

一种基于轻量化 CNN 的天文暂现源 智能识别方法*

李晓斌^{1,2} 薛长斌¹ 戴育岐^{1,2} 周莉¹

1(中国科学院国家空间科学中心 北京 100190)

2(中国科学院大学 北京 100049)

摘要 天文暂现源携带了关于天体本质及演化过程的丰富信息,对暂现源进行探测与研究具有极为重要的科学价值。天文暂现源的辐射峰值大多在 X 射线或伽马射线,天基望远镜对这些高能波段的观测优势是地基望远镜无法比拟的,更适合于暂现源观测。但由于星载计算机的性能约束,很难实现依托于地面强大算力的复杂检测算法。针对以上问题,提出了基于轻量化卷积神经网络(CNN)模型的天基暂现源检测算法,并在嵌入式 ARM 平台上实现了模型部署。实验结果表明,本文提出的轻量化 CNN 暂现源检测算法的模型复杂度和计算量不及 Deep Hits 算法的 1/4,准确率达到 96.52%,可应用于星载有限算力平台,实现未来的天基暂现源实时检测。

关键词 暂现源检测,轻量化 CNN 模型,星载有限算力平台,模型部署

中图分类号 P152, TP391

An Intelligent Detection Method of Astronomical Transients Based on Lightweight CNN Model

LI Xiaobin^{1,2} XUE Changbin¹ DAI Yuqi^{1,2} ZHOU Li¹

1(National Space Science Center, Chinese Academy of Sciences, Beijing 100190)

2(University of Chinese Academy of Sciences, Beijing 100049)

Abstract Astronomical Transients carry rich information about the nature and evolution of celestial bodies, and their detection and research have extremely important scientific value. Most of the radiation peaks of astronomical transients are in X-rays or Gamma rays. The observation advantages of space-based telescopes in these high-energy bands are unmatched by ground-based telescopes, and they are more suitable for transients observation, but due to the constraints of the performance of on-board computers, it is difficult to implement complex detection algorithms that rely on the powerful ground com-

* 中国科学院 GF 科技重点实验室基金项目资助(CXJJ-20 S017)

2021-12-16 收到原稿,2022-07-28 收到修定稿

E-mail: lixiaobinwf@126.com

puting power. In response to the above problems, a transient detection algorithm is proposed based on the lightweight Convolutional Neural Network (CNN) model, and the model deployment is implemented on the embedded ARM platform. The experimental results show that the model complexity and computational complexity of the lightweight CNN transients detection algorithm proposed are less than 1/4 of the Deep Hits algorithm, while the accuracy rate can reach 96.52%, and it can be applied to a space-borne limited computing power platform to realize real-time detection of space-based transients in the future.

Key words Transients detection, Lightweight CNN model, Space-based limited computing platform, Model deployment

0 引言

天文暂现源,特别是超新星、伽马射线暴(GRBs)、快速射电暴(FRBs)等极端天体爆发现象在宇宙中具有十分重要的科学研究价值,这些天体爆发现象与恒星的演化有关,因此暂现源对研究宇宙的物质演化具有重要作用^[1]。研究者通常利用大视场望远镜对天区开展系统性暂现源观测,并对感兴趣的区域进行针对性跟踪观测,以此获得大量有价值的观测数据。目前对天文图像暂现源探测的研究集中采用地面处理方法,利用地面强大的计算力实现复杂的检测算法。但是目前对天文暂现源的观测主要集中在高能波段,与地基望远镜相比,天基望远镜在高频段的观测更具优势,更容易获取高质量的天文暂现源图像。受限于星地通信链路,如何实现天基实时暂现源检测,将原始数据下传变为有效信息下传是目前亟待解决的问题。

常规的暂现源识别是利用图像相减(Difference Image Analysis, DIA)的方法实现的,这是一种直接比较天空中不同时间获取的相同位置处两幅图像的技术,根据两幅图像的特殊相减操作区分背景和目,之后确定该目标属于已知源、伪暂现源,还是新发现的暂现源^[2]。DIA 技术能够较为有效地识别暂现源,但是由于模板是领域专家通过特定方法确定的,其特征量较少,无法提取深层次的图像特征作为分类依据,因此暂现源检测的准确率较低。

