深度学习 | 话题

|期刊分享|深度学习|DNN模型实际应用分析

机器学习与人工智能
百家号 03-18 08:00

关注

编者推荐序：本文作者为2016年ENet的作者，从准确率、内存占用、参数、功耗等多角度全面对比分析了近4年ImageNet竞赛中表现优异的深度神经网络，主要结论有：功耗与batch大小及体系结构无关；准确率与推理时间呈双曲线关系；能量限制是准确率和模型复杂度的上限；运算次数可以有效评估推理时间。总言之，本文研究结论对实际应用中DNN模型的选择起到了很好的指导作用。

推荐指数☆☆☆☆☆

一、引言

自从第一个基于深度神经网络（Deep Neural Network，DNN）的AlexNet在2012年的ImageNet 竞赛上取得突破，为获得更好的性能，一些更复杂的DNN实例被提交到ImageNet 竞赛。

ImageNet 竞赛的终极目标是在多分类问题中获得最高的准确率（accuracy），而不管实际推理时间（inference time）。这样就导致了如下三个问题：a、参赛者会使用每副验证图像的多个相似实例去训练多个模型，通过模型平均或DNN集成，在付出推理时间或运算量显著提高的代价下，获得报告或论文中描述的足够高的准确率；b、不同提交作品使用不同次数的验证图像评估它们的模型，导致不同抽样方式或集成大小下模型准确率与报告中有偏差，即模型选择受到上述因素影响；c、实际应用中推理时间至关重要，影响着资源利用、功耗（power consumption）和延迟，而目前并没有加快推理时间的动机。

本文旨在就计算需要和准确率两方面，比较过去四年ImageNet竞赛中提交的当时最先进（state-of-the-art）的DNN架构。比较这些架构在实际开发中与资源利用率相关的多个指标：准确率、内存占用、参数、指令数（operations count）、推理时间和功耗（accuracy, memory footprint, parameters, operations count, inference time and power consumption）。文章主要目的是强调这些指标的重要性，因为它们是实际部署与应用中网络优化的硬性限制

二、比较方法

为了比较不同模型的质量，我们收集并分析了文献中的准确率，很快发现不同的抽样方法并不能直接比较资源利用率。比如，VGG-16和GoogleNet 的单次运行central-crop（top-5验证）误差分别是8.7%和10.07%，表明VGG-16性能优于googleNet；而采用10-crop抽样时，两者误差分别是9.33%和9.15%，VGG-16却比GoogleNet差了。正是基于如此，我们决定对所有网络统一采用单个central-crop抽样方法重新评估top-1准确率。

对于推理时间和内存占用，我们使用 cuDNN-v5和CUDA-v8配置的Torch 7来进行评估。所有试验使用的都是JetPack-2.3 英伟达 Jetson TX1板卡，该板卡内置了一块64位ARM A57CPU的视觉计算系统、一块 1T-Flop/s 的256核英伟达 Maxwell GPU和4GB的LPDDR4共享RAM。使用这种限量级的设备是为了更好地强调网络架构的差异，但是使用K40或Titan X等大多数最新的GPU可获得相似的结果。指令数使用我们开发的开源工具包评估。功耗评估使用的是 Keysight 1146B Hall电流探头，内置Keysight MSO-X 2024A 200MHz 数字显波器，采样周期2s，采样率50kSa/s，该系统由 Keysight E3645A GPIB数控直流电源供电。

三、测试结果

我们对比分析的DNN有：AlexNet（Krizhevsky, 2012）；批量标准化AlexNet（batch normalised AlexNet，BN-AlexNet，Zagoruyko, 2016）；批量标准化Network In Network（BN-NIN，Lin，2013）；ENet（Paszke，2016）for ImageNet（Culurciello，2016）；GoogLeNet（Szegedy，2014）；VGG-16和VGG-19（Simonyan & Zisserman，2014）；ResNet-18，ResNet-34，ResNet-50，ResNet-101和ResNet-152（He， 2015）；Inception-v3（Szegedy， 2015）和Inception-v4（Szegedy， 2016）。

为描述更直观起见，我们使用不同的颜色区分不同的架构和他们的作者，而同一个网络的色系则相同，如粉色系的都是 ResNet。

3.1 准确率

Fig1展示了提交给 ImageNet 挑战赛的各个网络架构的 1-crop 准确率，最左边的是 AlexNet，最右边的是 Inception -v4。最新的 ResNet 和 Inception 架构相比其他架构准确率至少高 7%。

