# 英伟达首席科学家 Bill Dally：摩尔定律已失效，“黄氏定律”成全新指标

■作者 Caleb

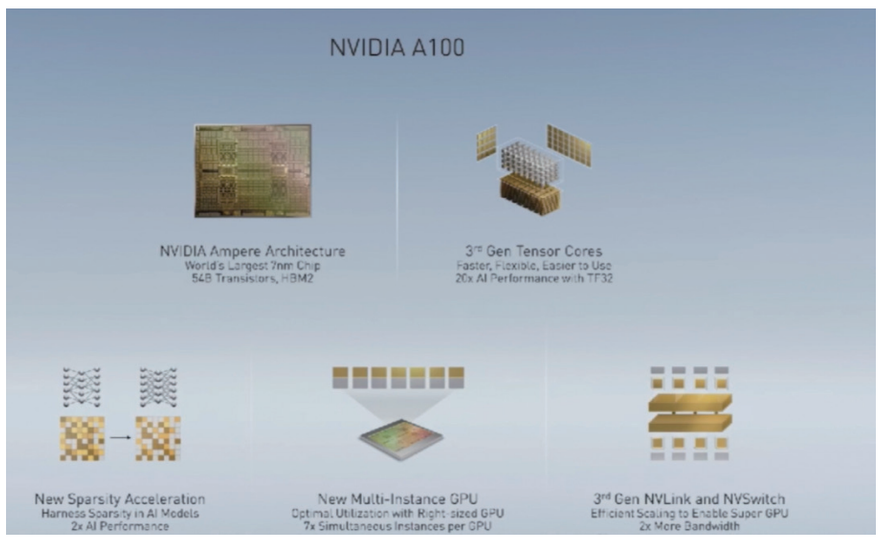
英伟达又来了！一年一度的英伟达 GTC 中国大会仍然以线上的形式和大家见面，不过这次现身的不是黄教主，而是英伟达首席科学家 Bill Dally。

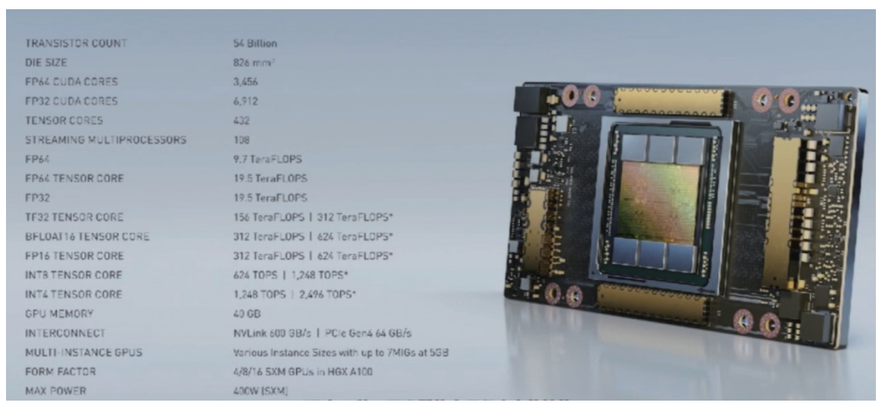
在视频中，Bill Dally 给我们介绍了英伟达在医疗、自动驾驶汽车和机器人等多个领域的身手，也分享了如何在具有更高带宽、更易于编程的系统中制造更快 AI 芯片的相关内容。

