ISSN: 3065-1220 Volume 1, Number 2, October 2024

Hans 汉斯

嵌入式技术与智能系统

Embedded Technology and Intelligent Systems

Qian Ru Shi Ji Shu Yu Zhi Neng Xi Tong

2024年10月1卷2期

https://www.hanspub.org/journal/etis

Editorial Board
编委名单

ISSN: 3065-1220
https://www.hanspub.org/journal/etis

Embedded Technology and Intelligent Systems, Vol.1, No.2, 51-91
嵌入式技术与智能系统, 1 卷 2 期
Published Online October 2024 in Hans (https://www.hanspub.org/journal/etis)

Hans 汉斯

TABLE OF CONTENTS
目　录

期刊信息
期刊中文名称：《嵌入式技术与智能系统》
期刊英文名称：**Embedded Technology and Intelligent Systems**
期刊缩写：**ETIS**
出刊周期：双月刊
语　　种：中文

出版机构：汉斯出版社(Hans Publishers, https://www.hanspub.org/)
编辑单位：《嵌入式技术与智能系统》编辑部
主　　编：何立民，北京航空航天大学教授
网　　址：https://www.hanspub.org/journal/etis

订阅信息
订阅邮箱：sub@hanspub.org
订阅价格：180 美元每年

广告服务
联系邮箱：adv@hanspub.org

Embedded Technology and Intelligent Systems 嵌入式技术与智能系统, 2024, 1(2), 51-56
Published Online October 2024 in Hans. https://www.hanspub.org/journal/etis
https://doi.org/10.12677/etis.2024.12006

Hans汉斯

OpenHarmony上利用Paho MQTT连接云平台

安皓楠

北京华清远见科技发展有限公司研发中心，北京

收稿日期：2024年6月28日；录用日期：2024年11月21日；发布日期：2024年11月29日

摘　要

在物联网设备与云端之间的通信中，MQTT作为一种轻量级的、基于发布-订阅模式的通信协议，具备了良好的适用性和灵活性，被广泛应用于物联网领域。在OpenHarmony的LiteOS内核上利用MQTT连接云平台是一项关键的技术任务，它涉及在轻量级操作系统上实现MQTT协议的客户端功能，并与云端平台进行稳定和高效的通信，因此需要选择合适的MQTT库，并进行有效的移植和优化，以保证在资源受限的环境下依然能够实现稳定可靠的通信连接。海思Hi3861芯片采用了LiteOS内核。文章探讨了在海思Hi3861芯片上移植和使用Paho MQTT库连接到华为云的实现过程和关键技术。文章首先介绍了MQTT的相关知识，然后详细讨论了嵌入式Paho MQTT库的内容，接着介绍Hi3861芯片相关功能及其移植Paho MQTT的方式，最后描述了使用移植好的程序连接华为云MQTT的步骤，包括设备鉴权方式和消息发布订阅的实现。实验结果验证了在Hi3861平台上使用Paho MQTT库连接到华为云的可行性和效果。文章的结尾探讨了项目未来的工作。

关键词

MQTT，云，物联网，Paho MQTT，鸿蒙

Using Paho MQTT to Connect to Cloud Platforms on OpenHarmony

Haonan An

Research and Development Center, Beijing Huaqingyuanjian Technology Development Co., Ltd., Beijing

Received: Jun. 28th, 2024; accepted: Nov. 21st, 2024; published: Nov. 29th, 2024

Abstract

In the communication between IoT devices and the cloud, MQTT, as a lightweight, publish-subscribe communication protocol, offers excellent applicability and flexibility, making it widely used in the IoT field. Utilizing MQTT to connect to cloud platforms on OpenHarmony's LiteOS kernel is a critical

文章引用: 安皓楠. OpenHarmony 上利用 Paho MQTT 连接云平台[J]. 嵌入式技术与智能系统, 2024, 1(2): 51-56.
DOI: 10.12677/etis.2024.12006

technical task. It involves implementing MQTT client functionality on a lightweight operating system and establishing stable and efficient communication with the cloud platform. Therefore, choosing an appropriate MQTT library and conducting effective porting and optimization are crucial to ensure reliable communication connections in resource-constrained environments. The Hisilicon Hi3861 chip utilizes the LiteOS kernel. This article discusses the process and key technologies of porting and using the Paho MQTT library to connect to Huawei Cloud on the Hisilicon Hi3861 chip. The article begins with an introduction to MQTT concepts, followed by a detailed discussion of the embedded Paho MQTT library. It then covers the features of the Hi3861 chip and how to port Paho MQTT to it. Finally, it describes the steps to connect to Huawei Cloud MQTT using the ported application, including device authentication methods and the implementation of message publishing and subscription. Experimental results validate the feasibility and effectiveness of using the Paho MQTT library to connect to Huawei Cloud on the Hi3861 platform. The article concludes with a discussion on future project directions.

Keywords

MQTT, Cloud, Internet of Things, Paho MQTT, OpenHarmony

1. 引言

近年来，物联网(Internet of Things, IoT)技术的快速发展改变了各行各业，让设备之间实现了无缝连接和智能互动。从传感器到智能家电，物联网设备需要与云服务进行高效通信，以传输数据、接收指令并实现远程管理。这一转变促使了专门针对物联网部署需求开发的专业物联网平台和通信协议，如 CoAP (Constrained Application Protocol)、AMQP (Advanced Message Queuing Protocol)、DDS (Data Distribution Service)、WebSocket 等。除上述协议外，MQTT (Message Queuing Telemetry Transport)因其适应资源受限设备、简单易用的设计、强大的消息传输能力和广泛的应用支持而成为物联网通信的常用选择。它不仅满足了各种物联网设备和应用程序之间实时通信的需求，还提供了灵活的部署和扩展选项，适应了不同规模和复杂度的物联网解决方案[1]。

Paho MQTT 是一个开源的 MQTT 库，由 Eclipse Paho 项目提供支持。它提供了多种编程语言的实现，包括 C、C++、Java、Python 等，旨在帮助开发者轻松地在各种设备和平台上实现 MQTT 通信协议。本文专注于将 Paho MQTT 客户端库移植到 OpenHarmony 项目中的 LiteOS 内核处理器上，如海思 Hi3861 芯片，实现与云平台的连接。在 OpenHarmony 项目的 LiteOS 内核处理器上集成 MQTT 不仅增强了设备与云端高效通信的能力，还为创建互联物联网生态系统的整体目标作出了贡献。

2. MQTT 协议

MQTT 协议是一种轻量级的、基于发布/订阅模式的消息传输协议，最早由 IBM 开发于 1999 年，用于传感器和施工机器之间的通信。随后，协议被开放，成为 OASIS (Organization for the Advancement of Structured Information Standards，结构化信息标准推动组织)标准，并在众多物联网应用中得到广泛应用。MQTT 协议在物联网领域起初设计用于传感器网络和设备间的低带宽、高效率通信。其轻量级和简单的发布/订阅模式使得设备能够节省能源和带宽，适用于智能家居、工业自动化和智能城市等多种场景。

2.1. MQTT 的架构

MQTT 的客户端是消息发布者或者消息订阅者，它们与 MQTT Broker 进行通信来实现数据的传输和接收[2]。客户端可以是各种设备或应用程序，包括传感器、嵌入式设备、移动应用等。MQTT 的客户端和 Broker 之间的关系如图 1 所示。

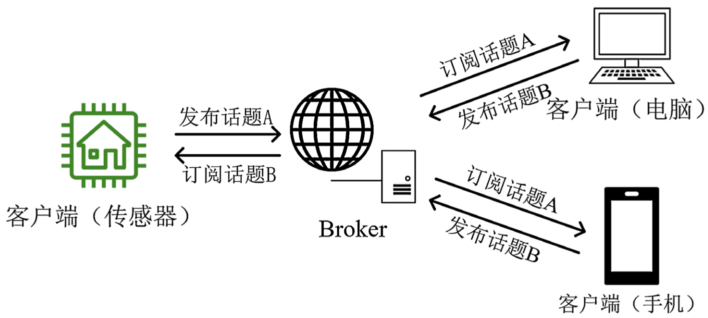


Figure 1. MQTT network architecture
图 1. MQTT 网络架构

客户端在 MQTT 中的主要功能包括：

发布消息：客户端可以向 Broker 发布(或发送)消息，消息可以是任何格式的数据，通常包含有用的传感器数据、控制命令等。

订阅主题：客户端可以向 Broker 订阅(或接收)特定的主题(Topic)，通过主题来过滤和接收感兴趣的消息。主题可以使用通配符来实现更灵活的订阅规则。

接收消息：客户端通过订阅主题，可以接收来自 Broker 转发的消息，并进行相应的处理和分析。

客户端可以选择不同的消息服务质量(QoS)级别来控制消息传递的可靠性和效率，包括至多传递一次(QoS 0)、至少传递一次(QoS 1)、恰好传递一次(QoS 2)。

MQTT 的 Broker 是中介服务器，负责管理客户端之间的消息传递。Broker 接收来自发布者(发布消息的客户端)的消息，并确保将这些消息按照订阅者(订阅消息的客户端)的需求正确分发。Broker 在 MQTT 协议中的关键作用包括：

消息路由和分发：根据订阅关系，将发布者发送的消息分发给所有订阅了相关主题的客户端。

连接管理：管理客户端的连接和状态，确保每个连接的可靠性和安全性。

QoS 管理：根据客户端设置的 QoS 级别，确保消息的按时传递和确认。

2.2. MQTT 的报文

为了让客户端和 Broker 之间进行通信，MQTT 协议定义了不同类型的消息(称为报文)，如 CONNECT、CONNACK、PUBLISH、PUBACK、SUBSCRIBE、SUBACK、UNSUBSCRIBE 等。报文包含固定报头、可变报头(部分报文包含)、负载(部分报文包含)。固定报头定义了报文类型、控制标志和后续数据长度，可变头部则依据报文类型规定了各自的详细信息，如后续消息内容、QoS 级别等。负载用于携带实际数据内容。这种设计使得 MQTT 能够在各种网络条件下高效传输消息，并支持灵活的通信需求和服务质量保证。

3. Paho MQTT 库

3.1. 简介

Paho 项目致力于提供开源的、可扩展的消息传递协议实现，支持 M2M (Machine-to-Machine)和物联

网的各种应用。它特别关注设备连接中的物理限制和成本问题。Paho MQTT C 是 Paho 项目的一部分，是专门为 C 语言开发的 MQTT 客户端库。Paho MQTT C 库旨在提供一个可靠、高效的实现，使开发者能够轻松地在 C 语言环境中实现 MQTT 的发布和订阅功能[3]。

3.2. 库文件内容

应用于嵌入式的 Paho MQTT C 库文件中，能够被移植的代码位于 MQTTPacket、MQTTClient-C、MQTTClient 文件夹。

MQTTPacket 文件夹包含了 MQTT 协议报文的解析和封装功能。这些文件提供了处理 MQTT 协议中不同报文(如 CONNECT、PUBLISH、SUBSCRIBE 等)的代码实现。它们负责将 MQTT 消息编码为字节流(序列化)，或者将接收到的字节流解析为可操作的消息(反序列化)。

MQTTClient-C 文件夹包含了 MQTT 客户端的 C 语言实现的核心部分。这些文件实现了 MQTT 客户端的连接、发布、订阅、断开连接等功能。它们是构建在 MQTTPacket 基础之上的更高级别的抽象，提供了一个易于使用和集成的 MQTT 客户端接口。在移植过程中，我们需要对这个文件夹中的一些网络接口进行修改，才能利用其中的函数正常地与服务器通信。

MQTTClient 是一个最初为 mbed 编写的 C++库，现已移植到其他平台上使用。该库基于并且需要 MQTTPacket。对于仅使用 C 语言的系统来说，可以忽略此文件夹。

表 1 说明了 MQTTPacket 文件夹中与客户端有关的文件的作用。

Table 1. The role of files in the MQTTPacket folder
表 1. MQTTPacket 文件夹中的文件作用

文件	描述
MQTTConnectClient.c	封装 MQTT 客户端连接相关的报文，如连接建立、参数设置等。
MQTTSerializePublish.c	将要发布的消息封装为 MQTT 报文格式，以便网络传输。
MQTTDeserializePublish.c	解析 MQTT 发布的消息，将报文数据解析为应用消息。
MQTTSubscribeClient.c	处理订阅请求和响应的报文。
MQTTPacket.c	包含了对 MQTT 报文的封装和解析的详细实现。
MQTTFormat.c	将 MQTT 报文的二进制字节流转换为易于阅读和分析的字符串形式。

由于这些文件只负责处理报文格式，不涉及与单片机有关的通信接口，因此在移植过程中直接复制到工程中即可，不需要修改。

MQTTClient-C 文件夹中使用的主要是 MQTTClient.c 文件。该文件是 MQTT 客户端库的核心实现，负责实现 MQTT 协议的各种功能，提供了连接 Broker、发布和订阅消息等高级抽象接口。它通过调用 MQTTPacket 文件夹中的程序来封装报文，并通过调用相关的网络接口将这些报文发送出去。在该文件中，初始化函数 MQTTClientInit()需要调用 Network 结构体来传入有关 Socket 网络接口。该结构体涉及到 Socket 接口编号、接收函数、发送函数。因此需要编写程序来初始化 Socket 接口，并编写接收和发送函数，以供 MQTTClient.c 文件中的函数调用。还需要编写供超时判断用的计时函数。

3.3. 移植说明

官方的 MQTTClient-C 文件夹中已经打包好了一些例程，它们适配于 FreeRTOS、Linux、CC3200。移植时需要依据这些例程编写一个程序文件(此处名称为 Hi3861_PahoMQTT.c)，为 MQTTClient.c 文件中

的函数配置好网络接口和计时函数。网络接口利用 Socket 通过 TCP 的方式连接到服务器并进行通信，计时函数用于通信超时判断。要编写的文件中的程序内容如表 2 所示。

图 2 通过发布消息的流程来说明在 Hi3861_PahoMQTT.c 文件中增加的程序的作用。

Table 2. Programs that need to be written during migration
表 2. 移植时需要编写的程序

函数	作用
NetworkRead()	利用 Socket 从网络中读取数据。
NetworkWrite()	利用 Socket 向网络中写入数据。
NetworkInit()	初始化网络结构体，包括 Socket 编号和读写数据的函数接口。
NetworkConnect()	利用 IP 地址和端口号连接到指定主机。
NetworkDisconnect()	断开网络连接。
其它计时函数	用于判断通信是否超时

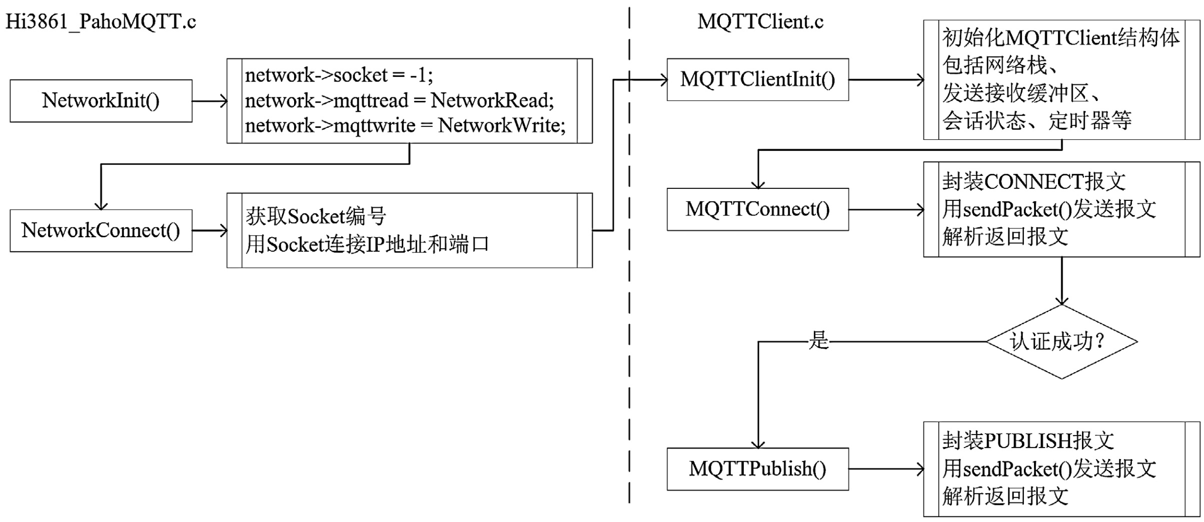


Figure 2. The process of publishing messages using the paho.mqtt.c library
图 2. 利用 paho.mqtt.c 库发布消息的流程

