


腾讯云总监手把手教你，如何成为AI工程师？



作者

星芒红哥 (/u/4cab6a616904)

+ 关注

2017.04.10 19:24

字数 5261

阅读 8

评论 0

喜欢 1

(/u/4cab6a616904)

推荐理由：

“人工智能”这个词在2016年经常被行业大佬提到，相信我们听的也挺过瘾，但真正什么是人工智能，如何成为AI工程师，小编我还是一脸懵逼；如何成为AI工程师似乎是梦想中的事，但今天我分享的这篇文章，是由腾讯云总监亲自带大家，这让我感到距离AI工程师还是可以想想的，至少有了清晰的思路 and 眼界；希望分享的这篇文章对大家也有所帮助。

以下为文章原文：

作者：朱建平 腾讯云技术总监，腾讯TEG架构平台部专家工程师

1.关于人工智能的若干个错误认知

人工智能是AI工程师的事情，跟我没有什么关系

大数据和机器学习(AI) 是解决问题的一种途径和手段，具有通用性，是一个基础的技能。当前我们工作中还有很多决策，是基于经验和预定的规则，未来这部分决策可以通过AI让我们做得更合理更好一些。

人工智能太厉害了，未来会取代人类

随着人工智能的发展，特别去年谷歌的AlphaGo围棋战胜代表人类的顶级棋手李世石，更是引爆了整个互联网。于是，网上不少人开始了很多担忧：机器人取代人类，有些人甚至在孩子高考填志愿时要求孩子填报艺术创作类似的方向，以避免未来与机器人或人工智能的竞争。

实际上，虽然目前人工智能在语音识别，图片识别近年来取得了突破，但人工智能还远未完善: 数学理论尚不完备，“智能”的取得建立在大量的人工前期工作基础上，缺乏无监督学习。

2. 传统开发 转行AI工程师的障碍

2.1 急于求成的心态

LR, SVM, 决策树，DNN,CNN, AlexNet, GoogleNet, Caffee,TensorFlow, 智能驾驶，AlphaGo, 个性化推荐, 智能语音，GPU, FPGA....

晕了没？ 没晕再来一波。。。。

2.1 传统开发转行 AI 工程师的障碍

腾讯云

急于求成的心态

LR, SVM, 决策树, DNN,CNN

AlexNet,GoogleNet, Caffee,TensorFlow 智能驾驶

AlphaGo, 个性化推荐, 智能语音, GPU, FPGA

$$J(\theta)=\frac{1}{m}\sum_{i=1}^m\frac{1}{2}(y^i-h_{\theta}(x^i))^2=\frac{1}{m}\sum_{i=1}^m\cos t(\theta,(x^i,y^i))$$
$$\cos t(\theta,(x^i,y^i))=\frac{1}{2}(y^i-h_{\theta}(x^i))^2$$
$$\frac{\partial J(\theta)}{\partial \theta_j}=-\frac{1}{m}\sum_{i=1}^m(y^i-h_{\theta}(x^i))x_j^i$$

5

这里面的水很深，不要太急躁很快能搞懂，事实上由于数学理论不完备，有些东西还真解释不清楚，比如在图像识别上ResNet 比GoogleNet识别率更高，ResNet是怎么推导出来的？

梳理好这些概念，结合实际应用，化整为零逐步理解和吸收，有的放矢，不可操之过急。

2.2 自底往上的学习方法，想要从基本概念学习

建议结合应用场景先动手实践，再逐步细化。

推荐《机器学习》周志华 清华大学出版社

2.2 传统开发转行AI工程师的障碍

腾讯云

自底往上的学习方法

不要从理论开始学起
不要去学机器学习所有的东西
不要在算法里浪费光阴

结合应用场景
先动手实践，再逐步细化

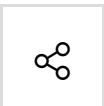


6

3.AI工程师的知识结构-机器学习的基础知识

3.1 人工智能<--> 机器学习<---->深度学习 的关系

这是现在大家经常混淆的概念，什么叫做人工智能？什么叫做机器学习？什么叫做深度学习？人工智能是最大的范畴，只要你用计算机做了一点智能的事情都可以称为做了人工智能的工作。真正的人工智能应该是让机器拥有人的智能，让机器跟人一样能看、能



