DOI: 10.13382/j.jemi.B2407775

一种地基云图分类算法及硬件加速实现*

冯 琳[1,2] 宋文强[1,2] 徐 伟[1,2]

(1. 南京信息工程大学气象灾害预报预警与评估协同创新中心 南京 210044;
2. 南京信息工程大学江苏省气象探测与信息处理重点实验室 南京 210044)

摘 要:地基云的自动观测和识别对分析大气运动趋势和天气预测具有指导意义。针对目前地基云图分类算法准确率不高、在嵌入式终端部署困难的问题,提出了一种基于残差网络结构的地基云图分类网络模型 GBcNet 及基于 ZYNQ 的硬件实现架构,PS 端用于加载模型的权重参数和云图数据,PL 端实现 DDR3 读写控制和 GBcNet 的硬件加速。设计了滑窗、卷积层、池化层、批量归一化层和全连接层等模块的加速 IP 核。实验在 CCSN 数据集上进行,结果表明,提出的模型在 PC 端的准确率达到 96.02%。采用现场可编程门阵列(FPGA)硬件加速后,准确率仍然保持在 94.5%。与 PC 端模型的识别率相比,各云类的识别精度损失均不超过 3%,整体精度损失小于 1.5%;FPGA 的最大资源占用不超过 48%,单张地基云图推理时间为 0.13 s。相较于现有地基云的识别方法,识别准确率高且推理时间较短。提出的识别模型和硬件加速方法为便携式地基云观测设备的研制提供了一种参考方案。

关键词:图像分类;地基云图;硬件加速;FPGA;残差网络

中图分类号:TP391;TN919.8 **文献标识码**:A **国家标准学科分类代码**:510.4050

Hardware acceleration implementation of a ground-based cloud image classification algorithm

Feng Lin[1,2] Song Wenqiang[1,2] Xu Wei[1,2]

(1. Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters Nanjing University of Information Science and Technology, Nanjing 210044, China; 2. Jiangsu Key Laboratory of Meteorological Observation and Information Processing, Nanjing University of Information Science and Technology, Nanjing 210044, China)

Abstract: The automatic observation and recognition of ground-based clouds have guiding significance for analyzing atmospheric motion trends and weather forecasting. To solve the problem of low accuracy of ground-based cloud image classification algorithms and difficulties in deploying them on embedded terminals, a ground-based cloud image classification network model GBcNet based on residual network structure and a hardware implementation architecture based on ZYNQ are proposed. The PS end is used to load the weight parameters and cloud data of the model, and the PL end implements DDR3 read-write control and GBcNet hardware acceleration. For the GBcNet network, accelerated IP cores corresponding to each module were designed, including sliding window, convolutional layer, pooling layer, batch normalization layer, and fully connected layer. Experiments were conducted on the CCSN dataset, and the results showed that the GBcNet model achieved an accuracy of 96.02% on a PC platform. After hardware acceleration, the accuracy remained at 94.5%. Compared with the recognition rate of the PC model, the accuracy loss for each cloud class did not exceed 3%, and the overall accuracy loss was less than 1.5%. The maximum resource usage on the FPGA did not exceed 48%, and the inference time for a single ground-based cloud image was 0.13 seconds. Compared to existing ground-based cloud recognition methods, this approach demonstrates higher accuracy and shorter inference time. The proposed recognition model and acceleration method provide a reference solution for the development of multi-node, portable ground-based cloud observation equipment.

Keywords: image classification; ground-based cloud image; hardware acceleration; FPGA; ResNet

收稿日期:2024-08-18 Received Date: 2024-08-18

*基金项目:国家重点研发计划政府间/港澳台重点专项(2021YFE0105500)、江苏省研究生科研与实践创新计划项目(SJCX24_0470)资助

0 引 言

云极大地影响了地球的辐射、水文循环和气候变化[1]。云的准确识别有助于分析太阳辐照度、水汽含量和大气运动,对光伏发电功率、天气趋势和恶劣天气变化预测具有重要意义[2]。世界气象组织根据云的形状、结构、高度和形成原因等方面将云划分为 3 族 10 属 29 类。然而,云具有种类多、变化快的特点,同一种类的云在不同的地域、天气条件下也会呈现一定的差异性,这些问题给地基云图的精细化识别带来了挑战。目前,地基云的观测除了利用雷达外,还可以利用地基云观测设备全天空成像仪[3-4](total-sky imager, TSI)或(all-sky imager, ASI),对云图进行人工判别,识别结果依赖于观测人员的专业知识,且人工观测的工作量较大,识别的准确率难以得到保证。

云图是一种形状不规则的特殊纹理图像,传统的机器学习方法由于纹理和模式的复杂性而难以提取云的特征,分类准确率不高[5]。近年来,深度学习方法在地基云图分类领域逐渐成为热点。神经网络作为分层特征提取框架表现出强大的性能,网络分类的自适应学习特征解决了人工识别带来的弊端,能够有效提取云图中的细节特征,以相对较高的准确率完成分类任务[6]。特别是一些大规模地基云图数据集的提出,使得地基云图在自动化分类方面取得了很大进展。

Zhang 等[7]创建了符合世界气象组织规定的卷云积云层云雨云数据集(cirrus cumulus stratus nimbus, CCSN),基于 AlexNet 网络进行改进提出了一种适用于地基云图分类的卷积神经网络模型 CloudNet,但准确率仅为 89.48%; Zhang 等[8]构建大规模公开云数据集 HBMCD,采用可分离卷积、膨胀卷积等技术搭建轻量级地基云识别网络 LCCNet,为地基云图识别算法推向应用提供了一套方案,但算法仅在 PC 平台实现,未进行嵌入式终端部署;Li 等[9]以卷积神经网络模型作为底层特征提取工具,利用 Transformer 模型的优势学习地基云图的全局特征,二者结合提出新的分类模型,在 CCSN 数据集上的准确率评估达到 92.73%,但 Transformer 模型中的自注意力机制计算复杂度为序列长度的平方级,大大增加了本地算法部署的难度,对服务器设备的内存要求极高; Li 等[10]以 DenseNet 架构为主干重新设计,搭建新的云密集模块以提高网络识别精度,在 CCSN 数据集上实现了 90.7%的分类准确率,但其结构中大量的稠密连接导致内存和计算资源的需求显著增加,因此对资源受限的边缘化部署不友好[11]。