从图 2 中可以看到，为了移植嵌入式的 Paho MQTT C 库而新增的几个函数的主要作用是初始化 Socket 接口、设置利用 Socket 进行接收和发送数据的函数，然后利用 Socket 连接服务器。配置完成后，设备后续与 MQTT Broker 进行连接认证、发布消息都依靠的是 MQTTClient.c 文件中的相关函数。这些函数通过调用 MQTTPacket 文件夹中的函数来进行报文的封装或解析，利用与 Socket 相关的函数进行数据的接收和发送。

4. 在 Hi3861 上移植 Paho MQTT

Hi3861 是海思推出的一款支持 WiFi 功能的处理器，其内核采用 OpenHarmony 架构下的 LiteOS 操作系统[4]。在官方提供的工程文件中，已经实现了在 LiteOS 系统下的 WiFi 连接接口和用于实现 TCP/IP 协议栈的 LwIP (Light weight IP)库。通过调用这些接口，Hi3861 可以连接到 WiFi 热点上面，获取到 IP 地址，并创建 Socket 套接字[5]。移植前需要将 Paho MQTT C 库相关文件复制到工程中，并设置好 BUILD.gn 项目构建文件。

在 LiteOS 操作系统中，setsockopt()函数用于配置套接字的选项，如协议级别、接收发送的超时时间、保活机制等，recv()用于从 Socket 接收数据，send()用于向 Socket 发送数据。因此，在第 3.3 节提到的 NetworkRead()函数中需要先通过 setsockopt()设置接收超时时间，然后调用 recv()函数进行数据的接收；在 NetworkWrite()函数中需要先通过 setsockopt()设置发送超时时间，然后调用 send()函数进行数据的发送。LiteOS 中的 socket()函数用来创建一个新的套接字(Socket)，connect()函数用于在客户端套接字上建立与远程服务器的连接。因此这两个函数需要在 NetworkConnect()函数中对这两个函数进行调用，从而连接到云平台的服务器上。

另外还需要移植用于超时判断的定时器相关函数，方法主要是通过内核 Tick 计数获取函数 osKernelGetTickCount()和系统定时器计数获取函数 osKernelGetSysTimerCount()，配合计数频率来计算秒数和微秒数，这涉及到对 MQTTClient.c 中的 TimerIsExpired()、TimerCountdownMS()、TimerCountdown()、TimerLeftMS()这几个函数的重定义。

配置好 MQTTClient.c 中的函数接口后，即可调用 MQTTClient.c 中的相关函数进行 MQTT 的认证、订阅、取消订阅、发布、接收相关操作了。

5. 实验部署

本文采用华清远见研发的 FS-Hi3861 开发板进行 Hi3861 工程的部署，使用华为云平台的设备接入 IoTDA 资源作为 MQTT 服务器。

首先需要在华为云的设备接入 IoTDA 资源中创建产品及相关属性、命令。属性为设备上报的数据，命令为向设备下发的数据，这两种数据的话题和格式是不一样的。然后注册一个设备，获取到设备的鉴权信息，包括客户端 ID、用户名、密码。

设备需要利用 MQTTClient.c 中的 MQTTConnect()函数包装好鉴权信息，与服务器进行认证。认证通过后，在华为云中会显示设备处于在线状态。然后，设备可以利用 MQTTPublish()函数按照华为云规定的格式向指定的话题发布属性数据。设备还可以利用 MQTTSubscribe()函数订阅指定的话题，并通过 MQTTRun()函数接收云平台下发的数据。

6. 未来工作

本文以 Hi3861 为例，介绍了在 OpenHarmony 项目的 LiteOS 内核上面通过移植 Paho MQTT 连接 MQTT 云平台的方法。由于云平台通常使用 JSON 格式进行数据传递，因此在未来的项目中，将通过移植 cJSON 库的方式进行数据的序列化和反序列化，从而更加方便地传输传感器数据并解析相关指令。另外，为了便于传感器网络的建立，未来的项目将探讨利用 Hi3861 进行 Mesh 组网，从而扩展传感范围的方式。为了适应低功耗需求，未来的项目也将探讨低功耗 WiFi 模式下 MQTT 的应用。

参考文献

[1] Vanani, K., Patoliya, J. and Patel, H. (2016) A Survey: Embedded World around MQTT Protocol for IoT Application. *International Journal for Scientific Research and Development*, **4**, 26-29.

[2] Light, R.A. (2017) Mosquitto: Server and Client Implementation of the MQTT Protocol. *The Journal of Open Source Software*, **2**, Article 265. https://doi.org/10.21105/joss.00265

[3] Eclipse Paho (2024) Eclipse Paho. https://eclipse.dev/paho/

[4] 海思官网. Hi3861LV100 产品简介[EB/OL]. https://www.hisilicon.com/cn/products/connectivity/short-range-IoT/wifi-nearlink-ble/Hi3861V100, 2024-06-27.

[5] 何进, 谢松巍. 基于 Socket 的 TCP/IP 网络通讯模式研究[J]. 计算机应用研究, 2001(8): 134-135.

Embedded Technology and Intelligent Systems 嵌入式技术与智能系统, 2024, 1(2), 57-63
Published Online October 2024 in Hans. https://www.hanspub.org/journal/etis
https://doi.org/10.12677/etis.2024.12007

AGI时代的电子及计算机工程师

周　娜[1]，何　铮[2]，何为民[3]

[1]海口经济学院中芯依智网络学院，海南 海口
[2]海南政法职业学院信息中心，海南 海口
[3]海南久迪物联网科技有限公司，海南 海口

收稿日期：2024年7月4日；录用日期：2024年11月21日；发布日期：2024年11月29日

摘　要

文章展望了AGI时代的特点及人才金字塔结构的分布，重点分析了AGI时代电子及计算机工程师的行业发展趋势及特点，并建言当今电子及计算机工程师如何应对AGI时代的来临。

关键词

通用人工智能AGI，人才金字塔，电子及计算机工程师

Electronic and Computer Engineers in AGI Era

Na Zhou[1], Zheng He[2], Weimin He[3]

[1] College of SMIC Network, Haikou Institute of Economics, Haikou Hainan
[2]Information Center, Hainan Vocational College of Political Science and Law, Haikou Hainan
[3]Hainan Jiudi Internet of Things Technology Co., Ltd., Haikou Hainan

Received: Jul. 4th, 2024; accepted: Nov. 21st, 2024; published: Nov. 29th, 2024

Abstract

This paper looks forward to the characteristics of AGI era and the distribution of talent pyramid structure. This paper focuses on the development trend and characteristics of electronic and computer engineers in AGI era. It also suggests how electronic, and computer engineers should deal with the arrival of AGI era.

Keywords

General Artificial Intelligence AGI, Talent Pyramid, Electronic and Computer Engineer

文章引用: 周娜, 何铮, 何为民. AGI 时代的电子及计算机工程师[J]. 嵌入式技术与智能系统, 2024, 1(2): 57-63.
DOI: 10.12677/etis.2024.12007

1. 引言

自 1956 年 AI (Artificial Intelligence)人工智能诞生到 2022 年，它一直是属于电子、计算机及哲学领域前沿基础理论研究的范畴。期间，AI 只是艰难曲折地缓慢向上发展，对社会各行业的影响都微乎其微。但 2022 年迎来一个重大转折，Open AI 生成式大模型 ChatGPT 发布，成为一个以生成式大模型为核心的人工智能新时代来临的新起点。AI 也逐步发展为通用人工智能 AGI (Artificial General Intelligence)，它的标志是可以实现自我学习、改进和调整，进而不需人为干预而解决任何问题，使人 AI “工具”开始有了“生命”。

自此。AI 技术进入 AGI 一个按指数型发展新领域。至今 ChatGPT 已发展到多模态的 GPT4o 及谷歌的 Gemini 等。模型参数已从千亿级发展到万亿级；所用算力从几十 PLOPS 扩展到几十 ELOPS。大模型已经学习了 5000~8000 万种人类有史以来创造的知识。大模型不仅能文字交互，而且能语音交互聊天，能看懂、生成图片，输出一定程度上符合环境规律的视频。

2. AGI 时代的特点

2.1. AGI 可渗透应用到所有的行业中

AGI 是文本、语音、图像视频多模态交互的通用大模型，它可渗透到当今世界的所有行业。当前，大模型已在很多应用端爆发出它的巨大威力。人们都认识到大模型将改变世界；世上所有行业都需要基于人工智能技术重新做一遍，所有的人群和工作都可归纳为“制造 AGI 工具的”和“使用 AGI 工具的”。虽然工作可取代，但人类的做决策、情感、想象力和创造力是人工智能短期无法取代的。

(1) 经验性、重复性的工作将被 AGI 替代

AI 应用首先逐步替代的工作是一些大量经验型、重复流水性工作。如文案，科技翻译、形象设计、码农等。但那些少数高端人脑创意性的工作则难以替代，大量的人力劳动(如建筑行业)也无法替代。没有唯一标准答案的文学翻译，好的美学作品，系统规划师等还需要一些高端的人才。在医院，急诊、各类手术科室的医生将加强，但门诊、社区及全科大夫将大量使用“AI 医生”，互联网医疗将进化为无人的 AI 医疗。

(2) 1 + 1 > 2

在 AGI 时代，AGI 工具的能力越来越强大。一个掌握了 AGI 工具人的工作效率将远远大于两个未掌握 AGI 工具人的工作效率。AI 翻译已 90%取代了科技翻译。随着读片 AI 化，CT 及 MR 的工作效率将大步提升。一个掌握 AI 代码工具的嵌入式工程师的工作效率能顶上数个传统的嵌入式工程师。

(3) AGI 多模态生成式大模型技术的三大方向

A、本身发展研究：类人神经算法进一步的研究，加大算力，减少能耗；学完已有数据并生成数据。

B、智能体(Agent)研究：如 AI PC、自主机器人、L5 无人驾驶等应用侧系统。

C、科学研究：在各自科学研究中，根据本专业建模，建立本行业、企业、部门的大模型，成为自身科学研究的工具，成为人类科学家的好助手。

2.2. AGI 时代的人才金字塔的变化(如图 1)

(1) AI 时代人才金字塔体量变小

AI 时代，掌握了 AGI 工具人的工作效率大幅提升，1 + 1 > 2。虽然增加了一些“训练师”、“标记

师”等新行业，但老行业裁员已成新的大趋势。所裁人员向人力产业和第三产业转移。但总体上是使用 AI 工具的个人，能力越来越强大，这部分人可以养活全球人。福利社会将到来，高端人士贡献大，生活将会更好，每周工作 3 天、4 天已成可能。因为社会的稳定，低保已能满足一部分无业底层人的基本生活。

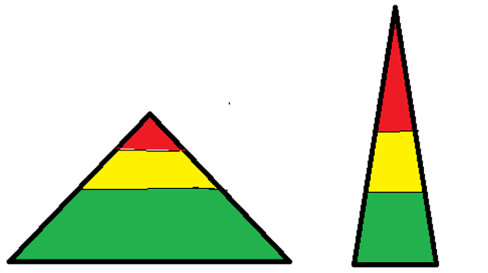


Figure 1. Talent pyramid of the AGI era
图 1. AGI 时代人材金字塔

(2) AI 时代人才金字塔更尖锐

在 AI 时代，人才高低端分离得更严重，人们可区分为“制造 AI 工具的人”和“使用 AI 工具的人”。在金字塔顶端，更需像奥特曼、黄仁勋、库克等这样的顶端大师，他们带领“制造 AI 工具的团队”，研究、推动 AGI 算法、算力的发展；生成多模态基础大模型及政府、行业、企业、个人的专属大模型。他们创造性思维带来的新工作远大于因 AI 消失的工作，它的体量会变大。在金字塔底端，“使用 AI 工具的人”因 AI 带来的高效率，消灭大量的凭经验及重复性的工作岗位。虽然它增加了训练师、标注师等工作岗位，但它总的体量会变小。

Table 1. Forecast of electronic and computer research and development team of small and micro enterprises for Internet of Things equipment
表 1. 物联网设备小微企业电子及计算机研发队伍预测

	系统规划师	软硬件开发工程师	测试工程师	结构工程师	总计
当前	3	12	4	1	20
AGI 时代	2	3	4	1	10

如表 1 所示，当前一个物联网设备小微企业，开发研究人员有 20 位(3 名系统规划师、12 名软硬件开发工程师，4 名测试工程师，1 名结构工程师)，到 AGI 时代可能只需 10 人(2 名系统规划师、3 名软硬件开发工程师，4 名测试工程师，1 名结构工程师)。其中创新思维，找到并提出解决方案的系统规划师仅减少 1 名，需手动非重复劳动的测试、结构工程师(非流水线上批量生产的)尚无法替代。而软硬件开发工程师由于上游提供大量标准电路及丰富的库函数，又有 AI 代写软件，其人数将大大压缩 3 分之一或更少。

所以 AGI 时代人才结构的金字塔会变窄、变高、变得更尖锐。

3. AGI 时代的电子及计算机工程师

作为 AGI 时代前沿的电子工程师和计算机工程师受 AGI 的影响最为直接。他们必须接受 AGI 强风暴雨的洗礼。

3.1. 硬件集成电路及微电子系统工程师

(1) 迎对摩尔定律的失效

集成电路芯片制程从 5 nm 走到 3 nm，生产成本翻了一番。摩尔定律“每 18~24 个月，集成电路上晶体管面密度会增加一倍，性能也将提升一倍(或同性能的成本减少一倍)”似乎已经走到了尾声，1 nm 几乎成为芯片大厂追求的极限。应对摩尔定律的失效，性能的提高，不仅是算力、带宽、速度，而且能耗也是越来越重要的考量因素。所以需要有根本性的突破，研究发展类脑神经形态系统级芯片、类人的非硅系芯片，光芯片[1]、量子芯片、DNA 计算芯片等。

“人脑是非常复杂庞大的神经网络系统，总功耗仅为 20 瓦，远小于现有的 AI 系统。”在算力比拼加速、能耗日益攀升的当下，借鉴人脑的低功耗特性发展新型智能计算系统成为极具潜力的方向。采用人脑“神经形态动态计算”概念，将人脑中的高抽象层次注意力机制应用于类脑芯片设计，进一步挖掘了类脑芯片在性能和能效上的潜力。Speck 在一块芯片上集成了动态视觉传感器和神经形态芯片，8 核具有极低的静息功耗。典型视觉场景任务功耗可低至 0.7 毫瓦，为人工智能应用提供了高能效、低延迟和低功耗的类脑智能解决方案[2]。

(2) 在摩尔定律的末端，榨取它的剩余价值

在摩尔定律的尾声，进一步提高性能，就不能仅靠提高晶体管的面密度，而是要推出新的晶体管架构、材料选择及先进制程、Chiplet 多核异构(CPU + NPU + GPU……)、光连接(TPU)、及多维连接等。