听、能说，能用自然语言跟人进行交流。这个涉及到计算机视觉、语音识别、自然语言处理、人机交互、语音合成等等，是常规的我们研究讨论的人工智能的主要发力点，在互联网公司有着广阔应用场景的。

机器学习可能是人工智能目前最火的领域，深度学习可能又是机器学习最火的子领域。什么时候需要人工智能？直觉上来讲数据越复杂，深度学习越可能起作用；数据很简单很明确，深度学习可能就不怎么起作用了。比如搜索领域，目前只有Google宣称他们用深度学习double了用户点击率，是指他们将深度学习运用在用户浏览过、搜索过的信息上，那是非常庞大非常复杂的数据。

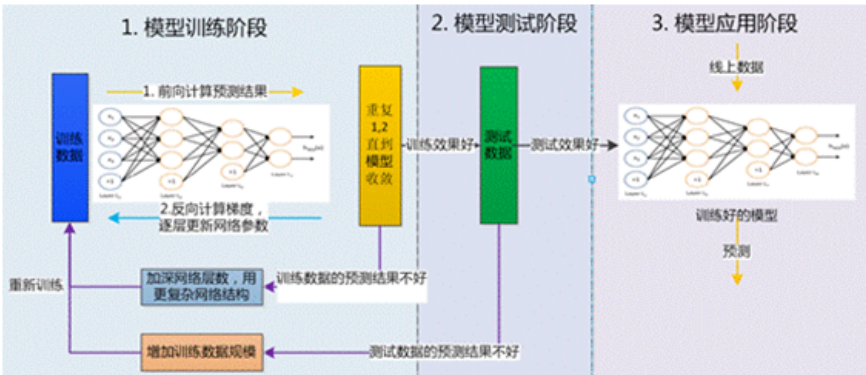
3.2 机器学习解决问题的基本步骤

一般应用机器学习实际问题分为4个步骤：

3.2 机器学习解决问题的基本步骤

腾讯云

- 1. 定义目标问题
- 2. 收集数据和特征工程
- 3. 训练模型和评估模型效果
- 4. 线上应用和持续优化



9

1) 定义目标问题

目前还没看到有一个机器学习模型适用于解决所有问题，不同问题有各自适用的模型，如图像相关问题有深度学习、推荐相关问题有专门的推荐算法、安全相关问题有异常检测模型等。脱离具体问题去讨论模型是没有意义的。

2)收集数据和特征工程

机器学习是面向数据编程，数据是机器学习的基础。训练模型时，一般会把样本数据拆成两部分，其中大部分(约7成)数据用于训练模型，称其为训练集；另外少部分数据用于测试“模型的好坏”（也称“泛化能力”），称其为测试集。

同一个机器学习算法，好的数据能让其表现更好，差的数据会让模型毫无用处。什么是“好的数据”？并没有统一定义，从结果看，能让模型表现良好的数据就是“好的数据”。一个可行的办法是想象“人”在解决该问题时，会依据哪些数据和特征做决策，然后挑选这些数据和特征作为机器学习模型的输入数据，这个过程就是特征工程。在应用机器学习时，可能需要反复做多次特征工程，特征工程是个试错的过程。

3)训练模型和评估模型效果

利用标注数据，训练模型数据，而一般的步骤是：

a. 从底层存储读取数据



- b. 对训练数据进行前向计算
- c. 计算训练误差
- d. 反向计算梯度，更新网络参数
- e. 重复a - d 步，直到模型收敛。

测试模型效果，一般测试数据集远小于训练集，这里主要是快速前向计算，一般合并在这一步中。

4)线上应用和持续优化

模型在训练集上性能达标，但在线上环境性能不达标，这一现象被称为“过拟合”。通常的原因是用于训练模型的数据中特征的分布与线上数据偏差太大，此时需提取更具代表性的数据重新训练模型。

