# 先端科学技術研究科　修士論文要旨

| 所属研究室<br>（主指導教員） | コンピューティング・アーキテクチャ<br>（中島　康彦　（教授）） | | |
|---|---|---|---|
| 学籍番号 | 2011032 | 提出日 | 令和 4年 1月 13日 |
| 学生氏名 | 稲益　秀成 | | |
| 論文題目 | CGRA統合型AIフレームワークの開発と評価 | | |

要旨


昨今，機械学習が広く活用されており，今後も自動運転等で需要は高まり続けると予想される．そこで課題となるのが，それらの演算にともなう消費電力である．JST低炭素社会戦略センターの報告書によると，2050年にはAIデータセンターの消費電力が供給可能電力を上回ると予測されている．持続可能な社会を目指すためにも消費電力低減が急務であり，特に電力消費の大部分を占めるCPUやGPU等の演算デバイスの消費電力低減が最も重要である．

そこで，低電力かつ高効率なデバイスとして，Coarse Grained Reconfigurable Array (CGRA) が注目されている．CGRAは演算器レベルの粒度で再構成が可能なアーキテクチャであり，演算器をメッシュ状に接続し，演算器から別の演算器へと演算結果を順々に流していくことでプログラムを実行するデバイスである．CPUやGPUといったノイマン型デバイスとは異なり，演算結果を次の演算器に直接転送するため，主記憶アクセスを減らすことができ，低消費電力で演算が行える．また，ゲートレベルの再構成が可能なField Programmable Gate Array (FPGA) に比べて，構成単位の粒度が粗いため，アプリケーションをマッピングするための処理が軽い，集積度や動作クロック周波数が10倍程度高い等の利点をもつ．予備評価の結果，AI演算デバイスとして主流なGPUと比べて，CNNプログラム実行時の面積あたりの性能が5.8倍であることを確認できた．

しかし，CGRAはAIライブラリが整備されていないという課題がある．AIアプリを実装する際，CPUやGPUでは，ユーザーはcuDNNやoneDNNなどのAIライブラリを用いることでハードウェアを意識することなく最適化済みのコードを実行できる．これに対しCGRAでは，ユーザーはハードウェアの特性を理解し，最適化したC言語やアセンブリ相当のコードを記述する必要があり，使用難易度が高い．今後CGRAが広く活用されていくためにも，CGRAを容易に利用できる仕組みが必要である．

そこで本研究では，CGRA統合型AIライブラリを提案する．ユーザーが既存のAIプログラムコードをほとんど改変せずとも，低電力かつ高効率なCGRAの恩恵を受けられることが目的である．まず，昨今AIアプリケーションはPythonで記述されることがほとんどである．そのため，Pythonでコードを記述し，CGRAを利用できるようにする．次に，CGRAの高い演算効率を維持するために，ソフトウェア面でのパフォーマンス低下を最小限にするべく，C言語で記述し，Pythonから呼び出せる共有ライブラリとして実装を行う．また，処理をCGRAで実行するべきか，CPUやGPUで実行するべきかは実行環境やプログラムの内容に依存するため，本ライブラリからはCGRAだけでなく，CPUやGPUも呼び出すことができるようにする．そして，様々なCGRAに対応できるよう，アプリケーション実行時に，CGRAハードウェアの仕様や，処理内容に合わせてプログラムを再マッピングし，実行できる仕組みを提供する．

評価の結果，提案手法のライブラリは，全てC言語で記述された場合と比較して，同等の性能を確認した．