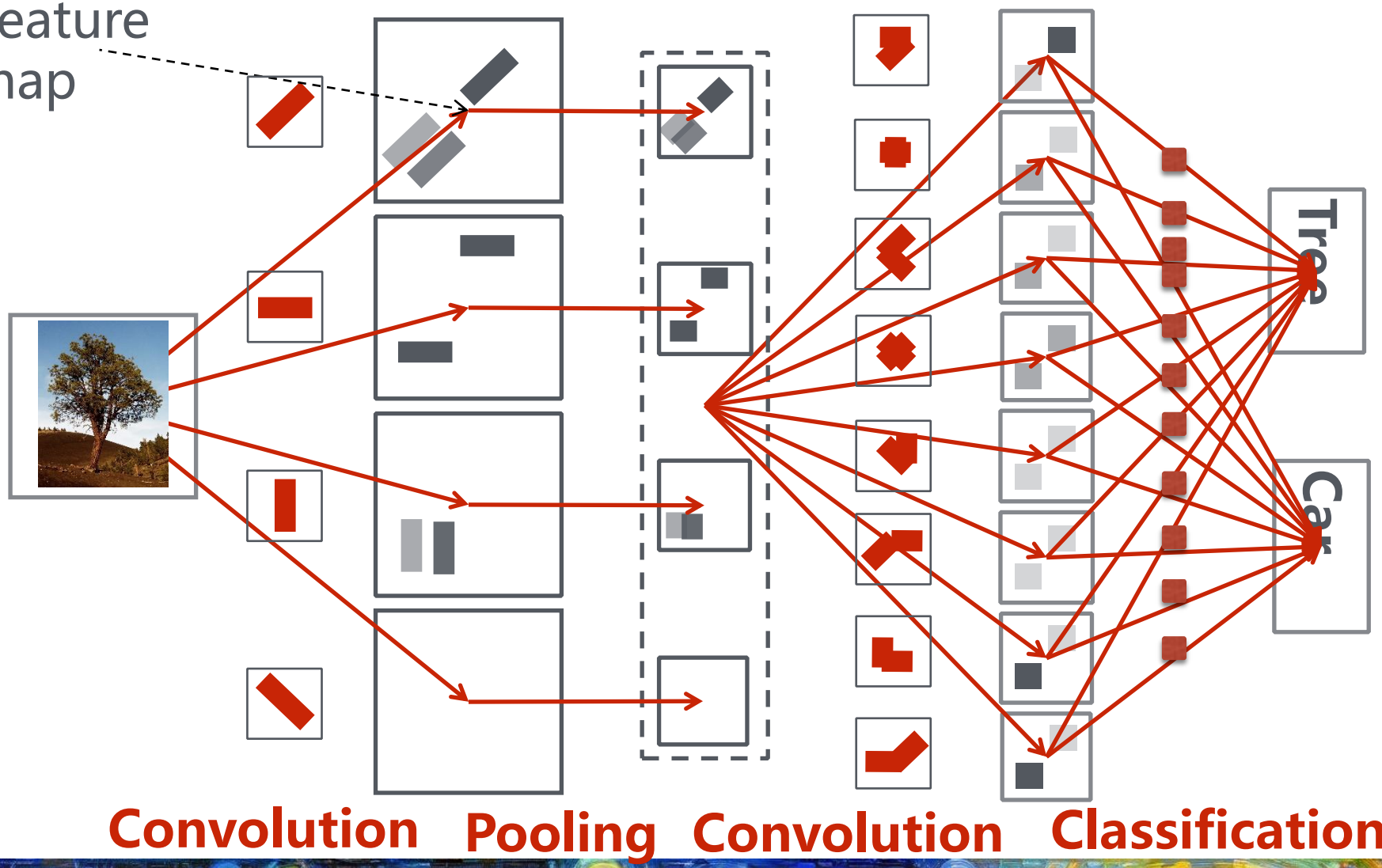
深而大的深度神经网络





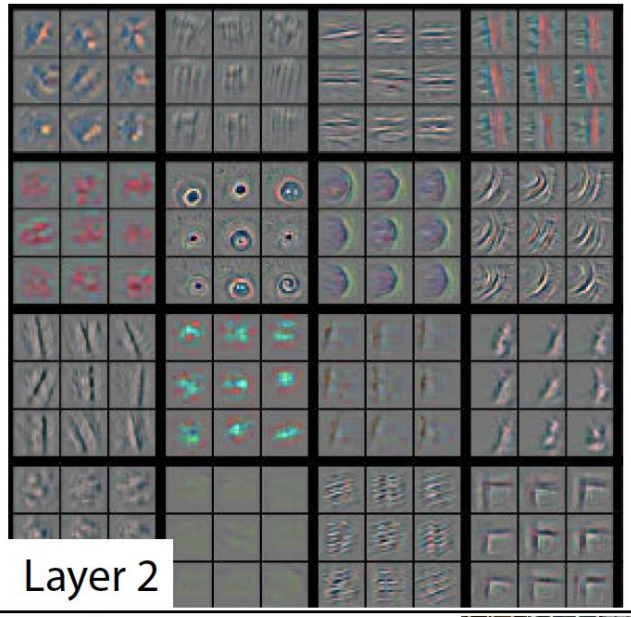
多层大规模人工神经网络

深度神经网络的组织方式

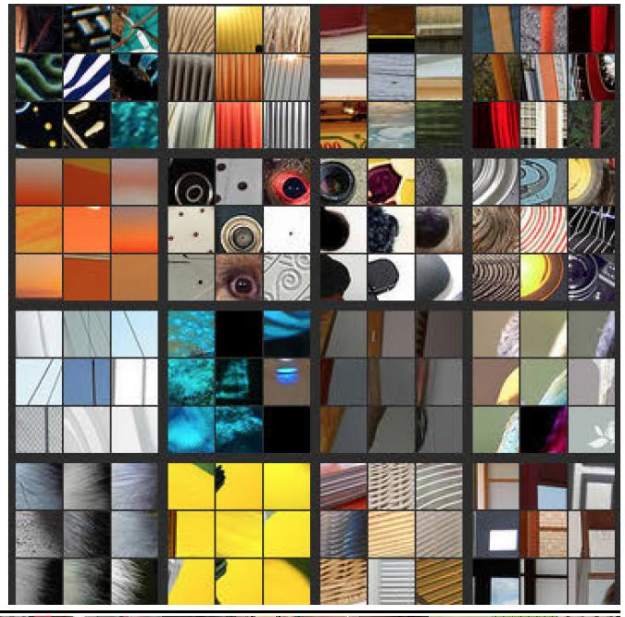


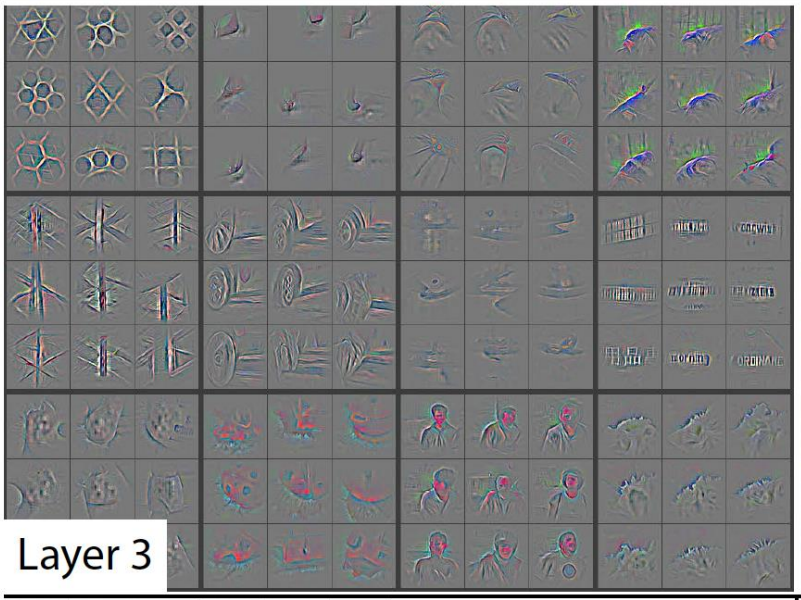
深度学习工作机理

Layer 1



Layer 2