详细的准确率值。最明显的是VGG，尽管它已经被广泛应用于许多应用中，但无论计算需求还是参数数量，都是迄今为止最昂贵。除VGG 的16层和19层的实现外，其他架构形成了一条斜线，到 Inception 和 ResNet 时，这条线开始变平缓。这表明这些模型在该数据集上达到一个拐点。在这个拐点上，计算复杂度的增加开始超过在准确率上获得的好处。

3.2 推理时间

Fig3展示了各架构在每个图像上的推理时间随batch大小的变化情况。从图中可以看出，VGG 处理一张图像所需时间约0.2秒，这限制了它在 NVIDIA TX1 上的实时应用；令人惊奇的是，AlexNet 随着batch大小从1变化到64，处理速度提高了3倍，这是由它的全连接层的弱优化导致（due to weak optimisation of its fully connected layers）。

深度学习 | 话题

3.3 功耗

Fig4展示了功耗随batch大小变化情况，可见大多数情况下功耗与batch大小无关。结合Fig3和Fig4可见，AlexNet （batch大小为1）和 VGG（batch大小为2）的低功耗与较慢的推理时间有关。

3.4 存储

Fig5展示了内存随batch大小变化情况，可见起初最大系统内存使用大小是不变的，内存占用随着batch大小的增大而增大。这是由于网络模型的初始内存分配以及批处理时的内存需求随着图像数量的增加而成比例的增大。由Fig6可见，对规模小于 100 MB的网络，初始内存分配不会小于 200 MB，而且随后呈现为参数大小的斜率为1.3的线性函数。

3.5 运算量

在神经网络加速器的自定义实现中，运算量对于预估推理时间和硬件电路大小是必要的。Fig7展示了运算量随推理时间和batch变化情况，当batch大小为16时，每个图像的运算量和推理时间之间存在线性关系。因此，在设计网络时可以控制运算量，以使处理速度保持在实时应用或资源有限的应用的可接受范围内。

3.6 运算量和功耗

Fig8展示了功耗和运算量的关系，可以发现不同架构之间没有特定的内存占用关系。当资源利用完全时，通常batch大小较大，所有网络的额外消耗大致为

系统功耗。如果功耗是我们要关注的点，那么可以选择满足最低功耗要求的最慢的架构。

3.7 准确率和吞吐量

Fig9展示了准确率与每秒推理数量的关系，可见准确率和每秒推理数量之间存在非凡（non-trivial）的线性上限。给定帧速率后，可以实现的最高准确率与帧速率成线性关系。准确率的线性拟合展示了所有架构的准确率与速度之间的关系。此外，给定一个推理时间，可以得出资源充分利用条件下理论上的最高准确率（chosen a specific inference time, one can now come up with the theoretical accuracy upper bound when resources are fully utilised）。当功耗固定时，我们甚至可以进一步得出能耗限定下的最高准确率，而这可以作为需要在嵌入式系统上运行的网络的基本设计因素考虑。

深度学习 | 话题

3.8 参数使用

DNN 在利用全部学习能力（参数数量/自由度）方面非常低效。Han在2015 年利用 DNN 的这个缺陷，使用权重剪枝（weights pruning）、量化（quantisation）和变长编码（variable-length symbol encoding）将网络规模减小了50倍。值得注意的是，使用更高效的架构能够产生更紧凑的表征。如Fig10所示，尽管VGG 比AlexNet准确率更高，但其信息密度更差，这意味着在VGG 架构中引入的自由度并没有带来准确率上的很大提升。此外，Fig10中可以看出，ENet分数最高，仅用了VGG 1/24的参数获得了state-of-the-art。

英文题目：AN ANALYSIS OF DEEP NEURAL NETWORK MODELS

深度学习 | 话题

原文连接：https://arxiv.org/pdf/1605.07678.pdf

互动：您觉得哪种DNN模型最好用，为什么？请留言探讨。

死磕自律，遇见更好的自己；认知升级，助你长出强两翼！

关注该百家号，一起创造奇迹。

相关推荐

AI杀入密码学：创造更恐怖的怪兽
小牛IT

三大银行（工行、建行、农行），新IT架构是啥样？
宇宙镜像宇宙

人工智能现在能“举一反三”学打游戏
东方头条

你家也有的古董PC处理器，如今炒到近万元
科技拌饭

查看更多

最新评论

还没有人评论，点击抢沙发 ~