通过增加模型复杂度或加深网络层数以提升分类效果,这导致模型计算的空间复杂度和参数量剧增。消耗大量的计算和存储资源,对服务器设备的有更高的要求,且延迟较高。通常采用中央处理器(central processing unit, CPU)和图形处理器(graphics processing unit, GPU)进行加速。但 GPU 功耗高且增加了嵌入式硬件成本[12]。

综上,目前地基云图自动分类识别方法主要存在分类准确率较低、网络规模大导致存储与计算资源开销大等问题。为提高地基云图分类算法的综合性能,需要算法和硬件联合优化[13]。一方面,基于残差结构提出一种适用于嵌入式系统中部署的地基云分类识别网络模型;另一方面,充分利用现场可编程门阵列(FPGA)强大的并行处理能力与计算效率,对网络模型设计对应的硬件加速核[14]。在保证分类准确的同时,降低模型的前向推理时间和网络带宽消耗,实现边缘化实时云图处理。

1 地基云图分类算法

1.1 地基云图数据集 CCSN

CCSN 数据集是符合世界气象标准的地基云图数据集,共 11 种类别,2 543 张云图,分辨率大小为 400×400 pixels,其中一种为飞机的尾迹云[15],图 1 展示了数据集的部分云属。

图 1 CCSN 数据集

Fig. 1 Presentation of the CCSN

为提高模型的鲁棒性和泛化性,利用图像增强技术将数据集扩充到 12 715 张,训练集与测试集按照 9∶1 进行划分,如表 1 所示。

1.2 基于残差结构的地基云图识别方法

1) 残差结构原理

地基云种类多、变化快、易与天空背景相融合,其边缘细节特征的提取决定了云图分类的结果。简单的浅层卷积神经网络能学习的云图特征有限,识别准确率不高。为便于后期进行边缘部署,在不增加网络算法复杂度的前提下,通过加深网络层数提取更多精细化的云图边缘

特征,实现高准确率的分类结果。

表 1　CCSN 数据集

Table 1　CCSN dataset

类别	扩充前	扩充后	训练	测试
积云	182	910	819	91
积雨云	242	1 210	1 089	121
层积云	340	1 700	1 530	170
层云	202	1 010	909	101
雨层云	274	1 370	1 233	137
高层云	188	940	846	94
高积云	221	1 105	994	111
卷云	139	695	625	70
卷层云	287	1 435	1 291	144
卷积云	268	1 340	1 206	134
尾迹云	200	1 000	900	100
共计	2 543	12 715	11 442	1 273

残差网络[16-17]是由何凯明等提出的一种卷积神经网络结构,由多个残差块组成,通过输入与输出相加的方式来引入残差连接,结构如图 2 所示。

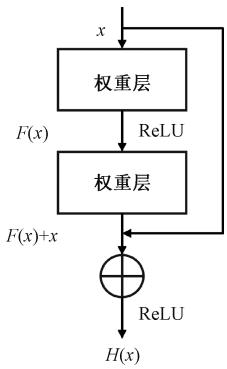

图 2　残差块结构

Fig. 2　Residual block structure

其中,x 和 $H(x)$ 分别为残差块的输入和输出,$F(x)$ 为学习的残差映射,整个残差块将输入 x 经过若干个卷积和 ReLU 激活函数后,通过捷径分支将输入与输出进行叠加。输入的数据可以从任意低层直接传播到高层,解决了由于网络层数加深而带来的梯度消失和网络退化的问题,同时大幅度抑制过拟合,提高训练速度和准确率。这种捷径分支相当于简单执行了同等映射,不会产生额外的参数,也不会增加计算复杂度。

2) 地基云识别网络模型的设计

基于残差块提出一种用于地基云图识别的网络模型 GBcNet[18],结构如图 3 所示。输入的云图经预处理后,分辨率降至 224×224 pixels。浅层网络使用大小为 7×7 卷积核提取云图浅层特征,利用最大池化层对其进行降采样,池化核大小为 3×3,输出特征图大小为 56×56。依次经过 5 个残差块结构,主分支采用大小分别为 1×1 和 3×3 的卷积核完成局部云图特征提取,部分残差分支上利用卷积核大小为 1×1 的点向卷积对特征图在通道维度上进行调整,以充分提取地基云图的纹理特征。利用自适应平均池化层将特征图降采样,最后通过全连接层和 Softmax 分类器对 11 种地基云图进行预测,输出分类结果。

设计的网络模型中每一个卷积操作之后都进行了批归一化(batch normalization,BN)处理,通过标准化每一小批次的输入,使得每层的输入特征数据具有零均值和单位方差,加速网络的收敛速度,提高网络的鲁棒性。激活函数选择 ReLU,它能够通过一次简单的比较运算学习到复杂的模式,降低计算复杂度和时间消耗。ReLU 函数的稀疏性特性意味着会输出大量的零,稀疏矩阵[19]的存储方式进一步优化内存使用,对于低功耗和资源受限的硬件部署非常友好。

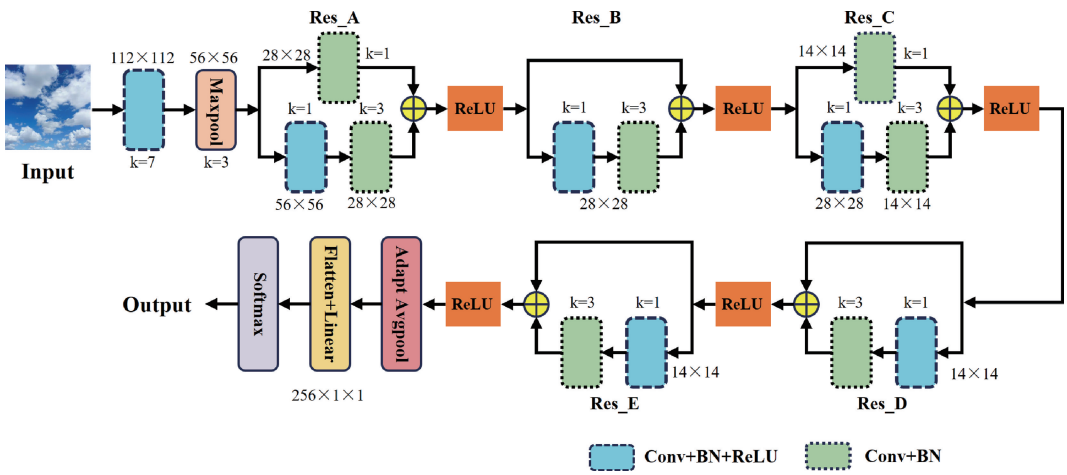

图 3　地基云图分类模型 GBcNet

Fig. 3　Ground-based cloud image classification model GBcNet

2 硬件架构设计

2.1 整体硬件架构

提出一种用于加速地基云图分类网络的硬件架构，主要由可处理系统(processing system，PS)端和可编程逻辑(progarmmable logic，PL)端两部分构成[20]，图 4 所示为整体网络硬件加速设计。