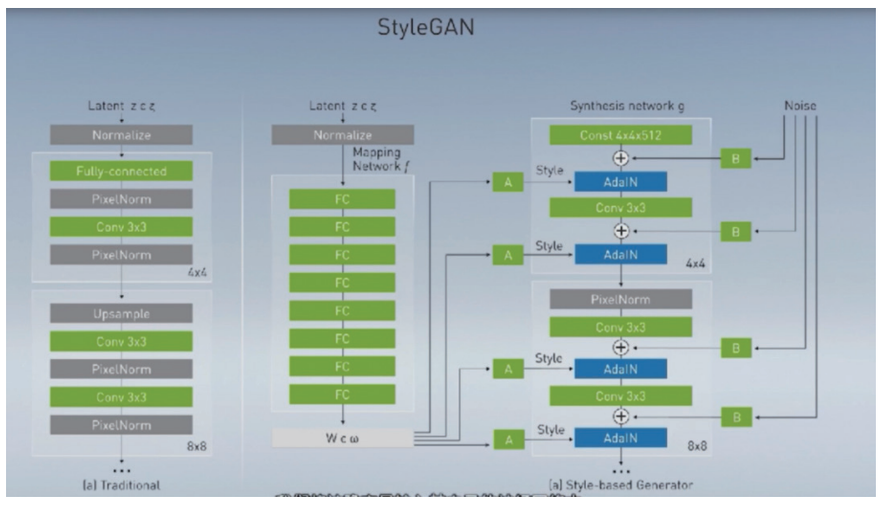
当然更多的还是关于安培，以及一些有趣的应用，比如当语音助手和 GAN 结合之后，能发生什么？

同时，Dally 称：“在‘摩尔定律’失效的当下，如果我们真想提高计算机性能，‘黄氏定律’就是一项重要指标，且在可预见的未来都将一直适用。”

想知道 GPU 如何在英伟达的各类产品中大展身手的吗？来和文摘菌一起看看吧。




## 安培如何合理运用稀疏性特征

我们都知道，英伟达的安培是世界上最大的 7nm 芯片，具有 540 亿个晶体管。

Bill Dally 表示，最让他激动的是，安培破解了如何利用神经网络的稀疏性获得更大的性能的问题。我们先复习一下安培的性能指数。

可以看到，对于高性能计算，安培具有双精度 Tenser Core，对于 FP64 运算，可以在执行矩阵乘法运算时维持 19.5teraFLOPS 的性能。

对于使用全新 TensorFLOAT 32 数据类型进行的深度学习训练，安培可以提供 156teraFLOPS 的性能。而对于深度学习推理，使用 Int8，安培可以提供 1.25petaops。

说到稀疏性，我们知道，大多数神经网络其实是可以修剪的，我们大可切断 70%-90% 的联系，从而达到压缩、释放内容、获得内存的效果，但是我们还无法充分使用这项特征。

现在，我们可以借助安培来合理运用这个特征。安培通过利用结构化稀疏性（允许 4 个权重中的 2 个为 0）解决了这个问题。因此，对于矩阵乘法指令，一旦将权重稀疏为 2/4 模式，就会实现双倍的性能。即使矩阵乘法只是整个应用场景的一部分，比如 BERT 的推理自然语言处理基准测试，安培仍然能达到 1.5 倍的性能。在深度学习构架中，这是一个巨大的飞跃。

同时，安培也简化了 AI 与科学应用之间的关系，你无需在一台计算机上完成一部分工作，再转移到另一台计算机上进行另一部分的工作，使用一台计算机就能完成两者。对于不少 AI 应用程序，很多人都在构建专门的加速器，但是这样的速度会更快吗？其实不一定。

早在 Kepler 时代，进行深度学习，最常用的指令是半精度浮点乘加，如果把它归一化成为技术，将这些都进行相同的比较，大约是 1.5 皮焦耳的能量，提取指令并对其解码，与执行该指令相关的所有开销约为 30 皮焦耳，开销超过了有效载荷，在开销上耗费的能量是有效载荷的 20 倍。

然后在 Pascal 时代，通过改进技术，采用半精度点积运算指令，对包含 4 个单元的向量执行点积运算。

如今，我们要做 8 个算术运算、4 个乘法运算、4 个加法运算，6 皮焦耳的能量，开销仅为 5 倍。虽然从结果上看依旧不是最理想的，但相比最开始，仍然优化了不少。TensorCore 的实际作用是为矩阵乘法累加提供专门的指令，在 Volta 中，采用半精度矩阵乘法累加（IMMA），一条指令所消耗的能力实际执行了 128 次浮点运算，因此完全可以摊还开销。

这样一来，开销只有 22%，在 Turing 中添加 IMMA 指令后，现在可以执行 1024 次 Int8 运算，有效负载所需的能量为 160 皮焦耳，开销仅为 16%。

