論文の内容の要旨

論文題目　ヘテロジニアス計算環境における深層学習基盤の研究

氏名　日髙　雅俊

機械学習モデルの 1 つである Deep Neural Network (DNN)は，画像認識，自然言語処理，ゲーム AI など様々なパターン認識・生成タスクにおいて高い性能を達成し，応用範囲の広がりを見せている．

深層学習や，学習された DNN モデルによる推論の計算量は大きく，深層学習を用いたシステムの実現には強力な計算資源が必要である．

計算資源として最も一般的なのは強力な演算装置を搭載したクラウドサービス上のサーバであり，データを端末からインターネットを通じて転送・処理する．しかし，サービス提供コスト，回線の速度や遅延，データがサービス提供者側にわたることによるプライバシー上の懸念が存在する．

上記のような課題の解決策として，消費者が保持するスマートフォンなどの情報端末や IoT デバイスの潜在的な計算資源を活用することが考えられる．

クラウドに代表される，強力かつ均質なコンピュータを用いた深層学習基盤はハードウェア・ソフトウェア両面において活発に開発・利用がなされている一方，情報端末など異種のハードウェア・ソフトウェアが混在するヘテロジニアス環境における計算基盤は未だ十分に整備されていない状況である．

本研究の目的は，ヘテロジニアス環境における DNN モデルの学習・推論を効率的に行える基盤を構築することである．
深層学習基盤の構成要素は大きく分けて 3 つある．DNN モデルは行列計算を繰り返すことにより計算を行う．行列計算を実行する機構と，行列計算の実行順序を制御する機構が必要である．さらに，コンピュータ一台で学習から推論までを完結することはまれであり，通信に関する機構が必要である．推論アプリケーションにおいては学習済みモデルの読み込み，学習システムにおいては分散計算での他の計算ノードとの協調が該当する．これらの要素をヘテロジニアス環境に適した手法を用いて実装する．

本研究では特に，従来科学技術計算に用いられることが少なかった一方，カメラの性能向上などにより DNN を用いたアプリケーション開発の需要が高まっているスマートフォン等の消費者向け端末をターゲットに具体的な要件検討，ソフトウェア実装を進める．
ハードウェアの差異を吸収し，またソフトウェアのインストールの煩雑さを解決するプラットフォームとして Web ブラウザを用いることができる．Web ブラウザ上で科学技術計算を行うことをブラウザコンピューティングと呼び，特に，深層学習に関する計算を高速に行うための GPU の活用の観点からソフトウェアの実装法を論じる．

本研究では大きく分けて，複数デバイスを協調させて深層学習を行う課題と，単一デバイスで推論を実行する課題に取り組む．前者の課題では，デバイス間の通信による計算結果の同期が必要な深層学習において，デバイスの性能差に着目し計算時間を均等化する分散計算アルゴリズムの提案および，単一の行列計算アルゴリズム実装をプラットフォームごとに異なる GPU の利用方法にあわせて変換する抽象化機構を提案する．後者の課題では，まず学習済 DNN モデルを圧縮し，通信環境にかかわらずアプリケーションの動作を迅速に開始できるようにする手法を論じる．次に，様々な学習済み DNN モデルについて，単一の中間表現へと変換し，演算の性質を活用した最適化を施すことにより統一的な高速化が達成できることを示す．

これらの組み合わせにより，DNN モデルの学習から利用までを高性能・均質なクラウド環境ではなく，ヘテロジニアスな端末上の計算能力を用いて実現可能となる．