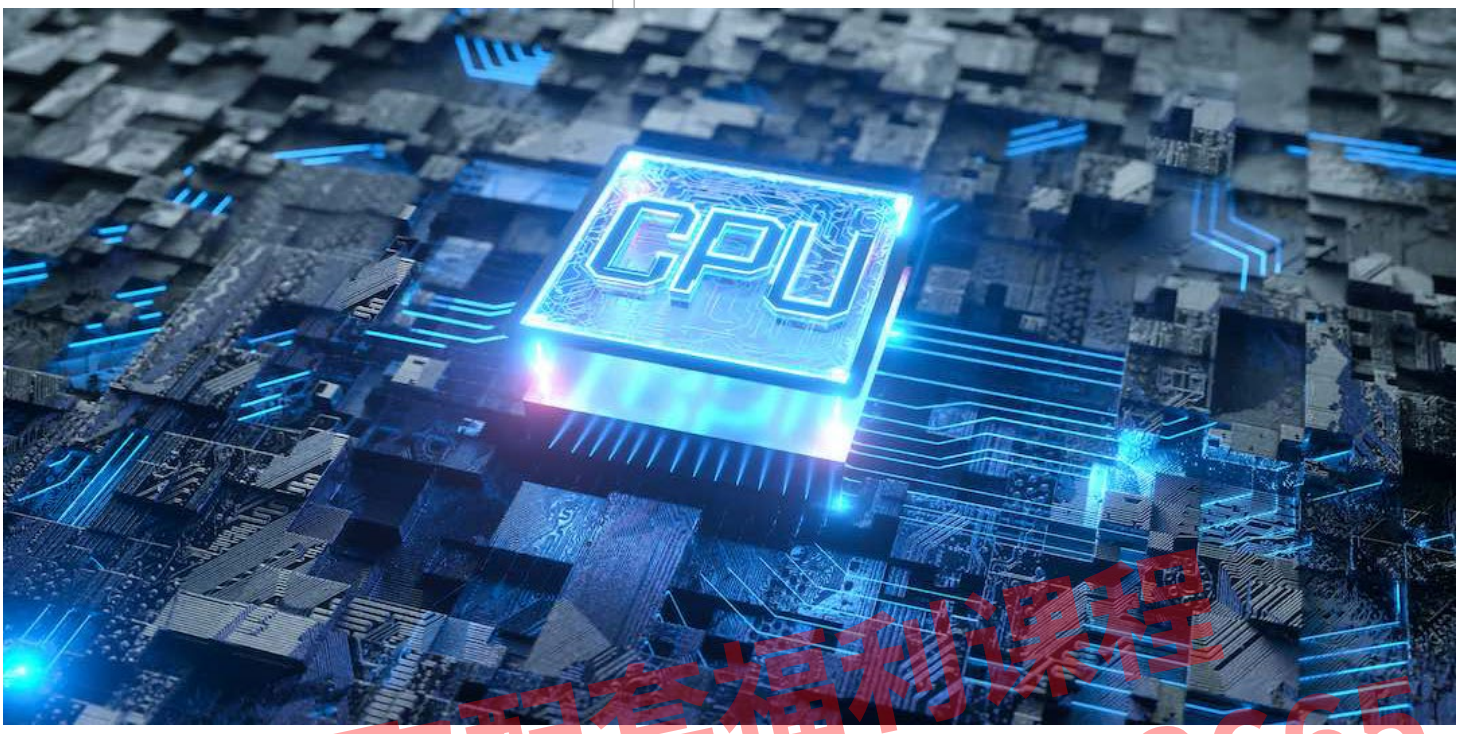

042｜芯片6：“存算一体”到底是什么意思？

卓克·科技参考2（年度日更）

02-18

042｜芯片6：“存算一体”到底是什么意思？
11分12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。

昨天我们说过，在冯·诺依曼结构中，存储和计算是分开的。可能有些人会把这句话误解为，CPU 和硬盘是分离的。

其实不是的。这句话中的“存”，指的是运算过程中必须要用到的“缓存”。想要计算什么，需先

要把数据从缓存中取出来，再放到计算单元里做运算。

对应纸带的操作就是，你得先把要算什么数的那些纸带准备好，放在旁边。轮到算它对应的数字的时候，再把它塞进打孔机里去。

而一个稍微复杂的计算设备，可不是只有纸带，还有很多设备需要和处理器连接，提供各种输入。比如，硬盘数据传输的管理，显卡运算结果的合成，键盘、鼠标、显示器的响应等待等，于是大家都挂在一条粗粗的线上。