近年来,在航天技术和人工智能技术飞速发展的背景下,以卷积神经网络^[3]为代表的计算机视觉方法在深空探测领域有着巨大的应用潜力,同时也为星上智能信息处理的研究和应用带来新的机遇。Cabrer-Vives 等^[4]提出了 Deep-HITS 模型,构建具有旋转不

变特性的深度卷积神经网络,实现天文暂现源的检测。Sedaghat 等^[5]利用改进的卷积神经网络实现图像差分,用于实时暂现源检测。Connor 等^[6]将多个神经网络进行集成,利用深度神经网络结合动态频谱等信息对快速射电暴单脉冲进行分类,判断其是否为真实天文暂现源。Huang 等^[7]提出一种基于随机森林(Random Forest)的天文暂现源检测方法,以 AST3-2 巡天望远镜 2016 年的观测数据为样本,结合图像匹配及相减技术,完成暂现源检测任务。

以上算法模型可以较好地实现暂现源检测任务,但是模型复杂度较高,计算量较大,均依靠地面强大算力实现暂现源检测,无法直接应用于星载有限算力平台。因此,需要设计轻量化模型算法,实现天基暂现源检测任务。

本文针对暂现源检测与星载有限算力平台的特点,提出基于轻量化 CNN 暂现源检测算法,并实现模型的终端部署。获取公开天文暂现源数据集,对其进行主要信息提取、维度变换以及数据增强等预处理操作;将数据输入到构建好的轻量化 CNN 模型中,进行模型训练以及优化;将模型部署到 ARM 嵌入式平台,实现终端的暂现源检测。提出的基于轻量化 CNN 暂现源检测算法在保证检测准确率的前提下,能够有效降低模型复杂度,更好地应用于星载有限算力硬件平台,对暂现源识别的未来空间应用起到促进作用。

1 轻量化 CNN 暂现源检测算法

暂现源检测任务本质上是检测分类问题,目前常用的卷积神经网络检测分类模型有 VGG^[8], ResNet^[9], GoogLeNet^[10], AlexNet^[11] 等,这些网络模型参数量较多,复杂度较大,若直接用于暂现源检测极易造成过

拟合问题,且算法效率和实时性较低,所需算力较大,无法应用于星载平台。针对以上问题,提出一种基于轻量化 CNN 暂现源检测算法,并将其部署到嵌入式 ARM 平台,验证其有效性。如图 1 所示,星地协同的在轨暂现源智能检测系统分为地面训练和天基部署两部分。地面训练阶段首先需要对数据集进行预处理以及轻量化 CNN 模型构建;将切分好的训练集以及对应的标签数据输入网络模型中进行训练;根据模型训练结果对模型结构参数进行优化,以获得更好的模型性能。天基部署阶段首先需要将地面训练完成的轻量化暂现源检测模型进行优化部署;将切分好的测试集输入网络模型中;网络模型进行预测推理,实现终端的暂现源检测。

1.1 地面训练

1.1.1 数据集预处理

研究采用的数据集是 HiTS(High cadence Transient Survey)^[4] 天文暂现源图像数据集,也是目前能够获取到的较为可靠的公开天文暂现源数据集。此数据集使用了 DECam 相机在 2013 年运行的数据,以 2 h 间隔观察 u 波段的 40 个场,持续 4 个晚上,共获得 802087 个负例天文图像和 802087 个正例天文图像,之后从每个样本集中各选择 717863 个数据构成最终的天文暂现源数据集。

为更好地使用数据集,需要将获取的数据集进行相关预处理,包括提取数据集的主要信息、维度变换以及数据增强等。首先,需要提取能够被用于模型训练的信息,主要包括模板图像数据(原始天区图像),科学图像数据(实际拍摄的天文图像),差分图像数据

