曳影 1520 NPU

用户手册

文档版本	1.0.0
保密等级	保密
发布日期	2023-08-26

Copyright © 2022 T-HEAD (Shanghai) Semiconductor Co., Ltd. All rights reserved.

This document is the property of T-HEAD (Shanghai) Semiconductor Co., Ltd. This document may only be distributed to: (i) a T-HEAD party having a legitimate business need for the information contained herein, or (ii) a non-T-HEAD party having a legitimate business need for the information contained herein. No license, expressed or implied, under any patent, copyright or trade secret right is granted or implied by the conveyance of this document. No part of this document may be reproduced, transmitted, transcribed, stored in a retrieval system, translated into any language or computer language, in any form or by any means, electronic, mechanical, magnetic, optical, chemical, manual, or otherwise without the prior written permission of T-HEAD (Shanghai) Semiconductor Co., Ltd.

Trademarks and Permissions

The T-HEAD Logo and all other trademarks indicated as such herein are trademarks of T-HEAD (Shanghai) Semiconductor Co., Ltd. All other products or service names are the property of their respective owners.

Notice

The purchased products, services and features are stipulated by the contract made between T-HEAD and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

平头哥（上海）半导体技术有限公司 T-HEAD (Shanghai) Semiconductor Co., LTD

Address: 5th Floor Number 2 Chuan He Road 55, Number 366 Shang Ke Road, Shanghai free trade area, China
Website: www.t-head.cn

Copyright © 2022 平头哥（上海）半导体技术有限公司，保留所有权利.

本文档的所有权及知识产权归属于平头哥（上海）半导体技术有限公司及其关联公司(下称“平头哥”)。本文档仅能分派给：(i)拥有合法雇佣关系，并需要本文档的信息的平头哥员工，或(ii)非平头哥组织但拥有合法合作关系，并且其需要本文档的信息的合作方。对于本文档，未经平头哥（上海）半导体技术有限公司明示同意，则不能使用该文档。在未经平头哥（上海）半导体技术有限公司的书面许可的情形下，不得复制本文档的任何部分，传播、转录、储存在检索系统中或翻译成任何语言或计算机语言。

商标申明

平头哥的 LOGO 和其它所有商标归平头哥（上海）半导体技术有限公司及其关联公司所有，未经平头哥（上海）半导体技术有限公司的书面同意，任何法律实体不得使用平头哥的商标或者商业标识。

注意

您购买的产品、服务或特性等应受平头哥商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，平头哥对本文档内容不做任何明示或默示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。平头哥（上海）半导体技术有限公司不对任何第三方使用本文档产生的损失承担任何法律责任。

平头哥（上海）半导体技术有限公司 T-HEAD (Shanghai) Semiconductor Co., LTD

地址： 中国（上海）自由贸易试验区上科路 366 号、川和路 55 弄 2 号 5 层
网址： www.t-head.cn

版本历史

版本	说明	作者	日期
V1.0.0	初始版本	平头哥	2023-08-26

©【平头哥上海半导体技术有限公司】版权所有

目录

©【平头哥上海半导体技术有限公司】版权所有

图表目录

©【平头哥上海半导体技术有限公司】版权所有

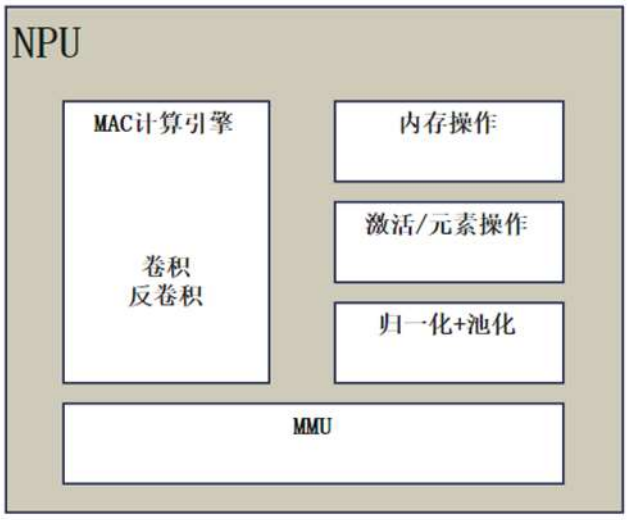
术语与缩略语

缩略语	英文全名	中文解释
NPU	Neural-network Processing Unit	神经网络处理单元

©【平头哥上海半导体技术有限公司】版权所有

1 概述

NPU 一种基于硬件的神经网络加速器，具有低功耗和高性能的特点。NPU 是 SoC 目标神经网络推理加速的关键组件，支持可变精度的数据和权重。NPU 支持权重压缩和灵活的低精度，使神经网络能够快速运行。NPU 的概要框图如图表 1-1 所示。



图表 1-1 NPU 模块功能图

©【平头哥上海半导体技术有限公司】版权所有

2 主要特性

NPU 的主要特性：

- 最常见的神经网络层的加速，如图表 2-1 所示
- 低带宽操作
  - 支持多种低精度数据格式
  - 将各层组合在一起以减少内存带宽
  - 无损权重数据压缩
- DRM 安全
- 互操作性
  - 支持多种内存格式，可与 CPU、GPU 或其他模块共享数据

图表 2-1 支持的层

层	通过软/硬件支持
卷积	
普通卷积	硬件
扩张/空洞卷积	硬件
分组卷积	硬件
逐深度卷积	硬件
卷积转置（反卷积）	硬件
完全连接	
完全连接	硬件
归一化	
批量归一化	硬件
局部响应归一化	硬件
L2 归一化	软件
激活	
ReLU	硬件
ReLU1	硬件