这条粗粗的线缆就叫作“总线”。又因为设备不同，所以还要解决设备和总线对接的问题，也就是“接口标准化”。

你看，等待送进 CPU 做运算的数据，总是无法避免来回搬运的问题，这就大大降低了运算效率。

而在存算一体的结构中，计算过程是通过忆阻器这个基本单元来完成的。这就不再有单独的计算单元和缓存了。

下面，我们就仔细说说忆阻器。

什么是存算一体？

这是一个在 20 世纪 70 年代才被发明的基础电路元件，也就是和电阻、电感、电容这些元件并列的一种东西。这个东西在电路中会表现出一个可变的电阻值。

具体电阻值是多少，得看测量之前有多少电荷流过了它。如果电流停止流过它的身体，它的电阻值就固定了。而如果电流反向流过它的身体，电阻值就会慢慢下降。

同时，由于它是阻性元件，于是就满足欧姆定律。换句话说，电流流过单个忆阻器的过程满足公式 $U = I \times R$，也就是电压等于电流乘以电阻。

一方面，你可以把它看作一个乘法器，这就是存算一体中“算”的部分。另一方面，如果了解 SSD 硬盘是怎么写入数据的，你就会知道，其实就是往固定的单元里存入电荷，于是这个过程也很类似于“存”。

总之就是一句话，存算一体其实可以理解为**“用存储电荷的方式实现计算”**。

我们想，在做乘法的过程中，有对乘数和被乘数做读取的操作吗？是没有的。而乘法算完后

的那个结果，有专门临时存放到存储单元里吗？也没有。因为当我们存电荷的动作结束的时候，结果就已经显示出来了，也就是那个忆阻器的阻值状态，根本不需要单独的存储。

你可能会问，这有什么用呢？用处可大了去了。

比如，一个列为 10 的行向量乘以一个 10 行 x 50 列的矩阵，结果应该是另外一个列为 10 的行向量。这个矩阵的运算过程你不必在意，但你要知道，在人工智能的运算中，99% 以上都是一系列的向量乘以矩阵或者矩阵乘以矩阵的运算。

冯·诺依曼结构的计算机中，这个运算过程很简单，但需要对大量数字做读取操作，于是整个过程极其耗费时间。

但如果用忆阻器，只需要码放一个 10 行 x 50 列这 500 个忆阻器组成的阵列，让每一列的输出都经过那一行的 50 个忆阻器就可以了。等电荷累计过后，那 10 列输出的电信号就是矩阵乘法的结果。

总之，不需要大量重复的读取操作，结果很快就出来了。

这还只是 10 × 50 的矩阵的情况，现在存算一体的阵列最高可以集成几百万个忆阻器。这就对应了几百万个元素的矩阵运算。

而在传统计算结构中，在几百万个元素里进行一次矩阵乘法，可能就要涉及上千万次的存取操作。而存算一体的结构只需要存一次、读一次，计算效率天差地别。

存算一体结构在运算上的优势

用存电荷的动作完成计算，就再也没有之前冯·诺依曼结构中存和算这两个动作速度不匹配的问题了。

这又是怎么回事呢？

互相匹配，就意味着不存在瓶颈。比如，如果运算速度太慢，那需要进行运算的数据哗哗哗的输入进来后，只好在存储器里排队等待，这就影响了效率。另外，你得配一套足够大的内存供排队的数据等待。

而这还是好的情况。比较让人惋惜的是，目前 CPU 的处理速度足够快，但即便存储设备开足了马力把任务喂给 CPU，CPU 还是吃不饱。这实在有点亏本。因为早知道存储器存在瓶

颈，我们当初何必花大价钱买这么贵的 CPU 呢？

实际上，自从以硅为基础的半导体芯片出现之后，这个令人惋惜的情况就一直存在。

在 2012 年之前，计算速度的增加一直符合摩尔定律，每 18 个月增加一倍。而存储单元的读写速度的增加就远远落后了，大约是平均每年只增加 10%。

于是，为了能让 CPU、GPU 这些运算设备吃饱，就不得不在靠近运算单元的位置安放一些造价昂贵的、速度特别快的存储器。我说的“靠近”，还真的是物理距离上的靠近，间距 1mm 就比间距 2mm 强。

