TIE: Energy-efficient Tensor Train-based Inference Engine for Deep Neural Network

Corresponding author

Chunhua Deng, Rutgers University

Fangxuan Sun, Nanjing University

Keywords

DNN hardware, tensor-train decomposition, compression,

推断、算法改进、硬件架构、仿真器、RTL EDA验证（28nm）

Summary

*Challenge*

**Tensor train format DNN** models inherently incurs massive amount of **redundant computations**, causing significant energy consumption. Thus, the straightforward application of TT decomposition is not feasible.

（原始算法冗余计算多，不高效，简单实现不可取）

*Contribution*

1. Develop a computation efficient inference scheme for TTI-format DNN. （算法改进）
2. Based on the scheme, develop an inference engine for TT format compressed DNN. （架构设计）
3. Implement 16 PE prototype using CMOS 28nm technology. （RTL实验测试）

*Result*

Compared with EIE, TIE achieves 7.22x-10.66x better area efficiency and 3.03x-4.48x better energy efficiency. Compared with CIRCNN, TIE achieves 5.96x and 4.56x high throughput and energy efficiency, respectively.