

加微信: 1716143665, 领取配套福利课程



038 | 芯片2: 人工智能遇到了什么瓶颈? (1450)



02-19

038 | 芯片2: 人工智能遇到了什么瓶颈?

09分11秒

| 卓克亲述 |

众筹新课联系微信: 1716143665, 你好。

欢迎回到《科技参考》,我是卓克。

今天,我们继续芯片的系列内容,讲讲人工智能发展遇到的瓶颈。

深度学习的巨大优势

这几年,AI 在以我们想象不到的速度渗透进人们的生活和工作中。

就拿面部识别功能来说吧。不知道你有没有担心过:自己最近正在减肥,而且减肥目标是如此的宏大。于是半年不见的朋友再次见到你时,都惊呼差点认不出来了,你简直是换了一个人!

那么问题来了,手机解锁、网上支付,也都是靠面部识别完成的。如果连你的朋友都认不出你来了,那这些软件中的识别功能难道还能认出你吗?

你可能觉得大幅度减肥是一个极端案例,99.9%的人都不会发生这么大的改变。但我们总会长大,总会衰老。一个孩子从12岁到16岁,也许胡子都长出来了,软件还能识别吗?

别担心,这些面部识别程序早就是人工智能控制的了。你每一次对着摄像头解锁的时候,摄像头就会捕捉你面部的最新特征,过往数据并不会消失,而新数据也会被纳入考量。

人工智能凭借这些数据判断出那个人就是你,甚至可以预测十年后你变老了是什么样。

深度学习是人工智能崛起的重要因素,它和传统的人工智能不一样。传统的人工智能和它比起来,可以算是"人工不智能",需要人为的写入很多条规则与逻辑。

而现实中的任务往往是复杂的, 岂是几百上千条规则就能统领一切的?

举一个最早的图像识别的例子吧。如果想训练机器从**X**光片中判断病人是否患了癌症,有两种方法——

在传统方法里,我们需要请放射科和肿瘤学领域的专家来,他们负责制定检查标准,比如在哪些部位容易出现不正常的阴影、纹理是什么样的、怎么区分脂肪组织和癌症组织等等。

我们假设,专家一共列出了 100 个特征,那么此后,这套系统就会根据这 100 个量化后的特征来判断患者的 X 光片。也许这个人工智能系统,能批量处理很多医学影像图,但可惜的是,它并不能取代任何一个专家。

因为总体统计下来,它只能处理最典型的 X 光片,在那 100 个量化特征之外的,或者量化特征不明显,但其实已经是癌症的 X 光片,它是认不出来的。

它的漏诊率、误诊率可能比一个只有两个月实习经验的医生还要高很多,所以不能用它代替任何一个真医生,否则医疗事故会让医院赔死。

但深度学习出现后,这个系统就灵活多了。它会尽可能多的发掘出所有的图像特征。这时, 也许就不是 100 个量化特征了,而是 100 万个。只要能在 X 光下呈现出特点的细节,一个 也不放过。

然后,在指定的数据中训练这个系统,系统会自行判断哪些特征更重要、哪些不太重要。一个使用大规模数据训练好的系统,判断有没有癌症的正确概率,可以轻松超过一个实习医生。也许还不能完全替代有经验的医生,但这个方向上的超越已经近在眼前了。

特征的抓取实在太有用了! 既可以用来判断手机摄像头前的人脸是不是主人, 也可以用来判断汽车行驶途中一闪而过的阴影到底是树荫还是一只猫。

还可以用来分析用户属性,为他们推荐定制化的广告。这类广告的吸引力非常大,往往直击用户的痛点,刚好缺什么就推荐什么。

同样的,它也可以用来给用户呈现定制化的内容,同一个软件,每个人打开,页面内容都不一样。

当然,也可以去挑战各种棋类、扑克比赛,甚至作为一个虚拟的游戏高手在网上和人对战,甚至还能在各种语言翻译中切换自如。

今天,在几乎所有的应用领域,深度学习都显示出了巨大的优越性。

深度学习面临的问题

但最近, MIT 的科学家对 1000 多篇深度学习的论文和数据做了调查后, 却提出了一个悲观的预言——深度学习会随着计算量的限制, 在到达某个性能水平后停滞不前。

