基于异构处理平台的机载实时图像识别

刘 鑫[1] 张瀚成[1] 王海翔[1] 赵 岩[1] 白林亭[1]
LIU Xin ZHAGN Hancheng WANG Haixiang ZHAO Yan BAI Linting

摘 要 近年来，深度神经网络模型在机载图像识别方面的应用场景不断扩大，为了满足神经网络模型对硬件资源更高的实时算力需求，适应机载场景下的设备运行条件，集合嵌入式异构处理器资源，选取 FT-D2000/8 作为主处理器进行算法调度，选取复旦微 FMQL100TAI 作为协处理器进行智能计算，并设计了具体的并行处理流程以加快计算过程。实验结果表明，相比于 CPU 单处理器，在 CPU+NPU 这样的异构处理平台下可以将图像识别帧率提高至 2.8 倍，同时识别精度误差控制在 1.035% 内，具有良好的性能功耗比表现。

关键词 深度神经网络；异构处理平台；机载；图像识别；并行处理

doi：10.3969/j.issn.1672-9528.2024.03.022

0 引言

随着人工智能技术的不断进步，机载领域中深度神经网络模型的应用实例不断增加，呈现出智能化的发展趋势 [1-5]。机载图像识别是机载领域内最为重要的技术手段之一，以 YOLO[6-7]、Faster R-CNN[8]、SSD[9] 等为代表的经典深度神经网络模型已被广泛应用于机载图像识别技术中，为上层决策提供专业的参考依据。现阶段，深度神经网络模型以及图像数据的规模不断扩大，对于图像识别的速度产生了直接影响，基于对机载图像识别技术的实时性要求，对高性能、高效率、低功耗的硬件计算资源需求也在日益提高。

基于上述背景，寻找合适的硬件计算资源搭建计算平台，以适应机载环境的计算场景，成为当下的技术难点之一。本文通过分析传统硬件计算资源的利弊，综合考虑性能与功耗等因素，选择了 CPU+NPU 形式的异构处理平台，并在异构平台上设计目标识别应用算法，可以在有效利用硬件资源的同时，满足机载平台图像识别的实时性与准确性需求。

1 异构处理平台设计

1.1 异构处理平台的发展必要性

传统的硬件计算资源主要包括 CPU、GPU、FPGA 等核心处理器。CPU 的优势主要体现为可以处理各类数据，拥有强逻辑判断的能力，解决单次复杂能力较强，但是能效比低，核心数量有限 [10]；GPU 在浮点运算、并行计算等方面可以提供更为强大的性能，但能耗较大且调度管理能力弱，不利于机载领域中数据传输频繁、多次少量场景中的计算加速 [11]；FPGA 芯片是一种可编程逻辑芯片，具有高度的灵活性和良好的性能功耗比，但是基本单元的计算能力有限，同时开发流程复杂，导致开发成本较高 [12]。对比以上传统核心处理器在深度神经网络计算中的不足，专门用于进行深度神经网络加速计算的 NPU 逐渐成为近年来人工智能领域的热门技术之一。NPU 在电路层模拟神经元，并且用深度学习指令集直接处理大规模的神经元和突触，一条指令完成一组神经元的处理，实现存储和计算一体化，以提高运行效率，具有计算效率高、平均性能强、体积小、功耗低等明显优势，但是需要 CPU 的协同处理才能完成特定的任务 [13]。

由此可见，不同处理器在智能计算中各有优势，同时伴随着较为明显的局限性，若只针对某一种处理器进行优化，算力与功耗的平衡问题依旧不可避免 [14]。为了更好地依靠现有芯片技术实现机载平台的神经网络模型部署与智能计算，同时在一定程度上平衡硬件的算力与功耗，以 CPU 和 NPU 两种类型的处理器搭建的异构处理平台成为代表性实现手段。通过综合 CPU 与 NPU 的处理优势，以 CPU 作为控制单元，NPU 作为计算单元，在确保逻辑运算与神经网络模型运算效率的同时进一步提升能耗比。

1.2 异构处理器选择说明

现阶段国产处理器不断发展，相关产品覆盖了嵌入式、个人计算机、服务器、高性能计算等诸多应用场景，在国产处理器领域基本实现了有芯可用。机载环境下选用的处理器需具备较高的自主可控程度并适用于嵌入式环境，应同时具备定点运算、浮点运算和向量运算能力。

