稀疏卷积神经网络加速器设计

李永博，王　琴，蒋剑飞

（上海交通大学　电子信息与电气工程学院，上海　200240）

摘　要：　为降低卷积神经网络推断时的时延和能耗，使用动态网络剪枝技术得到稀疏网络并设计出高能效比的稀疏卷积神经网络加速器.针对运算负载不均衡问题，提出适合稀疏运算的数据流；针对卷积运算高时延问题，采用 16×16 运算阵列提高运算并行度，设计索引单元避免无效运算，设计脉动输入层加强数据复用，采用乒乓缓存减少数据等待.综合结果表明，在 TSMC 28 nm 工艺下，芯片工作频率可达 500 MHz，功耗为 249.7 mW，卷积运算峰值算力达到 256 GOPS，能效比为 1.03 TOPS/W.

关键词：　稀疏卷积神经网络；阵列运算；加速器；高能效比

Design of sparse convolutional neural network accelerator

LI Yong-bo, WANG Qin, JIANG Jian-fei

(School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University,
Shanghai 200240, China)

Abstract：　In order to reduce the latency and energy consumption of convolutional neural networks, dynamic network surgery is used to get sparse networks and a high energy efficiency sparse convolutional neural network accelerator is designed. Aiming at the problem of unbalanced computing load, a dataflow suitable for sparse computing is proposed. To reduce the latency of convolution operation, a 16×16 process engine array is used to improve computation parallelism, index units are designed to avoid invalid operation,the systolic input structure is designed to enhance data reuse, and ping-pong buffers are introduced to reduce data waiting. The synthesis results showthat the frequency can reach 500 MHz, the power consumption is 139mW, the peak performance is 221 GOPS, and the energy efficiency is 1.59T OPS/W with TSMC 28nm process.

Key words：　sparse convolutional neural network; array computation; accelerator; high energy efficiency

作者简介：

李永博　男，（1995-），硕士研究生.研究方向为神经网络加速器设计、大规模集成电路设计. E-mail:lyb_sd@sjtu.edu.cn

王　琴　女，（1975-），副教授.研究方向为大规模集成电路设计、先进集成电路设计方法.

蒋剑飞　男，（1979-），助理研究员.研究方向为高速数字集成电路设计.