A、Chiplet 是一种芯片设计和集成的方法。它将一个大型 AI 芯片分解为多核异构多个独立的功能模块片段(称 IP 核或称 Chiplet)。每个 Chiplet 是已经过验证的、可以重复使用的具有某种确切功能的集成电路设计模块。各种 IP 核如图形处理单元 GPU (Graphics Processing Unit)、神经网络处理单元 NPU (Neural network Processing Unit)、视频处理单元 VPU (Video Processing Unit)、数字信号处理单元 DSP (Digital Signal Processor)、张量处理单元 TPU (Tensor Processing Unit)等。它使芯片设计更加模块化和灵活。不同的芯片片段可以独立设计和优化，然后通过集成技术组合成一个完整的芯片。这种模块化的设计使得每个芯片片段可以在独立的制造工艺下进行生产(不必都使用同一种最高级制造工艺。这样可以降低制造成本，并提高芯片的产量和良品率[3]。

B、多维集成芯片。当前计算机的计算单元和存储单元分离，存储带宽制约了计算系统的有效带宽，造成时延长、功耗高等问题。采用多维集成芯片，将多个芯片堆叠在一起，使存储与计算完全融合，以新的高效运算架构进行二维和三维矩阵计算，具有更大算力(1000TOPS 以上)、更高能效(超过 10~100 TOPS/W)、降本增效三大优势，有效克服冯•诺依曼架构瓶颈，实现计算能效的数量级提升。AMD 推出的 MI300 就采用了 3D 堆叠技术和 Chiplet 设计，配备了 9 个基于 5 nm 制程的芯片组，置于 4 个基于 6 nm 制程的芯片组之上。

C、互连就是电流在芯片中各个晶体管、存储器、处理单元和其他组件之间的连接方式。前期铜互联取代了铝互联。现在，业界一直在寻找替代铜互连更优材料，如碳纳米管(CNT)、单层石墨烯(SLG)、多层石墨烯(FLG))与钌(Ru)。台积电使用的石墨烯表现出高本征载流子迁移率和大载流能力，具有高导热性和抗电迁移的竞争稳健性，还可以制成原子级厚度，有助于减轻厚度对 RC 延迟的影响。IBM 则使用钌。钌可以扩展到 1 纳米及以上节点不需要衬垫，在导线的顶部通过减色图案化法形成的钌互连通孔，形成连续的导线和自对准通孔。从而减少互连寄生电容，助于实现更快、更低功耗的芯片[4]。

D、改变器件架构

采用堆叠式 CFET 场效应管架构。将晶体管组件垂直堆叠在一起，而不是横向堆叠，极大地增加了单个芯片上可以安装的晶体管数量。这种架构的集成密度进一步提升，将 n 型和 p 型 MOS 元件堆叠在一起，可以堆叠 8 个纳米片。

(3) AI PC [5]

AI 要为广大人群使用，就必须将商用 PC 机集成人工智能技术，演变为 AI PC。AI PC 必须满足五个

条件：配备个人智能体(AI Agent)；具备本地异构算力；具备本地化个人知识库；开放的 AI 应用生态以及设备级个人数据和隐私安全保护；可以离线运行。

要加大边缘芯片的算力，需在 AI PC 处理器芯片中加入 NPU (神经处理单元)等，采用 CPU + GPU + NPU 的多核异构的架构。以 TOPS 作为基础衡量指标。微软将自己的 AI-PC 命名为 Copilot + PC；处理器需拥有超过 40 TOPS 的算力。去年微软的首批 Copilot + PC 选择与高通合作，搭载了基于 Arm 架构的处理器。Arm 架构的功耗优势明显，更适合性能强大的 AI PC 落地普及。

AI PC 需要云端和应用侧混合运行。用户要求复杂，需要大型数据支撑，进行大型迭代，可依靠在云端运行的生成式 AI；但需快速反应、及有数据安全的要求时，则依靠自身的算力和应用侧端小型化的大模型运行。

(4)智能体芯片

智能体芯片和 AI 算力芯片一样，具有微处理单元、存储单元，可拥有自己的知识库和推理机，因而它能对自己的位置进行标识，自主地决定是否对外来信息作出响应或行为反应，而且具备与其它智能性通信的接口。它添加简单的外围部件就能构建成独立的智能体。它通常作为工业智联网的中间级或智能应用末端[6]。

3.2. 软件工程师

(1) 大模型算法 LLM(Large Language Model)研究

当前，LLM 算法研究基本都是源于类脑研究。文本生成、语音互聊、图像生成都已获得巨大进展，但无论在精细、速度、能耗等方面还欠缺不少，尤其是在多模态视频识别、视频生成方面还在摸索，在人工神经网络的可解释性、对齐和控制的领域仍处于起步阶段。亟待 MiniCPM-Llama3-V 2.5 等新模型的出现[7]。

(2) 推出算力运算平台

大模型所需的大算力需要上万块 AI 芯片堆叠运行。英伟达首推出了通用并行计算架构的 CUDA 算力平台，让全球超过 400 万开发人员依靠 CUDA 来构建 AI 及其他应用程序。CUDA 能够在运算基础不断增长的情况下，扩大生态系统，使成本不断下降。现在，随着计算膨胀和计算成本的提升，多种 AI 芯片的出现，推出各自(或共同)的 AI 芯片运算平台，对去垄断，多元化至关重要[8]。

(3) 跨行业，在各行业领域建模，用行业数据喂养，训练行业应用大模型

人类所有的行业都需要和值得用 AGI 重做一遍。这就需要跨界和行业交叉，在各个专业、行业、企业甚至个人根据自身的需求、条件及安全需求建模，利用开源系统，组建行业训练模型、标注数据队伍，建立各自的生成式大模型，应用于各自的工作、生活中。

(4) 为末端应用开发 API、SDK、APP 等

大厂不断推出了各种多模态生成式大模型，急于变现。非电子及计算机专业人士也急于大量应用 AGI。这就需要大厂为大模型的末端应用同步推出应用程序接口 API (Application Programming Interface)、软件开发工具包 SDK (Software Development Kit)及 APP 等二次开发应用端，以帮助二次开发商和电子及计算机人士能简便地安装应用在用户侧的手机及 PC 机上。

(5) 为应用端芯片开发库函数

AGI 时代，用户侧的电器都逐步演变成为具有标识、状态、行为和接口智能体。这些智能体本身的功能越来越强大。一个多模态家居智能体加上各种前端传感器和后端执行机构可能就强似现在家庭的智能家居 + 智能健生养护 + 家居智能办公、学习 + ……等。

这种用户侧的智能体芯片都具有功能强大化、标准化、积木化、环境适应化、应用简单化的特点。

智能体芯片厂商都需要少量软件开发者为其研发完善的开源应用函数库，以提供给智能体开发者，也提供作为 AI 代码大模型的训练数据。这样采用智能体芯片开发智能体的工作效率就很高，所需的研究开发人员大大减少。

3.3. 应用侧电子及计算机工程师

快速发展的人工智能主体都依赖在云端的生成式大模型、大算力和大数据。但为了信息安全、反应快捷及消除拥堵，应用侧的“边缘 AI”像“边缘计算”一样也得到迅速地发展。

(1) 大模型的小型化

生成式 AI 在云端运行，但正在迅速演进至能在末端运行。如很多行业的垂直应用领域(如 L5 无人驾驶汽车、AI 医生)及个人 PC 用户，有时需要断网在用户侧末端运行，这就需要在用户侧末端能装入并运行百亿级以上的小型大模型。末端算力、存储有限，这就急需针对应用领域将大模型小型化，使其能装入智能体，单独在应用侧端运行。

(2) AI 工业物联网电子及计算机工程师

由于当今的物联网已转化为工业互联网的末端。物联网嵌入式工程师将会逐步演化为“工业物联网应用侧电子及计算机工程师”[9]。他们从事工业物联网末端智能体的开发、生产、维护。他们与传统物联网嵌入式工程师不同，有了新特点。

此时，物联网模块均成为智能体，其芯片功能强大、接口多样、协议标准，并备有推荐电路及配套的开源库函数。这种单体能力的提高，标准化，大规模生产的低成本化。使得小微物联网企业面临被兼并、淘汰、转行。这部分被裁员的工程师就只能向金子塔的两端转化。底层的工作越来越少。工程师们需更多地向软件发展，实现硬软结合，硬软交叉。也更多与应用专业交叉。要根据应用专业的要求搭载前端传感器，后端执行单元。要用应用专业的数据标记和训练，这就需要这部分工程师向应用专业跨界。

4. 当今的电子及计算机工程师的应对

4.1. 欢迎 AGI 的到来，迅速掌握出现的 AI 工具

人类历史已经证明，新技术都是不可抗拒的双刃剑，蒸汽机、电力、流水线直到核武器等都是如此。科技本身是无国界的，但使用科技的人是有政治观点的。技术的发展的阵痛之余，工作效率的提升，会使每个人的生活变得比以前更好。在这股 AI 技术浪潮中，与其害怕，倒不如积极拥抱，占据 AGI 的制高点，力所能及地顺风而上。不要囿于“模型有种种缺陷”，甚至“胡说八道”，而要看到这是必然趋势，成长中的短暂错误。学习它的利远大于弊。不想成为被逐渐淘汰的“不会使用 AI 工具的人”，那就赶快跑入掌握 AI 工具赛道中的第一梯队。

4.2. 在继承、否定、发展中成长[10]

在学习、掌握 AI 工具之时，要善于学习、继承，但不要迷信，要勇于怀疑、否定，勇于试错。科学技术就是在怀疑、否定之否定中发展的。让自己也在学习、怀疑、否定、创新探索中向金字塔顶端攀登。

4.3. 对电子及计算机教学进行大刀阔斧的改革

AGI 进入到指数发展的赛道，知识爆炸、更新淘汰得更快，部分前沿技术更新周期已大大缩短到以月计和年计，远小于课程设置、教材大纲审订、教材编写、出版的周期。所以这些权利应该下放给学校、专业、教研室及教师本人。允许教师有几分授课内容、方式裁定权。教师在教学中除自己不断学习外，

而且要指导学生扎好基础，带领学生紧跟潮流、批判吸收。鼓励双学位和获取交叉专业学分。

5. 基金项目

2023 年海南省高等学校科学研究项目《基于大数据的海南特色分布式养老监护系统的研究》(项目编号：Hnky2023-50)。

参考文献

[1] Xu, Z., Zhou, T., Ma, M., Deng, C., Dai, Q. and Fang, L. (2024) Large-Scale Photonic Chiplet Taichi Empowers 160-TOPS/W Artificial General Intelligence. *Science*, **384**, 202-209. https://doi.org/10.1126/science.adl1203

[2] Yao, M., Richter, O., Zhao, G., Qiao, N., Xing, Y., Wang, D., *et al*. (2024) Spike-Based Dynamic Computing with Asynchronous Sensing-Computing Neuromorphic Chip. *Nature Communications*, **15**, Article No. 4464. https://doi.org/10.1038/s41467-024-47811-6

[3] 周娜, 何铮, 何为民. 人工智能的生态树及算力研究[J]. 单片机与嵌入式系统应用, 2023, 23(9): 7-10.

[4] 九林. 半导体工艺的极限 1 nm 之战[EB/OL]. https://www.eefocus.com/article/1646589.html, 2023-12-05.

[5] CNMO 手机中国. AI 将助力 PC 迎来新一轮换机潮?事情恐怕没这么简单[EB/OL]. https://baijiahao.baidu.com/s?id=1800147117699467870&wfr=spider&for=pc, 2024-05-27.

[6] The Blog of Bill Gates (2023) The Future of Agents AI Is about to Completely Change How You Use Computers and Upend the Software Industry. https://www.gatesnotes.com/AI-agents

[7] 缪青海, 王兴霞, 杨静, 等. 从基础智能到通用智能: 基于大模型的 GenAI 和 AGI 之现状与展望[J]. 自动化学报, 2024, 50(4): 674-687.

[8] 凤凰网. 打破英伟达 CUDA 霸权[EB/OL]. http://ishare.ifeng.com/c/s/v002k46UWetBHM1S1hsUgvMUHdqWBr24cECNRVvXbPGNSOA, 2024-03-27.

[9] 工业互联网产业联盟. 工业互联网体系架构(版本 2.0) [EB/OL]. https://www.miit.gov.cn/cms_files/filemanager/1226211233/attach/20238/7b6171f454f94a5e9a14f2fd3b5f1c4c.pdf, 2020-04-23.

[10] 徐光春. 马克思主义大辞典[M]. 武汉: 崇文书局, 2017.

Embedded Technology and Intelligent Systems 嵌入式技术与智能系统, 2024, 1(2), 64-76
Published Online October 2024 in Hans. https://www.hanspub.org/journal/etis
https://doi.org/10.12677/etis.2024.12008

基于深度学习的移动端水果识别

郭　健，吴　薇*

杭州电子科技大学电子信息学院，浙江　杭州

收稿日期：2024年7月4日；录用日期：2024年11月21日；发布日期：2024年11月29日

摘　要

超市水果识别主要依赖人工，计算机视觉成为一种解决方案。然而目前仍面临部分水果识别精度低、终端设备部署困难、误识别图片难处理等挑战。因此，文章基于深度学习对移动端水果识别进行研究，旨在替代人工识别。首先文章构建了包含49种水果的超市水果图像数据集DailyFruit-49。并针对细分类特征相似度高、包装遮挡、形状小量少的水果识别困难，以及低算力设备模型部署问题，筛选了满足部署要求的骨干模型。设计了新的注意力模块RMA，改进了ViT Block以增强模型的细节识别能力和深层语义特征整合能力，最终得到DenseRMA_ViT模型，并基于Focal Loss改进损失函数。并在公开数据集Fruits-262上进行消融实验验证模型改进的有效性。最后结合实际设备，实现水果识别系统，满足实际使用。基于与用户的交互行为对误识别水果图像进行收集，并基于误识别图像实现模型权重自动微调，随使用时间延长，系统收集更多图片，提升模型识别精度与泛化能力，以处理实际应用中误识别水果。

关键词

水果识别，数据集构建，改进注意力机制，ViT，系统设计，模型权重自更新

Mobile Fruit Recognition Based on Deep Learning

Jian Guo, Wei Wu*

School of Electronic Information, Hangzhou Dianzi University, Hangzhou Zhejiang

Received: Jul. 4th, 2024; accepted: Nov. 21st, 2024; published: Nov. 29th, 2024

Abstract

Supermarket fruit recognition mainly relies on manual processes, and computer vision has emerged as a solution. However, challenges remain, including low accuracy for some fruits, difficulties in deploying them on terminal devices, and handling misidentified images. Therefore, this paper

*通讯作者。

文章引用: 郭健, 吴薇. 基于深度学习的移动端水果识别[J]. 嵌入式技术与智能系统, 2024, 1(2): 64-76.
DOI: 10.12677/etis.2024.12008

researches mobile fruit recognition based on deep learning, aiming to replace manual identification. First, the paper constructs the DailyFruit-49 dataset, which includes images of 49 types of fruits. Addressing the challenges of recognizing fruits with high feature similarity, packaging obstructions, and small shapes, as well as the deployment issues on low-compute devices, the backbone model meeting deployment requirements was selected. A new attention module, RMA, was designed, and the ViT Block was improved to enhance the model's detail recognition and deep semantic feature integration capabilities, resulting in the Dense RMA_ViT model. The loss function was also improved based on Focal Loss. Ablation experiments on the public dataset Fruits-262 verified the effectiveness of these improvements. Finally, a fruit recognition system was implemented on actual devices to meet practical usage needs. The system collects misidentified fruit images based on user interactions and automatically fine-tunes the model's weights based on these images. Over time, as the system collects more images, the model's recognition accuracy and generalization ability improve, effectively handling misidentified fruits in real-world applications.

