論 文 内 容 の 要 旨

博士論文題目

Multi-Grained Reconfigurable Architecture Powered by Elastic Neural Network for Approximate Computing
（近似コンピューティングのための弾性ニューラルネットワークを搭載したマルチグレイン再構成可能アーキテクチャ）

氏 名 Kan Yirong

Beyond the boom of artificial intelligence, the next generation of computing architectures with high speed and low cost are always demanded. In this thesis, we proposed a multi-grained reconfigurable architecture for accelerating arbitrary functions in fully parallel with high speed and low cost. The proposed architecture is reconfigurable in fine-grained (arbitrary functions), mid-grained (flexible function feature, accuracy, and number of operands), and coarse-grained (organization of kernels). By implementing a large scale of novel bisection neural network (BNN) on hardware, the reconfiguration is conducted by partitioning entire BNN into any specific pieces without redundancy. Each piece of BNN retrieves the arbitrary function approximately. By reconfiguring the BNN topology in software, we can easily adjust dimensions of the computing kernel without rewiring, and achieve a wide range of trade-offs between accuracy and efficiency in hardware. In this manner, the multi-grained reconfigurable architecture is achieved. For proof-of-concept, a demo accelerator is built on FPGA. The processing element is designed in 16-bit fixed point scheme including two synapses and one neuron. In order to better support this architecture, we have also proposed a series of system-level optimization techniques, including design flow, on-chip interconnection, and configuration strategies, etc. Since the architecture is flexible in all grained levels, various configurations for each validation are demonstrated with rich options of performance-cost matrix. From the FPGA implementation results, compared with CPU baseline, proposed architecture achieves speedups of 5.1x to 30.3x. Compared with other traditional function approximation methods, our method provides fewer parameter storage requirements. The comparison against related works proves that our accelerator has reduced the area-latency product by at least 9.5% with a loss of accuracy by at most 8.9%.

氏　名	Kan Yirong

（論文審査結果の要旨）（Ａ４　１枚　１、２００字程度）

本論文は、任意の関数演算を多重粒度に並列化して高速化する、低コスト再構成可能アーキテクチャを提案している。本アーキテクチャは、細粒度（目標関数）、中度（関数の特徴、精度、オペランド数に対する柔軟性）、粗粒度（各演算コアとの接続）の再構成が可能である。独創的な大規模 Bisection Neural Network （BNN）をハードウェア上に実装し、BNN 全体を無駄なく任意の断片に分割して再構成を行うことができる。BNN の各要素は、任意の関数を近似的に実現する。BNN のトポロジーをソフトウェアにより再構成することで、配線を変更することなく、演算カーネルのサイズを容易に調整でき、ハードウェアレベルで精度と効率のトレードオフを幅広く探索できる。第 1 に、実証のため、FPGA 上に実証用のアクセラレータを構築している。Processing element（PE）は、2 つのシナプスと 1 つのニューロンを含む 16 ビット固定小数点方式で設計されている。第２に、本アーキテクチャを効率的に利用するための、自動トポロジー生成、および、オンチップメモリ相互接続など、一連のシステムレベルの最適化技術を提案している。第３に、本アーキテクチャは各粒度において柔軟であることが、性能-コストの様々な組み合わせとともに実証されている。FPGA への実装の結果、従来の関数近似手法と比較して、本手法はパラメータ量が少ないことが確認されている。また、関連研究との比較から、本アーキテクチャは、若干の精度低下を伴うものの、効果的に計算時間を短縮できることが証明されている。

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。