一种双向脉动数据流的全卷积神经网络加速器

涂凯杰 1，朱宜炳 2，许达文 1
（1 合肥工业大学 电子科学与应用物理学院，安徽 合肥 23009；
2 江西科技学院 信息工程学院，南昌 江西 330098）

摘　要： 全卷积神经网络近年来被应用于深度学习中的多个领域，其不仅能处理简单的图像分类任务，还能应用于例如物体检测、语义/图像分割以及基于生成式对抗网络的生成型任务.典型的全卷积神经网络中不仅包括了传统的卷积层，还有反卷积层，它们都是计算密集型的.现在大多数研究者大都关注卷积层的设计优化，而反卷积的加速优化很少.本文提出了一种双向脉动数据流的全卷积神经网络加速器，可以同时高效地处理普通卷积层以及反卷积层.实验中选取了多个具有代表性的全卷积神经网络模型，例如 DCGAN，Cascaded-FCN 等.相较于以往传统的未优化的加速方案，本文所设计的加速器平均可以达到 2.8 倍的加速比，并且能耗降低了 46.3%.

关键词： 全卷积；反卷积层；加速优化；双向脉动数据流

Fully convolutional neural networks accelerator for Bi-direction systolic data flow

TU Kai-jie1, ZHU Yi-bing2, XU Da-wen1
（1 School of Electronic Science & Applied Physics, Hefei University of Technology, Hefei 23009, China;
2 College of Information Engineering, JIANGXI Unversity of Technology, Nanchang 330098, China）

Abstract： Fully Convolutional Neural Networks (FCN) have been applied in various domains of deep learning in recent years. It can be used not only for simple image classification tasks but also for such as object detection, semantic/image segmentation and generative tasks based on Generative Adversarial Networks (GAN). Typical FCN includes not only traditional convolutional layers but also deconvolutional layers, which are computationally intensive. Most researchers are now focusing on the design optimization of the convolutional layer, but deconvolutional layer receives less attention. This paper proposes a Fully Convolutional Neural Networks accelerator using bi-direction systolic data flow, which can efficiently process the general convolution layer and deconvolution layer simultaneously. Several representative FCN models, such as DCGAN, Cascaded-FCN, etc., have been selected in the experiment. Compared with the traditional unoptimized acceleration scheme, the proposed approach can achieve 2.8X speed up on average, and the energy consumption is reduced by 46.3%.

Key words： fully convolution；deconvolutional layer; accelerated optimization；bi-direction 作者简介：
涂凯杰　男，（1994-），硕士研究生.研究方向为深度学习加速. Email: tukaijie@ict.ac.cn.
朱宜炳　男，（1982-），硕士研究生，讲师.研究方向为形式化方法与模型检测.
许达文　男，（1986-）博士，副教授.研究方向为 GPU 通用计算与嵌入式系统.