# Google Says Machine Learning Chips Make AI Faster and More Efficient 谷歌称机器学习芯片使人工智能更快更高效

原文链接：  
原文链接：

Google recently bared the inner workings of its dedicated machine learning chip, the TPU, marking the latest skirmish in the arms race for AI hardware supremacy.  
谷歌最近公开了其专用机器学习芯片TPU的内部工作原理，这标志着在争夺人工智能硬件优势的军备竞赛中出现了最新的小冲突。

Shorthand for Tensor Processing Unit, the chip has been tailored for use with Google’s open-source machine learning library TensorFlow, and has been in use in Google’s data centers since 2015. But earlier this month the company for the device.  
作为Tensor处理单元的简写，该芯片被定制用于谷歌的开源机器学习库TensorFlow，并自2015年起在谷歌的数据中心使用。但本月早些时候，该公司推出了这款设备。

The company says the current generation of TPUs are designed for inference — using an already trained neural network to carry out some kind of function, like recognizing voice commands through Google Now. On those tasks, the firm says the TPU is 15 to 30 times faster than contemporary GPUs and CPUs, and equally important, they are 30 to 80 times more power-efficient.  
该公司表示，当前一代的tpu是为推理而设计的，使用一个已经训练好的神经网络来执行某种功能，比如通过Google Now识别语音命令。在这些任务上，该公司表示，TPU的速度是当代gpu和cpu的15至30倍，同样重要的是，它们的能效是后者的30至80倍。

For context, CPUs, or central processing units, are the processors that have been at the heart of most computers since the 1960s. But they are not well-suited to the incredibly high computational requirements of modern machine learning approaches, in particular deep learning.  
对于上下文来说，CPU，或者说中央处理器，是自20世纪60年代以来大多数计算机的核心处理器，但是它们不太适合现代机器学习方法，特别是深度学习的高计算要求。

In the late 2000s, researchers discovered that graphics cards were better suited for the highly parallel nature of these tasks, and GPUs, or graphics processing units, became the de facto technology for implementing neural networks. But as Google’s use of machine learning continued to expand, they wanted something custom built for their needs.  
在21世纪末，研究人员发现图形卡更适合于这些任务的高度并行性，而gpu（即图形处理单元）成为实现神经网络的事实上的技术。但随着谷歌对机器学习的使用不断扩大，他们希望能根据自己的需要定制一些东西。

“The need for TPUs really emerged about six years ago, when we started using computationally expensive deep learning models in more and more places throughout our products. The computational expense of using these models had us worried,” lead engineer Norm Jouppi .  
“大约六年前，当我们开始在产品中越来越多的地方使用计算成本高昂的深度学习模型时，对tpu的需求才真正出现。使用这些模型的计算费用让我们担心，“首席工程师Norm Jouppi。

“If we considered a scenario where people use Google voice search for just three minutes a day and we ran deep neural nets for our speech recognition system on the processing units we were using, we would have had to double the number of Google data centers!”  
“如果我们考虑这样一个场景：人们每天只使用谷歌语音搜索3分钟，我们在使用的处理单元上为语音识别系统运行深层神经网络，我们就不得不将谷歌数据中心的数量翻一番！”

Nvidia, for its part, says the comparison isn’t entirely fair. Google compared its TPU against a server-class Intel Haswell CPU and an Nvidia K80 GPU, but there have been two generations of Nvidia GPUs since then. Intel has kept quiet, but Haswell is also three generations old.  
英伟达则表示，这种比较并不完全公平。谷歌将其TPU与服务器级的英特尔Haswell CPU和Nvidia K80 GPU进行了比较，但此后已有两代Nvidia GPU。英特尔一直保持沉默，但哈斯韦尔也有三代人的历史。

“While NVIDIA’s Kepler-generation GPU, architected in 2009, helped awaken the world to the possibility of using GPU-accelerated computing in deep learning, it was never specifically optimized for that task,” the company says in .  
“虽然2009年设计的NVIDIA的开普勒一代GPU帮助唤醒了全世界在深度学习中使用GPU加速计算的可能性，但它从未专门针对这一任务进行过优化，”该公司在报告中说。

To make their point, this was accompanied by their own benchmarks, which pointed to their latest P40 GPU being twice as fast. But importantly, the TPU still blows it out of the water on power consumption, and it wouldn’t be surprising that Google is already readying or even using a new generation of TPUs that improve on this design.  
为了说明他们的观点，他们自己的基准测试也与之相伴，这表明他们最新的P40 GPU速度是后者的两倍。但重要的是，TPU在耗电量上仍然让它水落石出，谷歌已经在准备甚至使用新一代的TPU来改进这种设计也就不足为奇了。

That said, it isn’t going to upend the chip market. Google won’t be selling the TPU to competitors and it is entirely focused on inferencing. Google still uses copious amounts of Nvidia’s GPUs for training, which explains the muted nature of the company’s rebuttal.  
也就是说，它不会颠覆芯片市场。谷歌不会向竞争对手出售TPU，它完全专注于推断。谷歌仍在大量使用英伟达的gpu进行培训，这就解释了该公司反驳的低调本质。

Google is also probably one of the few companies in the world with the money and the inclination to build a product from scratch in a completely new domain. But it is also one of the world’s biggest processor purchasers, so the fact that it has decided the only way to meet its needs is to design its own is a warning sign for chip makers.  
谷歌也可能是世界上为数不多的几家拥有资金并倾向于在全新领域从头开始开发产品的公司之一。但它也是全球最大的处理器采购商之一，因此，它决定满足需求的唯一途径是自行设计，这对芯片制造商来说是一个警告信号。

Indeed, that appears to be part of the idea. “Google’s release of this research paper is intended to raise the level of discussion amongst the machine learning community and the chip makers that it is time for an off-the-shelf merchant solution for running inference at scale,” .  
事实上，这似乎是想法的一部分。“谷歌发布这篇研究论文的目的是提高机器学习社区和芯片制造商之间的讨论水平，即现在是一个现成的商业解决方案在规模上运行推理的时候了。”。

This is probably not too far off, analyst Karl Freund . “Given the rapid market growth and thirst for more performance, I think it is inevitable that silicon vendors will introduce chips designed exclusively for machine learning.”  
分析师卡尔弗伦德（Karl Freund）说，这可能不会太远。“鉴于市场的快速增长和对更多性能的渴求，我认为硅厂商将不可避免地引入专门为机器学习而设计的芯片。”

Nvidia is unlikely to let its market leading position slip, and later this year Intel will release the first chips powered by the machine learning-focused . Even mobile players are getting in on the act.  
英伟达不太可能让其市场领先地位下滑，今年晚些时候英特尔将发布首款以机器学习为核心的芯片。甚至连手机玩家也开始参与其中。

Arm’s Dynamiq microarchitecture will allow customers to to bring native machine learning to devices like smartphones. And Qualcomm’s Project Zeroth has released a software development kit that can featuring its Snapdragon processors.  
Arm的Dynamiq微体系结构将允许客户将本地机器学习带到智能手机等设备上。而高通公司的“零号计划”也发布了一个软件开发工具包，可以使用Snapdragon处理器。

Google’s release of the TPU may be just a gentle nudge to keep them heading in the right direction.

Image Credit: [Shutterstock](http://www.shutterstock.com/)