也就是说，如果构建一个不具有任何可编程性的专用加速器，你将获得 16% 的优势。但同时，我们也不能忽视了，神经网络正在以惊人的速度发展，GPU 的可编程性迫使你跟上变化，新模型层出不穷，训练方法也日渐改善。想要利用这些资源，你就需要一台可编程性很强的设备。GPU 提供了一个完全可编程性的平台，通过构建 TensorCore，使用专门的指令分摊开销，你就可以得到与专用加速器无损的可编程性。

## 未来，GAN 也能语音助手化了

我们先来看一张图，可能大家也能猜到了，左边是生成的虚拟人物，中间是生成的风格化人物，右边则是生成的无生命实体。

最近，英伟达推出 StyleGAN，人们便可以在不同尺度、不同大小下独立控制各个特征，更轻松地分离隐变量，从而分离隐变量中控制图像不同特征的部分，例如控制某个人物是否微笑，是否戴眼镜，以及他们头发的颜色。

同时，在视频技术上，英伟达也有所发力。得到一个人的源图像，和一个人的动作视频，就能合成该说话者逼真的头部视频。在这一任务中，源图像主要负责编码人物的外观，视频则决定了人物的动作。这正是英伟达提出的一种纯神经式的渲染方式，即不使用说话者头部的 3D 图像，只在静态图像上训练生成的深度网络，从而进行头部动作视频渲染。

除此之外，未来，如果你希望自己能够变成一个蓝头发的卡通人物，这项技术也即将实现。

不得不说，GAN 已经渗透到了我们的日常生活中，但是有没有想过，当 GAN 和语音技术碰撞之后，会产生怎样的效果呢？

如果你说的话，被 Jarvis 提取，再转换成文本，输入到自然语言模型中查询、翻译、问答，最后能够生成一幅指定的画，比如你希望哪里有山、哪里有水，GAN 都会自动帮助填充。

GPU 成就了深度学习

在 AI 领域，深度神经网络、卷积神经网络、反向传播等，这些在上个世纪就已经出现的概念，一直要等到 2012 年 AlexNet 的出现，这场革命才真正开始。

那年，Alex Krizhevsky 在 AlexNet 上获得的性能提升，比此前在 ImageNet 上 5 年的工作成果总和还要多。可以说，GPU 成就了深度学习，但同时，也限制了深度学习的发展。

在自然语言处理网络的发展中，从 BERT 到 GPT-3，速度之快令人瞠目。

但是想要构建更大的模型，并在更大的数据集上进行训练，这就受限于在已有的 GPU 资源上可接受的时间内能训练到的程度。

我们再次搬出“黄教主定律”，可以看到，这 8 年里，英伟达将单芯片推理能力提高了 317 倍，这条曲线就是著名的“黄教主定律”，即实现推理能力每年翻倍。

## 自动驾驶：分场景的解决方案

自动驾驶的复杂程度不言而喻，其涉及到传感器、摄像头、雷达、激光雷达、实时计算等多种类型的技术。在实际应用中，还需要预测其他汽车、行人以及周围交通参与者的行为。

于是，英伟达选择利用 AI，打造 GPU 控制的自动驾驶汽车，毕竟 AI 驾驶员不会出现疲劳驾驶等情况。但这不是在汽车中布置一些 AI 技术那么简单，你需要解决的是从数据采集开始的端到端问题。首先，你需要通过各种传感器，包括摄像头、雷达、激光雷达、超声波设备生成大量带标记数据的数据集，然后接受所有的数据并进行筛选。

在将这些模型部署到汽车之前，需要通过硬件在环的仿真模拟进行测试。实际的 AI 硬件会模拟合成看到的信息，包括摄像头生成的合成视频流，激光雷达生成的合成激光雷达数据等，然后需要验证这些模型在仿真时是否正常工作。

除此之外，该神经网络还需要处理其他信息，比如天气、交叉路口的情况等。可以想见，这是相当大的计算负载，英伟达针对此，采用了专为边缘应用打造的基于安培架构的各种产品和解决方案。