模型在线上应用后，需持续跟踪模型的性能表现，机器学习是面向数据编程，如果线上系统上的数据出现了不包含在训练集中的新特征，需要补充新样本，重新训练迭代模型以保证预测效果。

3.3 机器学习的相关概念

3.1 AI 工程师的知识结构

腾讯云

机器学习算法	LR/SVM/决策树（传统的分类和聚类） DNN（深度神经网络） CNN（卷积神经网络）
CNN网络模型	AlexNet , GoogleNet , ResNet
框架	Caffee , TensorFlow
硬件	GPGPU , FPGA



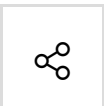
8

模型用途：分类、回归、聚类

主要区分在于output的描述是什么性质:分类是指output是整数（即多个类别标签）；回归是指output是一个实数，例如预测股票的走势，input是时间，output就是股票价格；聚类一般都是应用于非监督的状态下，对output完全不知道，只能对input数据本身进行统计分析，比如用户画像，通过数据之间的关系如关联程度将数据分成好几簇。

训练过程: 监督、半监督和非监督

机器学习是一个用数据训练的过程；监督是指input的每个数据样本，我们明确知道它的output（如类别标签）是什么；半监督是指我们只知道input数据样本中一小部分的output，另外大部分不知道；非监督是指所有input的数据样本，我们完全不知道它们的output是什么。



学习模型：LR/SVM/决策树（传统的分类和聚类）DNN(深度神经网络）CNN（卷积神经网络）

常用CNN模型：AlexNet, GoogleNet, ResNet

浅层和深层，以前的机器学习方法大都是浅层，浅层学习模型是从六十年代发展到现在；深度学习模型过去不怎么work，自2010年迄今有了非常大的突破，深层模型在大量（至少百万级别）的有标签的数据驱动下将input端到output端之间的映射做的更深更完善。

开源框架&平台：Caffee, TensorFlow（Google）,Torch (Facebook)

为什么有这么多深度学习框架，参考《Deep Learning System Design Concepts》
<http://mxnet.io/architecture/index.html#deep-learning-system-design-concepts>
(<http://mxnet.io/architecture/index.html#deep-learning-system-design-concepts>)

4.入门成为AI工程师的可行路径

虽然从垂直领域讲有语音识别，图像视觉，个性化推荐等业务领域的AI工程师，但从其所从事的研发内容来看，从事AI研发的工程师主要分为3类：

1)AI算法研究

这类人大都有博士学历，在学校中积累了较好的理论和数学基础积累，对最新的学术成果能较快理解和吸收。这里的理论是指比如语音处理，计算机视觉等专业知识。

AI算法研究的人主要研究内容有 样本特征，模型设计和优化，模型训练。样本特征是指如何从给定的数据中构建样本，定义样本的特征，这在个性化推荐领域中就非常重要。模型设计和优化是设计新的网络模型，或基于已有的模型机型迭代优化，比如CNN网络模型中AlexNet, GoogleNet v1/v2/v3, ResNet等新模型的不断出现，另外就是比如模型剪枝，在损失5%计算精度情况下，减少80%计算量，以实现移动终端的边缘计算等等。模型训练是指训练网络，如何防止过拟合以及快速收敛。