他们分析的深度学习涵盖了当前所有的热门领域,包括目标识别、图像分类、语言翻译、问答系统等等。为了让读者更快地了解这个危机,他们举了一个例子:

有一个图像分类的深度学习模型叫 AlexNet。这个模型建立于 2012 年,建立之初,只是一个由两张显卡支撑的系统。当时,只训练了 6 天的时间,它就可以完成表现尚可的图像分类任务,错误率大约 30%。

6 年后, AlexNet 扩充了训练规模,使用了几十张显卡,训练了好几个月,但图像分类任务中的错误率只比之前低了一半,从 30%降低到了 15%。

虽然这个例子比较极端,但也呈现出了深度学习领域性能提升的大致规则——得堆上巨量的算力。比如,在这个例子里,为了获得错误率低一半的成绩,训练用的计算量竟然增加了**1000** 多倍。

当然,增加的计算量也并不都是靠时间和金钱堆出来的,因为毕竟过了6年,芯片的性能还是有很大提高的。

但只可惜,由芯片性能提升而贡献的运算量增加只占不到 1%, 其余都是靠堆更多的显卡、堆更久的训练时间才收获的更低的错误率。

而人们期待的可实用的模型,错误率起码要降低到 5% 以下。MIT 的科学家测算,如果按照今天芯片性能的提升速度,把错误率训练到 5% 将是一场声势浩大的训练。

如果把完成这个训练任务所耗费的能源核算成碳排放,相当于纽约市一个月的排放量。而且,这样的计算量需求、能耗需求和碳排放量,还不仅限于图片分类这一类任务,而是所有类型的任务。

其实,在深度学习领域有这样一条规律——想提升 X 倍的性能,最少需要用 X^2 倍的数据 去训练模型。这就会导致提升性能的代价进一步加高。核算成计算量就是,想要提升 X 倍的性能,需要用 X^2 倍的数据去训练,且这个过程要消耗 X^4 倍的计算量。

如果这个规律中 X = 10 的话,也就是想提升 10 倍性能的话,代价并不是付出 10 倍计算量,而是 1 万倍计算量。

这篇文章里分析了,即便是 10 倍性能提升和 1 万倍计算量的提升,这样失衡的比例关系也仅仅是理论上最优的,现实中,提升 10 倍性能往往要搭上 10 亿倍的运算量。

而这些运算量所消耗的电,是可以直接核算出来的。以今天地球资源的状况看,想把 AlexNet 这个图像分类的模型错误率降低到 1% 以下,应该是没有任何希望了。

我们在去年《科技参考》中曾经讲过一个和去世的未婚妻对话的故事,那个故事背后的技术是一个叫作 GPT-3 的语言处理模型。

开发这个模型的公司 Open AI 自己都没有钱训练自己的娃,后来接受了微软 10 亿美元巨款的资助,又免费用了微软的计算中心的资源,才把这个模型的训练跑完。甚至后来发现了一个 bug 都不能修正,因为训练一遍的开销实在太大,只能积累着,和今后若干个重大的改进

一起进行。

Open AI 还是幸运的,因为它傍上了一个全球科技界排名前 5 的"大佬"微软,否则这辈子都不太可能有机会对 GPT-3 模型做第二次训练。

MIT的科学家预测,人工智能的发展会在不久后遇到瓶颈。

在刚刚出现的前几年,深度学习迅速展现出了算法上的巨大优势,配合今天还能支付得起的算力,把性能推向了新的顶峰。但目前不论是算法还是硬件,人们都已经开发过了,继续发展的唯一方向好像就是找更多的钱和更多的电。而这在全球碳中和的大计划下,必然会遇到阳碍。

所以,找更多钱、更多电这条路肯定走不通,最后还得在算法和硬件上想办法。

当前,虽然还有一些方法能进一步大幅提高算力,比如使用 FPGA 这种效率更高的把程序集成在硬件中的方法做计算。虽然牺牲了通用性,但貌似还是可行的。但问题是,FPGA 之后呢?硬件性能上的提升好像真的没有什么可挖的潜力了。

于是,今后在深度学习领域非常值得关注的就是新的算法,尤其是可以大幅降低训练成本的深度学习新算法。如果算法上没有本质的突破,深度学习下的人工智能的性能会卡在几年后的某一个水平上停滞不前。