Keywords

Fruit Recognition, Dataset Construction, Improved Attention Mechanism, ViT, System Design, Model Weight Self-Updating

1. 引言

中国自古以来重视农业发展，尤其在水果种植方面取得了显著成就，品种繁多，面积广泛覆盖，居全球前列。随着经济增长和生活品质提升，消费者更加注重饮食营养，水果已成为生活中不可或缺的一部分，使中国成为全球最大的水果消费地[1]。目前，水果商户主要依赖人工识别水果种类并进行称重和结算，这种方式不便且不环保。水果种类繁多，外观相似，增加了识别难度，人工处理速度慢且易出错。中国劳动力成本不断上升，传统服务模式浪费人力资源，影响顾客满意度和超市盈利。因此，探索自动水果识别结算方法十分重要[2]。

水果识别技术通过深度分析水果图像或视频，实现智能判别，广泛适用于水果商店和无人超市，提升购物体验，减少人力成本，优化运营效果。加快智慧农业发展，推动农业数字化、智能化发展，实现高效可持续的智慧农业。推动智能城市发展，在无人自助超市中，智能水果识别系统能自动完成消费流程，提供便捷高效的购物体验。

2. 相关工作

基于机器视觉的水果识别方法通常包括三个步骤。首先是图像预处理，包括调整图像的亮度和对比度、颜色空间转换、直方图均衡化、以及像素值归一化[3]。其次是特征提取，方法包括颜色直方图([4]、Gabor 滤波[5]、局部二值模式[6]、方向梯度直方图[7]、边缘特征如 Sobel [8]和 Canny [9]、尺度不变特征变换[10]以及主成分分析[11]。最后是分类，常用方法有支持向量机[12]、K 最近邻[13]、决策树(Decision Tree) [14]和随机森林[15]。

近年来，基于深度学习的水果识别方法取得了显著进展。Enciso 等人基于 AlexNet 模型在 320 张数据集上，对柠檬的新鲜和腐烂分类分别达到了 98.25%和 93.73%的准确率[16]。孟欣欣等人基于计算机视觉的香梨果实目标检测模型，使用 Resnet 模型在 9600 张水果数据上进行预训练，并构建了 Mask R-CNN

模型，在成熟香梨数据集上的平均分割准确率达到 98.02% [17]。Xue 等人提出了一种基于深度学习的混合水果分类方法，构建了 CAE-ADN 框架，结合了注意力模型和卷积自动编码器，实现了高效的水果图像分类[18]。Lu 等人基于 DenseNet201 模型和迁移学习技术，实现了番茄分类的高准确率，即使在图像质量受干扰的情况下，分类准确率仍达到 96.16% [19]。

此外，Chandel 等人基于 GoogLeNet 模型，提出了一种用于识别农作物水分胁迫条件的方法，在 1200 张包含玉米、秋葵和大豆的数据集上，表现优异[20]。Kang 等人基于 ResNet 模型，建立了一个能够区分水果新鲜度和种类的分类模型，收集了 18,000 张包含 7 类水果的图片，通过迁移学习技术进行模型训练，新鲜度分类的识别准确率达到 98.50%，种类分类的识别准确率为 97.43% [21]。Ismail 等人提出了一种基于 CNN 的实时水果等级分类系统，使用 EfficientNet 模型对苹果和香蕉的数据集进行验证，分类准确率达 99.2% [22]。Huang 等人提出了一种基于 ViT 的神经网络模型，建立了一个包含 9375 次对 15 种水果的触觉数据集[23]。这些研究展示了深度学习在水果识别中的广泛应用和高效性能，为进一步研究和应用提供了坚实的基础。

3. 方法

3.1. 基础模型选择

针对移动端的水果识别应用部署，本文选用内存为 2 GB 的 Rockchip RK3399-mid 芯片，因此需要考虑模型权重大小和模型计算量。因此本文对模型的筛选规则为模型计算量(FLOPS/10^6)小于 10,000，模型参数量小于 100 M。通过对目前主流的网络模型进行筛选，本文选用 DenseNet-169 [30]为本文的骨干网络，用于后续的研究，其模型结构如图 1 所示。

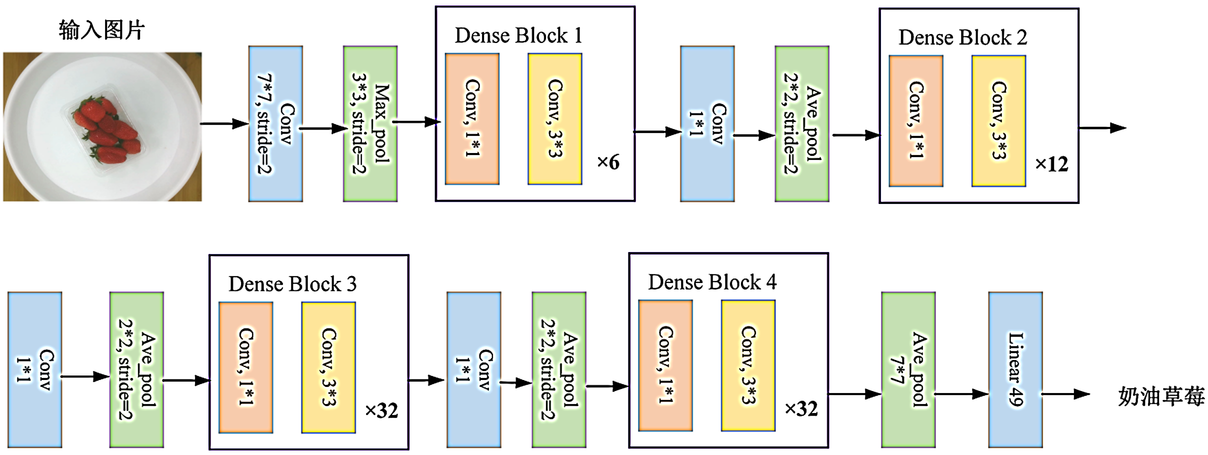


Figure 1. DenseNet-169 model architecture
图 1. DenseNet-169 模型结构

3.2. 注意力机制

为进一步提升模型性能。让模型更好地关注到水果的细节纹理，小目标水果特征，并缓解包装、摆放数量等干扰信息的影响，本文尝试在骨干网络中引入注意力机制。结合 DenseNet-169 模型自身特点，其优势在于稠密块连接，如在稠密块中添加注意力将会破坏稠密结构，因此，本文主要在其稠密块连接处进行添加注意力。

本文设计新的注意力模块，残差混合注意力(Residual Mixed attention, RMA)。同时考虑空间和通道信息，使用多尺度卷积特征融合形式提取通道信息，并添加残差连接，相比于单一尺度特征提取，多尺度

特征可以动态调整对目标的关注程度，使得模型对于尺度变化具有更好的适应性，综合利用水果的全局结构和局部细节信息提高模型对图像的识别准确性，从而提高模型的鲁棒性。

其模型结构如图 2 所示，首先分别通过 3 × 3 卷积核和 5 × 5 卷积核对输入特征进行卷积提取，得到两个不同尺度的特征图，然后通过相加操作同时结合两个不同尺度的特征图，输入至全局平均池化层得到权重向量，并通过一个线性层来映射全局特征向量，再通过 Sigmoid 激活函数将其归一化到 0 到 1 之间，得到全局特征向量，然后分别对所提取的多尺度特征进行加权输出，并通过 Concat 操作按照通道拼接，然后对特征使用 1 × 1 卷积核进行通道融合并输入空间注意力模块。

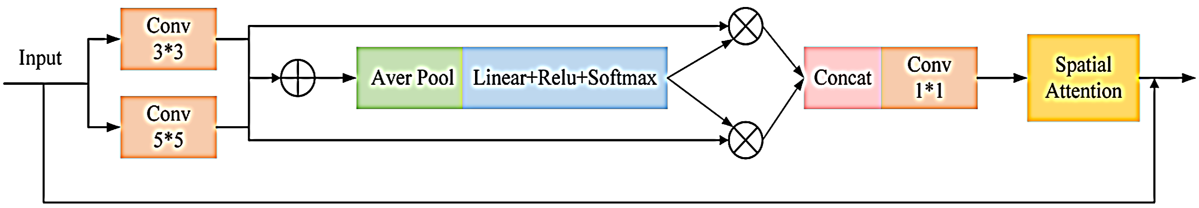


Figure 2. Residual Mixed Attention (RMA) module
图 2. 残差混合注意力 RMA 模块

3.3. 添加 ViT Block

深层特征是对输入数据进行多次非线性变换后得到的结果，具有更高级的语义信息，因此对深层特征的信息整合可以使得模型具有更好的泛化性，提升识别精度。ViT [24]架构使用自注意力机制来捕捉图像中的全局信息，能够更好地理解图像整体结构和上下文关系，具有更好的语义理解能力，可以增强模型的泛化能力。

因此为提升模型深层语义提取能力，增强泛化性。本文实现适合在 CNN 模型中添加的 ViT Block，其模型结构如图 3 所示，通过对三维输入特征按照形状依次展平变成二维特征，并进行线性映射实现通道特征融合得到新的长度为 D 的二维特征，然后对特征进行位置编码，输入 Encoder 块，进行深层特征抽取，并输出特征。

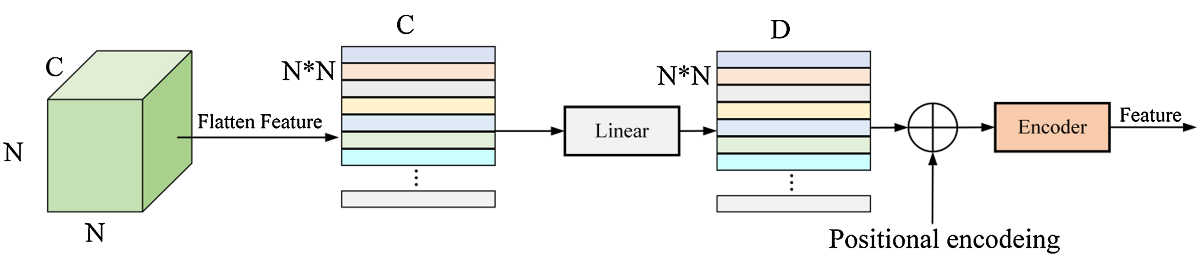


Figure 3. ViT Block module
图 3. ViT Block 结构图

本文尝试在 DenseNet-169 模型中添加基于 ViT 架构的语义提取模块。通过分析认为 CNN 在处理图像浅层特征时具有先验知识即归纳偏置(Inductive Bias)例如平移等变性、局部连接性，因此能够更好地利用图像本身的空间信息，而 ViT 结构缺少了这种归纳偏置，并且在前期 ViT 架构计算复杂度也较高。因此对于低级特征处理阶段 CNN 更具优势，而随着模型层数的加深，多重卷积操作使得模型特征逐渐深层化，使得特征更加具有语义性，且特征尺寸有所下降，此时 ViT Block 更具有优势。最终的模型结构如图 4 所示。

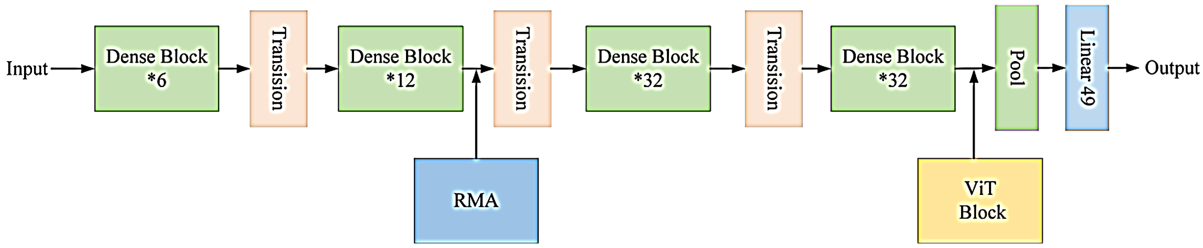


Figure 4. DenseRMA_ViT model architecture
图 4. DenseRMA_ViT 模型结构图

3.4. 改进损失函数

损失函数在分类模型中扮演着非常重要的角色，用于衡量模型预测与真实标签的差距，模型根据损失来调整参数，模型的训练过程旨在最小化损失函数，使得模型能够从训练数据中学到合适的特征和规律，以便正确分类样本。因此选择合适的损失函数可以帮助提高模型的性能、泛化能力以及对特定问题的适应性。常规分类损失函数为交叉熵损失(Cross-Entropy Loss)，计算公式如下：

$$L_{CE}\left(p,y\right)=-\sum_{i=1}^{K}y_i\log\left(p_i\right) \tag{1}$$

其中 K 为水果类别数，P 为预测概率类别向量，y 为真实标签独热编码向量。

通过分析模型训练损失曲线图和数据集特点发现，数据集样本中简单样本居多，困难样本相对较少。而交叉熵损失函数没有对简单样本和困难样本的损失进行区分。虽然困难样本的损失较大，但是简单样本数量多，这将会导致模型过多地关注简单样本的损失，模型难以学习困难样本的特征。同时本文水果识别系统在后续使用中会对识别错误的水果图片进行收集，并进行模型微调，然而在实际使用中，对于热销水果收集到的图片必定更多，这对于后续模型微调将会造成类别样本不均衡问题。基于上述两个问题，本文基于 Focal Loss [25]对损失函数进行改进，计算公式如下：

$$L_F\left(p_t\right)=-w_t\left(1-p_t\right)^2\log\left(p_t\right) \tag{2}$$

其中 t 为对应类别序号，w_t 为对应类别损失权重，在训练之前设定，统计训练样本类别数量，设定样本数最高类别的损失权重为 1，其他类别的损失权重为与样本数最高类别的比值的倒数，即数量越少权重越大，因此相对地提升小样本数量类别的损失权重。P_t 为对应预测类别概率。当 P_t 越接近 1，即样本为简单样本，$(1-P_t)^2$ 则相对越小，模型对简单样本的损失相对权重越小。当 P_t 越接近 0，即样本为困难样本，$(1-P_t)^2$ 则相对越大，模型对困难样本的损失相对权重越大。

4. 实验

4.1. 数据集

目前常见的水果数据集包括 FruitVeg-81 [26]、Fruits-262 [27]、Fruits-360 [28]等，其中数字代表数据集中水果种类数量，图 5 展示了部分上述水果数据集的图片。可以看出尽管 Fruits-360 的水果种类众多，但相对于其他数据集过于简单，不适合实际超市场景的应用。FruitVeg-81 数据集除了水果图片，还包含了蔬菜图片，并且果蔬类别数量相较于 Fruits-262 较少，因此在公开数据集中，Fruits-262 数据集更加适合作为水果识别系统模型训练数据集。

尽管 Fruits-262 包含大量水果种类和图片，但与本地超市的水果种类仍有差距。例如，本地超市的苹果种类包括树顶红富士、优选红富士、脆心甜红富士、甜心小苹果，而 Fruits-262 中有 apple (苹果)、elephant apple (象苹果)、malay apple (马来苹果)、otaheite apple (塔希提苹果)，相关图片如图 6 所示。虽然其他三类水果名称与 apple 相关，但与本地常见苹果差异较大，难以视为同一类。类似情况在其他水果中也存在。因此，Fruits-262 尽管种类多，但未细分类，也未覆盖本地超市的水果种类。因此，构建适用于本地超市的水果数据集对于水果识别系统至关重要。