今天，我们管这些存储设备叫“缓存”（cache）。计算机行业有句俗话——cache is cash（缓存就是金钱），也说明了缓存的昂贵。

比如，直接和 CPU 相连的 L1，也就是一级缓存，大小往往只有百分之几 MB。它使用 SRAM 这种昂贵的存储技术结构，搬运数据的能力大约是每秒几 TB。但这东西太贵了，要省着用。

比 SRAM 便宜得多的存储器是 DRAM，搬运数据的能力大约是每秒几 GB，和 L1 缓存比起来，速度只有千分之一了。

这个 DRAM 存储器在哪儿呢？其实就是 PC 机里的内存条、安卓手机里的 8GB 或者 12GB 内存、苹果手机里的 4GB 或 6GB 内存。一台电脑大约要配几万 MB 的内存，大小是 SRAM 的百万倍。

此外，电脑里还有硬盘。这是最廉价的存储设备了，容量又是内存条的几十倍，同容量价格只有内存条的几十分之一。

这三层结构看着复杂，实际上都是在解决同一个问题——存储设备喂不饱处理设备。

其实，如果都用 SRAM，那还是喂得饱的，但 SRAM 太贵，实在用不起。于是只能少少地用，通过各种算法预测 CPU 马上要用到的数据是什么，提前把这些数据放在 SRAM 中。

在实际情况中，大约 90% 的情况可以预测对，但还有 10% 的情况预测不对。这时候，需要的数据 SRAM 里没有，于是只能从慢得多的内存中去找。而即便在这样精心复杂的结构中，CPU 的算力依然被浪费了。

但前面说了，在存算一体的结构中，这个瓶颈是不存在的。

存算一体结构的应用和前景

那么，存算一体的结构都可以应用在哪些领域呢？

起码现在我们知道，在计算卷积、深度学习这些任务的时候，存算一体结构的能耗比特别好，大约是今天传统计算设备的百分之一。

于是，使用在可穿戴设备上，散热、供电等装置就可以大幅减少，设备的重量就能大大降低。如果可穿戴设备是手表的话，还好一些，人本来对手表的重量不那么敏感。但如果是眼镜的话，增加 2 - 3 克重量就已经足够让我们再也不想戴了。

而未来，AR 或者 VR 眼镜，正是人工智能发挥功能的领域，于是存算一体芯片可能就是智能眼镜特别看重的。

还有，像做监控的智能摄像头，摄像头的成本其实也就 200 块钱，但安装过程中，布线的成本比摄像头还贵，摊到每个摄像头上大约要 100 - 400 块。

而一旦人工智能芯片可以取代现有芯片，一块 50 块钱的锂电池就能让摄像头工作 1 年，于是也就不必布线了。这个布线的成本就也省了。

还有，今天很多 4K 电视、8K 电视里播的内容，其实都不是真 4K，4K 只是“比较清晰”的代名词而已。

因为真正能实时处理 8K 视频数据的芯片贵得很，没有 1000 块钱下不来，而一台液晶电视的利润才 100 多块钱，厂家是不会给电视里配这样的芯片的。

所以，如果要实时解码 8K 的视频，你只有用几千块钱的显卡做才可以。但如果用存算一体芯片来做这件事的话，刚才说过，它是非常适合处理这些矩阵乘法计算的，于是就可以把成本压低到几十块钱。

有人可能会继续问，说了半天，存算一体只能算乘法或者矩阵的乘法吗？

不，其实还能算很多，这里我就不一一举例了。存算一体结构的巨大优势，会在其它新算法开发出来之后吓我们一跳的。

算法层面的突破，我们可以脑洞一下：

比如一个最典型的任务，人脑是怎么学会语言的？这显然是个算法问题。对于人来说，一个孩子只需要 1 - 2 年时间，利用家庭环境里出现的几千上万个词，就能学会语言。

但今天，人工智能算法 GPT-3 要学会一门语言，它的训练集里的语言量差不多是人类文明使用过的所有文章、书籍、词典、诗歌、戏剧、电影里的素材。而这么大的训练量，训练结果也依然只是会把词语片段，连成一段读上去通顺的文章，而它本身并不理解文章含义。这不能不说是一种高昂的代价。

