加速器：

【论文】Nair A, Srinivasan P, Blackwell S, et al. Massively Parallel Methods for Deep Reinforcement Learning[J]. Computer Science, 2015.

【面临问题】

提出了针对深度强化学习的大规模分布式架构

【解决方案】

该架构利用四个组件来实现——产生新行为的并行节点，利用存储的内容进行训练的并行节点，函数值表示或行为策略的分布式神经网络，分布式的内容存储。利用新架构实现Deep Q-Network。

【论文】Chen T, Zhang S, Liu S, et al. A Small-Footprint Accelerator for Large-Scale Neural Networks[J]. Acm Transactions on Computer Systems, 2015, 33(2):1-27.

【面临问题】

针对大型CNN和DNN的机器学习加速器性能受限

【解决方案】

设计了一个用于大型CNN和DNN的高通量加速器，特别强调记忆对加速器的设计，性能和能量的影响。

【论文】Tallada M G. Coarse grain parallelization of deep neural networks[J]. 2016, 51(8):1-12.

【面临问题】

描述并实现以及分析DNN训练算法的网络不可知且收敛不变的粗粒度并行化。

【解决方案】

通过批次层级的并行机制实现粗粒度的并行化，并利用MNIST和CIFAR-10实验来进行验证粗粒度并行化的加速效果。

【论文】Braun G, Pokutta S, Zink D. Lazifying Conditional Gradient Algorithms[J]. 2016.

【面临问题】

有条件的梯度算法在实际的许多情况下，其实现的代价太高。

【解决方案】

提出了一种通用的松弛条件梯度算法，该算法在实际计算中会带来几个数量级的Wall-Clock的加速效果。

通信时间优化：

【论文】Zhang S, Choromanska A, Lecun Y. Deep learning with Elastic Averaging SGD[J]. 2015:685-693.

【面临问题】

在通信约束条件下实现深度学习并行计算的随机优化问题。

【解决方案】

提出了一种算法，使本地worker能够进行更多的探索，即允许变量可以发散性的从中心值进行探索，并同时减少本地worker与master之间的交互。该方法与常见的ADMM比较，得到更快的加速效果，通信效率也很快。算法中还提及了基于梯度动量的版本。

【论文】Iandola F N, Moskewicz M W, Ashraf K, et al. FireCaffe: Near-Linear Acceleration of Deep Neural Network Training on Compute Clusters[J]. 2015, 37:2592-2600.

【面临问题】

分布式算法的速度和可扩展性总是受到服务器之间通信开销的限制，如何减少通信开销并保证训练的DNN模型的精度不降低。

【解决方案】

提出了新的方法，其中有三个关键的地方。首先，选择在GPU服务器之间实现高宽带的网络硬件。其次，考虑其他的通信算法，发现reduction tree比参数服务器方法更加高效并且可扩展性也更好。第三，选择增加批量大小以减少DNN训练期间的通信总量，并且确定超参数，使我们能够在大批量训练的同时实现小批量的准确性。

【论文】Wang J, Kolar M, Srebro N, et al. Efficient Distributed Learning with Sparsity[J]. 2016.

【面临问题】

提高大规模并行计算中由于通信频繁带来的冗余开销。

【解决方案】

提出了一种新的，高效的分布式稀疏学习方法，该方法利用机器间的随机分块观测进行分布式稀疏学习。在每一轮迭代中，worker节点计算数据loss值的梯度，master采用L1正则化进行loss值最小化的计算。

【论文】Cho M, Brand D. MEC: Memory-efficient Convolution for Deep Neural Network[J]. 2017.