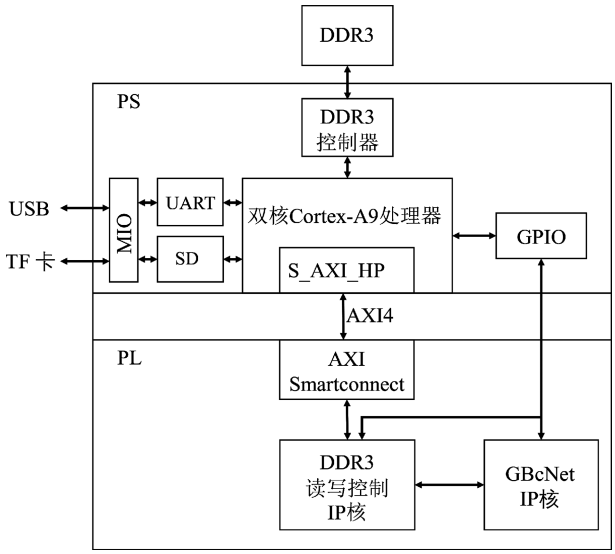

图 4 网络加速硬件架构

Fig. 4 Network acceleration hardware architecture

PS 端用于加载模型的权重参数和云图数据，以减少 FPGA 的资源占用。使用 FATFS 文件管理系统获取 TF 卡中的权重参数和云图数据，将读到的数据加载到第三代双倍数据率同步动态随机存取存储器(double-data-rate three synchronous dynamic random access memory，DDR3 SDRAM)中，便于 PL 端从 DDR3 中获取数据，数据写入 DDR3 后通过 GPIO 端口向 PL 端传输权重加载完成的使能信号。

PL 端主要实现 DDR3 读写控制 IP 核和 GBcNet IP 核的编写和封装，DDR3 读写 IP 核主要实现与 PS 端的数据交互，包括权重和云图的读取以及特征图的缓存。设计 GBcNet IP 核，利用 FPGA 并行运算的能力，实现网络模型 GBcNet 的硬件加速[21]。

2.2 DDR3 读写控制 IP 核设计

1) DDR3 内存空间分配

卷积神经网络的前向推理过程是逐层进行的，上一层的输出作为下一层的输入，层与层之间的计算存在数据依赖，需利用 DDR3 实现大量数据的缓存。卷积层、BN 层和全连接层均需要权重输入，为保证各项数据的互不干扰，将内存区域划分为了 5 个部分，包括云图数据、卷积层权重、BN 层权重、全连接层权重和所有层的特征图输入输出数据，如图 5 所示。

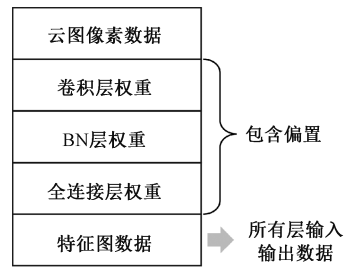

图 5 内存空间分配图

Fig. 5 Memory space allocation graph

云图像素数据和权重数据由 PS 端提前加载到 DDR3 中，数据读取时，PL 端 DDR3 读写控制 IP 核执行按照顺序读取。卷积层的权重按照层、输出通道、输入通道和卷积核的顺序依次排列，以一个卷积核所包含的权重为单位，存放的数据依次是第一个输出通道对应的每一个输入通道的卷积核的权重，第二个输出通道对应的每一个输入通道卷积核的权重。当输入通道数为 3，输出通道数为 64，卷积核尺寸为 7×7 时，该卷积层共有 9 408 个权重参数，在 DDR3 中的位置分布情况如图 6 所示。图 6 中左侧第 1 行 0~49×3×1-1 表示第 1 个输出通道对应的 3 个输入通道的权重参数的相对地址，共有 64 行，对应 64 个输出通道，每部分每行的 49 个数据对应一个 7×7 的卷积核。图中右侧表示该卷积层的所有权重参数对应的相对地址。

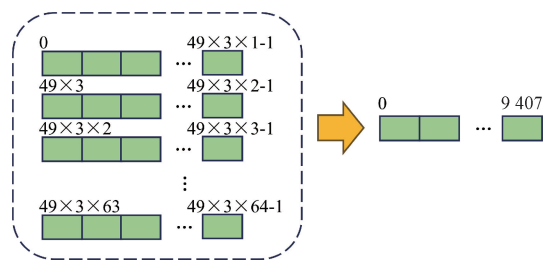

图 6 卷积层权重数据排列

Fig. 6 Convolution layer weight data arrangement

BN 层的权重按层、输入通道和输出通道顺序排列，所有权重全部排列之后，按照相同的顺序排列偏置，BN 层的偏置排列在权重之后。全连接层的权重按照输入通道、输出通道的顺序排列，偏置排列在所有权重之后。数据部分则按神经网络前向推理，依次将每一层的输入特征图的数据存储到 DDR3 中，前一层的输出作为后一层的输入，将整个网络前向推理过程中的层输入数据和输出数据都存储到 DDR3 中。

2) DDR3 读写控制

ZYNQ 系列 FPGA 的 DDR3 挂载到了 PS 端，需要 PL

端利用 AXI4 协议实现对 PS 端 DDR3 的读写控制。AXI4 协议是一种高性能、高带宽、低延迟的片内总线,主要由读地址通道、读数据通道、写地址通道、写数据通道和写响应通道 5 个独立的通道构成。5 个通道使用相同的握手机制传输数据和控制信息。主机产生 VALID 信号来指明接收数据和控制信息有效,而从机产生 READY 信号来指明已经准备好接收数据和控制信息。数据传输发生在 VALID 和 READY 信号同时为有效高电平时,AXI4 通信时序图如图 7 所示。

