

大規模グラフニューラルネットワーク推論性能の飛躍的向上 ー不規則なメモリアクセスの解消により、計算速度と効率化を両立ー

【ポイント】

- 大規模なグラフニューラルネットワーク（GNN）推論における大幅な高速化を可能にするスケーラブルな AI アクセラレータを実現。
- GNN 推論中の不規則メモリアクセスをほぼ完全に解決し、計算コストを大幅に削減可能。
- 自動運転や広告推薦システムのようなリアルタイムかつ大規模なアプリケーションへの活用に期待。

【概要】

東京科学大学（Science Tokyo）総合研究院 AI コンピューティング研究ユニットの藤木大地准教授と Jiale Yan（ジャロ・ヤン）ポスドク研究員（当時）らの研究チームは、大規模なグラフデータを効率的に処理できる新しい AI アクセラレータ「BingoGCN」を開発しました。

グラフニューラルネットワーク（GNN、用語 1）は、人と人のつながりや道路網のような、複雑な関係性を表す「**グラフデータ（用語 2）**」を扱うのに優れた AI 技術です。しかし、扱うデータが大きくなると、計算に必要な情報量が巨大になり、計算処理量・消費エネルギーが増大する課題がありました。

今回開発した BingoGCN は、この課題を解決するために、CMQ と **Strong Lottery Ticket（SLT）理論（用語 3）**という二つの新しい技術を搭載しています。CMQ は、グラフデータを小さな塊（パーティション）に分割して処理する際に、パーティション間でやり取りされる情報を「**オンラインベクトル量子化（用語 4）**」という技術で要約します。これにより、不規則なメモリアクセスを大幅に減らし、処理の精度を保ちながら効率を向上させます。また、SLT 理論では AI モデルの計算に必要な「重み」と呼ばれるパラメータを、必要な時に必要な分だけチップ上で生成します。これにより、メモリ使用量を削減し、計算効率を大幅に高めます。BingoGCN は、これらの技術を組み合わせることで、従来難しかった細かく分割されたグラフデータ上でも、高い効率で GNN の推論（用語 5）を行うことを可能にしました。

この成果は、ソーシャルネットワーク分析や自動運転など、私たちの身の回りのさまざまなサービスを支える GNN の性能を飛躍的に向上させるものです。

本成果は、2025 年 6 月 21 日～25 日に東京で開催される計算機アーキテクチャ分野の最高峰会議である「International Symposium on Computer Architecture (ISCA '25)」にて発表されます。

●背景

近年、ソーシャルネットワーク分析や広告推薦システム、自動運転など、複雑で非構造的な関係性を持つデータ(グラフデータ)を処理・解析する必要のあるさまざまな分野で、GNN という AI 技術の活用が急速に進んでいます。GNN は、こうしたグラフデータの関係性を捉え、分析・予測する時に非常に有効です。

しかし、この GNN を実際のシステムで活用する際、特に大規模なデータをリアルタイムに処理しようとする、いくつかの大きな課題に直面します。具体的には、GNN が扱うデータ(ノード特徴量(用語 6)や隣接行列(用語 7)といったグラフ構造を表す情報)は非常にサイズが大きくなることが挙げられます。そのため、データをメインメモリから読み書きする際の通信量や、チップ上で一時的にデータを保存しておくバッファのサイズが膨大になり、処理速度の低下やコストの増大を招きます。この問題は、メモリアクセスパターンが不規則である GNN において特に顕著に表れます。

