检索报告

计算机学院 温冬 19023077

选题背景

过去五年，卷积神经网络在多个计算机应用领域取得了极大的成功[1]-[3]。以AlexNet[4]、VGG-16[5]、ResNet[6]等为代表的卷积神经网络在图像识别、人脸识别、目标检测、语义分割和自动驾驶等领域得到了广泛的应用并成为当前最主流、使用效果最好的方法。尽管在很多领域都成为了最佳算法，但卷积神经网络（CNN）的计算同时具备了访存密集和计算密集的特征限制了它在很多场景中的应用。CNN的出现给传统的计算机硬件平台带来了诸多挑战，其访存密集的特点使得处理器对带宽的要求超出了现有内存接口的限制、其计算密集的特性导致CNN的部署将在嵌入式处理器和CPU上变得极其困难。上述特点使得CNN的训练和推理大多在图形处理器（GPU）上运行。CNN中90%的操作可以被规约为矩阵/向量乘累加（MAC）操作，GPU平台拥有比CPU和嵌入式处理器更多的流处理单元和运算器，在处理MAC时可以拥有更高的并行度和计算效率，然而GPU昂贵的价格、巨大的功耗和芯片面积使得其功耗消费比很低，因此在一些对功耗和成本敏感的场合中，使用GPU来加速CNN任务是极其困难的事情。对CNN进行优化并将其运行在更廉价、更轻量和更快的硬件平台上成为了一个很有必要的工作。

检索策略

考虑到计算机专业的几个著名协会与出版集团覆盖了每年95%以上的计算机领域顶级论文，因此我们的主要检索对象就可以设定为IEEE Xplore数据库和ACM Library数据库。通过对该两大数据库的检索即可得到绝大多数已经发表的计算机顶刊、顶会论文。此外，arXiv作为一个著名的理工科领域开源预印本网站，也会有作者选择将自己未正式发表的论文公布在arXiv上。考虑到期刊和会议论文发表的滞后性，因此arXiv也可以作为一个很好的补充。通过国防科技大学远程图书馆网站我就可以得到所有需要的资源。

检索过程

1. 首先分别在两大数据库ACM Library 和IEEE Xplore上以关键词“prune”“hardware”进行检索；
2. 先过滤掉明显不属于计算机领域的文章；
3. 再过滤掉比较发表时间较早和发表于水平较低的期刊与会议上的文章；
4. 对于一文两投（先投快捷的会议、再投翔实但慢的期刊）的内容，只选择内容细节更多的期刊文献；
5. 通过国防科技大学远程图书馆网站我得到所有需要的文本资源。

综述简介

深度卷积神经网络在很多任务中取得了优秀的表现。然而卷积神经网络属于计算密集型、访存密集型的任务，对现有的硬件平台提出了很大的挑战。在CPU的硬件资源无法高效完成卷积神经网络计算和GPU过高的功耗与高昂的价格限制了卷积神经网络应用的背景下，FPGA和ASIC这类可配置硬件平台成为新的选择。通过软硬件协同的深度卷积神经网络剪枝方法，在削减卷积神经网络存储规模与计算规模的同时，优化了其硬件友好特性，极大的扩展了模型的应用场景。

拟投期刊

《计算机工程与科学》

方向对口，且为国内高水平计算机期刊。

检索总结

科大远程图书馆除了访问境外站点慢一点外，真的啥啥都好使，真的太棒了！不过在科学上网工具的加持下，科大图书馆访问境外资源也是很快的。