審 査 の 結 果 の 要 旨

氏　　名　　日髙　雅俊

本論文は「ヘテロジニアス計算環境における深層学習基盤の研究」と題し，ハードウェア・オペレーティングシステム等が異なる計算機群をヘテロジニアス計算環境と呼称し深層学習を行うソフトウェアに必要な基盤技術の提案を行ったものであり，全７章からなる．深層学習基盤は主に３つの要素からなる．１）畳み込み層などの行列計算をGPU等のアクセラレータを利用して行う機構，２）多様なモデルを効率よく実行するためのDNN実行制御を行う機構，３）分散計算における同期，アプリケーションにおけるモデルの配布に必要な通信を行う機構である．既存のアルゴリズム，ソフトウェア実装はクラウド等で提供されている均質な計算機群を用いることを想定しており，ヘテロジニアス計算環境において新たに生じる課題に対処する手法を開発することが本研究の目的である．ヘテロジニアス計算環境にはスマートフォンやIoTデバイス等が考えられるが，本研究では具体的な環境として多くのOSで標準的にサポートされているWebブラウザ上で動作するソフトウェア実装を行うこととしている．アクセラレータ利用については複数の規格が並立しているGPUインターフェースに対し，一種類の行列計算アルゴリズム実装を自動的に変換する機構を提案し統一的な利用を可能とした．DNN実行制御においては，既存の各種深層学習フレームワークにおいて定義されたモデルが計算グラフのレベルで共通化できることを具体的な中間表現の実装を通じて示し，さらに属性情報のパターンマッチングにより実行速度を高速化した．分散計算における計算機間の同期の課題は，従来システムで主流である均一な計算負荷の配分では高速な計算機の計算能力を使いきれないという課題に対し，計算機の性能・台数によらず同期待ちが不要となるアルゴリズムを提案した．アプリケーション配布時の通信は学習済みモデルを圧縮して転送する場合に，通信回線の速度により精度とデータ容量の適切なトレードオフが変化するという課題に対し，完全なモデルを小容量モデルと補完のための追加情報に分解し逐次的に転送する手法を提案し解決した．

第１章「序論」では深層学習の計算資源として最も典型的であるクラウド環境を利用することの問題点を指摘し，従来使われてこなかったスマートフォン等の環境が均質でない計算資源を利用する意義を説明し，本研究の目的を示している．

第２章「Deep Neural Network」では後続の章のアルゴリズムの説明の補助とするため，深層学習の中核となるDeep Neural Networkモデルの基礎を説明している．

第３章「ヘテロジニアス計算環境における深層学習基盤」では深層学習基盤の構成および，それをヘテロジニアス計算環境において実現するための課題について論じたのち，具体的な実装環境であるWebブラウザ上で科学技術計算を実装する手段について解説している．

第４章「WebブラウザをDeep Neural Networkの分散計算」では機種・性能の異なる複数の計算機を連携させ深層学習を分散計算させるという目標を置き，異なるアクセラレータAPIを提供する計算機群に対して統一的な行列計算アルゴリズムの実装手法および計算機の性能差・台数差にロバストな計算負荷の配分手法を提案した．

第５章「WebブラウザのためのDeep Neural Networkの圧縮」では学習済みDNNを用いて推論を行うアプリケーションを様々な通信環境にある計算機へ配布する状況において，モデル転送の時間を短縮するという目標を置き，回線速度にロバストな圧縮手法を提案した．

第６章「WebブラウザにおけるDeep Neural Networkの推論高速化」では学習済みDNNを深層学習用に設計されていないハードウェア上においても高速に動作させるという目標を置き，DNNの計算グラフに対して最適化を加えることが可能な中間表現を提案し，最新の画像分類モデルが実用的な速度で動作することを示した．

第７章「結論と今後の展望」において提案手法を総括した上で，得られた知見および他の計算環境への応用可能性を提示している．

以上，本論文はヘテロジニアス計算環境における深層学習基盤という独自性の高い課題に取り組んでおり，提案された手法は新規の視点を持ち従来手法よりも高い性能を達成することに成功している．

よって本論文は博士（情報理工学）の学位請求論文として合格と認められる．