2) AI工程实现

这类人主要提供将计算逻辑，硬件封装打包起来，方便模型的训练和预测。比如：

精通Caffee/TensorFlow等训练框架源码，能熟练使用并做针对性优化；

构建机器学习平台，降低使用门槛，通过页面操作提供样本和模型就能启动训练；

通过FPGA实行硬件加速，实现更低延时和成本的模型预测；

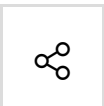
在新模型验证完成后，实现在线平滑的模型切换；

3) AI应用

侧重验证好的模型在业务上的应用，常见语音识别，图像视觉，个性化推荐。当然这也包括更多结合业务场景的应用，比如终端网络传输带宽的预测，图片转码中参数的预测等等。

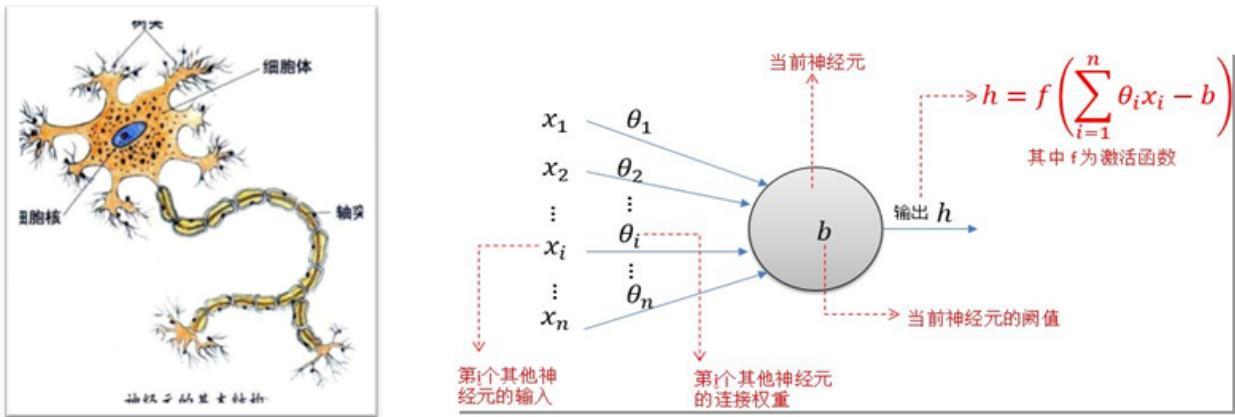
5.DNN 和 CNN网络

5.1 DNN原理



DNN深度神经网络是模拟人脑的神经元工作机制构建的计算处理模型。

3.3 DNN网络 - 神经元

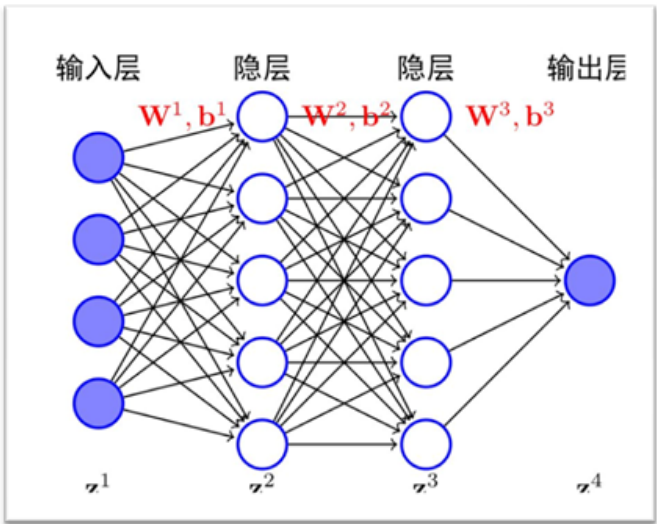


激活函数常用的有：sigmoid , ReLU等,比如 典型的sigmoid函数

3.3 DNN 网络



多个神经元分层组织起来构成了一个网络

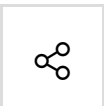


11