Figure 5. Sample Images from the public dataset
图 5. 公开数据集部分图片

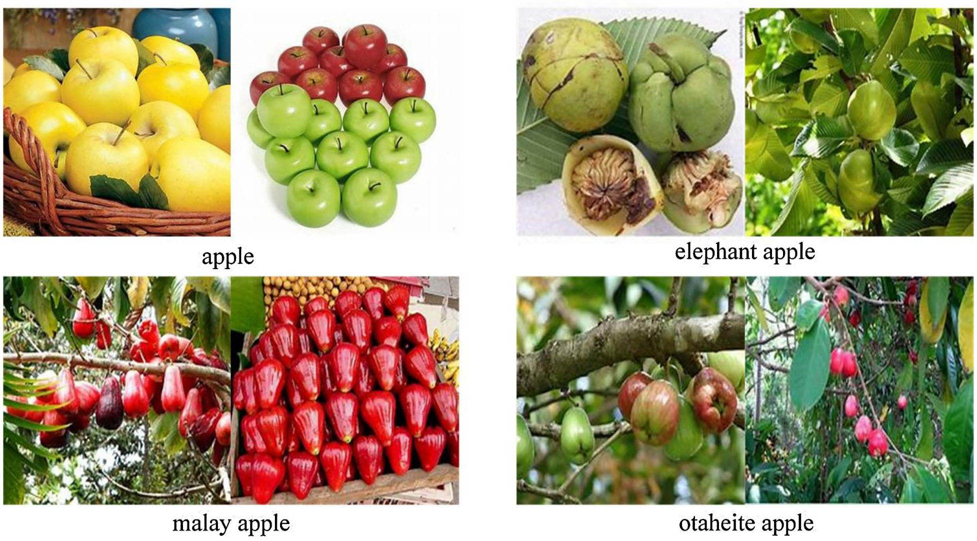


Figure 6. Images of some apple varieties from the fruits-262 dataset
图 6. Fruits-262 部分相关苹果种类图片

本文通过自制数据集用于训练模型得到权重，用于实际超市水果识别系统，并且使用公开数据集 Fruits-262 验证模型在算法上改进的有效性。在实际超市应用场景中，水果种类繁多同时存在较多细分类，因此为贴合超市的实际使用，在拍摄时覆盖超市所有种类水果，确保数据集的多样性。本文拍摄 49 种超市水果，在实际分类模型中，模型依照序号对类别预测概率进行输出。由于水果包装状态存在差异，包括袋装盒装和散装等，因此本文考虑到实际使用针对不同水果的不同包装进行拍摄，提高模型的泛化性。针对不同水果的状态拍摄，如图 7 为小番茄不同状态下的拍摄图片。

数据集扩增有助于提升模型泛化性和识别精度，防止过拟合。Meta 开源的 SAM (Segment Anything Model) [29]实现了无需标注即可对任意图像中的任何物体进行分割。本文基于 SAM 设计了水果数据集扩增方法。将图片传入 SAM，随机生成多个提示输入，对图像进行分割，得到分割对象。然后对分割对象

进行 1 到 4 次随机复制，并随机翻转、旋转、缩放，在背景图像上随机粘贴，完成数据扩增。如图 8 所示，对同一张草莓图像随机生成提示点进行扩充，左列为随机提示点，中间为分割实例，右列为最终扩充图。最终，每种水果图片扩增至 6000 张，总计 29.4 万张，命名为 DailyFruit-49。

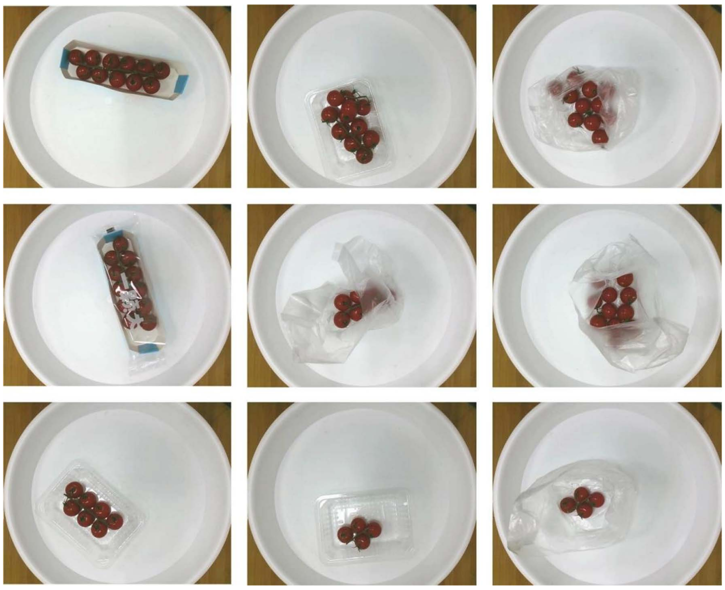


Figure 7. Photoshoot of cherry tomatoes in different states
图 7. 小番茄不同状态下的图片拍摄

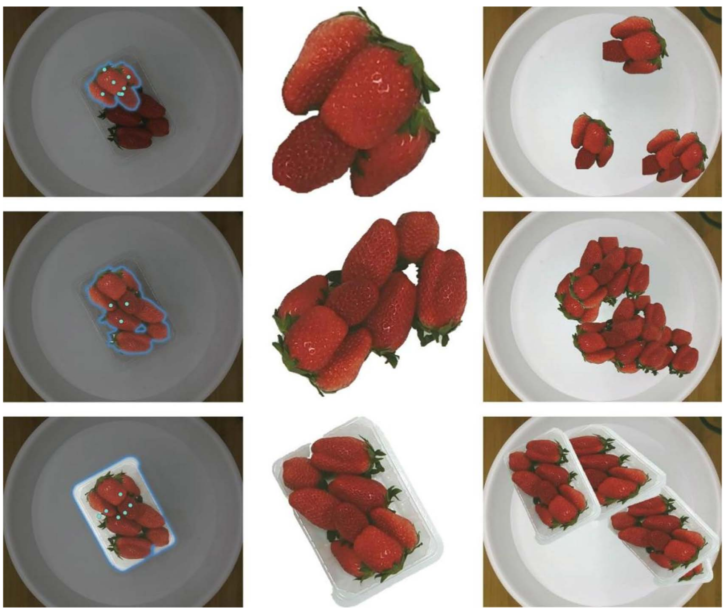


Figure 8. Fruit image augmentation concept based on SAM
图 8. 基于 SAM 的水果图片扩增示意图

4.2. 评价指标

为了全面客观地评价模型对水果识别的性能，本文采用多种图像分类评价指标，包括准确率(Accuracy)、精确率(Precision)、召回率(Recall)、F1 分数(F1 Score, F1)。准确率、精确率、召回率的数学

表达式如式(3)、(4)、(5)所示。表 1 解释了式中元素具体情况。

$$acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$pre = \frac{TP}{TP + FP} \tag{4}$$

$$recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1 = \frac{2 \times pre \times recall}{pre + recall} \tag{6}$$

Table 1. Table of precision and recall explanation
表 1. 精确率和召回率解释表

区分	预测 Positive	预测 Negative	合计
实际 Positive	TP (True Positive)	FN (False Negative)	TP + FN
实际 Negative	FP (False Positive)	TN (True Negative)	FP + TN
合计	TP + FP	FN + TN	TP + FP + FN + TN

F1 分数是精确率和召回率的调和平均，综合考虑了分类器的精确性和召回率，是一个综合性的评价指标，可以更加综合地反映了分类器的性能。其数学表达式如式(4)所示。

4.3. 消融实验

为验证改进模型的有效性，本文使用公开数据集 Fruits-262 进行消融实验，验证算法改进的有效性，尽管 Fruits-262 数据与实际应用数据相差较大，但是从算法角度，复杂数据依然存在参考价值。为证明本文对模型改进的有效性，对 RMA 模块、ViT Block、Focal Loss 进行消融实验，并统计结果如表 2 所示，由表可知相较于基础骨干网络，3 种改进方案的有效性，模型最终 Accuracy、Precision、Recall、F1 为 75.28%、75.84、75.37%、75.60%，相较于仅添加 RMA 模块，虽然精确率有所降低，但是召回率和综合性能 F1 有所提升，说明模型有效缓解了样本不均衡，提升了模型的综合性能，证明 Focal Loss 改进的有效性。同时对比原始模型 1 号 DenseNet-169，添加改进模块的模型性能均有所提升，证明 DenseRAM_ViT 改进的有效性。

Table 2. Ablation study results on the Fruits-262 dataset
表 2. 在 Fruits-262 数据集上的消融实验结果

编号	RMA	ViT Block	Focal	Accuracy/%	Precision/%	Recall/%	F1/%
1	-	-	-	73.63	74.94	74.21	74.57
2	√	-	-	74.64	**75.90**	74.75	75.32
3	-	√	-	74.63	75.77	74.76	75.26
4	-	-	√	74.02	75.12	74.48	74.80
5	√	√	√	**75.28**	75.84	**75.37**	**75.60**

同时为了直观展现添加 RMA 的效果对比，如图 9 使用梯度类激活热力图对 Fruits-262 数据集中部分图像各区域受关注程度做可视化展示，由图第一行对比可知，相对于原始模型 DenseNet-169，添加了 RMA

注意力模型减少了对背景的关注，并且加大了对水果目标的关注，由图第二行对比可知，改进模型增大了对水果目标的关注程度与范围，因此证明了本文改进注意力 RMA 的有效性。

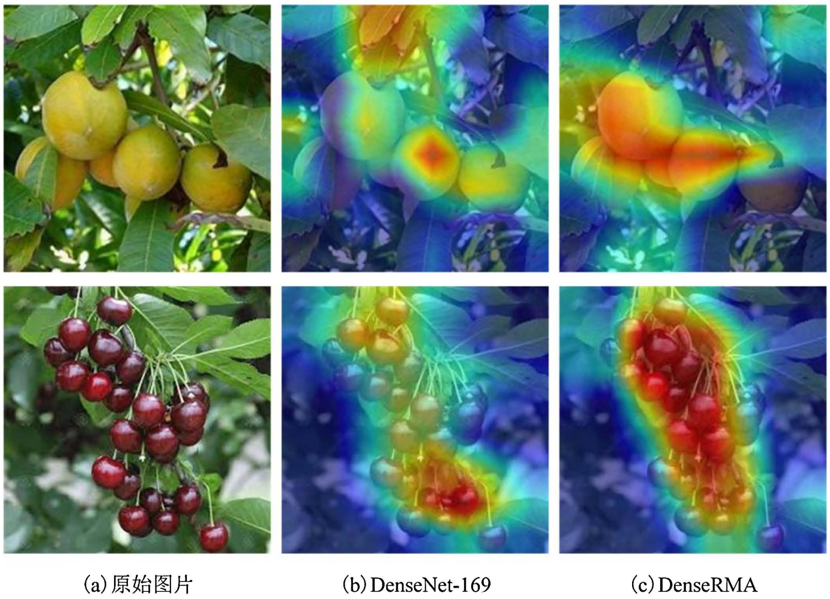


Figure 9. Comparison of heatmaps with and without the Introduction of the RMA module
图 9. 引入 RMA 模块热力图对比

4.4. 结果对比

为了说明本文模型设计的有效性，列举具有代表性的同类水果识别研究进行对比，结果如表 3 所示，尽管有些研究识别性能很好，但是他们的研究水果种类与数量较少，不具有代表性。在公开数据集 Fruits-262 上，本文最终实现模型 DenseRMA_ViT 性能相对于数据集发布者 Minut [28]有较大提升。在自制数

Table 3. Comparison of related studies on fruit recognition
表 3. 水果识别相关研究对比

文献	年份	数据集	水果种类	样本数量	Top1 Acc/%	Top5 Acc/%
Enciso [16]	2018	自制	1	320	98.25	_
Lu [19]	2021	自制	9	6100	96.16	_
Kang [21]	2022	自制	7	18000	97.43	_
Ismail [22]	2022	自制	2	8791	96.7%	_
Minut [28]	2021	Fruits-262	262	225,640	58%	_
DenseNet [30]	2017	Fruits-262	262	225,640	73.63%	_
DenseRMA_ViT (our)	2024	Fruits-262	262	225,640	75.28%	_
Mobile v2 [31]	2018	DailyFruit-49	49	294,000	83.65	92.64
ResNet [32]	2015	DailyFruit-49	49	294,000	85.75	93.57
DenseNet [30]	2017	DailyFruit-49	49	294,000	87.53	98.83
ViT [24]	2020	DailyFruit-49	49	294,000	87.95	97.07
DenseRMA_ViT (our)	2024	DailyFruit-49	49	294,000	92.24	99.73

据集 DailyFruit-49 上，本文对比了深度学习经典分类模型 Mobile v2、ResNet、DenseNet、ViT，由此证明模型的改进有效性，同时本实验覆盖的水果种类范围大，前 5 类识别精度达到 99.73%，识别率能够满足实际使用需求。综上可知本实验设计的水果识别算法在水果识别领域具有应用价值。

4.5. 系统实现

系统的终端为 Android 系统，深度学习模型基于 Pytorch 框架和 Python 语言实现，但无法直接在终端部署。因此，需要将模型转换为适合终端部署的格式。本文采用 TorchScript 进行模型格式转化。TorchScript [33]具有运行独立性、高性能、跨平台支持等优点，可以将 PyTorch 模型转换为独立格式，使其无需依赖 Python 解释器，可以在移动端和嵌入式设备上运行，并且无需重新训练模型。同时，TorchScript 的静态图表示可以优化模型性能和效率，如进行模型剪枝等。经过测试，水果识别终端设备不需要高性能 CPU 或 GPU，在瑞芯微的 RK3399 芯片上可以流畅运行，其他设备如摄像头也不需要高端配置。

本文系统设计框架如图 10 所示，分为四个模块：水果检测模块、前端交互模块、后台自更新模块、终端更新模块。四个模块相互连接，形成闭环，实现系统自更新。系统通过摄像头实时拍摄用户购买的水果图像，使用模型进行识别，并在终端界面展示供用户交互。用户可在终端界面更正识别错误的水果种类，并将正确信息上传至后台服务器。后台系统收集误识别图片，管理人员进行检测与修正。当误识别图片达到一定数量，后台启动迁移学习进行模型微调，并发布新权重。终端设备每日重启时会检查新权重并进行更新。整个系统实现了水果图像自动化识别与模型权重自优化，随着使用时间的推移，系统将获得更大的水果图像数据集和更优的模型权重。