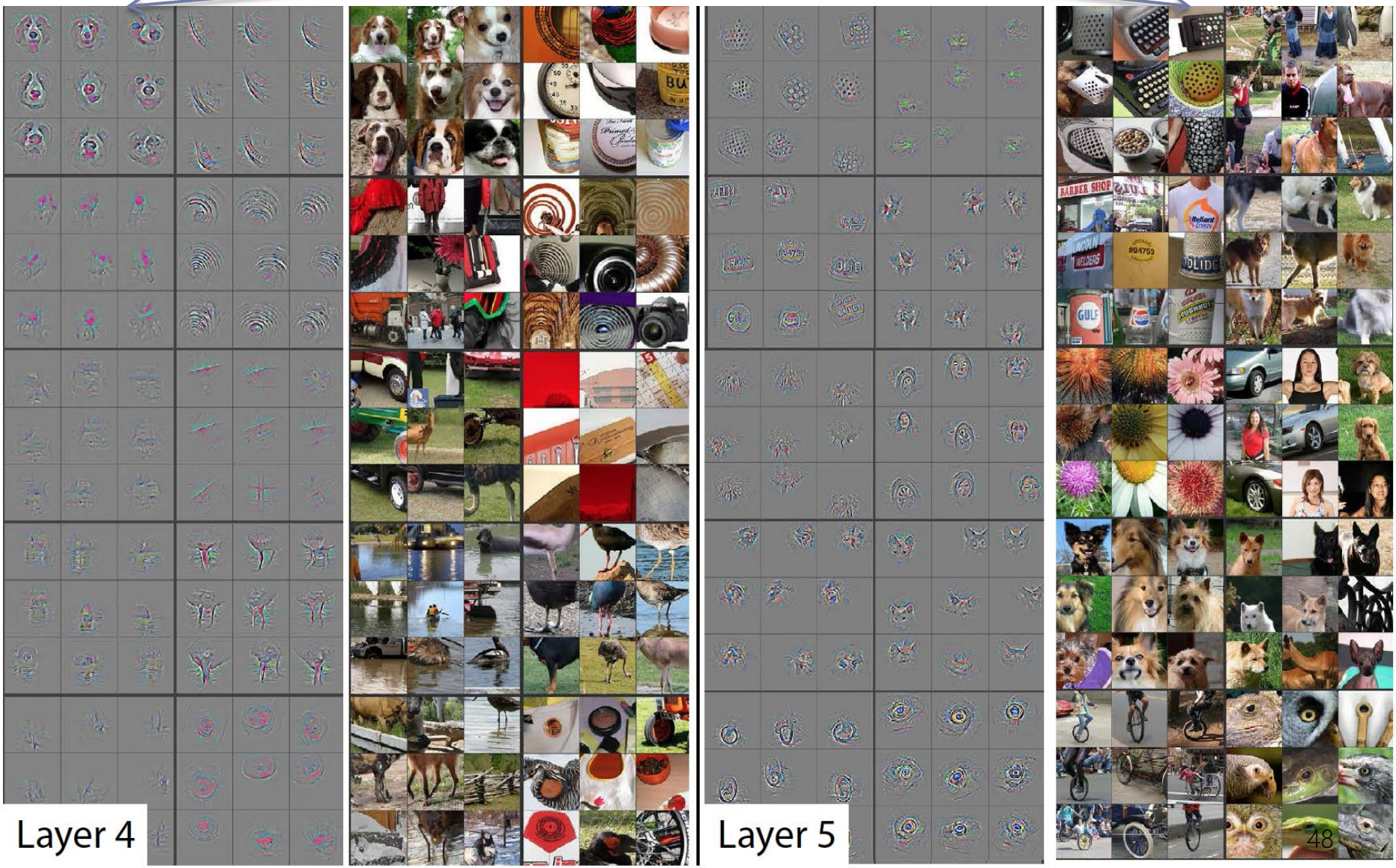




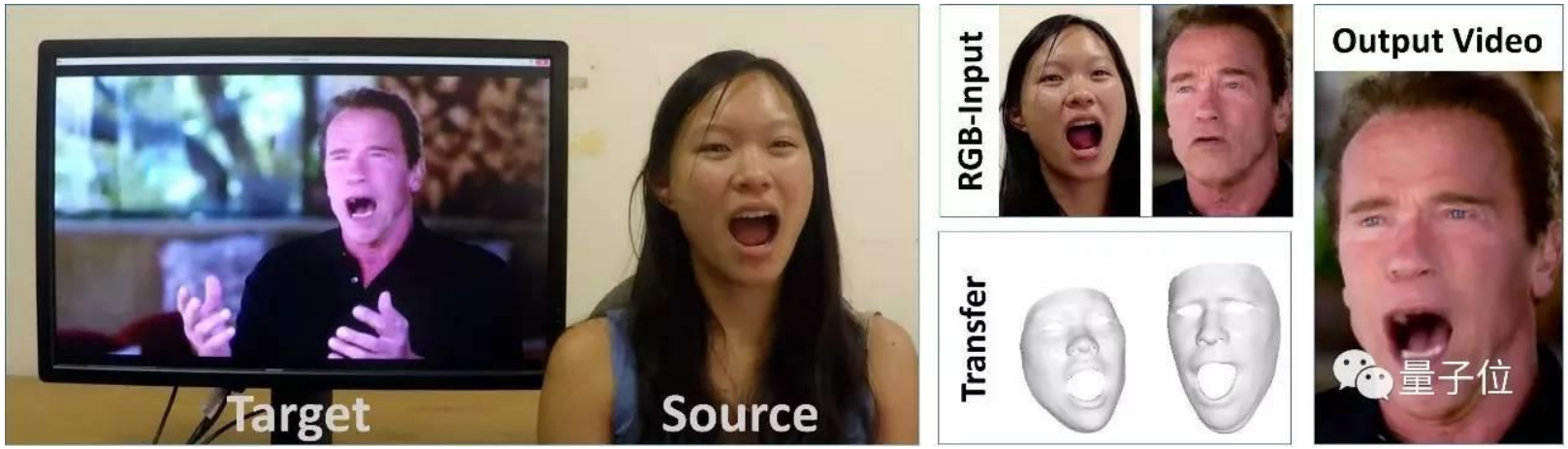
Layer 3

深度学习工作机理

分类器



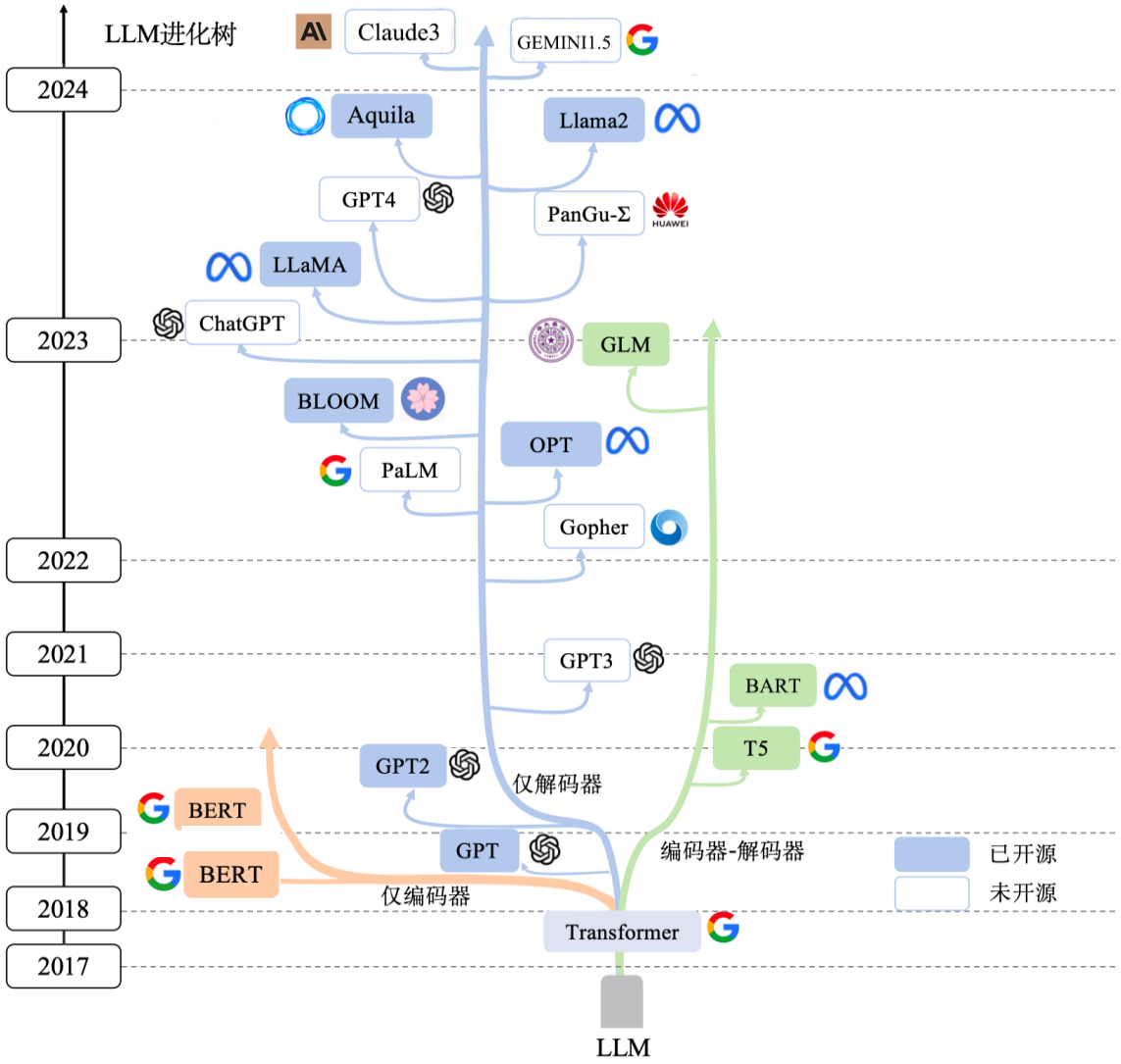
深度学习有意思的小应用



https://www.cnblogs.com/czaoth/p/6755609.html

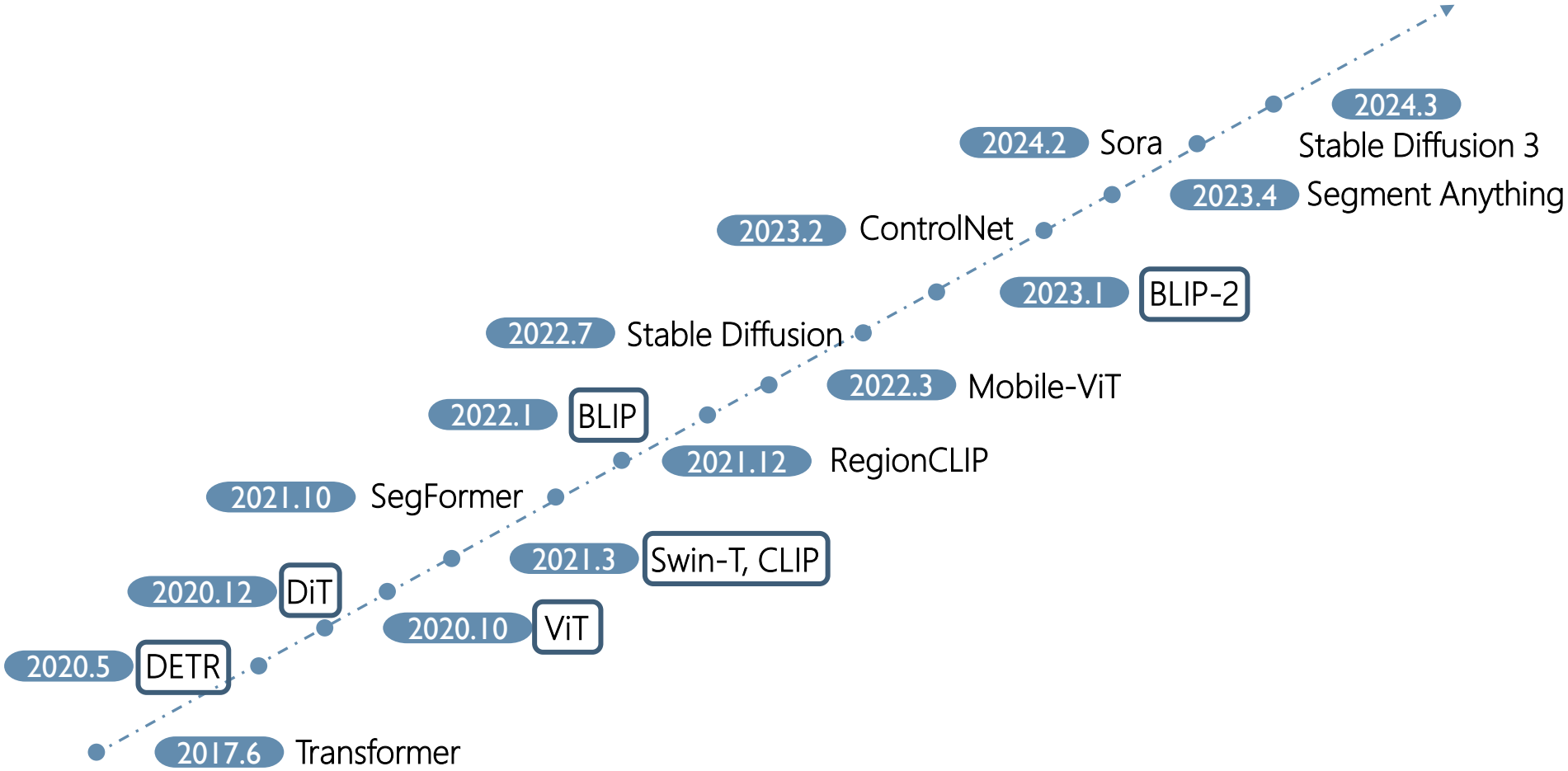
自然语言处理大模型

- 自然语言处理大模型主要分为三个分支：仅编码器、仅解码器以及编码器-解码器。



图像处理大模型

- 随着Transformer的提出，涌现出了一系列模型（包括多模态大模型）用于图像处理。



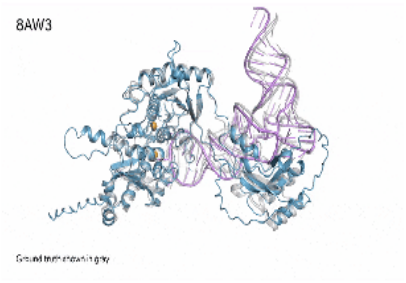
大模型应用

- 以GPT为代表的大模型具有规模性、涌现性、通用性等特征，已成为人工智能的重要发展趋势。大模型已经在数学、化学、医疗健康、国防安全等多个学科和领域展现出了巨大潜力。