好,这就是今天的内容。我是卓克,我们明天再见。

划重点

- 1. 深度学习和传统的人工智能不一样,是近年来人工智能崛起的重要因素。
- 2. 在深度学习领域有这样一条规律:想提升 X 倍的性能,最少需要用 X^2 倍的数据去训练模型,需要完成 X^4 倍的计算量。
- 3. MIT 的科学家预测,深度学习会随着计算量的限制,在到达某个性能水平之后停滞不前。

卓克•科技参考2

每天跟上全球科技新变化

版权归得到App所有,未经许可不得转载



收听更多课程微信: 1716143665



联系微信: 1716143665

0 / 5000

□公开

仅限群内使用! 严禁商业!

默认 最新 只看作者回复



卓克你好,我是一个毕业很多年的文科毕业生。

□ 关注

很久之前你说过普通人自学肯定达不到专业数学家的水平。我当时不服气,还 留过言说要看看能不能通过自学成为数学专家,证明你是错的,后来我花了1 万小时学数学。

这几年我系统学了数学专业的所有课程,也学了数学硕士的不少课程。

有一次我的一个数学微信群里一个教授问了一个问题,他手下的博士生都不会做,问谁能帮他解决的。我想了好几天竟然解决了。他激动的问我是哪个学校的博士,我说我是学文科的,也不是搞数学的。他赞叹不已。

不管怎么样,我发现自己还真的不是学数学的料,这个东西太吃天赋和努力了。 但是虽然没能成为专业数学家,但是也收获了不少数学之美。

展开

21

58

145

□分享

作者 回复:

给您鞠躬,服!



2020-2021年,人们因为疫情的关系在家上班,非接触式行为成为新常态,这也加速了人工智能技术从云端走向边缘。与此同时,由于越来越多云端 AI 的应用,开始追求更高准确度,和更多元更深化的应用,导致了处理运算越来越复杂,还需要用更多记忆体,也更耗能,因此更难在资源受限的 IoT 或嵌入式边缘装置上执行,成了人工智能落实的最大挑战。所以,一种微型机器学习TinyML(Tiny Machine Learning)的崛起,正是为了解决这个难题。

TinyML 是一种超低功耗的边缘机器学习应用模型,它采用谷歌的 TensorFlow Light 框架,配上高度省电设计,还支援全时待命,非常适用于 IoT 或边缘装置。它的硬体配置远低于现今的智慧型手机,譬如以 IoT 装置常用的低阶处理器为例,一颗微处理器(Microcontroller Unit,MCU)通常仅有 256 KB 记忆体和 1MB 的储存空间。

过去一年,企业对边缘 AI 需求持续增长,使得 TinyML 近来发展越来越受瞩目。多家厂商相系推出边缘端 AI 微处理器,譬如 Arm 的 Ethos-U55 微处理器将 TinyML 机器学习能力带进 IoT 硬体中,让 IoT 芯片可以支援 TinyML 应用,不

□关注

仅省电、AI 效能更获得百倍提升。市调机构 ABI Research 预估、采用 TinyML 芯片的边缘 AI 装置出货量将从 2020 的 1 千 5 百万个到 2030 年的 25 亿个。Gartner 调查则指出到了 2027 年,每 10 台边缘装置,就有 6 台具...

選募学习與拉往能力, 上的突破是新一代的神经网路架构 MCUNetV2, 这是 MIT 与 IBM 共同成立的 Watson AI 实验室在 2021 年 10 月 发表的 [1] 。这 篇论文提出一个 patch-based inference 机制,能打破深度神经网路(DNN) 初始层记忆体瓶颈的问题,模型每次仅在小比例区域 (比完整区域小 10 倍) 进 行计算,这可以有效降低峰值记忆体的占用。当完成该部分后,网路的其他具 有较小峰值记忆体的模组则采用常规的 layer-by-layer 方式执行。这个网路...

展开

6

□ 3

140

□分享



人工智能,有多少人工就有多少智能 这里的人工,包含数据和算力两个方面。 □关注

宏观视野来

2020 年全球算力总规模达到 429EFlops (每秒 10 ~ 18 次浮点运算)。

预计未来五年全球算力规模将以超过 50% 的速度增长, 到 2025 年整体规模将 达到 3300EFlops。

我国数据中心规模已达 500 万标准机架, 算力达到 130EFLOPS。

我国保守预计每年仍将以20%以上的速度快速增长。算力已成为国民经济...