(科学图像与模板图像进行相减得到的数据),信噪比图像数据(差分图像数据除以局部噪声,等效为每个局部信噪比分布图),图像对应的标签数据;其次,对获取的主要信息进行维度变换,因为数据集中的数据都被展开成一维,而图像数据是二维的,因此需要对原始数据进行维度变换,并且将不同图像合并,生成 4 通道图像以便进行模型训练;最后,对数据集进行切分,因为采用的数据集是百万级别的数据集,所以随机选择 98% 的总样本作为训练样本,1% 的总样本作为验证样本,剩下的 1% 作为测试样本,切分后训练集、验证集以及测试集的数目分别为 1407012, 14357, 14357。

经过对数据集的一系列预处理操作,将其转化为可用于 PyTorch 深度学习框架输入的 Tensor 数据类型,图 2 给出的是预处理后的数据集部分图像。

1.1.2 轻量化 CNN 暂现源检测模型构建

卷积神经网络一般由输入层、输出层、卷积层、池化层和全连接层等组成。输入输出层负责训练数据的输入以及最终结果输出。卷积层是卷积神经网络最重要的组成部分,具有局部连接和权值共享等特点,主要负责利用不同大小的卷积核提取输入数据的多种特征,其运算公式如下:

$$f(x) = \text{act} \left(\sum_{i,j}^n \theta_{(n-i)(n-j)} x_{ij} + b \right). \quad (1)$$

其中, θ 表示输入特征图, x 表示卷积核, b 表示偏置, act 表示激活函数, θ 与 x 逐点相乘加上偏置获取特征,再通过激活函数赋予非线性化因素。池化层主要负责提取图像的全局信息,有减少参数量及减轻模型过拟合程度的作用,常用的有平均池化和最大池化等。全连接层的主要作用是将学到的特征表示进行整合分类,输出最终分类结果。为了使模型具备更强的表达能力,通常还会引入非线性激活函数,激活函数会给模型赋予非线性因素,使其能够完成线性模型无法解决的问题,常用的激活函数有 RELU 函数、Sigmoid 函数以及 tanh 函数等,因为 RELU 函数可以很好地解决梯度消失的问题,有助于网络的训练优化,故常用 RELU 作为激活函数。RELU 函数形式如式(2)当 $x \geq 0$ 时,函数值为 x ; 当 $x < 0$ 时,函数值为 0,即

$$f(x) = \begin{cases} x, & x \geq 0; \\ 0, & x < 0. \end{cases} \quad (2)$$

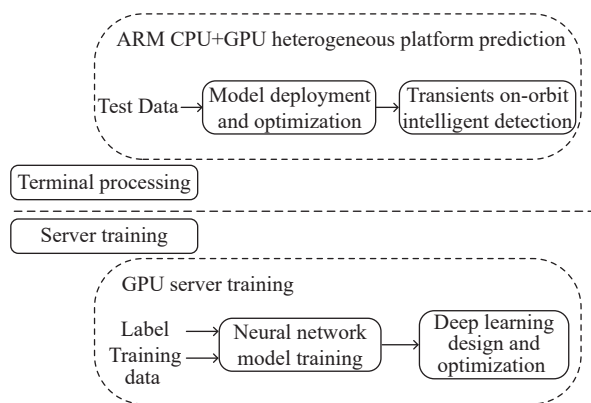


图 1 星地协同在轨暂现源智能检测系统

Fig. 1 Satellite-ground coordinated on-orbit transients intelligent detection system

