ABCI 3.0開発加速利用 （2025年度）成果概要（公開用）

課題名：
高効率AIチップを実現するHardware Friendly深層学習モデルの開発

実施時期： 2025/4/1~2026/3/31
所属機関名： 産業技術総合研究所
代表者氏名： 更田 裕司

成果概要：

本研究では、ハードウェア実装容易性を考慮したBERTモデルを提案する。従来のBERTモデルでは、演算に浮動小数点精度が必要とされるほか、softmax等の非線形関数の計算を含んでいた。ハードウェア実装の観点から見ると、これらは消費電力増加や面積効率低下の要因となる。そこで本研究では、非線形関数を用いず、整数精度の演算のみで処理可能なBERTモデルを開発した。さらに、FPGA上に本モデル向けのアクセラレータを実装し、少ないハードウェア資源で高い性能を実現できる事を示した。

成果のポイント：

BERTモデルは、自然言語処理分野で幅広く活用されている言語モデルである。Transformerアーキテクチャを基盤とすることで、文章内に存在する重要な関係性を高精度に捉えることができる。一般的に、この種のモデルの推論処理はデータセンターなどのクラウド環境で実行されることが通常だが、データ送信に伴う消費電力や通信遅延、さらにはセキュリティ面での懸念といった課題が指摘されている。このため、スマートフォンやPCなどのエッジ(端末)側でAI処理を行う、いわゆるエッジAIへの関心が高まっている。エッジAIにおいては、消費電力や実装コストが特に制約条件となることから、性能を損なうことなく、限られたハードウェア資源で動作可能なモデルが求められている。

そこで本研究では、ハードウェア実装容易性を考慮したBERTモデルを提案する。代表者らは、このようなハードウェア実装を考慮したモデルをHardware Friendlyと呼び、次の条件を満たすものとして定義する。(1) 完全量子化: データおよび重みは低ビットの整数精度で量子化されている(つまり、浮動小数点数精度の演算を必要としない)。(2) 計算の統一性: 行列演算のみに限定するなど、必要とされる演算種別が統一されている。(3) 演算の容易性: ビットシフトやルックアップテーブルなど、ハードウェア実装が容易な演算で構成されている。

従来のBERTモデルは、その処理に浮動小数点数精度が求められ、さらにsoftmaxなどの非線形演算が必要であった。そこで本研究では、softmax関数などの非線形演算を単純な行列演算に置き換え、さらに、8ビット整数精度のみで処理できる(重みは3値量子化)BERTモデルを開発した。本モデルは上記Hardware Friendlyモデル要件(1)~(3)を全て満たすので、効率的にハードウェア実装が可能である。これを実証する為、本モデル向けの推論アクセラレータをAMD社のZCU104評価ボードに搭載されたZynq Ultrascale+ FPGA上に実装した。その結果、提案アクセラレータのハードウェア資源効率は従来のFPGAベースのBERTアクセラレータに比べ81%高いことを示した。

なお、本研究で使用した学習済みモデルは[2]に公開されている。

成果についてより詳細な情報を提供しているWebページ、発表論文などの情報：

[1] H. Fuketa, T. Katashita, Y. Hori, and M. Hioki, “Fully Quantized Matrix Arithmetic-Only BERT Model and Its FPGA-Based Accelerator,” IEEE Access, vol. 13, pp. 107165–107174, Jun. 2025.
[2] https://doi.org/10.57765/2003355