如果存算一体的算法能利用它结构上的优势，导致人工智能的性能大大提升，进展到了能说人话的地步，那我们预测，在任何需要计算的领域，都相当于计算机被冯·诺依曼发明出来后，又被发明了第二次。

当然，除了存算一体之外，芯片领域当前的困境，其实也有一些短期解决方案。这些内容，我们明天继续说。

这就是今天的内容。我是卓克，我们明天再见。

划重点

1. 存算一体，是指利用存储电荷的方式实现计算。这个计算过程是在忆阻器里完成的，不再需要单独的计算单元和缓存。

2. 在传统计算结构中，如果在几百万个元素里进行一次矩阵乘法，要涉及上千万次存取操作。而在存算一体结构里，只需要存一次、读一次，计算效率天差地别。

3. 如果存算一体的算法能让人工智能进展到说人话的地步，那在任何需要计算的领域，都相当于计算机被冯·诺依曼发明之后，又被重新发明了第二次。

收听更多课程微信：1716143665

仅限群内使用！严禁商业！

默认 最新 只看作者回复

钰箭风
02-18

□ 关注 □

神经网络的计算，真的是忆阻器的用武之地。
卓老板仅从储存和计算的速度匹配角度，说明了忆阻器的优势。
我想补充一下另外两个视角

1，CPU 做现在的卷积神经网络等大型矩阵运算也还是吃力的，因为乘法不是 CPU 的强行 CPU 就算加法快

乘法除了专用乘法器，要转换成加法去慢慢算，这种转换有算法，但效率挺低的。
而 CPU 一般只会有 1 个专用乘法器，所以 10 成 50 的以为矩阵需要计算 50 次，这效率是不能忍的。…
所以卷积计算一般要用 GPU 或者 FPGA 的并行运算才有性能。
2，忆阻器这种特性，与人脑的神经连接是类似的，都是正反馈机制，就是用的越多的知识神经连接越强。
所以这种忆阻器的计算速度也能提高很多。
所以拿忆阻器来整神经网络，特别是第三代的脉冲神经网络，简直量身定做。

展开

5
3
128
分享

左星星
02-18

关注

实现存算一体有三种技术路线：

近存储计算（Processing Near Memory）：计算操作由位于存储芯片外部的独立计算芯片完成。

内存储计算（Processing In Memory）：计算操作由位于存储芯片内部的独立计算单元完成，存储单元和计算单元相互独立存在。

内存执行计算（Processing With Memory）：存储芯片内部的存储单元完成计算操作，存储单元和计算单元完全融合，没有一个独立的计算单元。

其中，「近存储计算」通过将计算资源和存储资源距离拉近，实现对能效和性能的大幅度提升，被认为是现阶段解决内存墙问题的最佳途径。

阿里巴巴的达摩院也是沿着「近存储计算」这一方向进行突破，发布了全球首款基于 DRAM 的 3D 键合堆叠存算一体芯片。

而三星走的是「内存储计算」路线。三星的技术称为 “Aquabolt-XL”，主要是基于其 HBM2 DRAM 技术做了内存储计算（HBM2-PIM）。

具体来说，Aquabolt-XL 在 HBM2 DRAM 内部集成了计算逻辑，因此拥有 HBM2-PIM 技术的 DRAM 既可以当作一块普通的 HBM2 DRAM 来用，也可以在写入和读出的时候同时让计算逻辑去做计算。

展开

4
10
102
分享

Stephen
02-18

存算一体，我为啥联想到了算盘。。。

关注

转发
2
40
分享

作者 回复：
这个比喻确实非常恰当，我当时也想用，不过后来担心有些人不知道算盘怎么用就作罢了

佛祖门徒
02-18

目前，存算一体的实现方式有近存储计算与存内计算。
近存储计算指的是计算操作由位于存储芯片外部的独立计算芯片完成。通过采用先进的 3D 封装方式把内存和计算单元封装在一起，可以达到几千根甚至上万根连线，两者之间的带宽增加，提高了数据搬运速度。近存储计算本质上来说还没有做到真正的存算 “一” 体。