多个神经元分层组织起来构成了一个网络，早期神经网络仅能做到浅层，在训练方法和计算能力获得突破后，深层神经网络DNN得到了更广泛研究和应用。

5.2 CNN原理

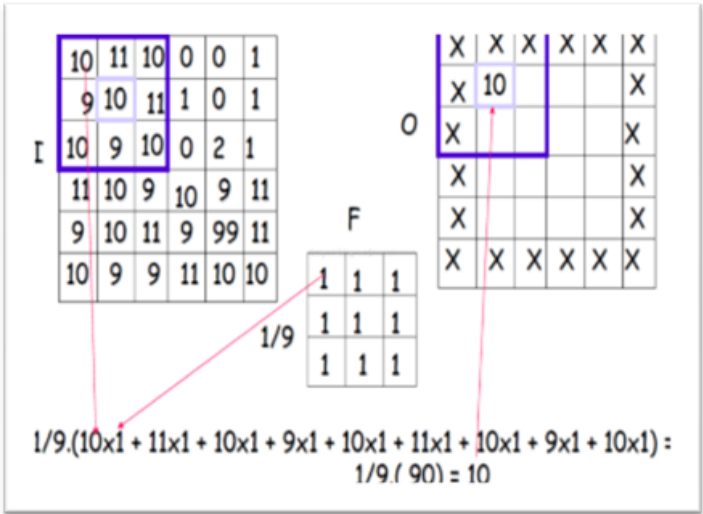
简化的计算过程：



3.4 CNN网络 - 卷积

腾讯云

简化的计算过程



12

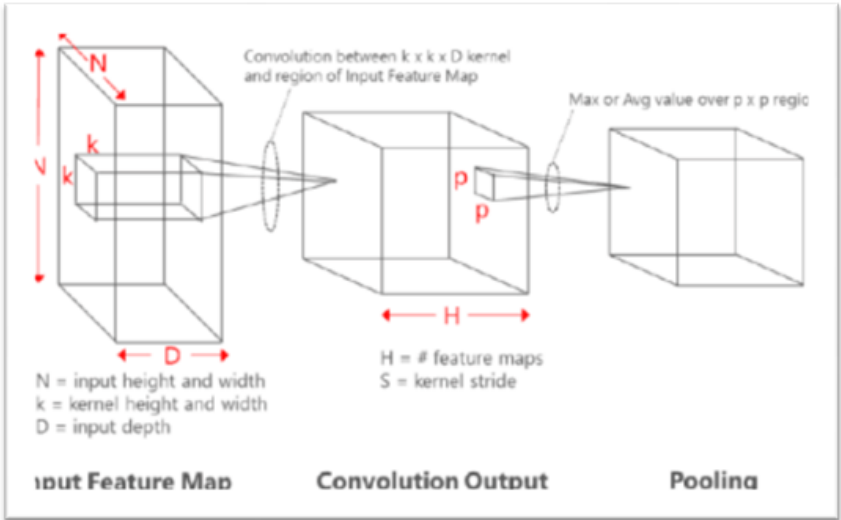
上图展示了一次卷积计算: 一个66的图片I 使用卷积核F进行卷积，得到输出图片O。输入图片中在patch范围内的元素和卷积核中对应的元素相乘，最后乘积结果相加。

真实的计算过程：

3.4 CNN网络 - 图片识别中的卷积

腾讯云

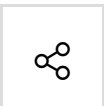
真实的计算过程



13

上图是三维卷积的展示，对于第一层来说卷积核是11x11x3，在输入立方体227x227x3上进行滑动，对应图表 2 中的k=11，N=227，D=3.卷积算法就是卷积核11x11x3和立方体227x227x3的重叠的每个值做乘运算，再把乘的结果做累加，最后得到一个值，数学公式为 $y = x[0]k[0] + x[1]k[1] + \dots + x[362]k[362]$ ，因为卷积核11x11x3共有363个值，所以我们可以看成一个1x363的矩阵乘以363x1矩阵。