1. 中国航空工业集团公司西安航空计算技术研究所
陕西西安 710065

目前在通用处理器领域，优势较强的是龙芯、飞腾、鲲鹏、海光这四大厂商，其中，飞腾以其高性能、低功耗的优势在机载环境下被广泛应用。同时，飞腾完整定义并实现了安全处理器平台架构规范，可以有效防止处理器出现安全短板，提升处理器的安全性。综合考虑到算力、内核数量等，主处理器选取 FT-D2000/8，该处理器集成 8 个飞腾自主研发的新一代高性能处理器内核 FTC663，兼容 64 位 ARMV8 指令集；最高主频 2.3 GHz；支持单精度、双精度浮点运算指令和 ASIMD 处理指令；支持硬件虚拟化；集成系统级安全机制，能够满足复杂应用场景下的性能需求和安全可信需求。

异构处理平台中协处理器 NPU 主要实现与主处理器之间的数据通信，完成图像 / 视频的智能处理算法应用工作。对比国内众多硬件厂商生产的 NPU，包括寒武纪 MLU370、华为 Atlas200、复旦微 FMQL100TAI、百度昆仑 R200 以及瑞芯微 RK3588 等，结合芯片的算力、功耗、软件栈成熟度以及对机载环境的适用性，协处理器 NPU 选择复旦微 FMQL-100TAI。FMQL100TAI 是复旦微电子研制的可编程融合芯片，该芯片集成了四核 A53 处理器的处理系统、视频接口处理单元、图形处理单元、可编程逻辑和 AI 加速引擎。基于 28 nm 工艺，配合相应开发软件，实现一体化软硬件平台。其中，AI 加速引擎支持加速卷积神经网络前向推理计算，主要包含 AI 加速硬核和 AI 加速软核两部分，硬核部分包含 MAC 计算单元和内部存储，支持卷积、池化和激活操作，最高可提供 27.5TOPS 的峰值 AI 算力。硬核使用时需调用 AI 加速软核，用于完成调度。软核部分包含数据通路和指令解析，主要使用 PL 逻辑实现。

本文中，FMQL100TAI 作为 PCIE 从设备，挂载于 FT-D2000/8 上，FT-D2000/8 实现图像识别整体计算过程的调度，FMQL100TAI 在其芯片内部主要实现 AI 加速引擎调用及智能处理过程。

1.3 异构处理平台模块功能介绍

目前 SAR 已成为各种军民用平台飞行器的标准配置之一，机载环境中 SAR 图像的实时处理已变得愈发重要。因此，本文选择机载 SAR 目标识别这一典型场景下的目标识别技术进行算法设计，完成在异构处理平台上的智能识别、数据传输、计算资源调度等主要工作，最终实现 SAR 图像中目标结果输出。为了模拟多数实际场景下 SAR 图像的大分辨率特点，本文将选取像素宽度为 70 000、高度为 8500 的超高分辨率 SAR 图像目标检测场景，利用 YOLOv5 算法进行图像目标识别，并实现目标经纬度解算以及检测效果显示。

在异构处理平台中，CPU 与图像生成机以及任务处理机之间建立 TCP 通信，CPU 与 NPU 之间建立 PCIE 通信，以实现异构计算资源之间高效可靠的数据传输。首先，CPU 接收图像生成机发送的单通道 SAR 切片图像与图像参数，主要的图像参数包括超高分辨率图像的左上角及右下角经纬度信息，以及切片图像左上角像素相对于原图像横纵向偏移量等。其次，在 CPU 中对图像进行前处理并写入 NPU 的接收 buffer 中，同时 NPU 加载 YOLOv5s 模型，对接收到的切片图像数据进行神经网络模型推理，在完成非极大值抑制（non maximum suppression，NMS）后处理过程后，将目标检测结果数据写入 CPU 的接收 buffer 中，最后在 CPU 中完成目标检测框绘制、目标经纬度解算、目标信息数据传输等后处理过程。具体的计算调度过程如图 1 所示。