居的重要基础设施的人工智能任务的微观视角来看,我搬来最具代表性的 **DeepMind** 斯在整 "东数西算",为解决算力的能耗问题提供着酿点。

2016 年那台打败李世石的 Alpha GO 的计算能力大约在 3.386 PFLOPS (10 ^ 15) 。

据说最少用到了 1202 个 CPU 和 176 个 GPU (显卡) , 外加 100 + 左右的计 算加速卡等等。...

第九票求是由两方面决定,算法复杂度和算法的运行量

算法复杂度,会随着神经网络算法的不断优化而下降,我觉得可以不那么悲观

人工智能训练是需要长时间和大数据,但这是可以积累的过程,某个具体需求的数据不断积累,不断迭代就好了,不需要一蹴而就。

还拿 DeepMind 的 Alpha GO 举例,他们在打败李世石和柯洁之后,继续优化算法,取得的成果是

升级版 Alpha GO Zero, 使用算力降低至十分之一。

- 3 天训练后达到 3500Elo 积分, 超过与李世石对战的 Alpha GO,
- 21 天训练后,达到 4800 Elo 分,超过对战科技的 Alpha Master
- 40 天训练后, 达到 5000Elo 分, 登顶。...

而最近,DeepMind 使用强化学习控制核聚变反应堆内过热的等离子体,宣告成功,这个研究也登上 Nature 杂志。

从这几个例子看,虽然人工智能需要数据,需要时间,但其瓶颈尚可接受,因 为等待和功耗都是值得的

并且这也是在滚动迭代的,人工智能自己算法历史上积累的数据和经验也可以不断传递给后续的人工智能应用。

不怕慢,不要停即可。

展开

4

2

_ 124

□分享



算力的提升可以归纳为两个原因:一个是「底部」的发展,即计算机部件的小型化,其受摩尔定律制约;另一个是「顶部」的发展,也就是软件、算法、硬件架构的发展。在后摩尔定律时代,提升计算性能的方法,虽然「底部」已经没有太多提升的空间,但「顶部」还有机会。

在软件层面

可以通过性能工程(performance engineering)提高软件的效率,改变传统软

□关注

件的开发策略,尽可能缩短软件运行时间,而不是缩短软件开发时间。另外,性能工程还可以根据硬件的情况进行软件定制,如利用并行处理器和矢量单元

在算法层面

在已有算法上的改进是不均匀的,而且具有偶然性,大量算法进展可能来源于新的问题领域、可扩展性问题、根据硬件定制算法。无论如何,算法上的进步相比服从摩尔定理的硬件的进步,能产生更多的效益。

在硬件层面

由于摩尔定律的制约,显然需要改进的是硬件的架构,主要问题就是如何简化处理器和利用应用程序的并行性。通过简化处理器,可以将复杂的处理核替换为晶体管数量需求更少的简单处理核。由此释放出的晶体管预算可重新分配到其他用途上,比如增加并行运行的处理核的数量,这将大幅提升可利用并行性问题的效率。

展开

4

5

105

□分享



正如人工智能领域顶级专家盖瑞·马库斯所指出的那样,关于人工智能的炒作总是甚嚣尘上,但要得到真正可信的 AI,却远比想象的要复杂得多,超级智能时代还远没有到来。创造真正可信的 AI 需要赋予机器常识和深度理解,而不是简单地分析数据。

深度学习作为人工智能的核心技术,关系到人工智能技术未来的发展,只有突破深度学习当前所面临的一系列瓶颈,克服其局限性,才能创造出真正可信的人工智能。

事实上,深度学习是一个好的开始,但是不能解决一切问题,它未来的发展,需要和更高维度的方式叠加。而跨学科的研究会在语义、知识图谱、机器记忆、想象、逻辑推理等类人脑的领域,弥补深度学习的一些短板。