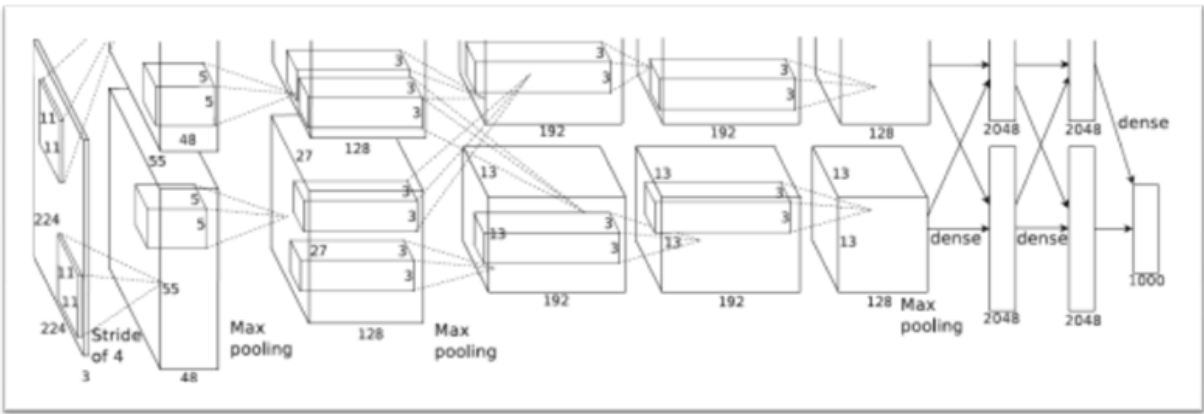
6.AI开发的典型场景



4.1 AI 实践案例分享 - 图片鉴黄FPGA加速

腾讯云

Alexnet 模型



16

训练采用caffe 单机框架，单机2卡K80 GPU，为充分发挥GPU，采用了数据并行，一次一个batch 256张图片输入，alexnet网络分为前5层卷积层，后3层为全连接层，主要的计算在卷积计算，我们将其用FPGA实现，全连接层采用CPU实现。

海量准确的样本也是个细致活，需要不断运营。

4.1 AI 实践案例分享 - 图片鉴黄FPGA加速

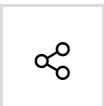
腾讯云

粗筛：基于Alexnet 对图片进行分类黄图（性感图）、正常图
精筛：针对黄图（性感图）采用Googlenet模型进行识别

层数	卷积核数	每个卷积核进行卷积次数	每个卷积核一次卷积运算量	浮点乘加个数
第 1 层	96	3025	(1x363)x(363x1)	96x3025x363=105M = 210M FLOP
第 2 层	256	729	(1x1200)x(1200x1)	256x729x1200=224M = 448M FLOP
第 3 层	384	169	(1x2304)x(2304x1)	384x169x2304= 150M =300M FLOP
第 4 层	384	169	(1x1728)x(1728x1)	384x169x1728=112M = 224M FLOP
第 5 层	256	169	(1x1728)x(1728x1)	256x169x1728 = 75M = 150M FLOP
第 6 层	1	4096	(1x9216)x(9216x1)	4096x9216 = 38M = 76M FLOP
第 7 层	1	4096	(1x4096)x(4096x1)	4096x4096 = 17M = 34M FLOP
第 8 层	1	1000	(1x4096)x(4096x1)	1000x4096 = 4M = 8M FLOP
总和				1.45G FLOP