数学定理证明

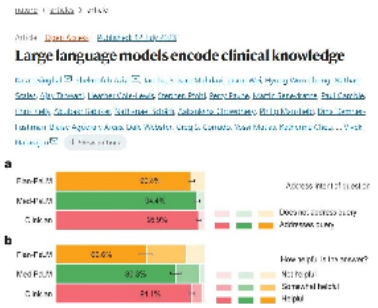
8天内完成了PFR的证明形式化，**实现定理证明效率数百倍提升**，有望推动形成新的科研合作范式

分子结构预测



AlphaFold3成功**预测了几乎所有生命大分子的结构和相互作用**，有望颠覆传统药物研发模式

自动医疗诊断



谷歌医疗大模型Med-PaLM在**美国医疗执照考试中达到“专家”水平**，已在美国梅奥诊所测试运行

军事指挥决策

Palantir使用大模型分析卫星图像、无人机画面和地面情报等，**大幅提升了乌克兰指挥官的决策效率**

深度学习的局限性

- 深度学习是一把梯子，而不是火箭

  - 泛化能力有限
  - 缺乏逻辑推理能力
  - 缺乏可解释性
  - 鲁棒性欠佳

提纲

- 为什么要开这门课
- 为什么来上这门课
- 人工智能
- 智能计算系统
- 驱动范例

什么是智能计算系统

智能计算系统是智能的物质载体

现阶段的智能计算系统通常是集成CPU和智能芯片的异构系统，软件上通常包括一套面向开发者的智能计算编程环境（包括编程框架和编程语言）

异构智能计算系统

- 为什么采用异构智能计算系统?

近十年来通用 CPU 的计算能力增长近乎停滞，而智能计算能力的需求在不断以指数增长，二者形成了剪刀差

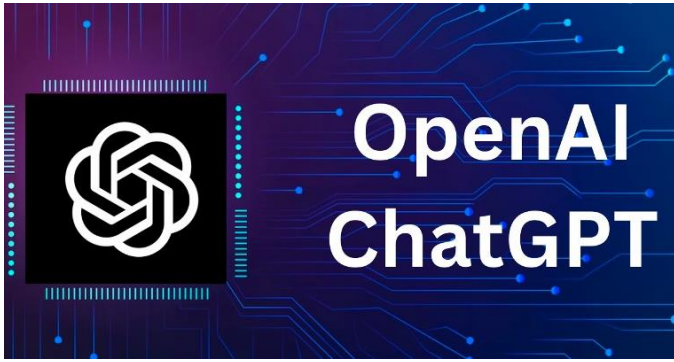
  - 例如，寒武纪深度学习处理器能够以比通用 CPU 低一个数量级的能耗，达到 100 倍以上的智能处理的速度

- 异构系统在提高性能的同时，也带来了编程上的困难
  - 一般会集成一套编程环境，方便程序员快速便捷地开发高能效的智能应用程序
  - 深度学习编程框架包括 PyTorch、TensorFlow 等
  - 深度学习编程语言包括 CUDA 语言和 BCL 语言等

为什么需要智能计算系统

1.6万个CPU核学一周识别猫脸的谷歌大脑

和李世石下一盘围棋电费数千美元的AlphaGo



1万个A100训练1个月的ChatGPT

人工智能必须有其核心物质载体

三代智能计算系统

- 第一代智能计算系统：1980年代，面向符号主义智能处理的专用计算机（Prolog机，LISP机）
- 第二代智能计算系统：2010年代，面向连接主义智能处理的专用计算机（深度学习计算机）
- 第三代智能计算系统：未来强人工智能/通用人工智能的载体

第一代智能计算系统

- 1975，MIT AI Lab的Greenblatt研制成功LISP机CONS
- 1978，MIT AI Lab发布CONS的后继，CADR
- 1980s，发展高峰
  - Symbolics (3600, 3640, XL1200, MacIvory)
  - Lisp Machines Incorporated (LMI Lambda)
  - Texas Instruments (Explorer and MicroExplorer)
  - Xerox (Interlisp-D workstations)
  - 日本，五代机
  - Prolog机，1983，David H. D. Warren Warren Abstract Machine
- 1980s末到1990s初，AI winter，第一代智能机市场坍塌

第一代智能计算系统

LISP机（MIT博物馆）

Symbolics 3640

第一代智能计算系统

- High-level language computer architecture
  - OS的编程语言和硬件“统一”化，如LISP
  - 只针对特定语言的优化
- 局限性
  - 没有太多的实际应用需求
  - 由于摩尔定律发展，性能比不上CPU
    - 贵，几十万美元一台

第二代智能计算系统

- 面向连接主义（深度学习）处理的计算机或处理器
- 第二代智能计算系统的优势
  - 深度学习有大量实际的工业应用，已经形成了产业体系，因此相关研究能得到政府和企业的长期资助
  - 摩尔定律在 21 世纪发展放缓，通用 CPU 性能增长停滞，专用智能计算系统的性能优势越来越大