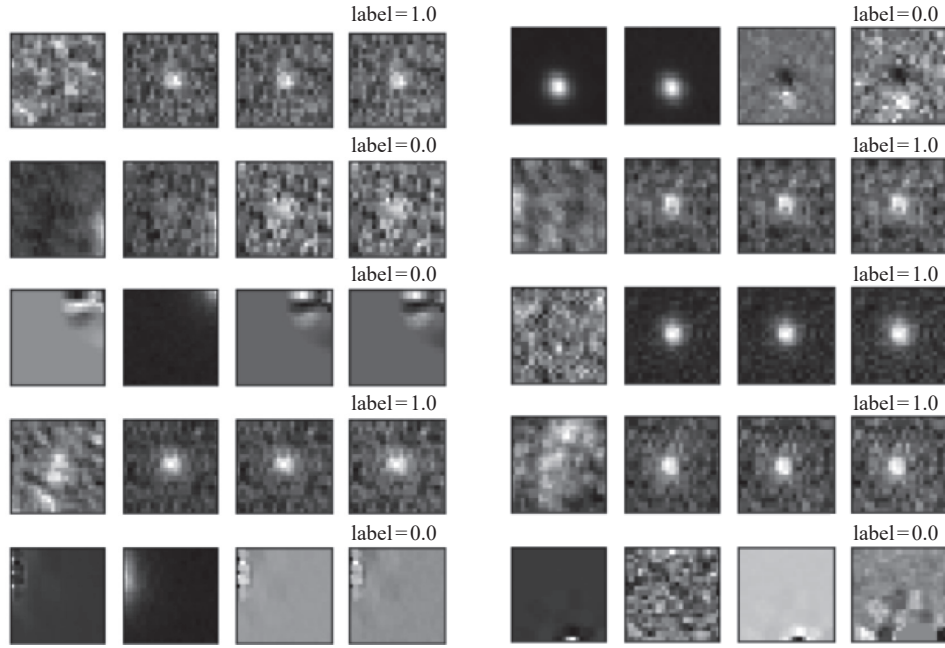


图 2 预处理后的图像数据。每一组图像从左到右依次为模板图像、科学图像、差分图像、信噪比图像。

label=1.0 代表真实暂现源, label=0.0 代表伪暂现源

Fig. 2 Preprocessed image data. Each group of images from left to right are template image, scientific image, differential image, signal-to-noise ratio image. label=1.0 represents the real transients, label=0.0 represents the pseudo transients

目前用于图像处理的 CNN 模型大多利用设计复杂网络结构、加深网络层数等方法提高检测分类效果,以处理一幅 $4 \times 1 \text{ kByte} \times 1 \text{ kByte}$ 的图像为例,对应到本数据集的数据格式应为 $2400 \times 4 \times 21 \text{ pixel} \times 21 \text{ pixel}$ 的图像,即处理 2400 幅 4 通道图像, VGG16 的模型复杂度为 138 MByte, 计算量为 306 GFlops; Resnet34 的模型复杂度为 22 MByte, 计算量为 74 GFlops。这些大模型具备很好的图像处理能力,但是存在模型复杂度过高、推理时所需算力较大、功耗较高等问题,仅适用于地面处理,不适用于星载有限算力平台。采用的硬件平台为 EAIDK-610, 其单精度浮点计算峰值为 38.4 GFlops, 无法满足上述复杂模型的算力需求,且星载平台通常多个任务并行,无法将全部算力提供给一个任务,需要构建复杂度低、计算量小的轻量化模型应用于星载有限算力平台。因此,提出构建轻量化 CNN 模型以适用于星载处理器的计算性能,实现未来天基暂现源实时检测任务。

构建的轻量化 CNN 网络模型如图 3 所示,其共有 10 层结构,包括 5 层卷积层、2 层池化层以及 3 层全连接层,其中 C 代表卷积层、P 代表池化层、Fc 代

表全连接层。选用 5 层卷积层是因为经实验验证,其完全胜任暂现源图像的特征提取任务,同时利用 2 层池化层进行图像全局特征提取,可使得模型获取的特征更为丰富。模型输入层数据的维度为 $4 \times 21 \times 21$, 输出层输出预测分类结果; C1 层采用 4×4 的卷积核,步长为 1, 填充为 3; C2, C3, C4, C5 层采用 3×3 的卷积核,步长为 1, 填充为 1; 池化层采用最大池化策略; 激活函数采用 RELU 函数。