17

腾讯云GPU云客户案例

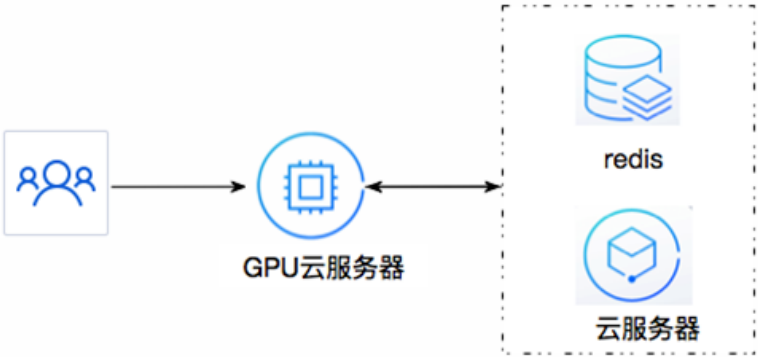


4.2 AI 实践案例分享 - 腾讯GPU云案例

腾讯云

应用案例

IN，图片社交分享
小红书
猎豹
微众银行
香港理工



18

香港理工项目

该实验项目是基于美国气象局提供的10年的气象数据，包括温度，湿度，风向，风速，降雨量，云层厚度，云图，空气浑浊度，日照等数据，对未来一段时间的天气进行预测。在该项目中，我们使用Google进行基于神经网络深度学习的Tensorflow框架，用Python2.7进行开发，并且在GPU上对深度神经网络进行训练。

7.CPU、GPU、FPGA区别

5.4 AI 底层硬件支撑 - CPU vs GPU vs FPGA

腾讯云

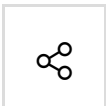
	CPU-E5 2697*2	GPU-P40*8	FPGA-UV9P*8
计算卡能力	0.4T FLOPS	12T FLOPS	5.2T
单节点计算能力	0.8T FLOPS	96T FLOPS	41.6T FLOPS
DDR内存带宽	N/A	24G-346GB/s GDDR5-384bit	16G-76GB/s DDR4-64bit
功耗	2*145瓦	8*250瓦	8*40瓦
研发效率	高	中上	低
处理延时	低，存在波动	高	低，稳定

大多数人可能有个大致的认识: 训练用GPU, 预测用CPU 或者 FPGA

CPU 开发门槛低，未来主要承载 高性能网络，计算分拆出来的逻辑复杂，不适合并行计算的部分。

GPGPU 最新的**P40,P100**系列，采用16nm工艺，因其Cuda开发环境比较成熟，学习成本低，灵活性高，将继续在AI的模型训练阶段发挥关键作用。

FPGA 最新的**YV9P(16nm) FPGA**，之前强调节能，单FPGA在数据中心的部署也是一个全新的课题，未来可能会加强 HBM2片上DDR内存容量和带宽的增长。未来在在线模型预测方面发挥重要作用，但IP不足，开发周期长是一个瓶颈。 FPGA卡，驱动开发，IP实现导致使用门槛较高，未来会在FPGA云上消除这些应用障碍。



8. 腾讯云的GPU云、FPGA云进展

8.1 腾讯云 Skylake CPU

2017年2月，腾讯云宣布在国内率先使用英特尔下一代至强®处理器（代号Skylake），推出国内最新一代云服务器。新一代云服务器在计算性能、内存带宽、网络时延等方面拥有显著优势，最高可提供96 vCPU，可满足企业对云服务器高规格高配置的广泛需求，尤其在人工智能等高性能计算领域将发挥更大价值。据介绍，目前腾讯云官网已开放新一代云服务器的试用申请，客户将花费更低的购买价格，享受到更高性能计算服务。

与过往采用至强系列处理器的云服务器相比，内置Skylake至强®处理器的新一代云服务器具有更高计算性能、更大内存带宽、更强存储I/O性能、更低网络时延等优势，能满足游戏行业、视频行业、金融行业等领域的更高计算需求。具体而言，Skylake至强®处理器具备的更优特性主要包括：

Skylake至强®处理器支持AVX-512指令，可支持更大数据宽度处理，能加速多媒体编解码、加解密数值运算，在机器学习、科学计算和金融分析等需要顶尖的浮点运算功能的场景提供更优质的处理性能。

Skylake至强®处理器支持Omni-Path 互联架构，有助于提供更快的数据访问效率、更低的延时服务。

8.2 腾讯云GPU

腾讯云推出基于NVIDIA最新企业级产品（M40和P40）的云产品GPU云服务器和GPU黑石服务器，其中，基于M40的GPU云服务器已于2016年底正式上线。今年上半年，腾讯云还将推出1机8卡的GPU云服务器，单机多卡整机性能更高，可以满足超大数据量超大规模机器学习算法需求，最大化提升计算效率。