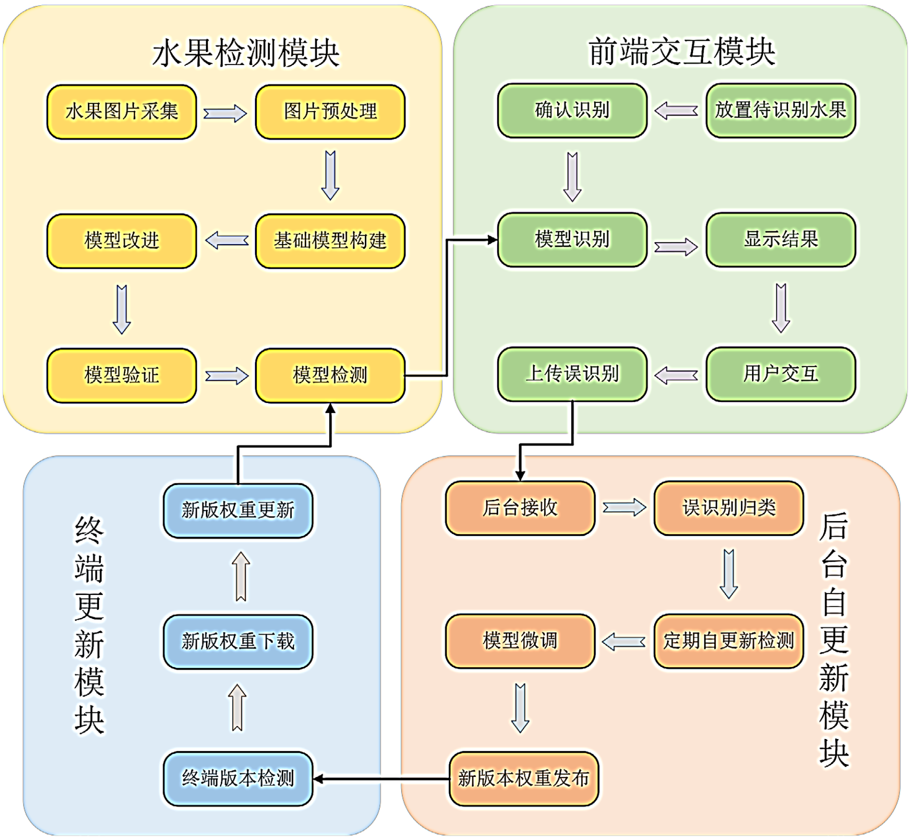


Figure 10. Framework diagram of deep learning-based fruit detection system
图 10. 基于深度学习的水果检测系统设计框架图

水果检测模块为主要功能，涉及整个装置，其整体结果如图 11 所示，该模块包括摄像头拍摄模块、水果展示台、终端处理模块。由用户将待购买商品放入展示台，并通过摄像头进行拍摄传入终端，终端设备部署识别程序，对水果进行识别。模型的部署包括对所得到的 Pytorch 模型权重 pth 文件进行转换为 TorchScript 格式，再结合 Android 应用程序进行模型嵌入转换好的模型完成部署。

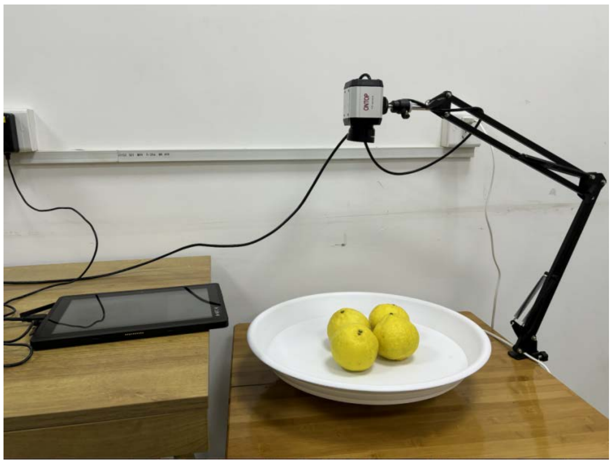


Figure 11. Fruit photography device
图 11. 水果拍摄装置

4.6. 实际超市水果验证

为验证 DenseRMA_ViT 水果识别算法的有效性，在本地超市进行了实地拍摄和识别测试。由于季节原因，部分水果未在售，对 28 种在售水果进行拍摄，每种 4 张，分别以袋装和散装 1:3 拍摄，共 112 张。拍摄过程中不固定光照、角度和距离，以增加难度。识别结果显示，Top5 全部正确，但有些不足，如奇异果的 Top1 识别错误。整体上，模型 Top1 识别准确率为 84.82%，Top5 为 96.43%。相比基础模型 DenseNet-169 的 Top1 准确率 79.46%、Top5 准确率 90.18%，改进模型分别提升了 5.36%和 6.25%。尽管存在误识别，系统可以收集误识别图像，更新数据集并微调模型权重，从而提升精度。

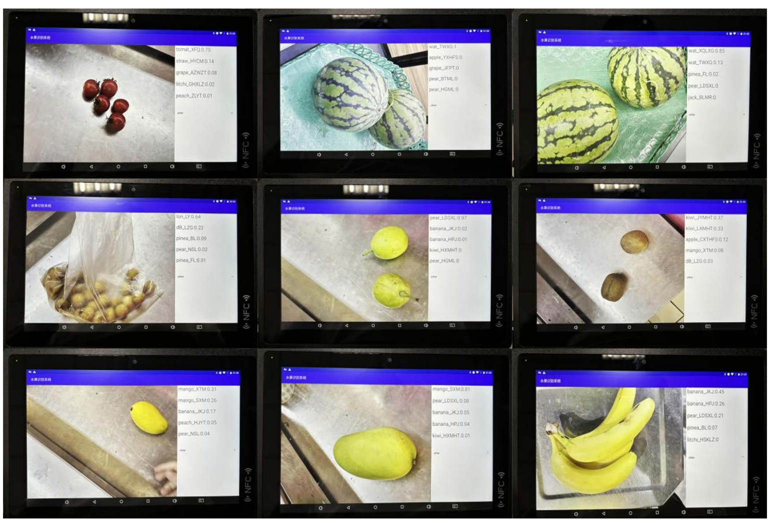


Figure 12. Supermarket real-world testing
图 12. 超市实拍测试

5. 结论

水果自动识别提升商店效益，减少标签，环保，是智慧农业关键。本文为本地超市构建了新水果数据集，设计了基于深度学习的识别系统，实现自动识别，加快智慧农业推广。拍摄超市各种水果，考虑盒装、袋装、散装等不同包装。首次应用 SAM 进行数据扩增，构建包含 49 种水果、29.4 万张图片的 DailyFruit-49 数据集。选择 DenseNet 为基础模型，结合通道与空间注意力机制，设计 RMA 注意力机制，改进 ViT Block 用于深层语义特征提取，得到 DenseRMA_ViT 模型。基于 Focal Loss 改进交叉熵损失函数，增强对困难和少样本水果的关注，提升模型泛化性。在低算力终端部署识别系统，并设计自动微调模型权重的机制，通过收集误识别图片不断优化模型，提升识别率，验证超市实际效果。

参考文献

[1] 吴中勇, 李延荣, 董中丹. 我国水果市场发展现状及对策研究[J]. 中国果菜, 2023, 43(11): 79-83+87.

[2] 中研普华公司, 2022-2027 年中国果蔬行业市场全面分析及发展趋势调研报告[R]. 深圳: 中国行业研究网, 2022.

[3] Jana, S., Basak, S. and Parekh, R. (2017) Automatic Fruit Recognition from Natural Images Using Color and Texture Features. 2017 *Devices for Integrated Circuit* (*DevIC*), Kalyani, 23-24 March 2017, 620-624. https://doi.org/10.1109/devic.2017.8074025

[4] Novak, C.L. and Shafer, S.A. (1992) Anatomy of a Color Histogram. *Proceedings* 1992 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Champaign, 15-18 June 1992, 559-605. https://doi.org/10.1109/cvpr.1992.223129

[5] Movellan, J.R. (2002) Tutorial on Gabor filters. *Open Source Document*, **40**, 1-23.

[6] Pietikäinen, M. (2010) Local Binary Patterns. *Scholarpedia*, **5**, 9775. https://doi.org/10.4249/scholarpedia.9775

[7] Tomasi, C. (2012) Histograms of Oriented Gradients. *Computer Vision Sampler*, **1**, 1-6.

[8] Gao, W.S., Zhang, X.G., Yang, L. and Liu, H.Z. (2010) An Improved Sobel Edge Detection. 2010 3*rd International Conference on Computer Science and Information Technology*, Chengdu, 9-11 July 2010, 67-71. https://doi.org/10.1109/iccsit.2010.5563693

[9] Ding, L. and Goshtasby, A. (2001) On the Canny Edge Detector. *Pattern Recognition*, **34**, 721-725. https://doi.org/10.1016/s0031-3203(00)00023-6

[10] Cruz-Mota, J., Bogdanova, I., Paquier, B., Bierlaire, M. and Thiran, J. (2011) Scale Invariant Feature Transform on the Sphere: Theory and Applications. *International Journal of Computer Vision*, **98**, 217-241. https://doi.org/10.1007/s11263-011-0505-4

[11] Wold, S., Esbensen, K. and Geladi, P. (1987) Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems*, **2**, 37-52. https://doi.org/10.1016/0169-7439(87)80084-9

[12] Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J. and Scholkopf, B. (1998) Support Vector Machines. *IEEE Intelligent Systems and their Applications*, **13**, 18-28. https://doi.org/10.1109/5254.708428

[13] Kramer, O. (2013) K-Nearest Neighbors. In: Kramer, O., Ed., *Dimensionality Reduction with Unsupervised Nearest Neighbors*, Springer Berlin Heidelberg, 13-23. https://doi.org/10.1007/978-3-642-38652-7_2

[14] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/a:1010933404324

[15] Song, Y.Y. and Ying, L.U. (2015) Decision Tree Methods: Applications for Classification and Prediction. *Shanghai Archives of Psychiatry*, **27**, 130.

[16] Enciso-Aragón, C.J., Pachón-Suescún, C.G. and Jimenez-Moreno, R. (2018) Quality Control System by Means of CNN and Fuzzy Systems. *International Journal of Applied Engineering Research*, **13**, 12846-12853.

[17] 孟欣欣, 阿里甫·库尔班, 吕情深, 等. 基于迁移学习的自然环境下香梨目标识别研究[J]. 新疆大学学报(自然科学版), 2019, 36(4): 461-467.

[18] Xue, G., Liu, S. and Ma, Y. (2020) A Hybrid Deep Learning-Based Fruit Classification Using Attention Model and Convolution Autoencoder. *Complex & Intelligent Systems*, **9**, 2209-2219. https://doi.org/10.1007/s40747-020-00192-x

[19] Lu, T., Han, B., Chen, L., Yu, F. and Xue, C. (2021) A Generic Intelligent Tomato Classification System for Practical Applications Using Densenet-201 with Transfer Learning. *Scientific Reports*, **11**, Article No. 15824. https://doi.org/10.1038/s41598-021-95218-w

[20] Chandel, N.S., Chakraborty, S.K., Rajwade, Y.A., Dubey, K., Tiwari, M.K. and Jat, D. (2020) Identifying Crop Water

Stress Using Deep Learning Models. *Neural Computing and Applications*, **33**, 5353-5367. https://doi.org/10.1007/s00521-020-05325-4

[21] Kang, J. and Gwak, J. (2021) Ensemble of Multi-Task Deep Convolutional Neural Networks Using Transfer Learning for Fruit Freshness Classification. *Multimedia Tools and Applications*, **81**, 22355-22377. https://doi.org/10.1007/s11042-021-11282-4

[22] Ismail, N. and Malik, O.A. (2022) Real-time Visual Inspection System for Grading Fruits Using Computer Vision and Deep Learning Techniques. *Information Processing in Agriculture*, **9**, 24-37. https://doi.org/10.1016/j.inpa.2021.01.005

[23] Huang, R., Zheng, W., Zhang, B., Zhou, J., Cui, Z. and Zhang, Z. (2023) Deep Learning with Tactile Sequences Enables Fruit Recognition and Force Prediction for Damage-Free Grasping. *Computers and Electronics in Agriculture*, **211**, Article ID: 107985. https://doi.org/10.1016/j.compag.2023.107985

[24] Dosovitskiy, A., Beyer, L., Kolesnikov, A., *et al*. (020) An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. arXiv: 2010.11929.

[25] Lin, T., Goyal, P., Girshick, R., He, K. and Dollar, P. (2017) Focal Loss for Dense Object Detection. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 2999-3007. https://doi.org/10.1109/iccv.2017.324

[26] Waltner, G., Schwarz, M., Ladstätter, S., Weber, A., Luley, P., Lindschinger, M., *et al*. (2017) Personalized Dietary Self-Management Using Mobile Vision-Based Assistance. In: Battiato, S., Farinella, G., Leo, M. and Gallo, G., Eds., *New Trends in Image Analysis and Processing—ICIAP* 2017, Springer International Publishing, 385-393. https://doi.org/10.1007/978-3-319-70742-6_36

[27] Minut, M. and Iftene, A. (2021) Creating a Dataset and Models Based on Convolutional Neural Networks to Improve Fruit Classification. 2021 23*rd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing* (*SYNASC*), Timisoara, 7-10 December 2021, 155-162. https://doi.org/10.1109/synasc54541.2021.00035

[28] Mureşan, H. and Oltean, M. (2018) Fruit Recognition from Images Using Deep Learning. *Acta Universitatis Sapientiae, Informatica*, **10**, 26-42. https://doi.org/10.2478/ausi-2018-0002

[29] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., *et al*. (2023) Segment Anything. 2023 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Paris, 1-6 October 2023, 3992-4003. https://doi.org/10.1109/iccv51070.2023.00371

[30] Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q. (2017) Densely Connected Convolutional Networks. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 2261-2269. https://doi.org/10.1109/cvpr.2017.243

[31] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L. (2018) MobileNetV2: Inverted Residuals and Linear Bottlenecks. 2018 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 4510-4520. https://doi.org/10.1109/cvpr.2018.00474

[32] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/cvpr.2016.90

[33] Reed, J., DeVito, Z., He, H., *et al.* (2022) Torch. fx: Practical Program Capture and Transformation for Deep Learning in Python. *Proceedings of Machine Learning and Systems*, **4**, 638-651.

Embedded Technology and Intelligent Systems 嵌入式技术与智能系统, 2024, 1(2), 77-84
Published Online October 2024 in Hans. https://www.hanspub.org/journal/etis
https://doi.org/10.12677/etis.2024.12009

在微控制器上实现在设备端训练的异常检测

宋 岩，许 鹏，张 岩

恩智浦(中国)管理有限公司北京分公司，北京

收稿日期：2024年9月5日；录用日期：2024年11月21日；发布日期：2024年11月29日

摘 要

在当前嵌入式系统与人工智能技术融合的前沿领域，文章聚焦于一种基于单类支持向量机(One-Class SVM)的异常检测算法，并提供了一套完整的MCU友好的工程实现，不需要依赖于动态内存分配以及文件系统，特别适合于在资源受限的边缘设备上进行高效、实时的训练与预测。我们的方法不仅可以实现在MCU上训练和高效存储机器学习模型，还支持增量学习，从而在几乎不增加计算负担的前提下，持续改进模型对实际工况的适应能力。我们的实验装置是安装了三轴加速度传感器的震动源(如风扇)，以模拟在工作期间发出振动的工业设备。文章的方法也可以通过替换传感器和特征计算的预处理算法来实现对其它设备的监控，以适应不同的工况环境和应用的需求。

关键词

微控制器单元(MCU)，设备端训练(ODT)，支持向量机(SVM)，人工智能应用，MCXN947

Anomaly Detection on Microcontroller with On-Device Training

Yan Song, Peng Xu, Yan Zhang

Beijing Branch of NXP (China) Management, Co., Ltd., Beijing

Received: Sep. 5th, 2024; accepted: Nov. 21st, 2024; published: Nov. 29th, 2024

Abstract

This paper introduces an anomaly detection algorithm based on one-class support vector machines (SVMs) and an MCU-friendly engineering implementation. It does not rely on dynamic memory and file systems, it is particularly suitable for efficient, real-time training and prediction on resource-constrained edge devices. Our method not only enables training and efficient storage of machine learning models on MCUs, but also supports incremental learning, thus continuously improving the model's adaptability to actual operating conditions without increasing the computational burden. Our

文章引用: 宋岩, 许鹏, 张岩. 在微控制器上实现在设备端训练的异常检测[J]. 嵌入式技术与智能系统, 2024, 1(2): 77-84. DOI: 10.12677/etis.2024.12009

experimental setup is a vibration source (such as a fan) with a triaxial acceleration sensor installed to simulate industrial equipment that emits vibrations during operation. The method in this paper can also be used to monitor other devices by replacing sensors and computing features.

