先端科学技術研究科　修士論文要旨

所属研究室 （主指導教員）	コンピューティング・アーキテクチャ （中島　康彦　（教授））		
学籍番号	2211336	提出日	令和 6年 1月 10日
学生氏名	KIM DOHYUN		
論文題目	リニアアレイ型 CGRA によるGCN の電力効率改善と評価		

要旨

Graph Neural Network(GNN)は、グラフ構造のデータを分析するためのニューラルネットワークの総称で、近年は分子やたんぱく質構造の分析といった自然科学はもちろん、論文同士の関係分析や商品との関係性、またネットワークの障害検知など、様々な分野においての活用が試されている。

しかし、先ほど例に挙げたネットワークの障害検知など、リアルタイム性かつエッジでの処理が求められる場合、消費電力の削減やその性能が課題となる。GNNは、入力データとモデルの特性から、多大なる計算資源を必要とされる。このような現状において、CPUやGPUのような既存の計算資源の場合、電力効率よくGNNの演算を行うのは困難であるため、
新たな計算基盤の採用が望ましいとされる。専用ハードウェアも選択肢の一つになり得るが、プログラミングによる他演算の実行が困難であるため、変化の激しいニューラルネットワークの実装に用いることは将来の運用の観点からすると望ましくない方法である。

その課題を解決するため、プログラミング可能で他モデルでも用いることができるCGRAによるGNNの実装を検討した。CGRAは、粗粒度再構成可能アーキテクチャのことで、複数のプログラミング可能なプロセッシングエレメント(PE)が繋がり、メモリに中間結果を書き戻すことなく連続して演算を行うことができる利点がある。

本研究では、リニアアレイ型CGRAにおけるGNNの一種であるGraph Convolutional Network(GCN)の高効率な実装方法を考案し、実装を行った。グラフデータを行列で表すと、その特性から疎行列になることが多く、GCNのアルゴリズムから特徴行列や重みが密行列になることが多い。そのことから、GCNの実装は疎行列−密行列積(SpMM)がメインとなる。

SpMMのリニアアレイ型CGRAにおける演算アルゴリズムを考案し、二通りのデータ格納方式を試し、最良とされる方法を採用した。Intel i9−10920X上でOpenMPによる実装とNVIDIA Jetson AGX OrinとRTX 3090上のcuSPARSE及びcuBLASによる実装を比較対象とし、実データセットにおける電力効率と単位性能当たりのメモリバンド幅を比較した。その結果、Jetson AGX Orinに対し、エッジで想定されるグラフデータにおいて3.64倍の電力効率を達成した。