目前针对 HiTS 数据集训练性能表现较好的是 Deep-HiTS 模型,其参数量为 1705.67 kByte, 计算量为 97.36 MFlops; 本文构建的轻量化 CNN 暂现源检测模型是针对天基暂现源检测任务而设计的,其参数量为 407.10 kByte, 计算量为 20.13 MFlops, 相比 Deep-HiTS 模型, 复杂度较低, 在保证检测准确率的前提下缩短了推理过程的时间, 降低了所需算力。

1.1.3 轻量化 CNN 暂现源检测模型训练优化

模型训练优化的目的是优化模型参数,提升模型性能,使模型具备更好的暂现源识别检测分类能力。模型训练的优化过程如下: 设置好相关的超参数, 设置初始学习率为 0.001, 训练批次大小为 128, 迭代次数为 1, 采用交叉熵损失函数和 Adam 优化函数; 输

入层分批次导入待训练的数据,依次经过卷积层、池化层、全连接层,随后进行反向传播更新参数;经过多次迭代,得到最终的训练结果,整个训练流程耗时 20 min。模型训练流程如图 4 所示。

1.2 ARM 平台部署

1.2.1 嵌入式 ARM 平台

NASA 在 2012 年提出用 8 核 A53 ARM 处理器 RAD5545 实现无人或载人任务的空间智能应用,ESA 提出用 4 核 A72 ARM 处理器 P78548 实现空间计算和智能处理。目前中国生产制作高等级 ARM 芯片机构主要有飞腾、华为海思、瑞芯微、申威等公司,以飞腾公司为代表,其具有全链条生态、全自主可控的麒麟操作系统条件以及地面成熟的嵌入式智能解决方案,目前其 FT2000 芯片已通过天智一号在轨飞行验证,未来将投入实际应用。综上可知,开发高等级 ARM 芯片实现星载智能处理是未来航空领域的重要发展趋势。

采用的嵌入式开发平台为 EAIDK-610,是全球首个采用 ARM 架构的嵌入式人工智能开发平台,主

要面向边缘计算。主芯片是 RK3399,具有 6 个内核,分别为 2 个 ARM Cortex-A72 核和 4 个 ARM Cortex-A53 核, GPU 采用四核 ARM 图像处理器 Mali-T860。选用 RK3399 芯片是因为其具有强大的算力,足以适用星上智能处理,并且 RK3399 对标 FT2000, FT2000 有三个内核,分别为 2 个 ARM Cortex-A53 和 1 个 ARM Cortex-A55,其 GPU 计算性能可以达到 8 GFlops。尽管 FT2000 的计算性能略低于 RK3399,但是能够完全满足提出的轻量化 CNN 暂现源检测模型的性能需求,同时两款处理器同为 ARM 架构,选用 RK3399 是为以后能够将模型更好地部署到 FT2000 中。

1.2.2 模型部署流程

部署模型之前需要在嵌入式系统平台上配置软件运行所需硬件环境,由于选用的 EAIDK-610 开发平台自带 Tengine 预编译库,无须手动编译。因此,开发平台的环境搭建只包括编译固件和固件烧录两部分。

开发平台环境搭建后,将训练好的轻量化 CNN

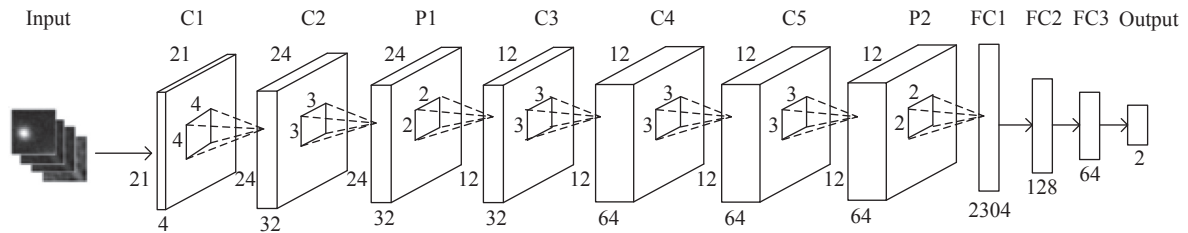


图 3 轻量化 CNN 网络模型架构 (C 代表卷积层、P 代表池化层、FC 代表全连接层,池化层采用最大池化策略,激活函数采用 RELU 函数)