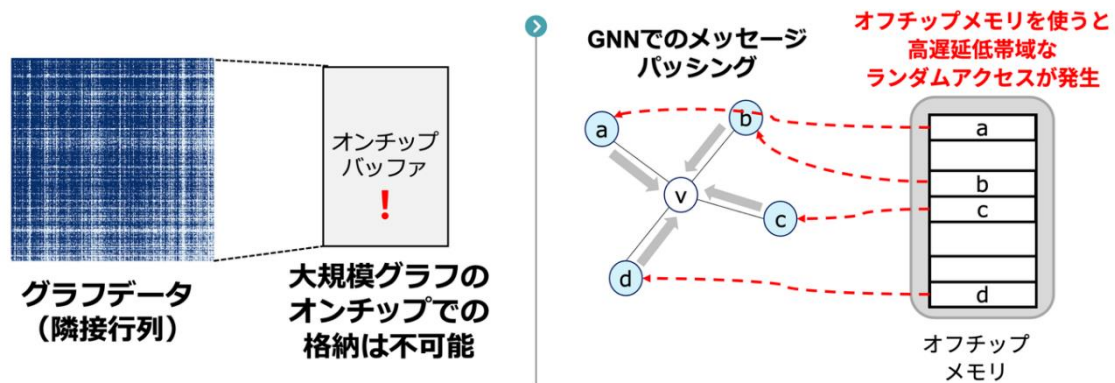


図1 大規模グラフにおけるリアルタイム GNN 処理の課題

グラフのデータサイズが膨大となる問題を解決するための一つのアプローチとして、「グラフパーティショニング」(グラフ分割とも呼ばれる)という手法が挙げられます。これは、巨大なグラフを結合が比較的密な小さな塊に分割して処理することで、メモリアクセスを局所化し、チップ上のバッファサイズを削減しようとするものです。しかし、この分割を細かくしすぎると(細粒度パーティショニング)、分割された塊同士を繋ぐエッジ(接続情報)の数が増え、結果的にチップ外のメインメモリへのアクセス回数が増加してしまい、全体の性能をかえって悪化させてしまうというジレンマがありました。

このような背景のもと、本研究では、大規模な GNN 推論を効率的かつスケーラブルに行うための新しいフレームワーク「BingoGCN」を提案しています。BingoGCN は、特に細粒度パーティショニングの課題を克服し、メモリアクセスの非効率性を解消することを

目指しています。

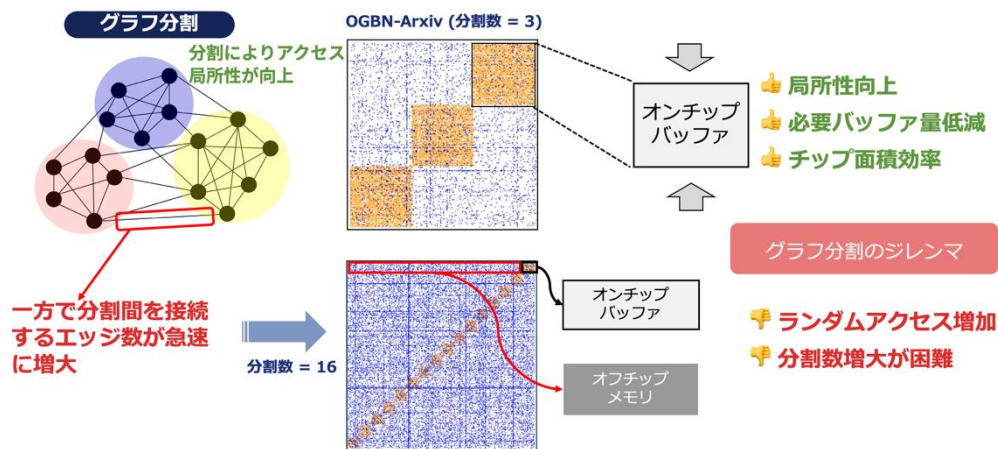


図2 グラフ分割（パーティショニング）手法とそのジレンマ

●研究成果

この研究では、「BingoGCN」を用いて、GNNをより速く、効率的に作動させることに成功しました。特に、大規模なグラフデータを扱う上での課題であったメモリ効率と処理速度を大幅に改善し、最新の研究と比較して最大65.7倍の高速化と最大107倍のエネルギー効率向上を達成しました。すなわち、メモリアクセスを大幅に削減しながらも、高い計算精度を維持できることが示されました。それぞれの成果の詳細について、下記の通り示します。

1. 不規則なメモリアクセスの大幅削減を実現する「Cross-Partition Message Quantization (CMQ)」

従来のGNNアクセラレーション（高速化・効率化）では、グラフデータを細かく分割して処理しようとする、分割された区画間で大量のデータ（メッセージ）交換が必要になり、これがチップ外の低速なメインメモリへのアクセスを頻発させ、大きなボトルネックとなっていました。特にこのメモリアクセスは不規則なパターンで行われるため、予測が難しく効率を著しく下げていました。

BingoGCNでは、この問題を解決するために「Cross-Partition Message Quantization (CMQ、クロスパーティションメッセージ量子化)」という独創的な技術を導入しました。CMQは、分割された区画を越えてやり取りされるメッセージを、多次元的にかつ動的に集約・要約するオンラインベクトル量子化技術です。具体的には、隣接する区画から送られてくる大量のノード特徴量をそのまま全て転送するのではなく、その特徴量を代表するような、より小さな情報（量子化されたメッセージ）に変換して伝達します。これにより、区画間の通信データ量を大幅に削減し、オフチップメモリ（チップ外メモリ）へのアクセ