因此，在可预见的将来，第二代智能计算系统还将长期健壮发展，持续迭代优化

第二代智能计算系统


物端设备


移动设备


客户端


汽车


服务器


超级计算机

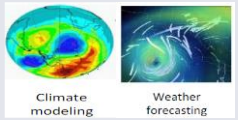

图像识别


语音识别


游戏竞技


自动驾驶

广告推荐

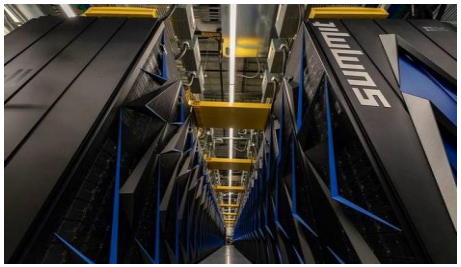


气象预报

TensorFlow caffe dmlc mxnet DRAGON ONNX PyTorch

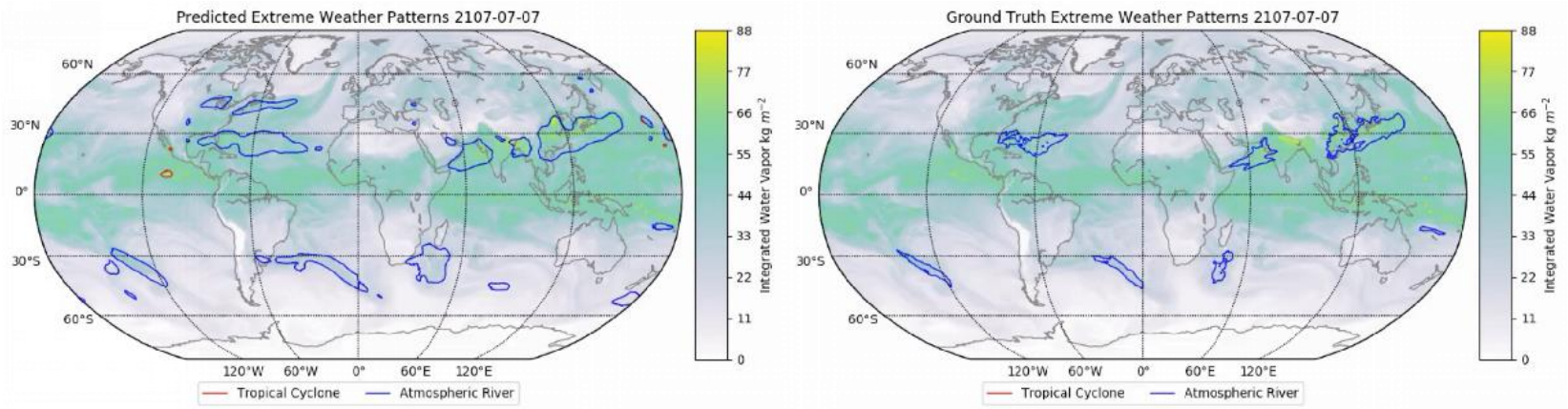
智能计算系统

第二代智能计算系统



美国智能计算系统代表“顶点”（Summit）
浮点运算速度峰值达每秒20亿亿次
（200PFlops）

中国智能计算系统代表“曙光7000”
浮点运算速度峰值达每秒30亿亿次
（300PFlops）



2018年**戈登•贝尔奖**由**劳伦斯伯克利国家实验室**和**NVIDIA公司**的联合研究团队使用**Summit智能计算平台**完成，获奖原因为“Employing **Deep Learning Methods** to Understand Weather Patterns”，该模型使用了混合精度进行训练，峰值算力达到了**1.13Eflops**

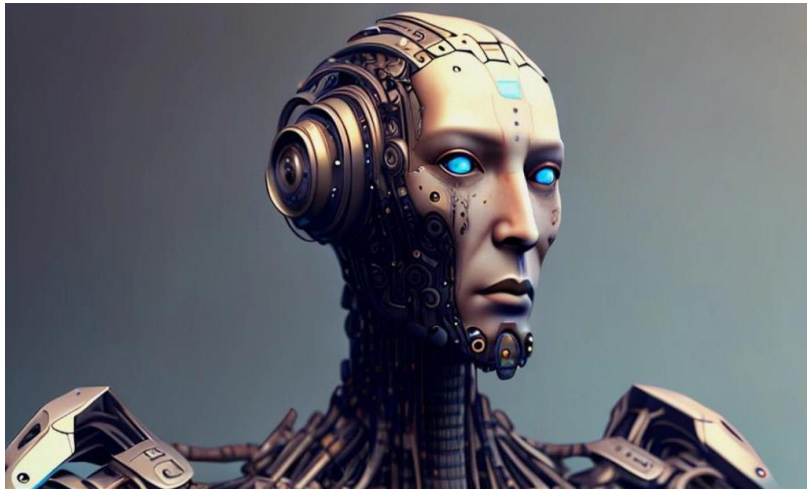
第二代智能计算系统

代表性深度学习处理器/计算机

时间	深度学习处理器/计算机	研制单位	特点
2013 年	DianNao [19]	中科院计算所	国际上首个深度学习处理器架构
2014 年	DaDianNao [20]	中科院计算所	国际上首个多核深度学习处理器架构
	cuDNN（深度学习库）	NVIDIA	升级 GPU 用于深度学习
2015 年	PuDianNao [21]	中科院计算所	国际上首个通用机器学习处理器
	ShiDianNao [22]	中科院计算所	端侧视频图像处理
2016 年	Cambricon [23]	中科院计算所	国际上首个深度学习指令集
	Cambricon-X [24]	中科院计算所	国际上首个稀疏神经网络处理器
2017 年	TPU [25]	Google	基于脉动阵列架构
	FlexFlow [26]	中科院计算所	动态数据流结构
2018 年	TPUv3 cloud	Google	基于 TPUv3 芯片的云计算
	DGX-2 服务器	NVIDIA	16 块 NVIDIA v100 显卡
	Summit 超级计算机	IBM	27684 块 NVIDIA v100 显卡
	MLU100	Cambricon	基于寒武纪云端智能芯片
2019 年	E-RNN [27]	Syracuse 大学	循环神经网络加速器
	Cambricon-F [28]	中科院计算所	分形冯诺依曼架构
	Float-PIM [29]	UCSD	支持训练的存内计算架构
2020 年	Azure	Microsoft	10000 块 NVIDIA 显卡，用于 GPT 系列研发
	DGX A100 Superpod	NVIDIA	140 个节点，1120 块 NVIDIA A100 显卡
2021 年	Frontier	Oak Ridge Leadership Computing Facility	8472 个节点，37888 块 AMD MI250X 加速器
2022 年	DGX H100 服务器	NVIDIA	8 块 NVIDIA H100 显卡
2023 年	DGX GH200	NVIDIA	256 块 NVIDIA Grace Hopper 超级芯片，900 GB/s 卡间互联

第三代智能计算系统展望

- 大模型的发展为第三代智能计算系统的发展提供了一种可能。
- 随着智能计算系统计算能力的逐步增强，深度学习大模型可以变得越来越大，甚至在规模上超过人脑，这将不仅仅是把个别弱人工智能问题做得更好，而是能逐步逼近强人工智能，从而像人一样在各种简单问题上表现出好的效果。
- 若我们能使大模型进一步拥有推理和涌现等高级认知智能，或许强人工智能有可能成为现实。
- 第三代智能计算系统应当具有超强计算能力，从而能涌现出强人工智能的系统。



第三代智能计算系统探索的思路

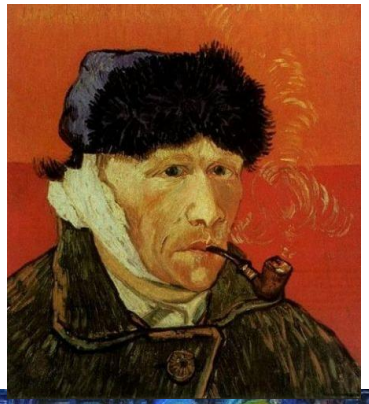
总体思路：具备全面感知能力和超大规模硬件的原始智人是怎样一步步获得智能的？ AI for System and System for AI

- 体系结构：面向海量并发认知智能计算线程和超大规模虚拟环境的计算机和芯片
- 算法：有限延迟的认知智能算法，能自主产生语言和文字，从本能之上建立起自己的知识图谱，打通感知到逻辑的鸿沟
- 编程框架，操作系统，网络等等都将为之巨变

提纲

- 为什么要开这门课
- 为什么来上这门课
- 人工智能
- 智能计算系统
- 驱动范例

如何解决一个AI的任务?