Fig. 3 Lightweight CNN network model architecture (C represents the convolutional layer, P represents the pooling layer, FC represents the fully connected layer, the pooling layer adopts the maximum pooling strategy, and the activation function adopts the RELU function)

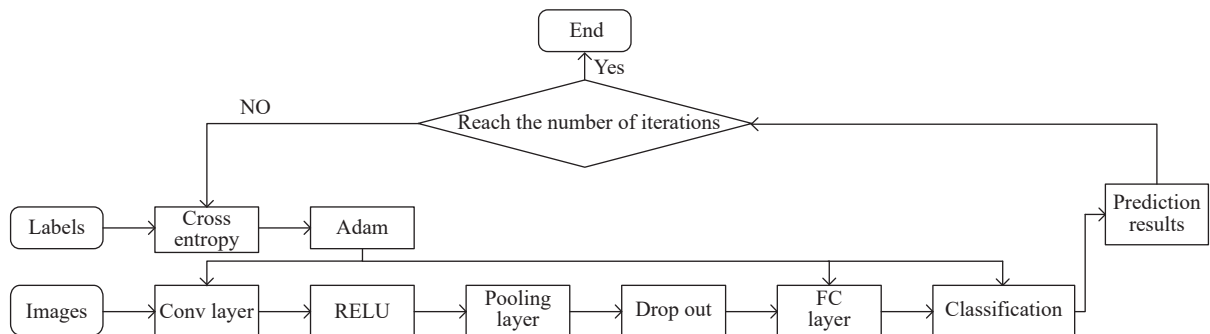


图 4 轻量化 CNN 网络模型训练流程 (损失函数采用交叉熵损失函数,优化函数采用 Adam 优化函数)

Fig. 4 Lightweight CNN network model training process (Loss function adopts the cross-entropy loss function, and the optimization function adopts the Adam optimization function)

模型转换为 ONNX 模型;对 ONNX 模型进行优化,使模型能够更好地适用于硬件平台,尽可能发挥硬件平台的最大算力;将优化后的 ONNX 模型转换为 Tengine 模型,部署到嵌入式 ARM 平台上,实现终端的推理功能。部署流程如图 5 所示。

开放神经网络交换(Open Neural Network Exchange, ONNX)是一种模型表达式,具有与框架无关的优点,将训练好的神经网络模型转化为 ONNX 模型后,不用再考虑硬件平台与多种框架兼容性的问题。ONNX 本质是一种文件格式,定义了模型(Model)、计算图(Graph)、计算节点(Node)、张量(Tensor)等层面的数据结构,并通过 Protobuf 存储神经网络模型的权重参数。同时,ONNX 可以对训练好的网络模型进行优化,在不改变模型结构参数前提下使其更易于部署到硬件平台。经试验验证,本文构建的轻量化模型部署后可以正常运行,并且模型准确率无明显变化,能够有效地实现终端暂现源检测。

2 暂现源检测结果分析

2.1 地面训练结果

计算机视觉图像分类任务的评价方法通常采用准确率(Accuracy, 定义符号 A),准确率表示正确预测的样本占总样本的比例,代表分类器整体性能的好坏。在二分类的情况下,模型预测结果通常分为 4 种情况,即真正例(TP, 定义符号 P_T)、假正例(FP, 定义符号 P_F)、假反例(FN, 定义符号 N_F)、真反例(TN, 定义符号 N_T),预测结果的混淆矩阵列于表 1。