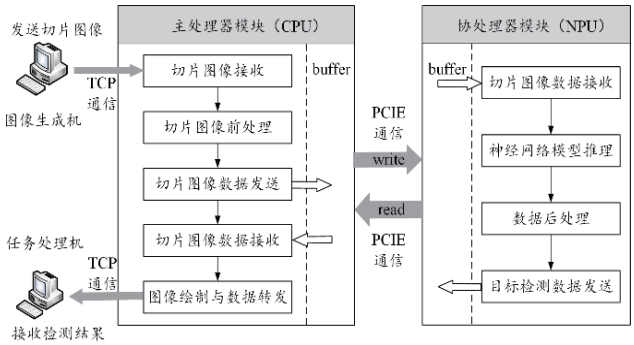

图 1 异构处理平台计算资源调度流程图

2 SAR 图像的实时处理方法

2.1 实时处理算法设计

针对 70 000×8500 的超大分辨率 SAR 图像，本文中将原始图像切分为 640×640 大小的图像，以适应 YOLOv5 模型的图像输入要求。为了避免目标漏检，同时满足图像识别的实时性需求，需要在规定时间内遍历所有切片图像进行目标识别，并依据已知的原始图像经纬度信息解算出每一个切片图像内的目标中心点经纬度，随后将目标经纬度信息以及目标识别结果图通过 TCP 传输至上层任务处理机进行下一步的决策规划。

规定 CPU 读取 NPU 回传的目标检测数据分别为切片图像上的目标检测框个数 N、目标检测框左上角的横向像素坐标 x 和纵向像素坐标 y、目标检测框宽度 w、目标检测框高度 h、目标类别 id 以及目标置信度 $conf$。为了实现第 2.2 节中提到的 CPU 与 NPU 组成的异构计算平台的计算调度流程，设计的基于异构处理平台的机载实时图像识别算法如下。

算法 1：基于异构处理平台的机载实时图像识别算法

输入：n 张 640×640 单通道 SAR 切片图像数据，原始图像经纬度信息，以及切片图像左上角像素相对于原图像横纵向偏移量

1：for k = 1, 2, …, n do

2： 切片图像单通道扩充为三通道；

3： 通过 PCIE 发送至 NPU 接收 buffer 处；

```
4:  NPU 调用 AI 加速引擎进行智能计算；
5:  CPU 读取检测数据 N，x，y，w，h，id，conf；
6:    if N > 0 do
7:      for i = 1, 2, ⋯,N do
8:        目标经纬度解算；
9:        目标检测框绘制；
10:       通过 TCP 将目标经纬度发送至任务处理机；
11:     end for
12:   end if
13: end for
```

2.2 并行处理流程设计

为了加快智能计算过程，保证 SAR 图像的处理实时性，本文以多线程的方式进行并行处理。FT-D2000/8 上运行的管理调度程序分为三个线程：主线程 CPU_main_thread、数据输入线程 CPU_input_thread 以及数据输出线程 CPU_output_thread。FMQL100TAI 上运行的智能计算程序分为四个线程：主线程 NPU_main_thread、数据输入线程 NPU_input_thread、AI 推理线程 NPU_forward_thread 以及数据输出线程 NPU_output_thread。具体的异构处理资源计算流程见图 2，主要的处理流程如下。

（1）FMQL100TAI 中的主线程 NPU_main_thread 先行启动，加载神经网络模型并创建 AI 运行时。

（2）启动 FT-D2000/8 中的主线程 CPU_main_thread 接收从图像生成机中传输的切片图像数据并发往 CPU_input_thread。

（3）CPU_input_thread 对图像数据进行前处理并发往 NPU_input_thread。

（4）NPU_input_thread 读入经过处理的图像数据并将数据发往 NPU_forward_thread。

（5）NPU_forward_thread 将调用 AI 运行时对图像数据进行前向推理，并将检测完成的数据发往 NPU_output_thread。

（6）NPU_output_thread 将检测数据做初步的 NMS 处理后，利用 PCIE 通信发往 CPU_output_thread。

（7）CPU_output_thread 进行目标经纬度解算以及目标检测框绘制等后处理操作，并将目标信息发往任务处理机，为后续的决策提供参考依据。

（8）CPU_main_thread 与 NPU_main_thread 等待各自子线程结束后分别结束，SAR 图像处理完成。

3 实验结果分析

结合原始 70 000×8500 的超高分辨率 SAR 图像内的目标平均像素尺寸，为了尽可能保证图像切分后目标的完整性，切片图像之间采取 10% 的重叠率，最终一张 SAR 图像形成的尺寸为 640×640 的切片图像数量为 1830 张。