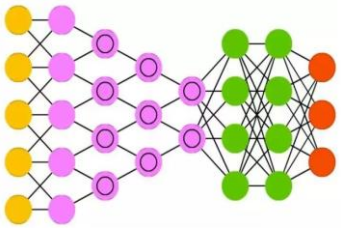


处理过程

输入

输入 ▶ 建模 ▶ 实现 ▶ 运行 ▶ 输出

建模

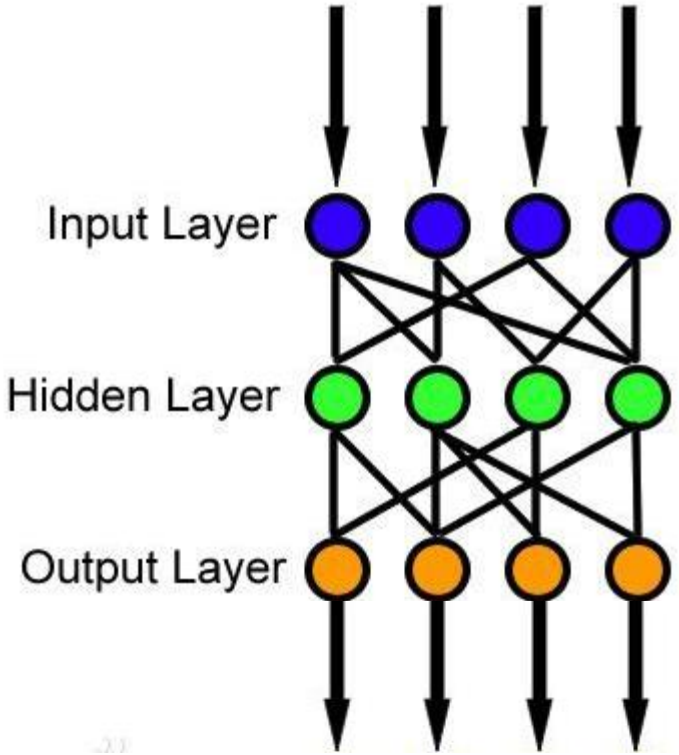


输入 ▶ 建模 ▶ 实现 ▶ 运行 ▶ 输出

深度学习基础

深度学习应用

深度学习基础



颜色 / 纹理/ 形状特征

监督 / 无监督

Sigmoid / ReLU / tanh

Forward Propagation
Back Propagation

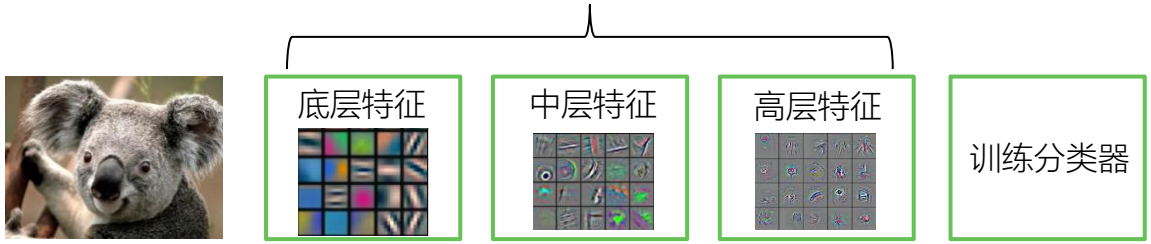
最优化

深度学习

- 传统模式识别

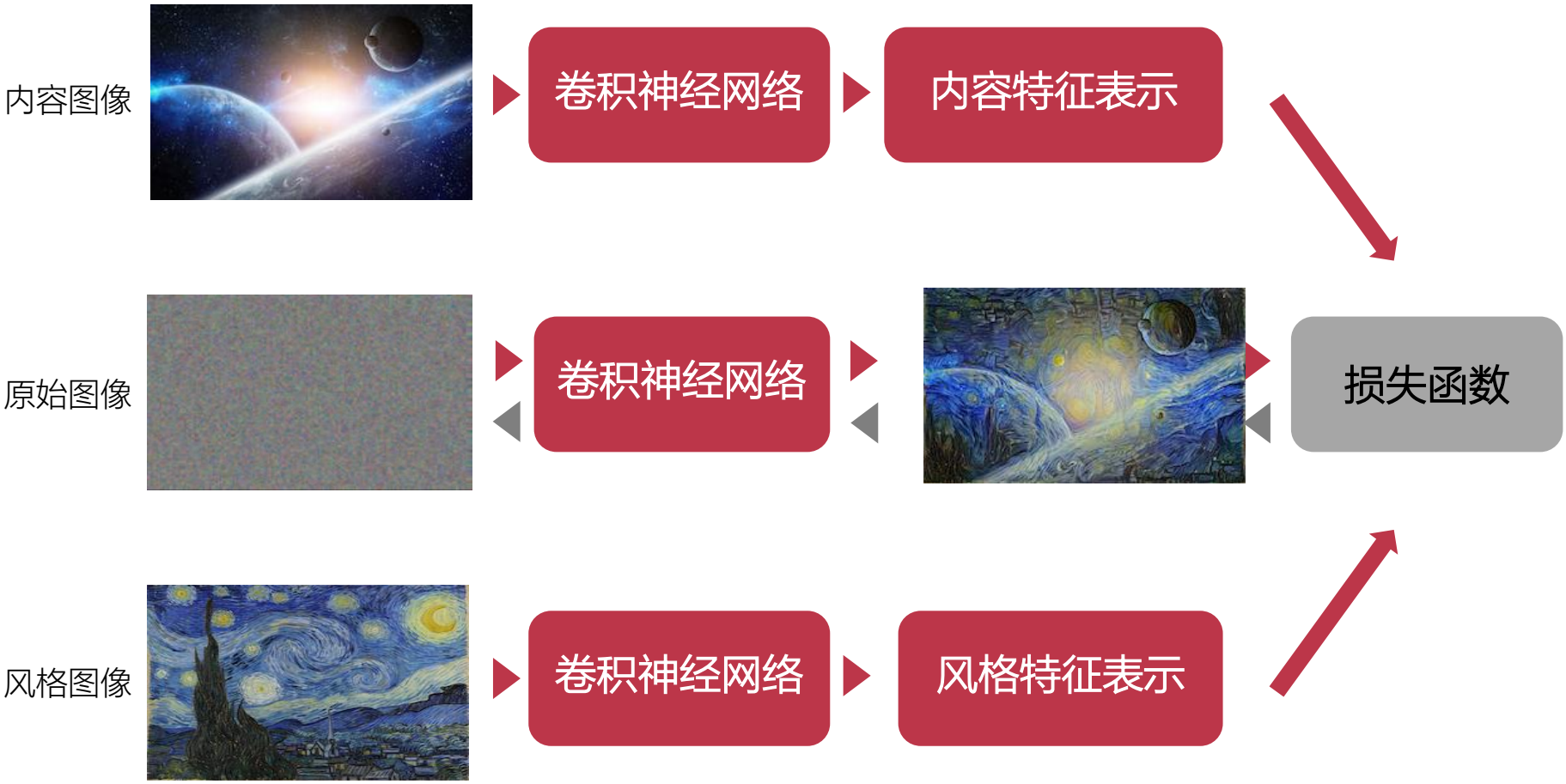
- 深度学习，就是多层人工神经网络

深度学习最重要的作用是**表征学习**，学习层级化的特征，“深度”这词指的就是很多层

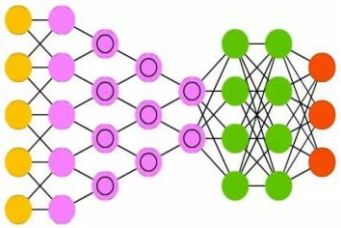


深度学习

- 算法



实现



输入 ▶ 建模 ▶ 实现 ▶ 运行 ▶ 输出

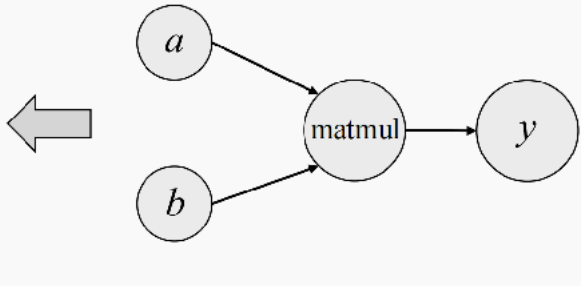
编程框架

智能编程语言

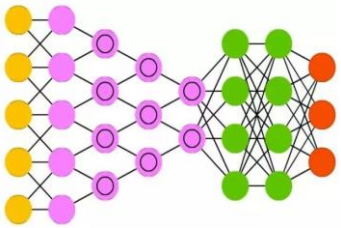