准确率表示正确预测的样本占总样本的比例,代表分类器的整体性能,其计算公式为

表 1 预测结果混淆矩阵
Table 1 Confusion matrix of prediction results

实际情况	预测结果	
	正例	反例
正例	6901(P_T)	299(N_F)
反例	151(P_F)	6956(N_T)

$$A = \frac{P_T + N_T}{P + N}. \quad (3)$$

其中, P 代表预测结果为正例的总数, N 代表预测结果为反例的总数, $P+N$ 代表样本总数。通过计算可得到模型的准确率为 96.52%。

采用 14357 个暂现源图像作为测试样本,进行轻量化 CNN 模型验证工作。首先对模型进行训练及优化,通过调整训练参数等方法使模型具备良好的效果,再进行模型推理预测。轻量化 CNN 模型的训练损失率及测试准确率分别如图 6 和图 7 所示。其中,模型损失率是指模型在训练过程中,通过预设定的损失函数计算模型输出标签与数据集样本实际输入标签之间的差异,以此评价模型性能,如果经过模型迭代训练,最终损失率收敛到一个较小的值则证明该模型性能良好。

由图 6 和图 7 可知,模型训练的损失率最终收敛

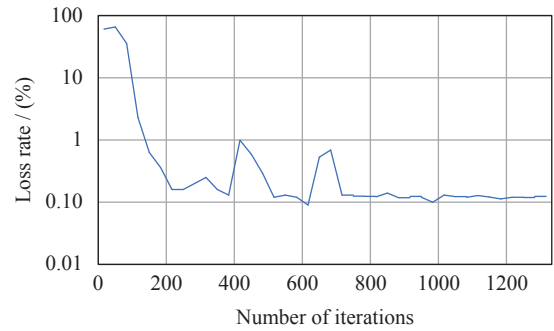


图 6 训练损失率

Fig. 6 Training loss rate

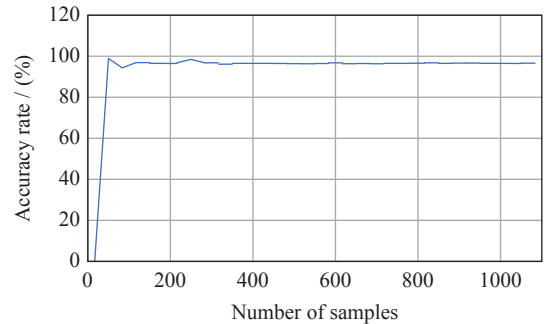


图 7 测试准确率

Fig. 7 Test accuracy rate

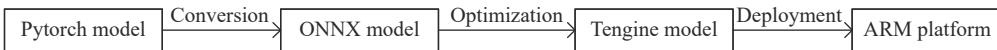


图 5 模型部署流程

Fig. 5 Model deployment process

到 0.12, 准确率达到 96.52%, 获得了较为理想的暂现源检测效果, 满足暂现源的检测任务需求。

为进一步验证模型的准确率以及轻量化程度, 针对准确率、参数量、计算量等指标, 将轻量化 CNN 暂现源检测模型与 Cabrera-Vives 等^[4]提出的 Deep-HITS 模型以及 Huang 等^[7]提出的基于随机森林+图像相减的天文暂现源检测方法(R-DIA)进行了对比(见表 2)。由于 Huang 等^[7]没有说明所构建的模型复杂度等信息, 因此这里只针对其准确率进行比较。

由表 2 可知, Huang 等^[7]提出的基于随机森林 + 图像相减的天文暂现源检测方法准确率较低, 仅为 90.89%。本文提出的轻量化 CNN 暂现源检测模型相比 Deep-HITS 模型准确率降低了 2.93%, 但是模型的参数量、模型大小以及计算量不及 Deep-HITS 模型的 1/4, 证明本文算法模型在保证暂现源检测准确率的基础上, 大大降低了模型复杂度。模型复杂度的降低是进行模型硬件部署的关键, 这使得轻量化 CNN 暂现源检测模型能够更好地部署到星载有限算力平台中, 实现终端的暂现源检测。