G2实例最多可提供 2 个 NVIDIA M40 GPU、56 个 vCPU 和 120GB 主机内存，以及双卡 48GB 的GDDR5 显存。GPU云服务器拥有高达6144个加速核心、单机峰值计算能力突破14T Flops单精度浮点运算，0.4T Flops双精度浮点运算。

在视频渲染、虚拟化桌面、深度学习等对计算能力要求极高的场景中，腾讯云GPU云服务器以及GPU黑石服务器都有广泛的应用前景，同时还能满足图形数据库、高性能数据库、计算流体动力学、计算金融、地震分析、分子建模、基因组学、渲染等领域对基础设施的高要求，且极具性价比。

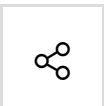
目前GPU云服务器已开放官网申请，腾讯GPU云申请 (<https://www.qcloud.com/product/gpu?fromSource=gwzwcw.57471.57471.57471>)

8.3 腾讯云FPGA

腾讯云在年前宣布推出国内首款高性能异构计算基础设施“FPGA云服务器”。已于2017年1月内测上线，以云服务方式将大型公司才能长期支付使用的FPGA推广到更多企业。腾讯云和业界厂商有良好的关系，提供了基于 Intel和Xilinx 两家的硬件平台和软件开发工具，方便开发者选择自己熟悉的开发模式，避免切换平台。

1) 硬件平台

腾讯云即将发布基于 Intel和Xilinx的单机4卡FPGA云服务器，推出多种规格的FPGA实例供您选择。单机多卡整机性能更高，可以满足超大数据量超大规模机器学习算法需求。也可选择单卡可节省计算效率，轻资产开发，降低项目研发期间的投入成本。



2) 腾讯云官方FPGA IP

Alexnet网络模型预测加速（已上线）--->用于图片鉴黄的粗筛

Googlenetv1 网络模型预测加速（今年上半年）---->用于图片鉴黄的精选

同步开放的还有 内部使用的图片压缩IP。

3) FPGA 生态建设

我们通过IP市场，以开放合作的心态引入更多第三方成熟的AI IP进来，为FPGA生态的发展注入新的生机。

Q&A机器学习对于模仿人的思考是怎么做到的？

现在机器学习模仿人的思考做的比较原始。目前主要还是提取人做某项决策时考虑的主要因素，在机器学习中我们叫样本特征来告诉模型，当遇到类似特征时应该输出什么。

相关阅读：

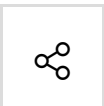
腾讯GPU云申请 (<https://www.qcloud.com/product/gpu?fromSource=gwzcw.57471.57471.57471>)

人人都可以做深度学习应用：入门篇 (<https://www.qcloud.com/community/article/451090001487836806?fromSource=gwzcw.57314.57314.57314>)

文章出自腾讯云技术社区

（埋文字链<https://www.qcloud.com/community/article/581518001490250627>
(<http://link.zhihu.com/?target=https%3A//www.qcloud.com/community/article/581518001490250627>))

推荐大家关注腾讯云技术社区微信 (<http://link.zhihu.com/?target=http%3A//lib.csdn.net/base/wechat>)公众号：QcloudCommunity





星芒红哥 (/u/4cab6a616904)

写了 40598 字，被 1 人关注，获得了 4 个喜欢 (/u/4cab6a616904)

+ 关注

相信自己，敢于挑战！

如果觉得我的文章对您有用，请随意打赏。您的支持将鼓励我继续创作！

赞赏支持

♡ 喜欢 (/sign_in)

1



更多分享

(http://cwb.assets.jianshu.io/notes/images/11182986/weibo/image_228e6541265d.jpg)



登录 (/sign_in) 后发表评论

评论

智慧如你，不想发表一点想法 (/sign_in)咩~