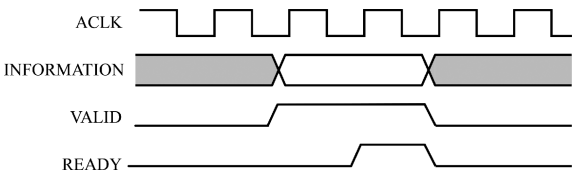

图 7　握手机制时序图

Fig. 7　Handshake mechanism timing diagram

DDR3 读写控制 IP 核内部采用了读写控制模块和 5 个先进先出数据缓存器(first in first out, FIFO)实现数据的传输,图 8 所示为 DDR3 读写控制 IP 核设计框图。特征图写 FIFO 和特征图读 FIFO 分别用于缓存神经网络中每一层的输出和读取网络中每一层的输入数据,卷积层权重读 FIFO、BN 层权重读 FIFO 和 BN 层偏置读 FIFO 用来缓存从 DDR3 中读出的权重参数。DDR3 读写控制模块主要采用状态机实现从 DDR3 中相应地址区域读取相应的数据并缓存在对应的 FIFO 中,同时将特征图数据写 FIFO 中的数据写入相对应的地址区域,利用状态机完成各个功能的仲裁,避免了数据之间混乱,提高了数据传输的效率。

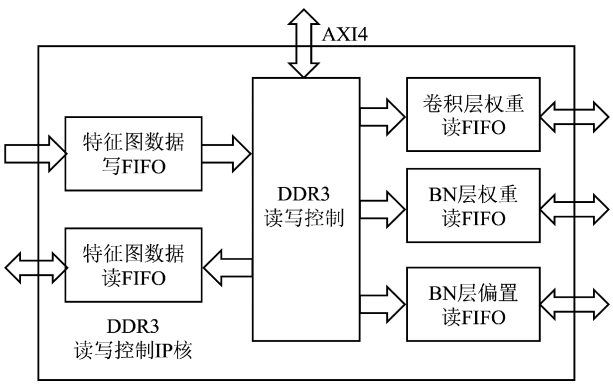

图 8　DDR3 读写控制 IP 核设计框图

Fig. 8　DDR3 read and write control IP core design block diagram

2.3　GBcNet IP 核设计

1) 滑窗模块设计

滑窗模块是卷积层和池化层的重要组成部分,采用滑窗模块形成特定尺寸的矩阵,进而实现卷积和池化操作。FIFO 作为数据缓冲器,具有先进先出的特点,因此采用 FIFO 实现数据滑窗。图 9 所示为一个 7×7 的特征图,数据按行进行读取。设置滑窗尺寸大小为 3×3 时,FIFO 读取数据方式如图 10 所示。

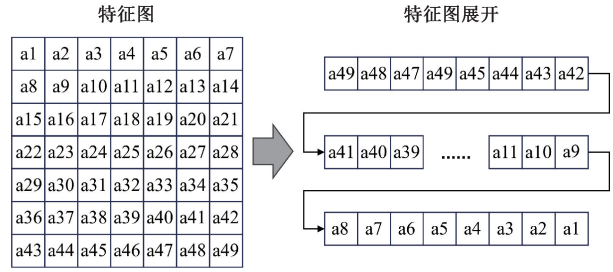

图 9　特征图串行输入

Fig. 9　Feature map serial input

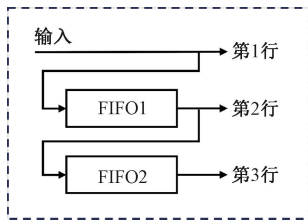

图 10　滑窗模块框图

Fig. 10　Slide window module block diagram

利用计数器对输入数据按行与列分别计数,当列计数器计数到特征图的列尺寸大小时归 0,行计数器加 1。输入的第 1 行数据作为滑窗的第 3 行数据,同时将输入的第 1 行数据按照输入顺序依次缓存到 FIFO1 中。当开始输入特征图的第 2 行数据时,输出 FIFO1 中的数据作为滑窗的第 2 行数据,输入作为滑窗的第 3 行数据,同时将 FIFO1 输出的数据按照输入顺序依次缓存到 FIFO2 中。当开始输入特征图的第 3 行数据时,输入为滑窗的第 3 行数据,FIFO1 的输出作为滑窗的第 2 行数据,FIFO2 的输出作为滑窗的第 3 行数据,此时特征图中的第 1 个矩阵形成,依次输入数据实现滑动矩阵。当需要实现不同步长的滑窗时,对输出信号采用使能控制,模型中 7×7 的滑窗可采用 6 个 FIFO 实现。

2) 卷积层模块设计

地基云图分类网络的计算量主要由卷积运算产生,卷积运算是 FPGA 主要的加速对象。GBcNet 网络第 1 层卷积层的卷积核大小为 7×7,输入通道数为 3,输出通道数为 64,将卷积层拆分为 64 次三维卷积循环,1 次三维卷积循环操作将输入的三通道特征图同时与对应的卷积核进行卷积运算。第 1 层卷积层将特征图大小从 224×224 降低到 112×112,3 个通道同时进行卷积运算,可以显著提升第 1 层的运算速度。考虑到 FPGA 片上资

源占用,将 1 次三维卷积运算循环 64 次,就能够实现第 1 层卷积层的所有卷积运算,不仅提升了计算性能,还节约了资源。

1 次三维卷积循环操作如图 11 所示,输入特征图尺寸为 7×7×3,若依次串行与对应的权重进行相乘操作,1 次三维卷积循环操作的一个卷积操作需要 49 个时钟周期。而利用 FPGA 的并行能力,部署 49 个乘法器同时进行乘法运算,同样的卷积操作仅消耗 1 个时钟周期,与一维串行卷积相比,1 个 7×7 的卷积运算节省了 48 个时钟周期,速度提升了 49 倍,1 个 7×7 的卷积层速度提升 147 倍。