2.2 硬件部署结果

轻量化 CNN 暂现源检测模型在完成训练优化后, 需将其部署到 EAIDK-610 硬件平台上, 实现基于终端的暂现源检测。完成模型部署流程后, 最终模型成功在 EAIDK-610 平台上正常运行。以 500 幅尺寸为 4×21 pixel×21 pixel 的 HiTS 图像作为测试样本, 将其输入到部署后的模型中, 模型预测分类准确率为 96.51%, 整体耗时 8.01 s。由部署后的模型性能可知, 构建的轻量化 CNN 暂现源检测模型可以很好地适用于硬件平台, 在运行速度较快的前提下保持较高的暂现源分类准确率, 满足未来天基暂现源实时检测任务的可靠性及实时性要求。

3 结论

通过分析天文暂现源以及星载有限算力平台的

特点, 提出了基于轻量化卷积神经网络的天文图像暂现源检测算法, 并在嵌入式 ARM 平台上实现了部署。在大大降低模型复杂度的同时实现了精确的天文图像暂现源检测分类。根据实验结果, 本文模型参数量仅为 Deep-HITS 模型的 1/4, 计算量仅为 Deep-HITS 模型的 20%, 识别准确率达到 96.52%, 能够更好地部署在有限算力的星载平台中, 对天基实时暂现源检测任务有着很高的应用价值, 可以促进暂现源识别的未来空间智能应用。在未来应用时, 应根据不同类型的暂现源数据集对模型进行微调训练, 使模型能够更好地适用于特定类型的暂现源分类。

参考文献

- [1] YU Hai. Applications of Transients and Other Observations in Cosmology[D]. Nanjing: Nanjing University, 2019
- [2] ZHAO Yifei. Astronomical Transient Source Recognition Based on Deep Learning and Raspberry Pi[D]. Taiyuan: Taiyuan University of Technology, 2019
- [3] PENG Bin, LI Zheng, LIU Yong, *et al.* Automatic code comments generation method based on convolutional neural network[J]. *Computer Science*, 2021, **48**(12): 117-124
- [4] CABRERA-VIVES G, REYES I, FÖRSTER F, *et al.* Deep-HITS: rotation invariant convolutional neural network for transient detection[J]. *The Astrophysical Journal*, 2017, **836**(1): 97
- [5] SEDAGHAT N, MAHABAL A. Effective image differencing with convolutional neural networks for real-time transient hunting[J]. *Monthly Notices of the Royal Astronomical Society*, 2018, **476**(4): 5365-5376
- [6] CONNOR L, VAN LEEUWEN J. Applying deep learning to fast radio burst classification[J]. *The Astronomical Journal*, 2018, **156**(6): 256
- [7] HUANG Tianjun. Research on Detecting Transients and Variable Sources in AST3-2 Survey[D]. Hefei: University of Science and Technology of China, 2019
- [8] CHEN Xiaotong. Research on aviation sheet metal parts missing detection technology based on VGG technology[J]. *Modern Industrial Economy and Informationization*, 2021, **11**(9): 219-220
- [9] XU L Y, GAJIC Z. Improved network for face recognition based on feature super resolution method[J]. *International Journal of Automation and Computing*, 2021, **18**(6): 915-925
- [10] RAN H H, WEN S P, SHI K B, *et al.* Stable and compact design of memristive GoogLeNet neural network[J]. *Neuro-computing*, 2021, **441**: 52-63
- [11] TAN Y H, JING X J. Cooperative spectrum sensing based on convolutional neural networks[J]. *Applied Sciences*, 2021, **11**(10): 4440

表 2 模型复杂度对比

Table 2 Comparison of model complexity

模型名称	参数量 /kByte	模型大小 /MByte	计算量 /MFlops	准确率/(%)
Deep-Hits ^[4]	1705.67	6.86	97.36	99.45
轻量化CNN	407.10	1.57	20.13	96.52
R-DIA ^[7]	—	—	—	90.89