存内计算指的是通过在存储器颗粒上嵌入算法，使得存储芯片内部的存储单元完成计算操作，存储单元和计算单元完全融合，没有独立的计算单元。在这种方式下，数据不需要单独的运算部件来完成计算，而是在存储单元中完成存储和计算，消除了数据访存延迟和功耗，是一种真正意义上的存储与计算融合。同时，由于计算完全依赖于存储，因此可以开发更细粒度的并行性，获得更高的性能和能效，存算一体对于符合的应用会带来较高的性能收益和能效收益…

关注

展开

1

评论
32
分享

陈谦
02-18

关注

AI 时代下的数据量激增会导致 “存储墙” 问题愈发地凸显，而存算一体架构能直接利用存储单元进行计算，极大地消除数据搬移带来的开销，解决传统芯片在运行 AI 算法上的 “存储墙” 问题，可以数十倍甚至百倍地提高运算效率，降低成本。随着设计能力不断提升、工艺不断成熟、成本算力能效持续优化，未来存算一体芯片可用在大多数 AI 应用场景，发展空间巨大。

#

1
评论
28
分享

张洋
02-18

关注

存算一体芯片很适合 AI 的应用，因为其技术可以减少搬运存储（memory）的功耗。

但这样做的成本是非常高的，存储工艺会比逻辑（logic）慢三倍，单元面积（cell）会大 4 倍，而且只有 3-4 层金属（metal），综合这些就需要走新路（前端 + 后端，此路可通，但难度也不小）。

相关技术力晶可以做（好像是收购了三菱的整套技术），包括三星也在做这方面的研发。

1
评论
26
分享

乔昕
02-18

关注

卓老师，还有一个问题，存算一体因为读的是电阻值，是不是就不适合精确计算了哪？比如对同样的矩阵乘法计算两次，会不会出现第一次结果是 11，而第二次结果是 12 的情况哪？

1
4
25
分享

作者 回复：
是的，不过幸好这类运算也往往不要求太高精度

乔昕
02-18

关注

卓老师，您好！
说到计算机的功耗，我想问一下，假设计算机正常运行的时候是 100w，那么这 100w 的能耗最后都转化成什么了哪？因为在计算的前后改变的只有硬盘上的数据，而不管是机械硬盘还是 SSD，改变数据都仅消耗极少极少的能量就够了，再就是有一点点能量以电磁辐射的形式消耗了，剩下的能量是不是全都转变成热能了哪？那么假设超导计算机实现了，是不是计算就几乎不用耗能了…

1
1
25
分享

作者 回复：
有一个兰道尔原理描述了改变一个比特位所需的最小能量，除此之外的能耗都会变成热。

沛文沛语
02-18

关注

在听完卓老板这几节课之后，重读一下最初冯诺依曼对于人脑和计算机的思考《计算机与人脑》，有了一种豁然开朗和技术轮回的感觉，也许以后非冯诺依曼体系的计算机会大放异彩，但是当初冯诺依曼对于计算机的思考并不过时，只是当时的技术条件基础走上了现在的这条路线。冯诺依曼在书中也谈到人脑的语言并不是数学的语言，他对计算问题本质的洞悉真的是穿越时代。
“神经系统是基于两种类型的通讯方式的。一种是不包含有算术形式体系的，…
展开

1
评论
23
分享

阿哲Panda
02-18

关注

我也说一下我对存算一体的理解（In-memory Computation），主要是在 AI/ML 的应用领域

存算一体，包括模拟计算（analog computing）在芯片研究领域也算是 “古已有之”，研究使用各种介质的都有，只不过一直没有那么大的关注。但是近年来爆炸的 AI/ML 的应用场景极大地加强了人们对它的兴趣，迅速成为学界的显…
展开

2
1
23
分享

金戈铁马
02-18

关注

SRAM+DRAM + 硬盘这样的存储结构，其实就是对于计算机存算速度不匹配这一问题，所做出的一种无奈之下的处理。虽然说这样的方法，对于如今大多应用场景，都勉强够用。但只要冯・诺依曼结构这一计算机结构的底层逻辑不改变，运算和存储之间速度不匹配的问题就始终存在。这其实也意味，芯片、特别是高端芯片的强大算力，很有可能会被白白浪费掉。而基于忆阻器的存算
…