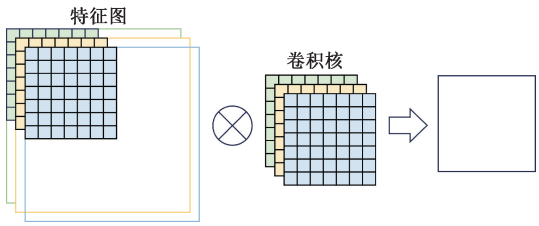

图 11 一次三维卷积循环

Fig. 11 A three-dimensional convolution cycle

GBcNet 网络模型的残差模块中使用了大量的 3×3 卷积,可将 3×3 的卷积模块进行复用,降低 FPGA 资源占用,其中 3×3 的卷积采用与 7×7 卷积相同的卷积运算方式。假设输入通道数为 N,输出通道数为 M,则将卷积层拆分为 $M\times N/16$ 次 16 维卷积循环,1 次 16 维卷积循环操作的 1 个卷积操作仅消耗 1 个时钟周期,和串行卷积相比,1 个 3×3 的卷积运算节省了 8 个时钟周期,速度提升了 9 倍,1 个 3×3 的卷积层速度提升 144 倍。残差模块的输入通道数有 128 和 256 两种,若采用和第 1 层卷积层相同的以空间换时间的速度优化方法,即将通道数为 128/256 的特征图同时进行卷积运算,在资源有限的 FPGA 平台上是不适用的,仅适用于通道数较少的情况,对于通道数较多时,采用 16 维循环卷积的方法更加合适。

卷积核大小为 1×1 卷积的卷积层不需要利用滑窗模块构造矩阵窗口,串行输入的数据与权重直接相乘即可。1×1 卷积的卷积层输入通道数为 128 和 256,全部采用并行卷积的方法能够大幅度提升卷积的速度,但 FPGA 资源占用过大,为使资源的合理化利用,1×1 卷积的卷积层采用与 3×3 卷积的卷积层相同的方式,只需将 3×3 的卷积换成 1×1 的卷积即可。

3) 池化层模块设计

GBcNet 网络模型中的池化层采用了两种池化方案,最大池化和自适应平均池化。由于池化层的输入通道数较大,若每一个通道都采用并行的方式进行运算,FPGA 的片内资源消耗较大。采用串并复用的方式实现池化操作,将单个特征图内的池化操作进行并行运算,对于所有通道的运算采用串行循环的方式,这样既能提升计算速度,又节约 FPGA 的资源。

最大池化模块主要由 4 个求最大值的模块构成,最大值模块计算流程如图 12 所示。该模块主要实现对输入的 3 个数据求最大值,需消耗 2 个时钟周期。最大池化模块的输入为滑窗模块形成的 3×3 的矩阵,将其每一行同时输入到 3 个并行的求最大值模块,得到矩阵每一行的最大值,消耗 2 个时钟周期。然后将求出的每一行的最大值输入到最后一个求最大值模块,最终输出即为所求 3×3 矩阵的最大值,消耗 2 个时钟周期。整个 3×3 矩阵求最大值共消耗 4 个时钟周期,若使用串行的方式求矩阵最大值时,需消耗 8 个时钟周期,可见速度提升了 2 倍。

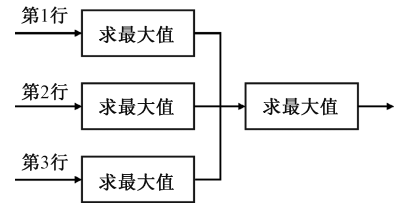

图 12 最大池化模块框图

Fig. 12 Block diagram of maximum pooling module

网络中的自适应平均池化为 14×14 的平均池化,由于输入的特征图大小为 14×14,整张特征图的平均池化只需使用平均池化模块计算一次。平均池化模块设计采用和 3×3 平均池化一样的方案,共由 15 个求平均值的模块构成,如图 13 所示。求平均值模块对输入的 14 个数据进行平均,消耗 13 个时钟周期。平均池化模块将 14 行的像素数据分别并行输入给 14 个求平均值模块,计算出每一行的平均值,消耗 13 个时钟周期,接着将 14 个求平均值模块输出的平均值输入到下一个求平均值模块,最终输出一整张特征图的平均值,为自适应平均池化的结果。最后一次求平均值消耗了 13 个时钟周期,共消耗了 26 个时钟周期,相较于串行方式求平均池化,速度提升了 7.5 倍。

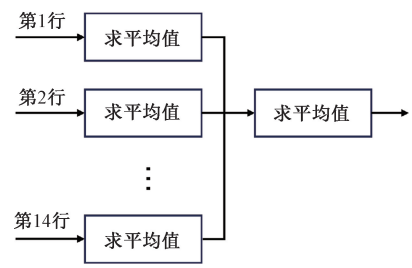

图 13 平均池化模块框图

Fig. 13 Average pooling module block diagram

4）BN 层模块设计

BN 层的运算原理与卷积核大小为 1×1 的卷积运算类似，不需要使用滑窗模块构造矩阵。输入特征图大小为 4×4 的 BN 层运算流程如图 14 所示。

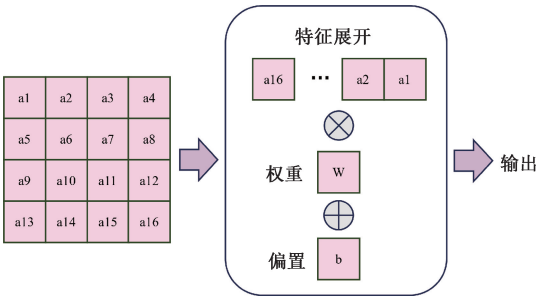

图 14　BN 层示意图

Fig. 14　BN layer diagram

特征图为串行输入，图像素点与预先缓存的权重相乘，再与偏置相加，消耗 1 个时钟周期。为减少特征图向 DDR3 的写入次数，将卷积层的输出直接与 BN 层的输入相连，卷积层输出通道输出的串行特征图数据直接进入 BN 层进行运算。

5）ReLU 激活函数模块设计

ReLU 激活函数名为线性整流函数，如式(1)所示。

$$f(x)=\begin{cases}x, x \leqslant 0\\ 0, x > 0\end{cases} \tag{1}$$

当特征图中的像素点大于等于 0 时，输出与输入相等，当像素点小于 0 时，输出为 0。ReLU 激活函数输出示例如图 15 所示。ReLU 激活函数模块与 BN 层的输出直接相连，计算完成后将输出缓存到 DDR3 中。