©【平头哥上海半导体技术有限公司】版权所有

层	通过软/硬件支持
ReLU6	硬件
PReLU	硬件
Clamped ReLU	硬件
Leaky ReLU	硬件
Tanh	硬件
Sigmoid	硬件
Logistic	硬件
池化	
最大池化	硬件
平均池化	硬件
最小池化	硬件
元素操作	
取消	硬件
加/减	硬件
乘	硬件
最大/最小	硬件
内存操作	
重新排列	硬件
转置	硬件
重塑	硬件
挤压	硬件
压平	硬件
空间到批量	硬件
批量到空间	硬件
深度到空间	硬件
空间到深度	硬件

©【平头哥上海半导体技术有限公司】版权所有

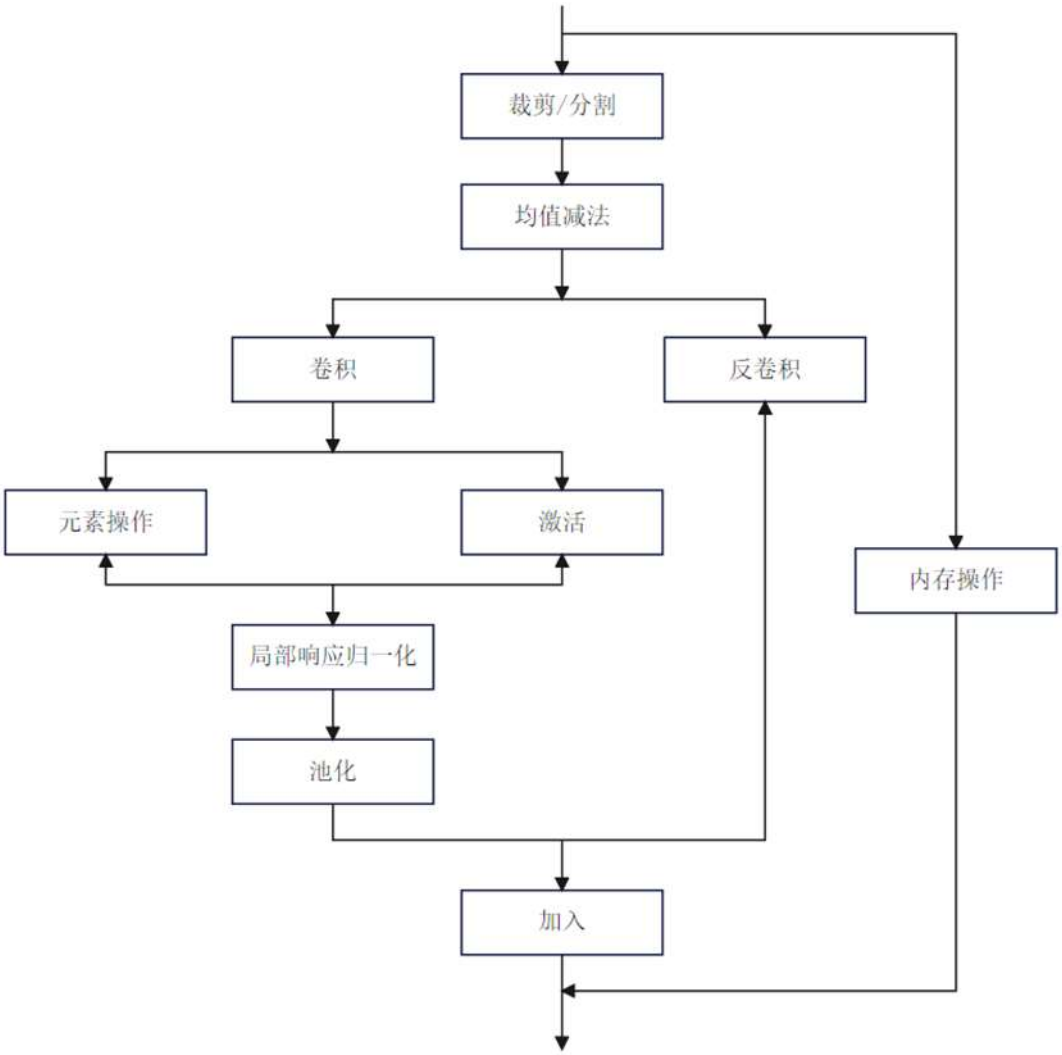
层	通过软/硬件支持
空间大小调整操作	
填充	硬件
裁剪	硬件
双线性调整大小	硬件
最近邻调整大小	硬件
预处理	
均值减法	硬件
后处理	
Softmax	软件

©【平头哥上海半导体技术有限公司】版权所有

3 功能描述

3.1 NPU 处理顺序

图表 3-1 显示了 NPU 处理层的顺序。可以通过硬件单次组合在一起的不同层称为“层组”。如果目标网络中的处理顺序与图表 3-1 所示的顺序不匹配，操作将被分为不同的层组。例如，要在池化后执行局部响应归一化，第一个层组将执行池化层，而第二个层组将包含局部响应。



图表 3-1 NPU 流程图

©【平头哥上海半导体技术有限公司】版权所有

4 使用

核的上电步骤如下：

1. 释放 NPU 复位。
2. 打开 NNA 电源。
3. 启用 NNA 时钟。
4. 等待至少 16 个周期。
5. 复位 NPU。
6. 等待至少等待 8 个周期。
7. 启动 NPU。

©【平头哥上海半导体技术有限公司】版权所有