一体结构，很有可能会带来计算机行业的一次革命性突破，无论是硬件还是…
件领域，都会呈现出完全不一样的景象。而这样的变化，也很有可能带动整个
人工智能行业的发展水平，迈上一个新的台阶。想来也确实让人无比期待。

21

分享

无关风月

02-18

关注

泼个冷水，可穿戴设备的功耗还有个大头是传感器，而我理解这块跟存算一体没什么关系。比如摄像头，很多功耗就是花在那些感光元器件上，还有云台电机上。

另外，关于 GPT-3 的训练集问题，我理解需要优化的是算法吧？硬件只能加速训练的速度和降低功耗而已

1

评论

17

分享

沛文沛语

02-18

关注

从非冯诺依曼体系的架构来看待今天芯片产业遇到的困境，已经在这些困境中当初曾经提到过解决问题的假设，会有不一样的答案。

1：比如“为什么不把芯片尺寸造的大一些呢？这样晶体管的数量不就更多了么？”在冯诺依曼结构中的约束是大量的数据在运算器和存储器之间的通信，导致的时延问题和能耗问题。如果采用忆阻器，就不会有这方面的问题，所以我大胆猜测一下，用忆阻器作为计算基本原件，芯片的尺寸（面积、体积）可…

展开

1

评论

17

分享

萧枫
02-18

卓克老师，能否聊聊卷积神经网络算法？这个算法是如何模拟人脑思考，如何自我学习的原理

关注

转发
2
17
分享

作者 回复：
之后有涉及，是从神经网络到卷积神经网络的一小点内容

AI架构师易筋
02-18

忆阻器（英语：memristor /ˈmɛmrɪstər/），又名记忆电阻（英语：memory resistors），是一种被动电子元件。如同电阻器，忆阻器能产生并维持一股安全的电流通过某个装置。但是与电阻器不同的地方在于，忆阻器可以在关掉电源后，仍能 “记忆” 先前通过的电荷量。两组的忆阻器更能产生与晶体管相同的功能，但更为细小。最初于 1971 年，加州大学伯克利分校的蔡少棠教授根据电子学理论，预测到在电阻器、电容器及电感元件之外，还存在电路的第四…基本元件，即是忆阻器 [1][2]。目前正在开发忆阻器的团队包括惠普、SK 海力士、HRL 实验室（英语：HRL Laboratories）。

关注

之后从 2000 年始，研究人员在多种二元金属氧化物和钙钛矿结构的薄膜中发现了电场作用下的电阻变化，并应用到了下一代非挥发性内存 - 阻抗存储器（RRAM 或 ReRAM）中 [3][4][5][6][7]。2008 年 4 月，惠普公司公布了基于 TiO2 的 RRAM 器件，并首先将 RRAM 和忆阻器联系起来 [8][9][10][11]。但目前仍然有专家认为，这些实作出的电路，并不是真正的忆阻器。— 维基百科

展开

2
1
16
分享

不周山
02-18

关注

语言学习可不仅仅是语音或者文字的事情，实际上婴儿学习语言那几年几乎不会接触实际的文字，而多出来的部分是和其他人的互动，视觉、触觉、味觉、动作反馈…
现在把那些信号合理的转换成机器能处理的有逻辑的数字就已经够麻烦了，更何况要指望机器理解、再和听到的语言做关联。
在听存算一体介绍时很疑惑一件事，单元状态的读取跟存的电荷量有关，可…
展开

1
评论
15
分享


炎熇兵燚
02-18

关注

請問存放一體的缺點是什麼?

转发
1
11
分享

作者 回复:
缺算法，精度低


易得
02-18 编辑

关注

大脑就是存算一体的湿硬件啊!
巧的是大脑计算的能耗约是电脑计算能耗的 x%。
且大脑的逻辑就是无限自指递归，尤其善于处理识别，没准用的就是卷积算法。
还有因果类比算法可能是以贝叶斯算法变形而来。

算法的生理因素是受多巴胺控制。…
硬件受存算一体架构控制。
以下为引用
英国皇家学院院士，伦敦大学的著名神经科学家曾带领科研组做过实验，人类大脑执行运算的次数是每秒 4 千 4 百亿次。而大脑的功率，大概 20 瓦上下，和家里的台灯能耗差不多。而我们一台普通的家用电脑，功率也要在 300 瓦左右，更不要说超级计算机，它们的功率，都是惊人的天文数字。