Keywords

Microcontroller Unit (MCU), On-Device Training (ODT), Support Vector Machine (SVM), Artificial Intelligence Application, MCXN947

1. 引言

在嵌入式系统中，特别是包含电机和较多机械结构的系统中，异常检测是确保系统稳定运行的关键。异常检测用于判断被监控的系统的当前工作状态，其最基本的需求是给出系统的状态是正常还是异常，常常以检测离群点的方式实现[1]。有很多研究对异常检测的方法和应用做了介绍，以传统机器学习和深度学习为主要方法[2] [3]。如果能评估出系统“健康值”的当前结果和历史变化，就可以此为依据，来进行异常评估。这需要采集系统运行时的状态，继而使用数据分析方法来检测出可能的异常。基于机器学习的 AI 方法近年来在异常检测中脱颖而出，在行业中，NXP、ST、ADI 等头部半导体公司在异常检测方面都多有耕耘，并且提供示例、参考方案、甚至是商业化的完整产品。

在微控制器上部署基于机器学习的算法大多需要进行离线训练，即在高性能计算平台(如个人 PC、服务器或云)上预先完成机器学习模型的训练，随后将训练好的模型部署至目标设备(如微控制器)。这种工作流程对于深度学习尤其常见，因为微控制器计算力与存储有限，难以支撑模型训练所需，而且，直接训练会大幅增加功耗。

尽管离线训练在微控制器机器学习中占据主导地位，但在设备端进行训练也有其独特的优势，尤其是在应用类别五花八门而工作条件变化多端的工业异常检测领域，在这种场景下，设备端训练使模型能够根据实时收集的数据动态调整，提高对新异常的检测能力。这对于环境变化频繁或未知异常类型的情况尤为重要。除此以外，每台设备的工作条件和环境可能不同，设备端训练能够针对特定设备进行模型优化，提高异常检测的精确度和针对性。

为了实现设备端的训练和推理，软件框架和示例具备以下特点：

1) 少样本训练。最好是 100 笔以内的样本即可训练，以满足 MCU 上的内存限制。

2) 自动标注或无需标注。传统的监督学习需要标注，但是嵌入式系统一般不包含用于标注的工具。

3) 小巧精悍和易于解释的 ML 算法，如 SVM。

4) 轻量级的训练引擎。传统的训练引擎需要的代码量大，依赖动态内存和文件系统，使用容量大的数据类型，往往需要避免使用这些大的数据类型或者有必要对其进行优化。

5) 简练而高效的数据处理。包括利用传感器的 FIFO，通用和易于实现的特征提取算法，使用环形缓冲存储预处理后的特征而非原始数据等。

2. 硬件装置和整体框图

基于上述的目标和约束，我们选择使用单类支持向量机(One-Class SVM, OCSVM) [4] [5]机器学习算法，

并且深度重构和优化 Libsvm 库而得到对 MCU 友好的 Libsvmcu。并且使用可调转速的风扇为装置，来模拟在工作期间可能出现多种正常的振动模式的工业设备。我们搭建了异常检测装置用于验证，如图 1 所示。

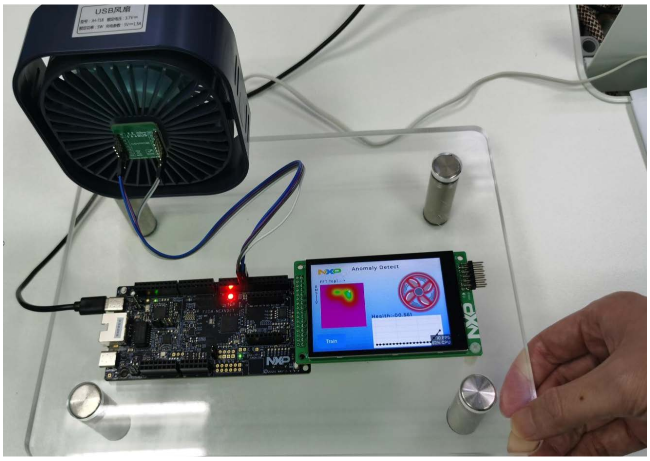


Figure 1. Anomaly detection device
图 1. 异常检测装置

上图中扮演工业装备的是一个有 3 个转速档位可调节的 USB 风扇。在扇叶罩外表面贴装了小型三轴加速度传感器，它使用 I2C 接口和 MCU 主板通信。MCU 主板是 FRDM-MCXN947，它上面包含 NXP 的 MCXN947 微控制器，并且扩展了 3.5 英寸 LCD 屏幕作为图形用户界面，还使用一个三色 LED 来显示当前是正常还是异常状态。整个系统的结构框图如图 2 所示。

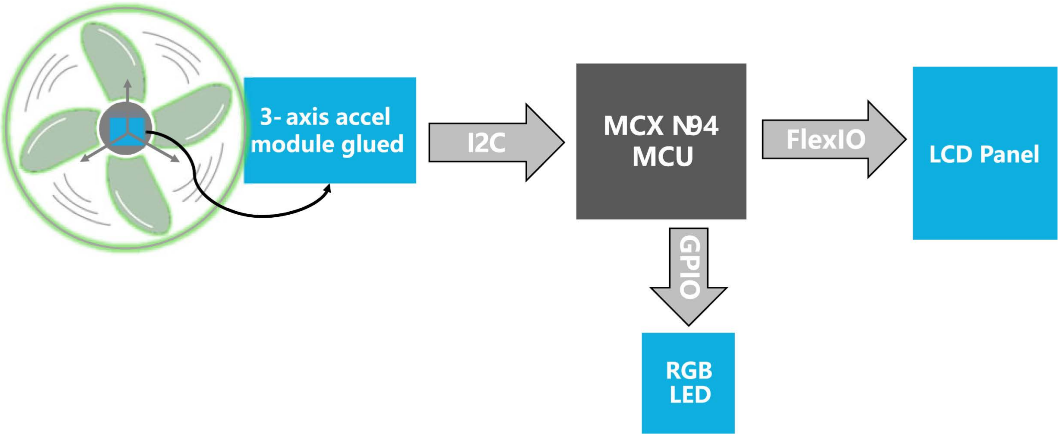


Figure 2. Block diagram of anomaly detection
图 2. 异常检测框架图

3. 训练和推理框架-Libsvmcu

为了实现基于支持向量机(SVM)的在设备上训练和预测，我们基于著名的开源支持向量机库 Libsvm [5]，进行多项重构和定制化，实现了 Libsvmcu 库。Libsvm 在 MATLAB，sckit-learn 等众多知名框架中得到了广泛应用，但是部署到 MCU 上还相对有些重。我们改版的 Libsvmcu 和原版的 Libsvm 相比做了以下修改。

1) 去掉对堆内存的依赖。原版 Libsvm 中的所有动态生长的数据结构、软件缓存、以及临时不定长

DOI: 10.12677/etis.2024.12009

度的 buffer 都使用 C++的 new/delete 或 malloc/free。我们通过大量调整分配和释放的顺序以符合后进先出的方式，去掉了对堆内存的依赖，使它们可以存储在单块静态预分配的内存中。这对于可靠性、确定性、平均性能都有明显的改进。

2) 实现支持就地执行(XIP)的模型序列化机制。在确保所有动态产生的数据块都位于单块内存的基础上，我们适当调整 Libsvm 中模型对象类中的字段顺序并添加辅助字段，把所有指针和二级指针字段都收集在一起。以此为基础，实现了对模型对象的重定位，使得位于内存中的模型对象可以直接序列化到微控制器内置的 99999++Flash 存储器中，并且支持在 Flash 上直接就地使用(XIP)，无需再加载到 RAM 中。这种技术甚至支持把无需使用封装格式打包的模型对象通过网络传输到另外的机器上，再经过简单的重定位后即可直接在 RAM 或烧写到 Flash 的地址中使用。有了这套基于重定位的序列化/反序列化技术，我们直接移除了 Libsvm 中原先使用的基于文件 I/O 的模型存储和加载机制。

3) 支持稠密向量。Libsvm 原先只支持稀疏化存储的训练样本的标量基本元素，以便支持高维数据的稀疏性，例如对于 100 维样本中只有几个非零元素的情况。然而，这样的代价是需要占用额外 4 字节保存各标量基本元素的位置。这对于在嵌入式系统中更常出现的低维且各维数据都有意义的情况却是没有必要的额外开销。我们添加了对常规数据的支持，减少了一半的存储空间来保存训练样本和学习出来的支持向量。

4) 减少核矩阵缓存的内存开销。Libsvm 使用双向链表实现一个基于最近最少使用(LRU)替换策略的核矩阵缓存(Kernel Cache)，这个核矩阵是各训练向量在经过核函数高维映射后得到的在相应的再生核希尔伯特空间中的距离矩阵，缓存线的容量和数量都是训练向量的个数。原来的 LRU 缓存使用标准的双向链表来存储缓存线的头部并且动态申请和释放内存。我们根据在 MCU 上 on-device training 时大部分情况不超过 256 个训练样本的特点，以及结合前文提到的栈式内存管理机制，改造成基于 8 位索引和一次性分配缓存最大容量的策略。

5) 支持 32 位和 16 位浮点数。Libsvm 原先只支持 64 位双精度浮点数。我们把它改造成支持 32 位浮点数，并且在小范围使用 16 位浮点类型。Libsvm 采用 C++实现，C++的运算符重载对于我们实现 16 位浮点类型提供了极大的语法帮助，使 Libsvm 主体代码几乎无需改动。

经过我们裁剪优化后的 libsvmcu 运行库，实现 One-Class SVM 仅需要十余 KB 程序存储器的空间，并且一般训练时间少于 1 秒。

4. 为应用 SVM 算法的特征设计

SVM 是传统机器学习方法，需要基于原始信号合理设计特征方可使用。在我们的装置中，是模仿很多工业设备在工作期间会产生的机械振动。我们利用三轴加速度传感器采集数据，通过深入分析加速度信号，提取了根均方值(RMS)和傅立叶变换后辐值最高的频率所对应的位置(FFTTop1)作为关键特征。这些特征不仅能够有效反映设备运行状态，还具备较强的鲁棒性和计算效率。

为了计算上述的(RMS, FFTTop1)特征，我们先收集一定长度的原始传感器读数，当前的设定是 256 笔(a_x, a_y, a_z)信号组成的形如(3, 256)的原始信号，其中 3 表示 3 个方向轴。然后，对每轴分别计算 RMS 和 FFTTop1，最后把各轴的 RMS 和 FFTTop1 取平均。

在计算 RMS 时，为了抵消掉重力加速度的影响，我们对各个轴上的数据做了滑动平均处理。

这里我们想强调，这种双特征的设计方法只是众多特征设计中的一种，经过实验观察发现对于我们的硬件装置还是很好用的。但是，SVM 算法可以支持远远更高维的特征空间，用户如果使用不同的传感器，或者如果发现默认的这两个特征不足以表达更复杂的数据模型，还需要具体问题具体分析地设计相应的特征。在包括 SVM 的经典机器学习算法中，特征工程也是非常重要的环节。

5. 应用层软件架构设计

在实现完整的在设备端训练的异常检测应用时，我们使用了 FreeRTOS 来简化应用逻辑的划分和调度。整体来看，使用了 3 个任务。下图 3 展示了任务的划分与关系。

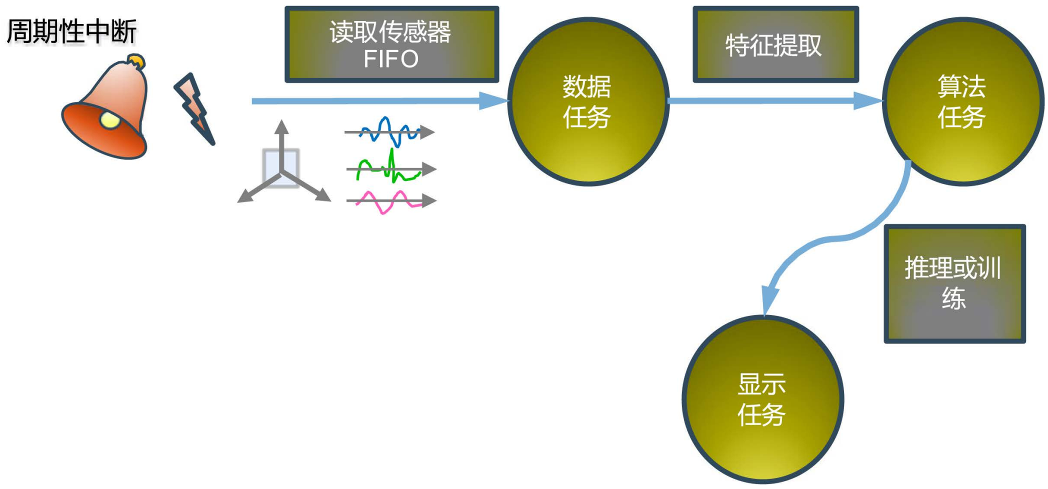


Figure 3. Software block diagram
图 3. 软件框图

数据任务：在它的主循环中，每次在循环开始时获取任务开始的时钟计数。接着，检查传感器的 FIFO 缓冲区数据量，计算要读取的数据量，并尝试从传感器读取数据存入临时缓冲区。对于读取到的数据，遍历不同维度进行处理，累加到当前均值变量并存储到另一个缓冲区，同时记录已填充的数据量。当该数据量达到特定长度(满足一个 FFT 窗口的长度)时，进行一系列数据处理操作，包括计算每个维度的均值并更新直流偏移指数移动平均，对数据进行减去直流偏移后计算均方根值(RMS)。接下来，对这段信号应用汉宁窗，然后执行快速傅里叶变换，找出幅值最大的频率所在的位置(FFTTop1)。处理后的 RMS 和 FFTTop1 就是提取的特征，并存入特征数组，并通过 FreeRTOS 的队列发送出去给算法任务。之后，对数据进行滑动处理，将数据向左移动一个固定的长度并更新相关变量。最后，获取任务结束时钟计数，根据任务起始执行时间与指定时间的比较结果调整任务延迟时间，确保数据任务的执行以固定的节奏进行。数据任务的周期必须短于传感器的 FIFO 装满所需的时间。

算法任务：在算法任务的主循环里，其功能涉及数据收集、模型训练和基于模型的预测及状态控制等多个方面。

首先，代码处于一个无限循环中，不断地从一个队列中接收数据。每次循环开始，通过 FreeRTOS 的队列 API 函数从来自数据任务的特征队列中获取特征数据并保存，然后设置标志表示有新的特征数据可用。

接下来，根据应用的不同状态进行不同的操作。当处于收集状态时，会逐步将接收到的特征数据存入一个特征数组中。如果数据积累到一定数量，会进行模型训练。倘若开启了增量训练，则首先获取已有模型中的支持向量，然后使用新的数据和已有的支持向量重新训练模型。训练完成后，更新应用状态为等待返回。

当处于预测状态且有训练好的模型时，首先获取当前风扇状态。然后通过调用模型预测函数使用模型，对接收的特征数据进行预测，得到预测结果。根据预测结果更新系统的当前健康值，并根据健康值与特定阈值的比较来判断风扇所处的状态。如果风扇状态发生变化，还会控制红绿指示灯的亮灭，以直