ス回数そのものを劇的に減らすことができます。さらに、この量子化処理によってデータのパターンが整流化されるため、より効率的なメモリアクセスが可能になり、不規則なオフチップメモリアクセスを効果的に排除することができます。

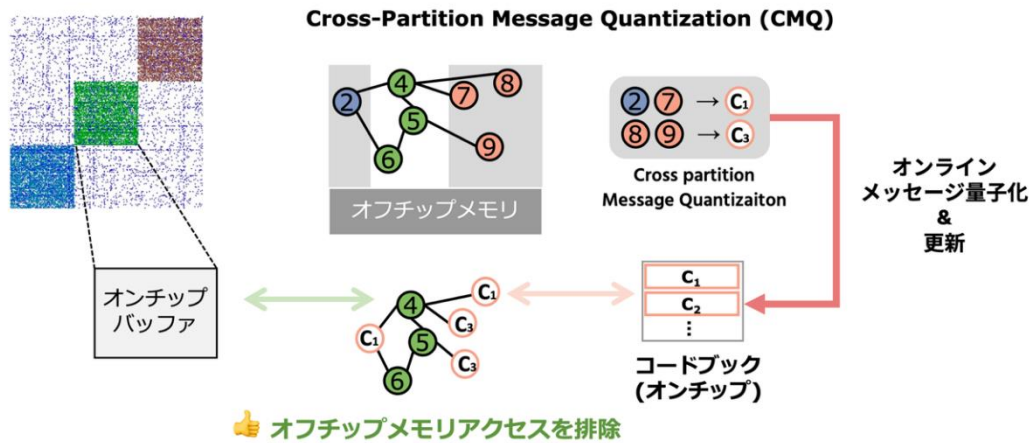


図3 Cross-Partition Message Quantization (CMQ)

2. 重みが不要で高速処理を可能にする「Strong Lottery Ticket (SLT) 理論」の応用

通常、GNN が高い性能を発揮するためには、事前に大量のデータを用いてモデルを学習させ、最適化された重み（パラメータ）を使用して計算を行う必要があります。このプロセスには多大な計算資源と時間が必要となることが一般的です。

BingoGCN は、この課題に対し、「Strong Lottery Ticket (SLT) 理論」というアプローチを採用しました。SLT 理論は、大規模なニューラルネットワークにおいて、初期のランダムな重みのままでも、まるで「当たりくじ」を引いたかのように、学習済みのネットワークと同等の性能を発揮する小さなサブネットワークが存在するという考え方です。BingoGCN は、この SLT 理論を活用することで、学習済みの重みを使用することなく、ランダムに生成された重みを用いた場合でも、高い GNN の推論精度とモデルのスパース性（計算に関与するパラメータが少ないこと、つまり効率が良いこと）を両立できることを示しました。これにより、ボトルネックがメモリから計算に移った場合でも、効率的な処理基盤を構築することが可能になります。

3. 高い精度を維持しつつ、メモリ効率を飛躍的に向上

上記の CMQ による通信量削減と、SLT 理論の応用によるモデルの軽量化・効率化を組み合わせることで、BingoGCN は GNN のタスクにおける分類精度などの性能を損なうことなく、メモリアクセスの総量を大幅に削減することに成功しました。私たちの研究では、既存の手法と比較して、メモリアクセスを著しく削減しつつ、同等以上の計算精度を達成しています。これは、限られたメモリ容量やメモリ帯域幅を持つハードウェア上でも、より大規模で複雑なグラフデータを扱えるようになることを意味し、エッジデバイス（セン