大脑不止硬件上特别省电，在算法上，似乎也比人工智能要高明许多。谢诺夫斯基就在《深度学习》中总结出一个「100 步法则」。

也就是，人工智能中使用的算法在运行了数十亿个步骤之后，却常常得不出一个正确的结论，而大脑只需要经历大约 100 个步骤，通常就会得出一个正确的结论。

展开

#

转发
1
10
分享

Minsky
02-18

芯片 4 那讲里提到类脑芯片处理视觉信息时能只处理移动的物体信息，突然想到了在此之前一直不理解的特斯拉自动驾驶坚持使用机器视觉而不是雷达，到 model s 用 AMD 高性能芯片，突然感觉特斯拉在下一盘大棋

关注

转发
评论
10
分享

Shan
02-18

个人觉得存算一体对于参与运算的参数多，但是运算流程少的情况下比较适用

关注

，矩阵运算是比较典型的参数多，运算流程简单的问题。另外一类问题是参数少，但是运算流程多的情况下，还是传统的芯片有优势，例如斐波那契数列的简单实现。将来两类芯片还是要配合使用。

请问卓老板，GPT-3 需要大量语料是由深度学习本身决定的，和具体的硬件关系不大。存算一体只能对存和算的速度有影响，对矩阵运算的速度有较大提升，但是如何能够降低训练语料？

转发
1
8
分享

作者 回复：
这还不行，软件上的提升依赖于其他智慧

假装独立思考
02-18

请问卓老板，谷歌的 TPU，还是冯诺依曼架构么？

关注

转发
1
7
分享

作者 回复：
硬件当然还是冯诺依曼构架，但算法是DNN了

戚志光
02-18

5 讲听下来，这一讲最让人兴奋，未来的计算设备还有巨大的提升空间，即便摩尔定律失效了，人们还有办法推动 IT 行业高速发展。希望这些技术早日成熟，更够造福我们的生活。

关注

转发

评论
7
分享

老工人
昨天

如此优秀的模型。居然也只有七十年的生命周期，可见我们对世界的认识能力是一个逐渐提升的过程，或许上帝就是这样安排了人类的学习节奏，就让我们在不断的压力下，寻找他老人家事先放好的答案

关注

转发
评论
5
分享

我是谁，我要去何方
02-18

目前存算一体芯片的主要研发集中在传统非易失存储：SRAM,DRAM 以及非易失存储：PRAM,PCM,MRAM 与闪存等，比较成熟的是 SRAM 和 MRAM 为代表的通用近存计算架构。
1.SRAM 二值存储器 = XNOR 累计运算，可用于二值神经网络运算。核心思想：网络权重存储于 SRAM 中，激励信号由额外字线给出，最终利用外围电路实现 XNOR 累加运算。计算结果通过计数器或模拟电流输出。 难点：实现大…
展开

关注

1
评论
4
分享

假装独立思考
02-18

请问卓老板，谷歌的 TPU，还是冯诺依曼架构么？

关注

转发
1
7
分享

作者 回复：
硬件当然还是冯诺依曼构架，但算法是DNN了

戚志光
02-18

关注

5 讲听下来，这一讲最让人兴奋，未来的计算设备还有巨大的提升空间，即便摩尔定律失效了，人们还有办法推动 IT 行业高速发展。希望这些技术早日成熟，更够造福我们的生活。

转发
评论
7
分享

老工人
昨天

关注

如此优秀的模型。居然也只有七十年的生命周期，可见我们对世界的认识能力是一个逐渐提升的过程，或许上帝就是这样安排了人类的学习节奏，就让我们在不断的压力下，寻找他老人家事先放好的答案

转发
评论
5
分享

我是谁，我要去何方
02-18

关注

目前存算一体芯片的主要研发集中在传统非易失存储：SRAM,DRAM 以及非易

失存储：PRAM,PCM,MRAM 与闪存等，比较成熟的是 SRAM 和 MRAM 为代表的通用近存计算架构。
1.SRAM 二值存储器 = XNOR 累计运算，可用于二值神经网络运算。核心思想：网络权重存储于 SRAM 中，激励信号由额外字线给出，最终利用外围电路实现 XNOR 累加运算。计算结果通过计数器或模拟电流输出。 难点：实现大…
展开