展开

□关注

□2 □28 □分享 康师傅当学徒(♂) 22-19 算法这个东西让我想到了大刘《三体》中那个锁死人类科技的智子,而人工智能是未来科技发展的方向与趋势,人类要想实现真正的人工智能,算法也是我们必须要跨过的砍。	± □
□ 28 □ 分享 康师傅当学徒(♂) ②2-19 算法这个东西让我想到了大刘《三体》中那个锁死人类科技的智子,而人工智 能是未来科技发展的方向与趋势,人类要想实现真正的人工智能,算法也是我 们必须要跨过的砍。	± □
□分享 康师傅当学徒(♂) \$2-19 算法这个东西让我想到了大刘《三体》中那个锁死人类科技的智子,而人工智能是未来科技发展的方向与趋势,人类要想实现真正的人工智能,算法也是我们必须要跨过的砍。	± □
康师傅当学徒(♂) ②2-19 算法这个东西让我想到了大刘《三体》中那个锁死人类科技的智子,而人工智 能是未来科技发展的方向与趋势,人类要想实现真正的人工智能,算法也是我 们必须要跨过的砍。	<u> </u>
第法这个东西让我想到了大刘《三体》中那个锁死人类科技的智子,而人工智 章法这个东西让我想到了大刘《三体》中那个锁死人类科技的智子,而人工智能是未来科技发展的方向与趋势,人类要想实现真正的人工智能,算法也是我们必须要跨过的砍。	± □
能是未来科技发展的方向与趋势,人类要想实现真正的人工智能,算法也是我们必须要跨过的砍。	± □
们必须要跨过的砍。	
☐ 27	
王木子	
92-19	
个人猜想,比人工智能耗能低而且性能还高的,那应该就是智能生物。但如果	± □
弄不好,人类很可能驾驭不了。	
# 我爱问卓克	
□ 1 □ 1	
□ 1□ 1□ 24	
□ 1□ 24	
□ 1□ 24	
□ 1□ 24	

深度学习原来已经有了天花板,没想到会这么快,喂养数据,消耗能量都是指 ^{□ 关注} □ 数级别的,,原先看到的只是深度学习的前身,永远都要记住,事情是一个整





我找到了原文:

□关注

The computational limits of deep learning arxiv.org/pdf/2007.05558.pdf

 $\prod 1$

_ 2

_ 20

□分享



卓克老师这讲说到人工智能的算力消耗,这不但是个真问题,还是一个老问题 ^{□ 关注} □

1997年,导师安排我做某钢厂冶炼工艺人工智能项目。现在看就是 6 输入、8 输出、4 层结构的 BP 神经网络,数据样本 500 左右。

当时为了调节网络节点和初始条件,就花了1个多月。为什么呢?算力不够。

为了提升计算速度,首先就是优化算法和计算计划,以免计算长期不收敛;其次是选用计算更快的 C 语言,为此还舍弃了较为熟悉的 BACSIC 和FORTUNE;最后就是硬件 -- 一台满配 386,这可是中关村能装能找到的最高配电脑。

就这样,计算一次也需要一整天。为了防止计算机被误动中断运算,就在键盘上和屏幕上贴上提示,"正在运算,不要动键盘和鼠标!"

这样规模的神经网络模型,现在可以通过调用函数,迅速得到结果。然而,没有想到,现在的算力依然面临挑战,以至于我们提出了"东数西算"的大布局。

人类的需求真是无穷的啊,算力也不能幸免。

展开

 $\prod 1$

□评论

17

□分享



新算法和技术的追求意味着学科上的基础理论必须先行,这又是另一个需要时间沉淀才可能看到的"涌现"吧~

□关注

此外,对算力的追求让我想起巴别塔的寓言,"无论通天塔建成之后究竟会给人类带来什么,建造通天塔的事业本身就是一场灾难,而且必定会因为一场更大的灾难而仓皇中辍",人类还有许多要努力的雄心目标,人类也要有与之配套的敬畏之心~

 \Box 1

□评论

17

□分享



卓老板提到的:深度学习规律 —— 想提升 X 倍的性能,最少需要用 X^2 倍的 ^{□ 关注} 数据去训练模型",且这个过程要消耗 X^4 倍的计算量。

我的理解如下4点,请卓老板帮忙点评一下是否正确。

①寻找更准确的模型类似于一个 N 维空间寻找最优函数,如果能够从 N+1 维空间来看这个模型,准确度会更高。模型训练的过程相当于不断提高空间维度

。… 展开

 \Box 1

_ 1

_ 16

□分享

作者 回复:

分析的到位



从来没有一劳永逸的办法。

□关注

□关注

虽然深度学习一度霸占了 AI 行业的先锋,但随着输入产出比不再划算,边际 效应递减不期而遇。

所以,从来不要在一条道路上走到黑。...