在自动驾驶汽车中，如果只需要驾驶员辅助功能，则可使用基于 Orin Ampere 架构提供每秒 10 万亿次运算，且耗能仅为 5 瓦的嵌入式芯片来处理该任务。

对于 L2 级自动驾驶，可能更需要 45 瓦能耗，每秒 200TOPS 的 Orin AGX 来处理该工作负载。当然，对于 L5 级别的自动驾驶，该计算机采用了一对 Orin 和一对 A100 算力高达每秒 2 千万亿次运算，功耗为 800 瓦。这种双重的计算机可提供冗余，如果一部分系统失效，另一部分系统可以继续工作处理传感器信号，至少确保在安全停车前汽车的驾驶是安全的。

## 并行模拟机器人，合理应对各类路况

如今，越来越多的工厂都配备了机器人，它们都能做到毫米级精度的精准定位，但很多也都缺少与人的交互能力。

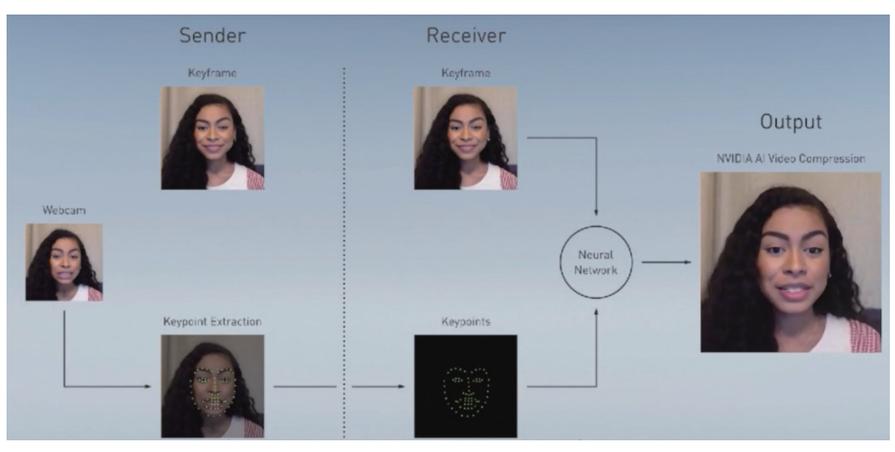
通过利用深度学习的能力，英伟达的技术人员开发了一项名为“黎曼运动策略”的新技术，本质上从数学角度进行简化后，可以实现机器人与人的互动。比如递球给机器人，除了在模拟环境中设置块状物，还人为设置了一些障碍，但是实验证明，不管球滚到哪里，机器人都能快速计算出一条路径绕开障碍物并抓住球。如何操控未知目标，需要针对机械手进行一系列的泛化训练。英伟达在模拟环境中训练了大量四足机器人，从什么都不懂，到遇到各种障碍物，再到合理应对每种表面、上下楼梯，这些技能在真实环境下也能充分应用起来。

这一切，也都得益于英伟达利用 GPU 的并行性进行的模拟。

## 优化图形光源和照明

在好莱坞大片中，我们经常能看到非常逼真的 CG 技术的运用。根据 Bill Dally 介绍，这种离线的计算机图形通常使用的是一种称为基于物理性质的路径追踪渲染技术，对每个像素投射数万条光线，每一帧都需要花费数小时。最近，英伟达的技术团队推出了一种以每秒 60 帧或者更快的速度实时处理照片渲染画面的技术，从效果上看，不管是球体之间的反射，还是球体对光源的反射，都做到了十分逼真的程度：

同时，Bill Dally 表示，这其实是在单个英伟达 GPU 上以每帧 60 秒的速度渲染的效果。这要得益于英伟达在图形领域方面的持续贡献，首先就是 RTXDI 技术，正如上面的照片所展示的：传统图形在阴影投射上表现不够令人满意，但是通过 RTXDI，每个光源都会将其光线投射到其相邻的表面上，这才是逼真阴影效果的奥秘，即光线和物体之间的关系。其次，在间接照明上，RTXDI 使用光探测器将光线从一个表面投射到另一个表面，就能看到一个表面将点亮另一个表面，第二个表面将点亮第三个表面，如此循环。




BIG DATA DIGEST 大数据文摘

（本版由《大数据文摘》杂志授权转载）