概要

ソーシャルネットワークを始めグラフ構造で表現可能な様々な関係に対して, クラス分類／回帰問題等を扱うグラフニューラルネットワーク (GNN) およびグラフ畳み込みニューラルネットワーク (GCN) が注目されている. ハードウェア実装して推論／トレーニングの速度を改善したり, 消費電力の削減をするためには, 入力として与えられるグラフを行列形式で表現した際に, 非ゼロの要素のデータ分布が非常にアンバランスである点, および, 行列サイズが非常に大きい点を考慮した設計をする必要がある. GNN ／ GCN の要素技術についてこれまで数々の研究がなされてきたが, GNN ／ GCN のハードウェア実装に関する知識の整理が不完全であるという課題がある. 本研究では, GNN ／ GCN のハードウェア実装に特に注目して, 関連するこれまでの研究について, 分類および有効な最適化手法を明らかにすることにより知識の整理を行った.

GNN ／ GCN のハードウェア実装は, 推論レイテンシの削減, 消費電力の削減, メモリアクセスの効率化等, 様々な最適化目標を持っている. また, 対象とする GNN の処理フェーズやデータセット, ネットワーク等, 前提とする条件も論文により大きく異なる. 既存研究におけるこれらの実装の傾向を明らかにするために, 各手法において対象とするフェーズ, 目的, データセット, ネットワーク, デバイス, および量子化精度に基づいて整理した. GNN ／ GCN のハードウェア実装の対象フェーズは, ①推論 (推論のみに対応) , ②トレーニング (トレーニングのみに対応), ③推論とトレーニング (推論とトレーニングの両方に対応) の 3 種類に分類される. ハードウェア実装の全体的な傾向として, フェーズについては, 推論フェーズのみを対象とする論文が 75 本, トレーニングフェーズのみを対象とする論文が 20 本, 推論／トレーニングの両方のフェーズを対象とする論文が 5 本であり, 推論フェーズを対象とする論文数が非常に多い. これは, 推論は訓練済みモデルを実際のアプリケーションで使用する段階であり, 多くの実際のアプリケーションで実行するにあたって低遅延の処理が求められていて, 推論の高速化に対する需要がトレーニングと比べて高いことが要因として考えられる. 目的については, 大半の研究がレイテンシと消費電力の削減の 2 つを挙げているが, 他にも, オフチップメモリアクセスの削減やチップ面積の削減, オフチップメモリ帯域幅使用量の削減を目的とする既存研究もある. データセットについて, ①推論の対象データセットでは, 上位 4 つの Cora, Reddit, Citeseer, Pubmed を対象としている論文は 40 本以上ありこれらが大半を占めていた. ②トレーニングのみを対象とするハードウェア実装について

は, ①と同様 Reddit と, Yelp, OGBN-products, Amazon が上位 4 つであった. ③推論／トレーニングの両方を対象とするハードウェア実装についても, Reddit, Cora, Citeseer, Pubmed 等, ①／②と同様のデータセットが採用されている. ネットワークについては GCN が極めて多く採用さていて, デバイスについては大半が ASIC あるいは FPGA である. 量子化精度については, 論文中で明示されていないものの中では特徴ベクトル／重みベクトルにおいて 32bit 固定小数点／ 32bit 浮動小数点が多く採用されている.