発表概要

深層学習フレームワークにおけるマルチコアCPU向け計算グラフスケジューリング

樋口 兼一[1,a)]　田浦 健次朗[1,b)]

2019年7月26日発表

Chainer を始めとする多くの深層学習フレームワークは，ニューラルネットワークに含まれる層ごとの処理をノード，各層間の接続関係をエッジとした計算グラフを内部的に構築し，各ノードを逐次に実行することによりネットワークの学習を行う．そのようなフレームワークにおけるホットスポットは各ノード内部の処理であり，既存の研究・実装はノード単位の高速化に焦点を当ててきた．しかし，広く用いられているネットワークモデルの多くでは各処理が軽量であるために，マルチコア CPU 上のすべてのコアを効率的に利用できず，結果として使用するコア数の増加に対して実行速度の向上は乏しい．そこで CPU 上の複数のコアを複数のグループにまとめ，互いに独立な複数のノードを各グループに配分して実行することにより，ノード内・ノード間それぞれで並列処理を行う手法を提案する．使用可能な総コア数を超えないよう同時に実行するノード数に制限をかけつつも，可能なものから積極的に実行開始するというスケジューリングを行う．並列処理可能な分岐を持つ計算グラフが生成される深層学習モデルは ResNet を始めとして複数あり，そのようなネットワークにおける推論・学習の高速化が期待される．実装は Chainer を用いて行い，ノード単位の高速化も含めた総合的な性能向上に対する評価を複数のモデルに対して行う．

Presentation Abstract

Scheduling Computation Graphs of Deep Learning Frameworks for Multi-core CPUs

Tomokazu Higuchi[1,a)]　Kenjiro Taura[1,b)]

Presented: July 26, 2019

Many deep learning frameworks, including Chainer, train a neural network by processing a calculation graph with layers in the network as nodes and connections between them as edges sequentially. Performance hotspots of such frameworks are calculations in each node and the existing work has focused on speed-up them. However, nodes in widely used network models cannot efficiently utilize all available cores on a multi-core CPU because calculations in such nodes are too lightweight for such a CPU. As a result, an improvement in execution speed is relatively poor against an increase in the available number of cores. Therefore, we propose a method to parallelize several node executions in addition to an internal node parallelization by allocating executable several nodes to each core group consisting of multiple cores on the CPU. A scheduler positively tries to assign as many executable nodes as possible to each core group with a restriction of the whole number of cores. There are multiple deep learning models, such as ResNet, in which computation graphs with nodes that can be processed in parallel are generated. Then we expect there is a speeding up of inference and learning in such networks. The system is built on Chainer, and evaluations for overall performance improvement including speeding up on each node are conducted in several models.

This is the abstract of an unrefereed presentation, and it should not preclude subsequent publication.

[1] 東京大学大学院情報理工学系研究科電子情報学専攻
Information and Communication Engineering, Graduate School of Information Science and Technology, The University of Tokyo, Bunkyo, Tokyo 113–8656, Japan

[a)] thiguchi@eidos.ic.i.u-tokyo.ac.jp
[b)] tau@eidos.ic.i.u-tokyo.ac.jp