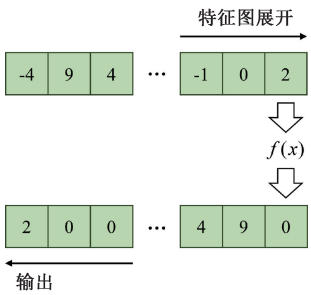

图 15　ReLU 激活函数输出

Fig. 15　ReLU activation function output

6）全连接层模块设计

全连接层的输出为地基云图的识别结果，采用 11 个并行的求概率模块计算所有类别的概率，可将速度提升 11 倍，框图如图 16 所示。将自适应平均池化层的输出结果直接与全连接层模块相连，将输出同时输入到 11 个求概率模块，计算出 11 种地基云的识别概率，然后将 11 个概率值输入到求最大概率模块，求出最大概率值及其对应的类别，最后输出识别结果。

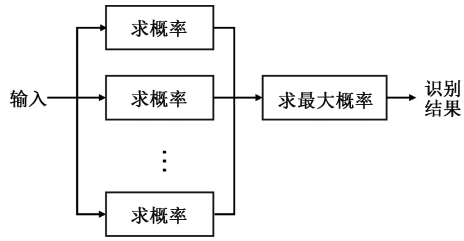

图 16　全连接层模块设计框图

Fig. 16　Fully connected layer module design block

求概率模块如图 17 所示。由于求概率模块的输入是按照通道数排列的特征图数据，不需要使用滑窗模块构造矩阵窗口。将输入的 1×1 的特征图依次与对应的权重相乘，然后将 256 个通道特征图的计算结果相加，最后与偏置相加，输出即为求得的概率值。

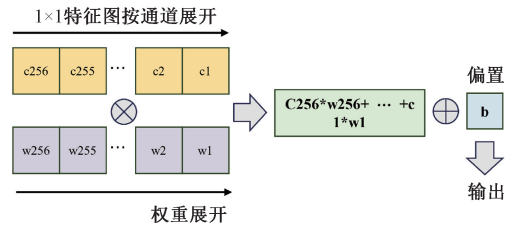

图 17　求概率模块框图

Fig. 17　The probability module block diagram

3　实验与分析

3.1　实验环境

PC 端网络模型框架基于 PyTorch 1.13.0，GPU 型号为 NVIDIA TELSA P100 16 G，硬件环境为 CUDA 11.4，Python 版本为 3.7.12。初始学习率设置 0.001，批处理大小设置为 128，采用 Adam 优化器进行优化。

FPGA 硬件平台型号为 ZYNQ 7020，是一款异构 SoC 芯片，集成了一颗双核 ARM Cortex-A9 处理器和一块 FPGA，充分集成了 ARM 和 FPGA 的优势。

3.2　GBcNet 模型性能分析

经过 300 轮训练迭代后，GBcNet 模型对每一类云的分类精确率结果如表 2 所示。各云类别的识别精度都在 96%左右，模型对所有类别都有较高的识别率。

将提出的 GBcNet 模型与 AlexNet、VGG16[22]、ResNet34、CloudNet 以及 Li 等[9]提出的网络模型进行对比实验。采用相同的图像预处理技术和训练方式，选择宏观准确率指标对各模型分类效果进行比较，实验结果如图 18 所示。与现有的研究中的 CloudNet 模型和 Li 等[9]提出的网络模型相比，地基云图的识别准确率分别提高了 7.62%和 3.52%，证明 GBcNet 模型在地基云图识别方面具有更好的泛化性。

表 2 各类别云图分类精确率

Table 2 Classification accuracy of cloud images for each category

类别	符号	精确率
卷云	Ci	0.955
卷层云	Cs	0.951
卷积云	Cc	0.955
高积云	Ac	0.963
高层云	As	0.957
积云	Cu	0.946
积雨云	Cb	0.951
雨层云	Ns	0.971
层积云	Sc	0.965
层云	St	0.972
尾迹云	Ct	1

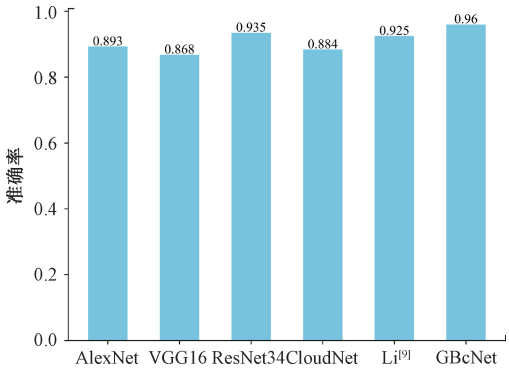

图 18 各模型准确率对比

Fig. 18 Comparison of accuracy of each model

3.3 基于 ZYNQ 的模型性能测试与分析

利用 ZYNQ 系列 FPGA 的 PL 端完成 GBcNet 模型的部署,充分利用 FPGA 强大的并行运算能力实现对模型的加速,其中,采用串并结合的方式使资源占用和速度达到平衡,搭建系统框图如图 19 所示。

利用和 PC 端相同的测试集完成模型的验证,将 PC 端预训练好的最优模型权重参数和测试集存储到 SD 卡,FPGA 首先将一张地基云图和所有权重参数加载到 DDR3 中,然后 PL 端的 DDR3 读写控制 IP 核和 GBcNet IP 核开始工作,最后模型识别结果通过 PS 端的串口打印,权重参数只需向 DDR3 中加载一次,每一轮测试只需重新向 DDR3 中加载云图数据即可。

1) FPGA 与 PC 端模型性能对比分析

将 PC 端模型测试用到测试集存储到 SD 卡中,对 FPGA 端模型进行测试。通过对 PS 端串口打印测试结果进行整理,得到各类别的识别准确率,并与 PC 端模型的测试结果对比,如表 3 所示。FPGA 端模型的大部分类别的召回率在 94% 左右,Ct 类别的召回率达到了 99%,与 PC 端模型的召回率相比,误差维持在了 3% 以内,其中 Ci 类别的召回率有所提升,Ct 类别保持不变。FPGA 端模型整体的准确率为 94.5%,与 PC 端相比,准确率下降了 1.52%。FPGA 端模型的整体识别性能和不同类别的识别性能均表现优异,其产生误差的主要原因为权重参数由 32 bits 浮点数转化为 16 bits 的定点数会造成的。