根据算法 1 中的算法设计，主处理器 CPU 将一张超高分辨率的所有切片图像数据加载到内存，并按照接收顺序依次将图像数据通过 PCIE 通道发送至 NPU 中完成智能计算。在发送之前，图像前处理只包含将单通道的图像数据扩充至三通道，后续的图像数据归一化等处理过程在 NPU 中完成。

本文中重点针对切片图像的识别帧率以及检测数据的精度设计对比实验，以 CPU 中的智能计算结果为基准，将异构处理平台计算得到的目标检测框数据与 PC 端实验数据结果进行分析与误差计算。其中，用于对比的 PC 端 CPU 型号

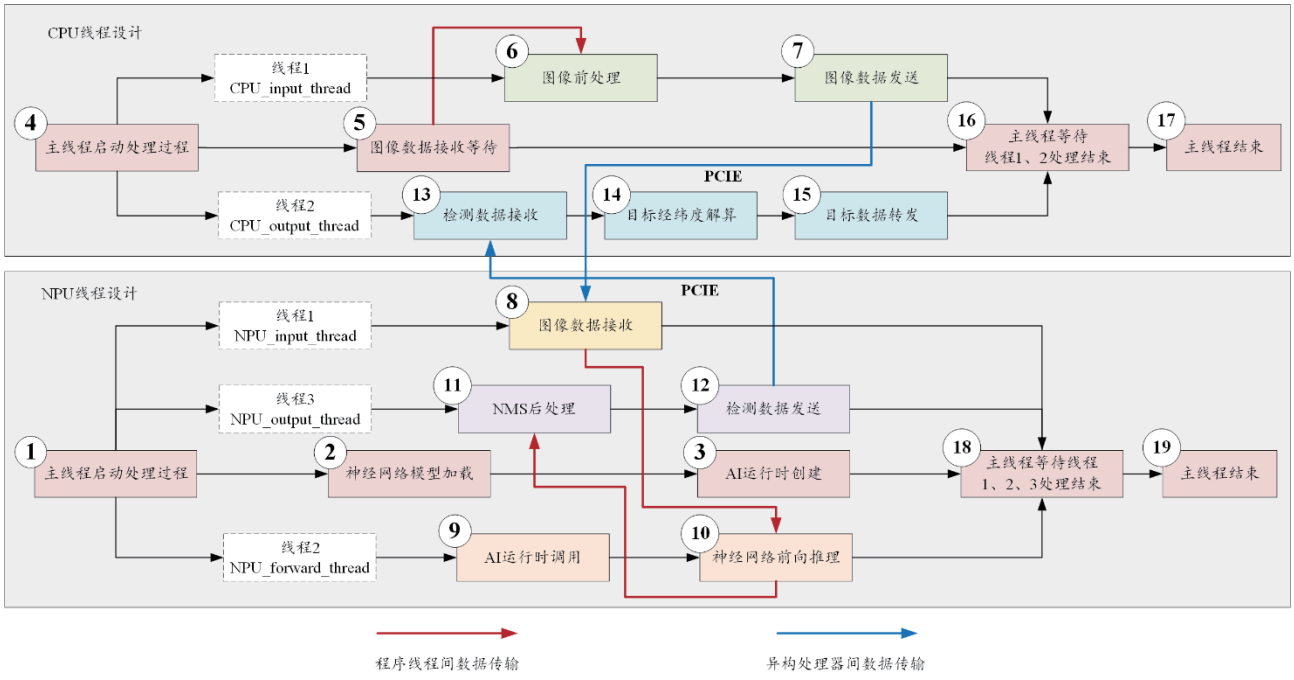

图 2 基于多线程的异构处理平台并行处理流程图

为 Intel i5-12500，操作系统为 Ubuntu20.04，深度学习框架为 PyTorch 1.9.0。

对于接收到的 1830 张切片图像，规定每张切片图像的识别时间包括前处理部分、推理部分以及后处理部分。在图像识别帧率方面，PC 端平均每张图像识别速度为 25.64 ms，识别帧率为 39 帧 /s；异构处理平台端平均每张图像识别速度为 9.09 ms，识别帧率为 110 帧 /s，速度上升 2.8 倍。此外，在图像识别精度方面，将目标检测框的位置数据以及目标置信度进行对比分析，最终异构处理平台端相对 PC 端的平均相对计算误差为 1.035%，属于可接受误差范围内。具体的图像识别效果图例对比见图 3，左为异构处理平台端结果，右为 PC 端结果，可以得到异构处理平台端与 PC 端的识别结果中的检测框位置以及目标置信度数值相近，可视化效果也基本无差。因此，利用该异构计算平台以及设计的具体计算方法，可以在平衡硬件性能功耗比的前提下达到满足实际使用需求的计算结果。