编程框架

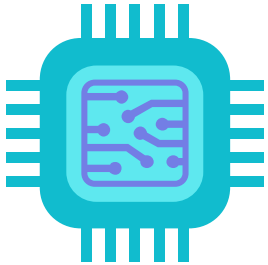
- 将深度学习算法中的基本操作封装成一系列组件，帮助研究人员更简单的实现已有算法，或设计新的算法。这一系列深度学习组件，即构成一套深度学习框架
- 以Pytorch为例


```
import torch

a = torch.randn(1, 2)
b = torch.randn(2, 1)

y = torch.matmul(a, b)

print(y)
```



运行



输入 ▶ 建模 ▶ 实现 ▶ 运行 ▶ 输出


架构原理

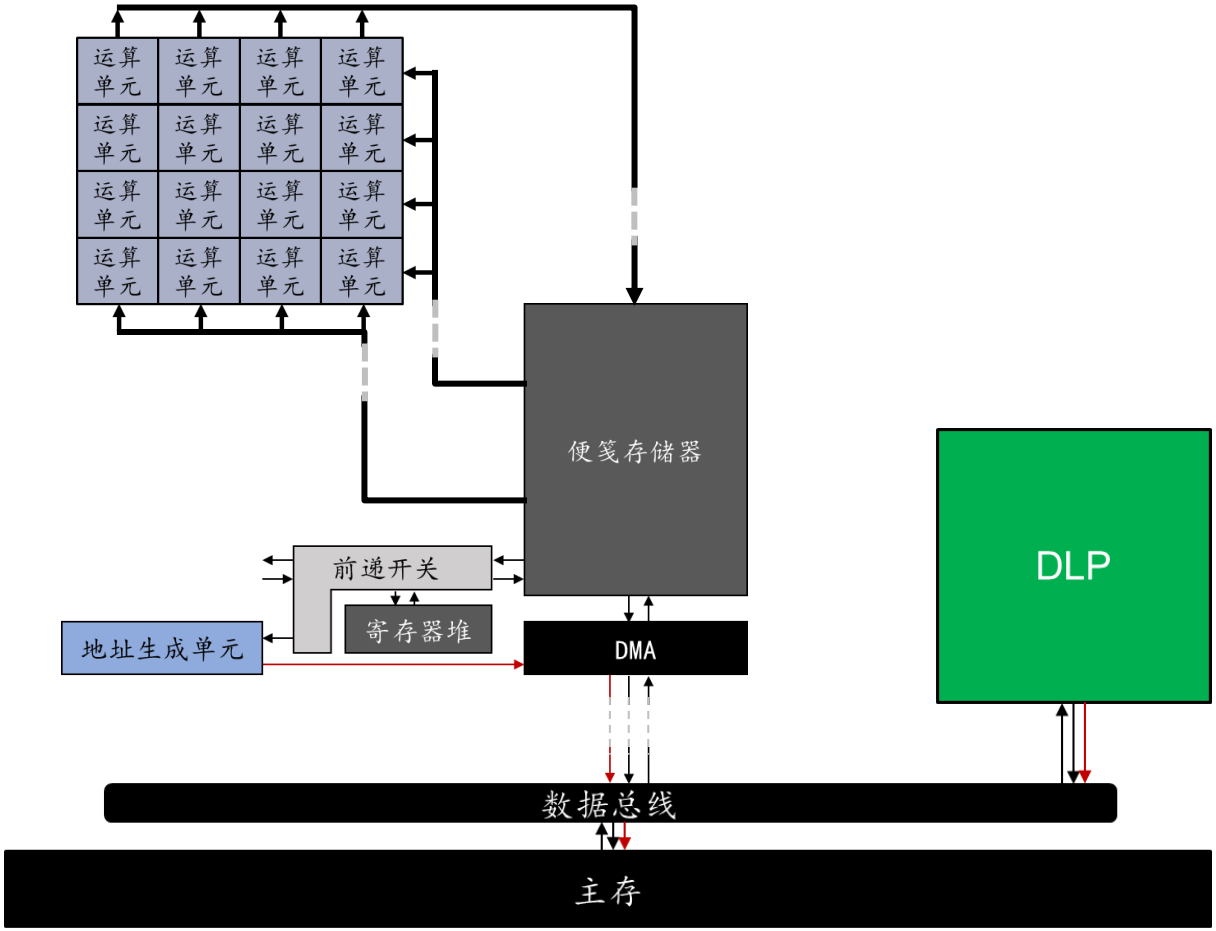

架构设计

架构原理

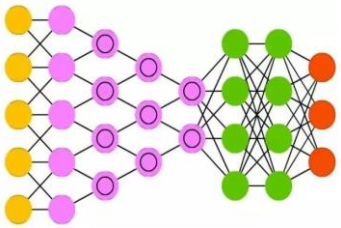
- 人工智能处理器的意义
  - 人工智能算法的新需求：神经网络规模不断增加
  - 通用处理器的适用范围：适用小模型、少量数据、延迟和成本敏感的推理场景
  - 通用处理器的局限性
    - 谷歌大脑（百亿突触）：1.6万CPU核一周训练猫脸识别模型
    - 不可能扩展至人脑规模（百万亿突触）
    - 性能和能耗问题
- 人工智能处理器发展简史
  - 硬件化：计算和访存模式的适配
  - 算法优化：降低存储和计算量
  - 软硬件协同：面向硬件的算法优化