【面临问题】

卷积已是当下深度神经网络中的关键组件，但无论何种卷积方法，都带来了巨大的内存开销，从而导致性能下降并且通常的做法是在性能和内存损耗之间提供较差的折中。

【解决方案】

提出了一种内存高效的卷积（MEC），能够减少内存的开销，实质上加速了卷积的过程，MEC以一种有效的方法减小了输入的矩阵（降低内存开销），然后并行地执行多个小矩阵乘法以此完成卷积操作。MEC可以显著降低内存消耗，同时提高加速性能。

梯度评估与版本控制：

【论文】Asynchronous Parallel Greedy Coordinate Descent

【面临问题】

现有的异步梯度下降方法由于其机制的问题，导致模型准确率收敛比较慢。

【解决方案】

提出了一种对有界约束的光滑函数最小化的异步并行贪心坐标下降算法。最后实验证明该算法具有良好的加速性能，并且优于现有的多核支持向量机的解决方案。

【论文】Chen C Y, Choi J, Brand D, et al. AdaComp : Adaptive Residual Gradient Compression for Data-Parallel Distributed Training[J]. 2017.

【面临问题】

高性能计算平台上的大规模分布式深度神经网络训练受到严重的通信限制。为了克服这个限制，需要新的梯度压缩技术。

【解决方案】

介绍一种新技术——自适应残差梯度压缩方法，AdaComp可以基于本地的梯度残差，并根据本地的活跃性调整压缩率，从而自适应地进行梯度压缩。并且证明该方法在全连接层和复发层的压缩率达到200倍，卷积层的压缩率达到40倍，伴随着梯度的压缩，模型的准确率没有出现下降。

【论文】Lian X, Zhang H, Hsieh C J, et al. A Comprehensive Linear Speedup Analysis for Asynchronous Stochastic Parallel Optimization from Zeroth-Order to First-Order[J]. 2016.

【面临问题】

异步并行优化可以给我们带来多大的加速收益。

【解决方案】

提供一个全面而通用的分析，以研究从零阶到一阶方法的各种异步并行随机算法的加速性能，其结果改善了对特殊情况的现有分析，为理解异步并行行为提供了更多见解，并首次提出了一种新颖的异步并行零阶方法。

【论文】Han J, Liu Q. Bootstrap Model Aggregation for Distributed Statistical Learning[J]. 2016.

【面临问题】

在分布式或隐私保护学习中，需要利用一组概率模型估计来自不同的本地存储库，并要求将它们合并成一个模型，以提供有效的统计估计。一种简单的方法是对局部模型的参数进行线性平均，然而，对于非凸模型或具有不同参数维度的模型，这些参数趋向于退化或不适用。另一个更实用的策略是从局部模型生成自举样本，然后基于组合自举集合学习联合模型。 不幸的是，自举程序引入了额外的噪音，并可能显著恶化性能。

【解决方案】

提出了两种方法来降低自举噪声，其中包括一个加权的M估计器，其在统计上是有效的并且效果十分明显。

【论文】Reddi S J, Hefny A, Sra S, et al. On Variance Reduction in Stochastic Gradient Descent and its Asynchronous Variants[J]. Evidence Report/technology Assessment, 2016, 31(183):2647-2655.

【面临问题】

研究基于方差的随机梯度下降算法，通过SAG,SVRG,SAGA等算法的发展，这些算法已经被证明在理论上和实践上都优于SGD。然而这些算法的异步版本还没有得到广泛研究。缺少统一的实验框架是主要问题。

【解决方案】

提出了一种基于框架的异步算法，并证明了它的快速收敛性。

【论文】Jiang J, Cui B, Zhang C, et al. Heterogeneity-aware Distributed Parameter Servers[C]// ACM International Conference. ACM, 2017:463-478.

【面临问题】

在异构环境下的分布式机器学习，现有的分布式随机梯度下降系统在同构环境中工作得很好，但在异构环境中，他们的同步协议不适合异构设置，所以掉队者很常见，因此它们的性能可能会下降，有时甚至达到10%。

【解决方案】

提出一种异构感知算法，它在进行全局更新前不间断地进行训练，抑制陈旧节点对收敛性的影响。以及考虑到每次更新的陈旧信息采用更为复杂的学习速率自适应机制。理论上证明该两种方法的有效融合。比原型系统快了2-12倍。