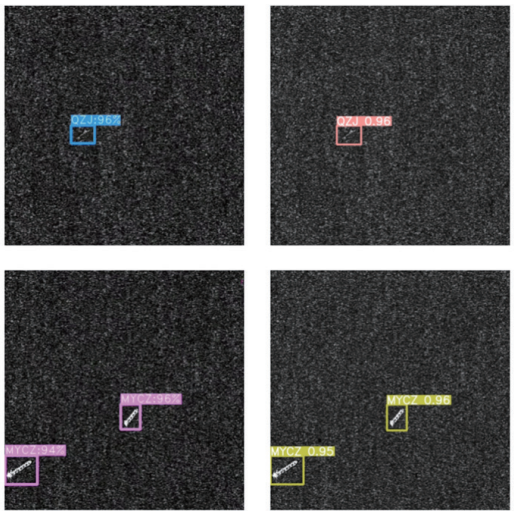

图 3 图像识别效果对比

4 总结与展望

从实验结果来看，本文中搭建的 CPU+NPU 的异构处理平台可以充分发挥相关硬件的计算优势，其中 CPU 主要提供资源分配、调度管理以及部分运算能力，NPU 进行深度神经网络模型推理以及部分后处理功能的实现。这样的异构处理平台能够在保证计算精度损失较少的情况下显著提高图像识别速度，提升硬件资源的性能功耗比，缓和算力与功耗之间的矛盾。相对于传统计算平台更具计算优势，有效验证了全国产化计算资源的机载领域适配性。

后续工作中若想继续提高整体的图像识别速度，可以考虑在主处理器 CPU 上挂载多个 FMQL100TAI 作为 PCIE 从设备，利用 CPU 调度多个 FMQL100TAI 进行并行智能运算，每个 FMQL100TAI 负责部分切片图像的识别，能够进一步提高图像识别帧率。

参考文献：

[1] 何涛，李大亮，曹兰英 . 基于深度学习的机载 SAR 典型目标识别算法 [J]. 现代雷达 ,2022,44(12):87-92.

[2] 程思远 . 基于深度学习的空对地红外面目标识别算法研究 [D]. 北京：中国电子科技集团公司电子科学研究院 ,2023.

[3] 汪珩，李鹏，文鹏程 . 基于机载场景的昇腾 310 智能芯片评估 [J]. 信息技术与信息化 ,2021(6):210-212.

[4]KIM S, CHOI H. Convolutional neural network-based multi-target detection and recognition method for unmanned airborne surveillance systems[J]. International journal of aeronautical and space sciences, 2019, 20(4):1038-1046.

[5] GU Y, WU J, FANG Y, et al. End-to-end moving target indication for airborne radar using deep learning[J]. Remote sensing, 2022, 14(21):5354.

[6] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas.Piscataway:IEEE,2016:779-788.

[7]REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR).Piscataway:IEEE,2017:7263-7271.

[8] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detecto[C]//European Conference on Computer Vision, Amsterdam.Berlin:Springer International Publishing,2016:21-37.

[9]REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis & machine intelligence, 2015, 39(6):1137-1149.

[10] 王荣阳，曲国远，童歆，等 . 面向机载应用的领域专用加速器研究 [J]. 航空电子技术 ,2022,53(3):1-8.

[11] 文鹏程，白林亭，高泽，等 . 机载智能计算技术工程实践思考 [J]. 航空计算技术 ,2021,51(2):130-134.

[12] 文鹏程，白林亭，高泽，等 . 适用于机载环境的智能计算处理器分析研究 [J]. 航空科学技术 ,2020,31(10):81-86.

[13] 杨焕永 . 面向移动端的高通量异构卷积神经网络推理系统优化设计与研究 [D]. 西安：西安电子科技大学 ,2021.

[14] 万朵，胡谋法，肖山竹，等 . 面向边缘智能计算的异构并行计算平台综述 [J]. 计算机工程与应用 ,2023,59(1):15-25.

【作者简介】

刘鑫（1997—），女，陕西咸阳人，硕士，助理工程师，研究方向：智能算法及其硬件部署方法。

（收稿日期：2023-12-25）