サーやスマートフォンなど、ネットワークの末端に位置する機器）での高度な GNN 応用への道を開くものです。

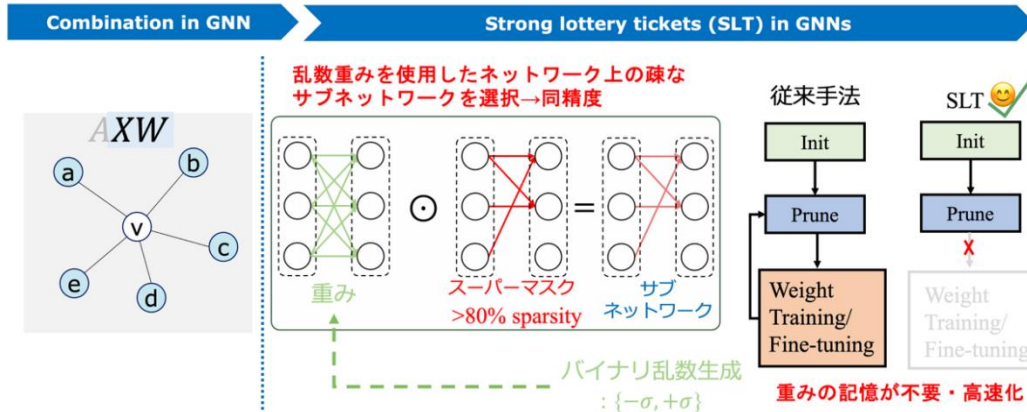


図4 Strong Lottery Ticket (SLT) 理論の応用

4. FPGA への実装による実用性の実証

提案された BingoGCN のアーキテクチャは、理論上の有効性だけでなく、実際のハードウェアにおける実用性も検証されています。研究チームは、BingoGCN の主要な機能を FPGA (Field-Programmable Gate Array) 上に実装しました。FPGA は、設計者が後から回路構成を書き換えられる集積回路であり、特定の用途に特化したアクセラレータのプロトタイピングや実装に適しています。この FPGA 実装を通じて、BingoGCN が実際にハードウェア上で効率的に動作し、期待される性能向上（高いスループットや低い処理待ち時間など）を達成できることが示されました。これは、BingoGCN の設計が現実のシステムに組み込まれ、実世界の問題解決に貢献できる可能性を具体的に裏付ける重要な成果です。

これらの成果は、GNN をより身近で強力なツールとして、ソーシャルネットワーク分析、推薦システム、交通流最適化、創薬といった多岐にわたる分野で活用していくための基盤技術として、今後の発展が大きく期待されます。

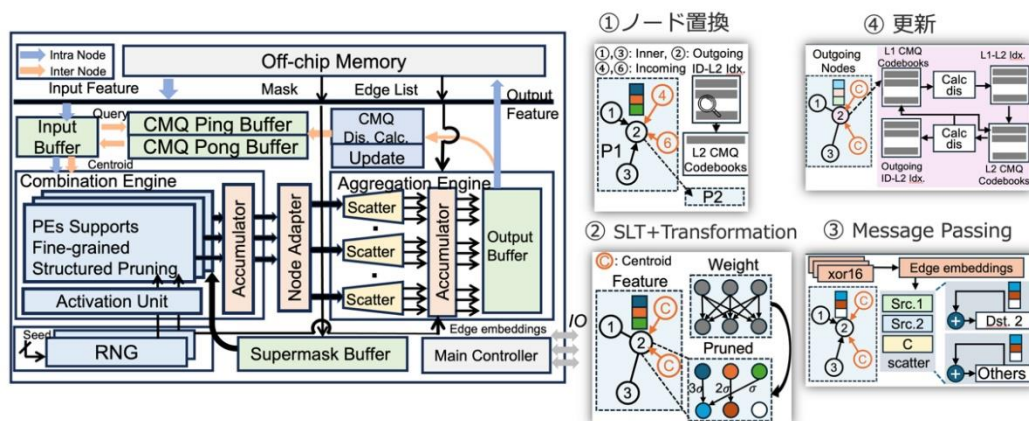


図5 BingoGCN アーキテクチャ

●社会的インパクト

BingoGCN はグラフをもとにしたリアルタイム推論を通してさまざまな応用分野に貢献することが期待されます。

例えば、Facebook、X（旧 Twitter）、LINE のようなソーシャルネットワーキングサービスでは、ユーザー間のつながりや情報の拡散といった複雑な関係性をグラフデータとして扱います。BingoGCN のような効率的な GNN アクセラレーション技術は、これらの巨大なグラフデータを高速に分析し、例えば、コミュニティの発見、パーソナライズされた推薦、影響力のあるユーザーの特定、フェイクニュースの拡散パターンの解析などに役立ちます。これにより、より健全で活発なオンラインコミュニティの形成や、経済活動の活性化、情報操作への対策強化などが期待できます。

また、自動運転システムでは、車両のセンサーが収集する周囲の環境情報（他の車両、歩行者、道路標識、信号など）や、それらの間の関係性をリアルタイムに認識・予測する必要があります。これらの関係性はグラフ構造としてモデル化でき、GNN による解析が有効です。BingoGCN によって GNN の推論が高速化・効率化されれば、自動運転車がより複雑な交通状況を瞬時に理解し、安全な判断を下す能力が向上し、交通事故の削減や交通渋滞の緩和といった社会的な課題解決に貢献する可能性があります。

更に、タンパク質の構造解析、化合物と病気の関連性予測、遺伝子ネットワークの理解など、医療や創薬の研究開発では複雑な関係性を持つデータが数多く存在します。GNN はこれらの分野で新たな発見を促進するツールとして期待されています。BingoGCN により、より大規模で複雑な生体関連グラフデータを効率的に処理できるようになれば、新薬の開発期間短縮や個別化医療の進展に貢献する可能性があります。

