基于神经元容错度分析的神经网络裁剪与近似计算技术研究

王　旭　　1　，王　晶　　1，2　，张伟功　　1,3

（1 首都师范大学 信息工程学院，北京 100048；

2 中国科学院计算技术研究所 体系结构重点实验室，北京 100086；

3 电子系统可靠性北京市重点实验室，北京 100048）

摘　要：本文将神经元裁剪和近似计算技术相结合，首先提出基于统计排序的神经元容错能力量化方法.然后，为了识别神经元的裁剪度，根据神经元的容错能力提出神经元重要程度排序算法.其次，引入轻量级的重训练，提出循环裁剪法，以探寻最优裁剪率.最后，根据神经元的容错能力，在神经网络运行过程中使用近似计算技术进一步降低功耗开销.本文通过两个实验，证明了该技术的有效性，其中以 MNIST 为例，在精度损失小于 5%的情况下，压缩率达到 50%，节能 1.35 倍.

关键词：神经网络；神经元容错能力；节点裁剪；近似计算

Research on neural network pruning and approximate computing technology based on neuron fault tolerance analysis

WANG Xu　　1　，WANG Jing　　1，2　，ZHANG Wei-gong　　1，3

（1 Capital Normal University Information Engineering College, Beijing 100048, China;

2　Institute of Copmuting Technology, Chinese Academy of Sciences, Beijing 100086, China;

3 Beijing Key Laboratory of Electronic System Reliability and Prognostics,

Capital Normal University, Beijing 100048, China）

Abstract：This paper proposes to use neuron node pruning and approximate computing simultaneously. First, we propose a method to quantify the fault tolerance capability of neurons based on statistics. Then, to identify whether the neuron can be pruned, an importance ranking algorithm is proposed based on the fault tolerance capability. Next, introducing retrain and cyclic pruning to find the optimal pruning rate. Finally, approximate computing technique is used to further reduce power consumption during neuron network execution. The effectiveness of above technique is proved by two experiments. In the case of MNIST dataset, the compression rate is 50% and the power saving is 1.35× when the output accuracy loss is less than 5%.

Key words：neural network; neuron fault tolerance capability;node pruning; approximate computing

作者简介：

王　旭　男，(1993-)，硕士研究生.研究方向为计算机系统结构、容错计算.

王　晶(通信作者)　女，(1982-)，博士，副研究员.研究方向为计算机系统结构、高性能计算、容错计算、智能芯片设计.　E-mail:jwang@cnu.edu.cn.

张伟功　男，(1967-)，博士，研究员.研究方向为高可靠计算机体系结构及 SoC 设计、高速高可靠总线技术、计算机容错技术.