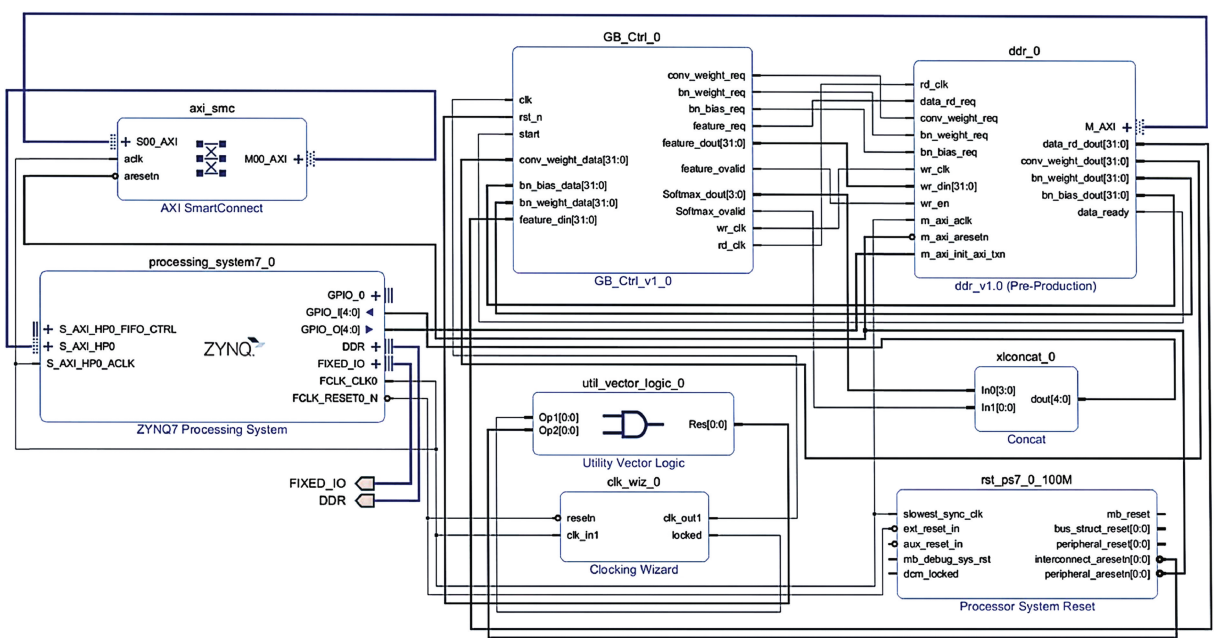

图 19 硬件系统连接框图

Fig. 19 Hardware system connection block diagram

表 3　模型测试结果对比

Table 3　Comparison of model test results　　(%)

类别	FPGA 实现		PC 端实现		误差	
	准确率	召回率	准确率	召回率	准确率	召回率
Ci		94.2		93		1.2
Cs		93.1		94.5		−1.4
Cc		94.8		96		−1.2
Ac		93.6		94.6		−1
As		93.6		95		−1.4
Cu	94.5	94.5	96.02	97	−1.52	−2.5
Cb		94.2		96.7		−2.5
Ns		94.2		96.4		−2.2
Sc		94.7		97.6		−2.9
St		94.1		97		−2.9
Ct		99		99		0

2）各模块计算速度测试分析

模型工作在 100 MHz 时钟频率下，利用 Modelsim 软件对模型进行仿真，得到模型各模块的计算速度，如表 4 所示。7×7 卷积层模块采用 3 个 7×7 卷积模块并行运算，消耗时间为 31.97 ms。残差模块中所有的卷积运算模块均采用 16 个并行的 3×3 卷积或者 1×1 卷积模块并行运算，残差模块中的各子模块消耗时间分别为 24.1、16.06、26.15、16.1 和 16.1 ms，全连接层模块消耗 0.028 ms。大量的并行运算显著提升了模型的推理速度，经测试，GBcNet 模型在 FPGA 上完成一张地基云图的识别任务仅需要 0.13 s。

表 4　各模块的计算速度

Table 4　Calculation speed of each module

模块名	时间/ms
7×7 卷积层模块	31.97
残差块 Res_A	24.1
残差块 Res_B	16.06
残差块 Res_C	26.15
残差块 Res_D	16.1
残差块 Res_E	16.1
全连接层模块	0.028

3）FPGA 资源消耗分析

在实现模型加速的同时，充分考虑了资源消耗，模型在 FPGA 上的资源消耗情况如表 5 所示。LUT 资源消耗 35.08%，FF 资源消耗 11.59%，BRAM 资源消耗 13.21%，DSP 资源消耗 47.27%，BUFG 资源消耗 9.38%。其中 DSP 资源占用最高，是由于模型中需要用到大量的乘法运算导致的。模型的识别速度和 FPGA 资源消耗达到了较好的平衡。

表 5　FPGA 资源占用情况

Table 5　FPGA resource usage

资源	使用量	可用量	消耗率/%
LUT	18 664	53 200	35.08
FF	12 333	106 400	11.59
BRAM	18.50	140	13.21
DSP	103	125	47.27
BUFG	3	32	9.38

4　结　论

提出的基于残差网络的地基云图识别模型 GBcNet 和基于 ZYNQ 硬件平台的网络模型加速方案，解决了目前地基云图模型识别准确率较低、网络型复杂而难以在嵌入式系统中部署应用的问题。采用串并联合运算的方式设计了网络模型中每个模块对应的 IP 核。模型部署到 ZYNQ 系统后，GBcNet 模型对各类云图的识别性能依然达到 94.5%，单张云图的前向推理时间仅仅需要 0.13 s。且在 FPGA 中，LUT、FF、BRAM、DSP 等资源占用率均不超过 50%。

未来可在模型的轻量化上做进一步改进，采用模型剪枝、参数共享等技术，在保证网络性能的前提下压缩网络参数，提升推理速度。

参考文献

[1] TANG Y ZH, YANG P L, ZHOU Z M, et al. Improving cloud type classification of ground-based images using region covariance descriptors [J]. Atmospheric Measurement Techniques, 2021, 14(1): 737-747.

[2] DUDA D P, MINNIS P, KHLOPENKOV K, et al. Estimation of 2006 northern hemisphere contrail coverage using MODIS data [J]. Geophysical Research Letters, 2013, 40(3): 612-617.

[3] MANANDHAR P, TEMIMI M, AUNG Z. Short-term solar radiation forecast using total sky imager via transfer learning[J]. Energy Reports, 2023, 9: 819-828.