展开

 \Box 1

□评论

14

□分享



这真是盆让人失望的冷水。不过想想人类貌似总是在未曾预料的领域取得无法。 想象的成就,这往往是从无到有,而可以看到的目标却迟迟达不到,像可控核 聚变,在人类登月之后登陆火星,癌症的普遍治疗......是否人类在不同领域的 科技存在一个难以逾越的上限?

100 年前的人应该绝对想象不到今天的世界,但之前大多科幻作品对若干年后 未来的设定基本上是失误的。嗯 比如我就不相信人类能做到星际旅行 甚至...

□转发

| 评论

14

□分享



□关注 其实根本瓶颈还是在芯片上,摩尔定律失效,我们真的撞到科技发展天花板了

我一直在想:类似这样的超大规模并行计算的问题,到底有没有可能作为区块 链 POW 算法? 本质上有点难, 因为 POW 要求难算易验证 (AI 是有验证集数

据的,可惜不是无限供应的,也不易调节难度),但万一真能突破,那么经济驱动的力量会跨数量级提升算力,也能让虚拟货币拜托浪费能源的指责。 □1 □评论 □13 □分享		
Sherry 感觉,上节课提到的"存算一体"的芯片可以解决,课程中提到的人工智能的"瓶颈"。至少,这个芯片可以减低,训练模型所需的时间和电力消耗。 □转发 □评论 □13 □分享	□ 关注	
张洋 2-19 AI 芯片概念发展数年来,基本被分成了两类,一类是用于云端服务器的芯片, 一类是用于终端的 AI 芯片。	□关注	
两者对比,云端芯片的门槛更高,无论是对半导体工艺、封装技术和配套软件都有很高要求,一旦成型,其他企业亦很难进入,可以形成竞争壁垒。同时,这个领域更为通用,前景更广,但云端芯片也需要更多的资金和生态支持,高门槛使得大多数创业公司几乎加入无望。尤其是在华为与阿里巴巴两大巨头下场,先后推出云端芯片。		
但是这个方向有不小的困难,芯片从业者分析就说,寒武纪的确拥有不错的研发技术能力,但是做云端服务器的条件是拥有生态资源。"比如你需要有数据库资源,要训练自己的芯片,这对于大厂是天然的,可是创业公司没有,要花巨资去购买。"		

□ 转发□ 评论

□8□分享		
老工人 22-19 所有的发展, 都是能量的消耗, 和策略的改进为基础。 □ 转发 □ 评论 □ 5 □ 分享	□ 关注	
逆天的人 ②-19 目前的人工智能都是靠堆训练量上去,所需要的能量惊人。这种训练是没有限制的运算,再由人工智能自己生成一套方法,不能人为干涉。目前看来光是计算资源是有上限,不亚于虚拟货币消耗的电量。这也暗示了人工智能致命缺陷,除非更高的算力出现,还有无节制的资源消耗。 □转发 □评论 □5	□关注	
注: 注: 注: 注: 注: 注: 注: 注: 注: 注:	□关注	



个人观点,深度学习确实是现阶段不错的方式,更多模仿人脑神经网络不停做 ^{□ 关注} 卷积运算,能够在一定范围内实现比较好的人工智能,只要能够满足人类一部 分需求就足够了。毕竟不能让人工智能和爱因斯坦大脑一样,那样感觉人的情 性会放弃自己大脑,不去进化了。 □转发

□关注

□评论

2

□分享



通过这几年的学习,我逐渐明白一个道理,任何事情的成立,都有一个边界。 对于深度学习, 我以前没有看到他的边界, 通过今天的学习, 我知道它的边界 : 能源。回到吴军老师的洞察,这个世界的两个核心变量: 一是能源,另一是 信息。

□转发

□评论

_ 2

□分享

加微信: 642945106 发送"赠送"领取赠送精品课程 发数字"2"获取众筹列表