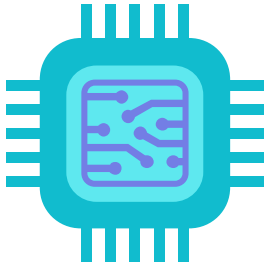
架构设计

- 计算
- 存储
- 通信
- 设计优化



输出




输入 ▶ 建模 ▶ 实现 ▶ 运行 ▶ 输出


运行环境搭建


运行与调试

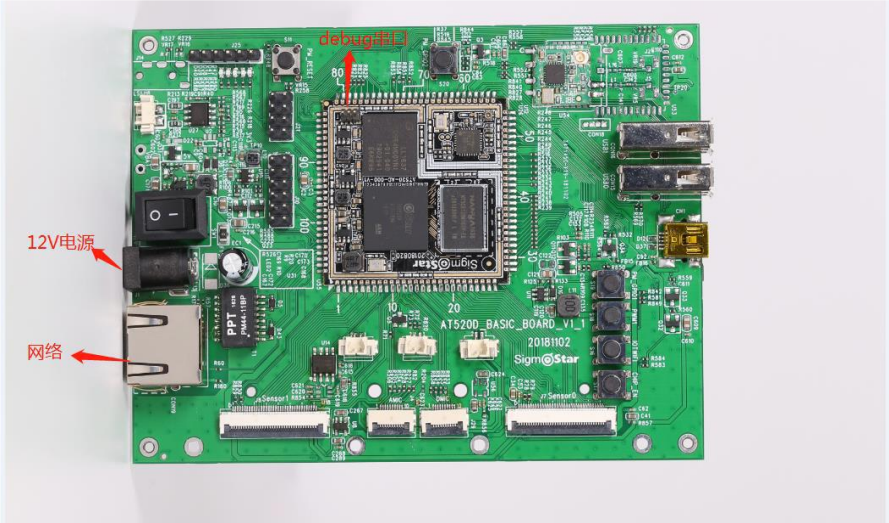

应用与开发

运行环境搭建

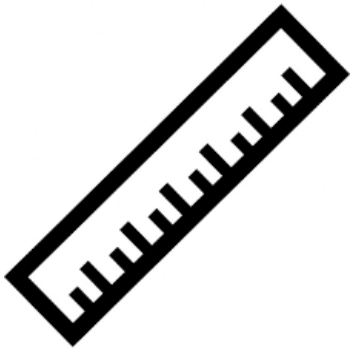
- 硬件环境
  - 采用寒武纪系列智能加速卡
    - 国内首款自主知识产权的智能处理器
    - 支持所有现存的人工智能算法（包括但不限于CNN/DNN/DBN/RNN/LSTM/SOM/RCNN/Faster-RCNN/DeepID/YOLO等）。
    - 相比传统的通用处理器（CPU），能效提升100倍，广泛应用在手机和服务器中

运行与调试

- 代码的开发及编译
  - 串口调试
  - 配置网络文件系统
- 结果测试



模型训练



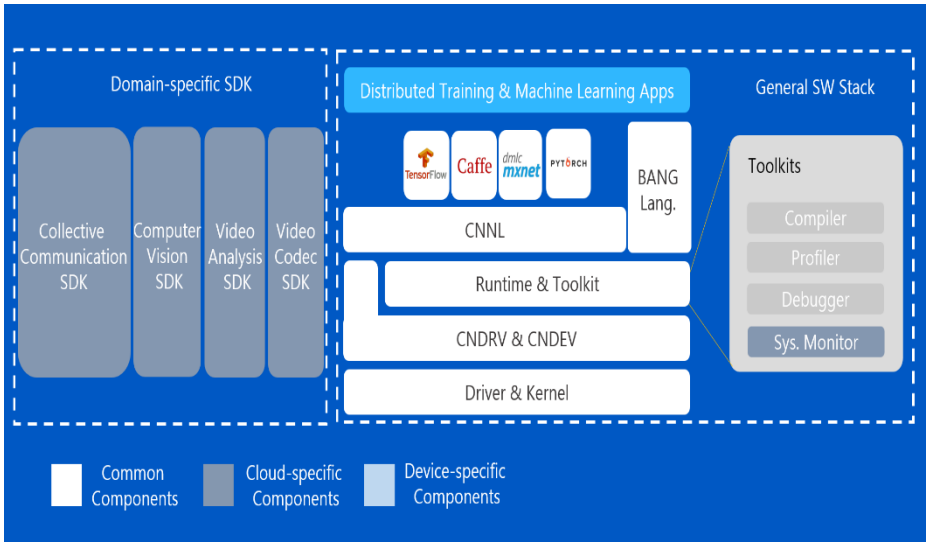
性能剖析

系统监控

应用与开发

- 智能应用依赖库的开发

Cambricon CNNL（寒武纪人工智能计算库）是一个基于寒武纪 MLU 并针对人工智能网络的计算库。Cambricon CNNL 针对人工智能网络应用场景，提供了高度优化的常用算子，同时也为用户提供简洁、高效、通用、灵活并且可扩展的编程接口。



- 数据预处理
- 人工智能网络运行
- 运行结果的后处理

```
./gen_all_models.sh
#编译结果
...
I1218 10:05:30.578752 100149 subnet.hpp:169] subnet[1] fusing...
I1218 10:05:30.578883 100149 fusion.cpp:58] [Fusion] setFusionIO (size: 1, 3)...
I1218 10:05:30.578904 100149 fusion.cpp:45] [Fusion] compiling...0x3ebf220
I1218 10:05:43.869375 100149 net.cpp:426] Offline model generated!
************************ Offline model information BEGIN
**************************
file name : offline_models/zf_faster_rcnn/zf_faster_rcnn.cambricon
model name: offline_models/zf_faster_rcnn/zf_faster_rcnn
model details as follow:
[On CPU] subnet[0] layers : 0(input)
[On MLU] [call via func "subnet0"] subnet[1] layers : 1(conv1) 2(relu1) 3(norm1) 4(pool1) 5(conv2) 6(relu2) 7(norm2) 8(pool2) 9(conv3) 10(relu3) 11(conv4) 12(relu4) 13(conv5) 14(relu5) 15(rpn_conv/3x3) 16(rpn_relu/3x3) 17(rpn_cls_score) 18(rpn_bbox_pred) 19(proposal) 20(roi_pool_conv5) 21(fc6) 22(relu6) 23(drop6) 24(fc7) 25(relu7) 26(drop7) 27(cls_score) 28(bbox_pred) 29(cls_prob)
                    func "subnet0" inputs: data,
                    func "subnet0" outputs: rois, bbox_pred, cls_prob,
************************ Offline model information END
**************************
End zf_faster_rcnn offline models!!!!!
```

生成离线模型

```
执行单个模型测试 3 样例
cd ../mobilenet_v2
./run.sh
##测试结果分析
----------detection for ./jpg/98.jpg----------
0.7651    n01753488 horned viper, cerastes, sand viper, horned asp, Cerastes cornutus
0.1929    n01756291 sidewinder, horned rattlesnake, Crotalus cerastes
0.0406    n01740131 night snake, Hypsiglena torquata
0.0013    n01729322 hognose snake, puff adder, sand viper
0.0006    n01744401 rock python, rock snake, Python sebae
top1 hit num : 62 and 62.6263%        #top1 命中率***
top5 hit num : 81 and 81.8182%        #top5 命中率***
 iter 98 execution time: 94686.0000        #98 照片所用的执行时间***
warining! image ./jpg/99.jpg size is wrong!
input size should be :3 224 224        #输入尺寸大小
now input size is :3 256 256        #实际输入尺寸大小
img is going to resize!        #img 将调整大小
[cnrtInfo]:hardware time: 28917.000000 us        #硬件时间***
after cnrtSyncStream
```

运行应用程序

相关教材

- 陈云霁、李玲、赵永威、李威、郭崎、文渊博、张蕊，智能计算系统——从深度学习到大模型 第2版，机械工业出版社，2024.
- 陈云霁、李玲、李威、郭崎、杜子东，智能计算系统，机械工业出版社，2020.
- 李玲、郭崎、陈云霁等，智能计算系统实验教程，机械工业出版社，2021.

谢谢大家！