論 文 内 容 の 要 旨

博士論文題目

Efficient Implementations of Neural Network Powered by Spike Coding and Scalable Bisection Spanning

（スパイクコーディングとスケーラブルな二分木構造を利用したニューラルネットワークの効率的実装）

氏 名 Man Wu

Deep neural networks (DNNs) have demonstrated state-of-the-art performances in the broad field of AI applications. However, these algorithms have intensive computational and colossal memory, making it challenging to develop on hardware platforms with limited computational resources. To address this challenge, this dissertation focuses on constructing efficient elastic neural networks (NNs) with spike coding and scalable bisection spanning to make NN inference less complicated, less redundancy, and more efficient to support the fully parallel reconfigurable NN platforms.

We present directly-trained spiking neural networks (SNNs) with ternary weight to achieve a good trade-off between complexity, latency, and performance to improve capacity extension of fully parallel and reconfigurable NN platforms by reducing its parameter computation and time steps. The proposed approach achieves 98.43%, 89.07%, 65.24% accuracy on N-MNIST, CIFAR-10, and CIFAR-100 with 4 time steps, respectively, and achieved up to 16x model compression. On the other hand, we develop and evolve a spatially scalable bisection NN architecture for fully parallel and reconfigurable NN platforms by improving the efficiency during the reconfiguration, which also supports on-demand array partitioned and reconfiguring (seen as "DiaNet"), and processes multiple applications in fully parallel through multi-grained reconfigurable architecture (MGRA). A byproduct of the model bisection paradigm, the proposed scalable bisection NN minimizes computation and memory costs by reducing the number of operations and model size. This proposed DiaNet can achieve 90.86% parameters reduction with no loss of accuracy on MNIST. Moreover, we demonstrate that MGRA can process multiple applications in parallel and achieve a 10.8% power reduction simultaneously. Finally, we introduce the spike coding into spatially scalable bisection NN architecture to achieve temporal-spatial combined bisection NN architecture. The combined NN takes full advantage of the robustness and low power of SNN and ultra-sparse and multi-grained reconfigurable of DiaNet. In this sense, the model becomes linearly separable while processing information in temporal and spatial domains, better yet, reducing the computational and memory costs. Finally, the model achieves 96.10% accuracy with a 90.86% compression ratio with 8 time steps on MNIST and 98.15% accuracy with 69.38% parameters reduction with 6 time steps on N-MNIST.

氏　名	Man Wu

（論文審査結果の要旨）（Ａ４　１枚　１、２００字程度）

本論文は、非ノイマン型コンピュータ向けアルゴリズムに関する研究である。現在、ディープニューラルネットワーク（DNN）は、AI 応用分野で優れた性能を発揮している。しかし、これらのアルゴリズムは、膨大な計算量と大量のメモリを必要とするため、従来のノイマン型コンピュータをベースとする、大規模並列ハードウェアプラットフォーム上では、メモリボトルネックが問題となっている。本研究は、メモリボトルネック解消のために、非ノイマン型コンピュータに着目している。本研究は、ニューラルネットワークの時間領域および空間領域の両方の並列性を利用している。第１に、時間領域では、３値を使用するスパイキングニューラルネットワークとモデル圧縮を提案している。この高効率圧縮手法と SNN モデルは、複雑さ・レイテンシ・性能の間の自在なトレードオフを形成することができる。検証のために、N-MNIST と CIFAR-100 において、若干の精度低下で、最大 16 倍のモデル圧縮を達成することが可能であった。第２に、空間領域では、大規模二分木ニューラルネットワーク（DiaNet）を利用して、さらなる実装コスト低減のために、複数の DiaNet 発展版を考案した。改良後の DiaNet は、MNIST において精度を落とすことなく、90.86%のパラメータ削減を達成することが可能であった。さらに、DiaNet が他のアプリケーションについても並列処理が可能であり、10.8%の電力削減を達成できることを実証している。第３に、DiaNet と SNN の利点を効率的に利用するために、低電力スパイクパラダイムを空間的に拡張した DiaNet に発展させ、時空間結合 NN 構造を実現している。これにより、時間・空間領域の情報を処理しながら、線形分離可能なモデルを構成することが可能となり、計算コストとメモリコストを削減することができている。

以上、本論文は学術上、実際上寄与するところが少なくない。よって、本論文は博士（工学）の学位論文として価値あるものと認める。