[4] HASENBALG M, KUHN P, WILBERT S, et al. Benchmarking of six cloud segmentation algorithms for ground-based all-sky imagers[J]. Solar Energy, 2020, 201: 596-614.

[5] GUO B, ZHANG F, LI W W, et al. Cloud classification by machine learning for geostationary radiation imager [J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-4.

[6] JEPPESEN J H, JACOBSEN R H, INCEOGLU F, et al. A cloud detection algorithm for satellite imagery based on

deep learning [J]. Remote Sensing of Environment, 2019, 229: 247-259.

[7] ZHANG J, LIU P, ZHANG F, et al. CloudNet: Ground-Based cloud classification with deep convolutional neural network [J]. Geophysical Research letters, 2018, 45(16): 8665-8672.

[8] ZHANG L, JIA K, LIU P, et al. Cloud recognition based on lightweight neural network [C]. 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2020: 1033-1042.

[9] LI X T, QIU B, CAO G L, et al. A novel method for ground-based cloud image classification using transformer [J]. Remote Sensing, 2022, 14(16): 3978.

[10] LI SH, WANG M, SUN SH, et al. CloudDenseNet: Lightweight ground-based cloud classification method for large-scale datasets based on reconstructed DenseNet[J]. Sensors, 2023, 23(18): 7957.

[11] 欧阳一鸣, 王奇, 汤飞扬, 等. MRNDA:一种基于资源受限片上网络的深度神经网络加速器组播机制研究[J]. 电子学报, 2024, 52(3): 872-884.
OUYANG Y M, WANG Q, TANG F Y, et al. MRNDA: A multicast mechanism for resource-constrained noc-based deep neural network accelerators [J]. Acta Electronica Sinica, 2024, 52(3): 872-884.

[12] CHEN K, EBRAHIMI M, WANG T, et al. NoC-based DNN accelerator: A future design paradigm [C]. Proceedings of the 13th IEEE/ACM International Symposium on Networks-On-Chip, 2019.

[13] 张萌, 张经纬, 李国庆, 等. 面向深度神经网络加速芯片的高效硬件优化策略[J]. 电子与信息学报, 2021, 43(6): 1510-1517.
ZHANG M, ZHANG J W, LI G Q, et al. Efficient hardware optimization strategies for deep neural networks acceleration chip [J]. Journal of Electronics & Information Technology, 2021, 43(6): 1510-1517.

[14] ZHANG X F, WANG J S, ZHU CH, et al. DNNBuilder: An automated tool for building high-performance DNN hardware accelerators for FPGAs[C]. 2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), 2018.

[15] 张敬林, 张国宇, 杨全, 等. 飞机尾迹云识别及其辐射强迫的研究进展[J]. 大气科学学报, 2018, 41(5): 577-584.
ZHANG J L, ZHANG G Y, YANG Q, et al. Review of recognition of aircraft contrails and their radiative forcing[J]. Transactions of Atmospheric Sciences, 2018, 41(5): 577-584.

[16] HE K M, ZHANG X, REN SH Q, et al. Deep residual learning for image recognition [C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.

[17] 慕晓冬, 魏轩, 曾昭菊. 基于注意力残差网络的航天器测控系统故障诊断[J]. 仪器仪表学报, 2022, 43(9): 81-87.
MU X D, WEI X, CENG ZH J. Fault diagnosis method of spacecraft tracking telemetry and control system based on the attention residual network[J]. Chinese Journal of Scientific Instrument, 2022, 43(9): 81-87.

[18] 宋文强, 徐伟, 冯琳. 基于残差网络的地基云图识别方法研究[J]. 电子测量技术, 2024, 47(2): 185-192.
SONG W Q, XU W, FENG L. Research on ground-based cloud recognition method based on residual network [J]. Electronic Measurement Technology, 2024, 47(2): 185-192.

[19] 李宁, 肖昊. 基于 FPGA 的稀疏卷积神经网络加速器设计[J]. 电子测量技术, 2024, 47(5): 1-8.
LI N, XIAO H. Accelerator design of sparse convolutional neural network based on FPGA [J]. Electronic Measurement Technology, 2024, 47(5): 1-8.

[20] 戴伟杰, 王衍学, 李昕鸣, 等. 面向 FPGA 部署的改进 YOLO 铝片表面缺陷检测系统[J]. 电子测量与仪器学报, 2023, 37(9): 160-167.
DAI W J, WANG Y X, LI X M, et al. YOLO aluminum profile surface defect detection system for FPGA deployment[J]. Journal of Electronic Measurement and Instrumentation, 2023, 37(9): 160-167.

[21] 彭宇, 姬森展, 于希明, 等. 语义分割网络的 FPGA 加速计算方法综述[J]. 仪器仪表学报, 2021, 42(9): 1-12.
PENG Y, JI S ZH, YU X M, et al. A review of FPGA-accelerated computing methods for semantic segmentation network[J]. Chinese Journal of Scientific Instrument, 2021, 42(9): 1-12.

[22] SIMONYAN K. Very deep convolutional networks for large-scale image recognition[J]. ArXiv preprint arXiv: 1409.1556, 2014.

作者简介

冯琳,2020 年于苏州大学应用技术学院获得学士学位,现为南京信息工程大学硕士研究生,主要研究方向为图像分类与深度学习。

E-mail:fll13913279903@163.com

Feng Lin received her B. Sc. degree from Applied Technology College of Soochow University in 2020. Now she is a M. Sc. candidate at Nanjing University of Information Science and Technology. Her main research interests include image classification and deep learning.

宋文强,2021 年于黄淮学院获得学士学位,2024 年于南京信息工程大学获得硕士学位,主要研究方向为信号处理及 FPGA 应用。

E-mail:songwq0105@163.com

Song Wenqiang received his B. Sc. degree from Huanghuai University in 2021 and M. Sc. degree from Nanjing University of Information Science and Technology in 2024. His main research interests include signal processing and FPGA applications.

徐伟(通信作者),2004 年、2007 年、2019 年于南京信息工程大学获得学士、硕士、博士学位,现为南京信息工程大学教授,主要研究方向为气象观测方法及仪器。

E-mail:xw@nuist.edu.cn

Xu Wei (Corresponding author) received his B. Sc. degree, M. Sc. degree and Ph. D. degree in Nanjing University of Information Science and Technology in 2004, 2007, 2019, respectively. Now he is a professor. His main research interests include meteorological observation methods and instruments.