●今後の展開

本研究で提案した技術は、特に大規模なグラフデータをリアルタイムで処理する必要があるさまざまな分野での社会実装が期待されます。この実現のため、今後は本技術のさらなる効率化と適用範囲の拡大を模索していきます。

例えば、現在の手法では、グラフをあらかじめ分割しておくオフライン分割処理が必要ですが、リアルタイム性が求められるストリーミングデータや、刻々と構造が変化する動的グラフへの対応を強化するため、オンラインでの効率的なグラフ分割手法の開発を進めていきます。また、BingoGCN のアイデアは、計算処理にボトルネックをシフトさせることでスケーラビリティを向上させるものであり、これは他のアーキテクチャにも広く応用可能です。特に、複数の計算機で処理を分担する分散 GNN 環境において、通信コストを最小化する効率的なフレームワーク構築への展開が期待されます。

●付記

本研究は、科学技術振興機構（JST）戦略的創造研究推進事業 さきがけ（課題番号：JPMJPR22P7）、同事業 ALCA-Next（先端的カーボンニュートラル技術開発）（課題番号：JPMJAN24F3）、日本学術振興会（JSPS）科学研究費助成事業（課題番号：JP23H05489）などの支援により実施されました。

【用語説明】

- (1) **グラフニューラルネットワーク (GNN)**：グラフデータを効率的に学習・処理するために設計されたニューラルネットワークの一種。
- (2) **グラフデータ**：点（ノード）とそれらを繋ぐ線（エッジ）で構成され、物事の関係性や繋がりを表すデータ形式。
- (3) **Strong Lottery Ticket (SLT) 理論**：大規模なニューラルネットワークの中には、初期状態のままでも高い性能を発揮する小さな部分ネットワーク（当たりくじ）が存在するという理論。
- (4) **オンラインベクトル量子化**：データ処理中に、複数の数値の組（ベクトル）を代表的な値に置き換えることで情報量を圧縮する技術。
- (5) **推論**：新しいデータに対して予測や分類などの判断を下す処理。
- (6) **ノード特徴量**：グラフデータにおける各点（ノード）が持つ情報や属性。
- (7) **隣接行列**：グラフデータにおいて、どのノード同士が繋がっているかを行列で表したもの。

【論文情報】

学会名：The 52nd Annual International Symposium on Computer Architecture (ISCA '25)

日時：2025 年 6 月 21 日～25 日

論文タイトル：BingoGCN: Towards Scalable and Efficient GNN Acceleration with Fine-Grained Partitioning and SLT

著者：Jiale Yan, Hiroaki Ito, Yuta Nagahara, Kazushi Kawamura, Masato Motomura, Thiem Van Chu, Daichi Fujiki

DOI：10.1145/3695053.3731115

【研究者プロフィール】

藤木 大地（フジキ ダイチ） Daichi FUJIKI

東京科学大学 総合研究院 AI コンピューティング研究ユニット 准教授

研究分野：計算機アーキテクチャ、機械学習

ジャロ・ヤン Jiale YAN

東京科学大学 総合研究院 AI コンピューティング研究ユニット ポスドク研究員(当時)
研究分野：計算機アーキテクチャ、機械学習

本村 真人 (モトムラ マサト) Masato MOTOMURA

東京科学大学 総合研究院 AI コンピューティング研究ユニット 教授
研究分野：集積回路、機械学習

ティエム・ヴァン・チュ Thiem Van CHU

東京科学大学 工学院 情報通信系 准教授
研究分野：リコンフィギャラブルシステム、機械学習

【お問い合わせ先】

(研究に関すること)

東京科学大学 総合研究院 AI コンピューティング研究ユニット 准教授
藤木 大地

Email : dfujiki[at]artic.iir.isct.ac.jp

TEL : 045-924-5658

(報道取材申し込み先)

東京科学大学 総務企画部 広報課

取材申し込みフォーム : <https://forms.office.com/r/F3shqsN7zY>

Email : media[at]adm.isct.ac.jp

TEL : 03-5734-2975 FAX : 03-5734-3661



科学技術振興機構 広報課

Email : jstkoho[at]jst.go.jp

TEL : 03-5214-8404 FAX : 03-5214-8432

(JST 事業に関すること)

科学技術振興機構 戦略研究推進部 ICT グループ

櫻間 宣行

Email : presto[at]jst.go.jp

TEL : 03-3512-3526 FAX : 03-3222-2066