观地显示风扇整体是处于正常(绿灯)还是异常(红灯)。

数据任务和算法任务的配合流程如同生产者和消费者，它们的整体工作步骤如下图 4 所示。

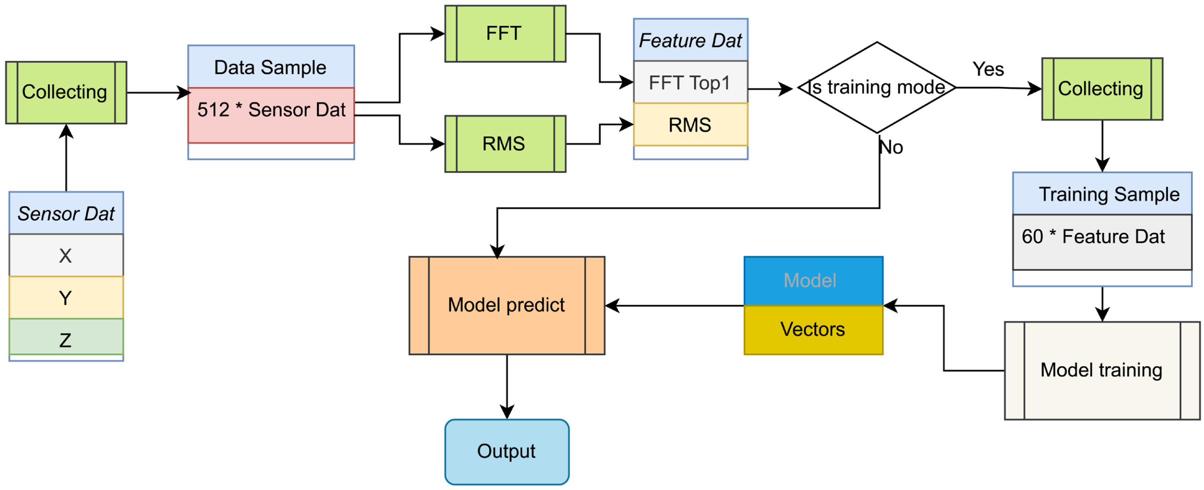


Figure 4. Workflow regarding to data task and algo task
图 4. 数据任务与算法任务的工作流

显示任务：显示任务的功能比较繁杂，它实现了一个基于 LVGL 的 GUI 应用。主要包含了检测界面、训练界面，以及参数设置界面。显示任务定期执行，它直接通过共享变量读取当前最新的状态和数据，并与其它任务没有通信。显示任务的功能不再展开叙述，唯有一个注意事项是它的优先级必须低于其它应用任务，避免它导致数据任务不能及时得到调度执行，这可能会丢失传感器数据。在我们的调试期间这曾经是一个晦涩的 bug。

6. 更多讨论

6.1. 增量学习

在很多系统中都包含多个正常的子状态。例如，我们的装置中包含了一个可以调速的风扇，在风扇处于关机、1 档、2 档、3 档的状态下，都属于正常。于是，我们为了考虑这种应用的实际需要，也支持一种增量训练的机制。

在每次训练时，如果系统中已经包含了之前训练得到的支持向量，我们会把它们读取出来和新获得的样本一起重新训练，并产生一个新的支持向量的集合，并且用新模型替换旧模型。这样可以支持多个正常子状态。

6.2. One-class SVM 中的超参数调节

在本文使用的 one-class SVM 用于异常检测时，有两个对结果影响非常大的超参数，Gamma 和 Nu。Gamma (γ)是核函数的一个参数，它决定了数据映射到高维空间后的分布情况。在径向基函数(RBF)中，较大的 Gamma 值意味着较小的决策边界，使模型更关注训练数据的局部特征；而较小的 Gamma 值意味着较大的决策边界，从而产生一个更平滑的模型，对训练数据中的局部波动不太敏感。Nu(ν)是一个用户定义的参数，表示错误数据点比例的上限和边界的下限。这个参数有助于控制支持向量的比例以及决策边界的宽松度。简单来说，较小的 Nu 值使模型更倾向于忽略更多的异常值，导致决策边界更宽；相反，较大的 Nu 值使模型更倾向于包含更多的数据点，导致决策边界更窄。

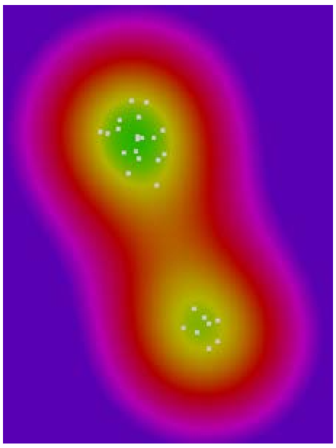
在大多数演示设置中，默认值(Gamma: 50, Nu: 0.1)通常在敏感性和容忍度之间实现合理的平衡。

Nu 值较大时，模型对数据变化更敏感，但非常大的 Nu 值(如≥0.4)可能会使模型将一些远离平均值的训练样本视为异常。因此，如果想要高敏感性和快速响应的演示，可以使用较大的 Nu 值(如 0.1 到 0.4)；但如果希望演示对随机或意外干扰更具容忍度，则应使用较小的 Nu 值(如 0.03 到 0.1)。

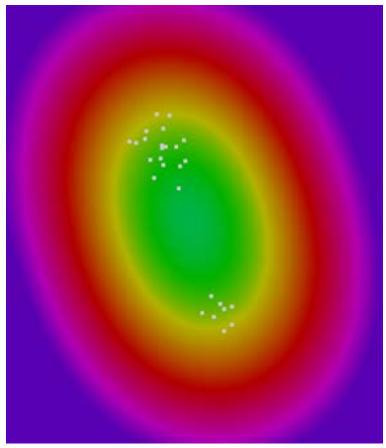
Gamma 值较大时，模型倾向于将训练数据处理为多个聚类(每个训练样本的有效范围变小)。所以，如果正常状态包含多个子状态(例如风扇关闭和风扇开启都是正常状态)，那么应该使用较大的 Gamma (如 20 到 200)；但较大的 Gamma 值也会使模型对训练数据中的随机性不太稳健，因此，如果演示环境受到严重干扰(如正在工作的计算机的振动、附近风扇引起的空气振动、附近走动的人)，则应使用较小的 Gamma 值(如 5 到 20)。

为了便于演示与调节超参数，我们在液晶显示屏上对模型的决策边界进行二维彩色等高线可视化，横轴表示 FFTTOP1 幅值所对应的频率，纵轴表示 RMS。等高线中，绿色-蓝色的颜色区域是正常区域，而黄色、红色和紫色的颜色区域是异常区域。表 1 显示了超参数对异常决策边界的重大影响。

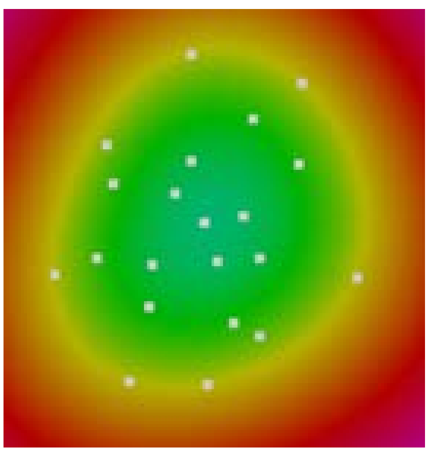
Table 1. Hyperparameter tuning (Gamma and Nu)
表 1. 超参数的整定(Gamma 和 Nu)



Gamma = 50, Nu = 0.1
Optimal Gamma and Nu that makes the model can detect 2 sub normal states.

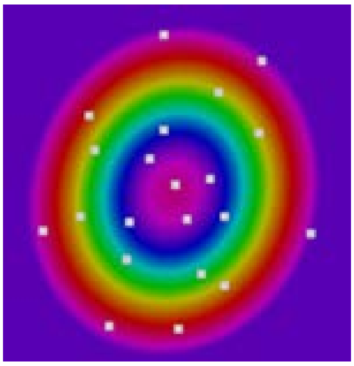


Gamma = 10, Nu = 0.1
Gamma is so small that the decision boundary is too smooth to distinguish the two sub normal states.

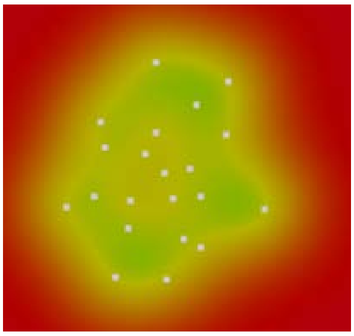


Gamma = 50, Nu = 0.1
Model can tolerate some random noises.

续表



Gamma = 50, Nu = 0.4
Nu is too large that the model output big range of numbers, and wrongly treats some points that are far from center as abnormal.



Gamma = 300, Nu = 0.1
Gamma is too large that the model clusters too much, makes the decision boundary is not smooth and tends to overfit.

7. 结语

在微控制器上部署异常检测的常见方法是使用有监督机器学习在 PC 上预训练模型。然而，一些异常检测的问题受个体差异影响大，或者难以获得异常情况的样本。为了解决这一问题，本文介绍了一种基于单类支持向量机(SVM)算法的设备上训练的异常检测方法。该方法利用了 SVM 算法在小样本和高维数据上的优越性能，通过深度重构 Libsvm 库，如内存管理和模型序列化等，得到最终的 MCU 友好的 Libsvmcu 库，并开发了一套软件应用框架在设备上进行训练，实现了对风扇异常的有效检测。同时，该方法还支持增量学习，使得模型能够随着时间的推移不断更新，以适应嵌入式系统的变化。

本文中配合使用的示例已开源：https://github.com/nxp-appcodehub/dm-on-device-training-fan-anomaly-on-mcxn947。

参考文献

[1] Chandola, V., Banerjee, A. and Kumar, V. (year) Anomaly Detection: A Survey. *ACM Computing Surveys*, **41**, 1-15.

[2] Pimentel, M.A.F., Clifton, D.A., Clifton, L. and Tarassenko, L. (2014) A Review of Novelty Detection. *Signal Processing*, **99**, 215-249. https://doi.org/10.1016/j.sigpro.2013.12.026

[3] Hodge, V.J. and Austin, J. (2004) A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, **22**, 85-126.

[4] Chang, C.-C. and Lin, C.-J. (2024) LIBSVM—A Library for Support Vector Machines. https://www.csie.ntu.edu.tw/~cjlin/libsvm/

[5] Tax, D.M.J. (2001) One-Class Classification. Doctoral Thesis, Delft University of Technology.

Embedded Technology and Intelligent Systems 嵌入式技术与智能系统, 2024, 1(2), 85-91
Published Online October 2024 in Hans. https://www.hanspub.org/journal/etis
https://doi.org/10.12677/etis.2024.12010

Hans汉斯

巷道掘进中孔中地震高精度预报系统

陈家烨[1]**，陈杰炜**[2*]

[1]哈尔滨工业大学机电工程学院，黑龙江 哈尔滨
[2]福建平潭旭坤实业有限公司，福建 福州

收稿日期：2024年8月15日；录用日期：2024年11月21日；发布日期：2024年11月29日

摘　要

巷道掘进中孔中地震高精度预报系统是完全自主开发的高性能产品。该预报系统主要是由“井下”和“地面”两大部分组成的。其中，井下部分主要是由1个无线主机、3个无线探头、1个无线触发器、1个震源铜锤、1根触发信号线以及其它配件(如：蜂鸣器、锤垫等)组成。主要功能是进行现场数据采集和存储，如果无线主机安装有分析软件，就可在现场解析出探测结果。井下设备都是本质安全型设计，并且通过了国家煤矿安全机构的防爆性能检测和安全认证。地面部分主要是由PC机、仪器电源适配器(充电器)和分析软件组成的，其主要功能是对所采集的地质数据进行转储、深度解析、分析处理和形成成果报告文件，亦即预报结果。该系统与同类产品相比精度高、准确率高和施工方便的优势。

关键词

高精度预报系统，无线探头，信号处理，孔中地震勘探，MEMS检波器

High Precision Earthquake Prediction System in Roadway Excavation

Jiaye Chen[1]**, Jiewei Chen**[2*]

[1]School of Mechanical and Electrical Engineering, Harbin Institute of Technology, Harbin Heilongjiang
[2]Fujian Pingtan Xukun Industrial Co., Ltd., Fuzhou Fujian

Received: Aug. 15th, 2024; accepted: Nov. 21st, 2024; published: Nov. 29th, 2024

Abstract

The high-precision tunnel seismic forecast system for boreholes in roadway tunneling is a completely self-developed, high-performance product. This forecasting system comprises two main components: the “underground” section and the “surface” section. The underground section includes one wireless

*通讯作者。

文章引用: 陈家烨, 陈杰炜. 巷道掘进中孔中地震高精度预报系统[J]. 嵌入式技术与智能系统, 2024, 1(2): 85-91.
DOI: 10.12677/etis.2024.12010

host, three wireless probes, one wireless trigger, one seismic source copper hammer, one trigger signal line, and additional accessories such as buzzers and hammer pads. Its primary function is to collect and store on-site data. If the wireless host is equipped with analysis software, it can also process and display detection results on-site. The underground equipment is designed with intrinsic safety and has passed explosion-proof performance testing and safety certification by national coal mine safety institutions. The surface section consists of a PC, an instrument power adapter (charger), and analysis software. Its main function is to transfer, deeply analyze, and process the collected geological data, ultimately generating result reports, *i.e.*, forecast results. Compared to similar products, this system offers higher precision, greater accuracy, and ease of use.

Keywords

High-Precision Forecast System, Wireless Probe, Signal Processing, Borehole Seismic Exploration, MEMS Detector

1. 引言

目前，相应的在国内外使用的巷道超前预报预测设备主要有：国外的有瑞典生产的 TSP203 探测系统[1]、美国生产的 TRT7000 探测系统[2]，俄罗斯生产的 TGS-360Pro 探测系统[3]等；国内的有北京同度工程物探技术有限公司生产的 TST 探测系统[4]、北京市水电物探研究所生产的 TGP206 探测系统[5]和云南航天工程物探检测股份有限公司生产的 AGI-T3 探测系统[6]，以及还有在煤矿井下巷道掘进前方作预报预测的 MSP 探测设备[7] [8]。这些设备在使用过程中预报预测的准确性和可靠性还有待提高，主要存在的问题如下：在现有设备中，基本采用传感部件与采集部件分开设计，也有把传感部件安装于钻孔中，例如，KDZ1114-6B30 矿井巷道地质探测仪中的孔中传感器是由三个正交的单分量速度传感器安装在钢管里制作而成[9]。单分量速度传感器选择动圈传感器，频率响应不超过 400 Hz，响应频带窄，工作时，适应于竖直安装，而置入孔中时，动圈传感器并不是竖直安装，这大大降低了传感器灵敏度，有效的微弱信号无法感知。另外，施工时孔中传感器置于预先打好的钻孔中，利用打气装置对其打气把孔中传感器压紧在孔壁，采集装置放在孔外使用信号线缆连接孔中的传感装置，施工极其不便，而且有效信号损失比较严重。所以，现有巷道地质超前探测设备在使用过程中无法接收极微弱信号，对断层、陷落柱、老空等较大构造的反射信号反应较为明显，对较小断层、软弱结构、溶洞和含水层反映不明显，特别在电磁波干扰较大、地形复杂的环境中，其探测精度和准确率更是无从谈起。

2. 预报系统组成和功能

2.1. 组成和功能

巷道掘进中孔中地震高精度预报系统是由“井下”和“地面”两大部分组成的。其中，井下部分主要是由 1 个无线主机、3 个无线探头、1 个无线触发器、1 个震源铜锤、1 个锤垫、1 根触发信号线以及其它配件(如：蜂鸣器等)组成。主要功能是进行现场数据采集和存储，井下设备都是本质安全型设计，并且通过了国家煤矿安全机构的防爆性能检测和安全认证。地面部分主要是由 PC 机、仪器电源适配器(充电器)和分析软件组成的，其主要功能是对所采集的地质数据进行转储、深度解析、分析处理和形成成果报告

文件，亦即预报结果。预报系统组成设计如图 1 所示。

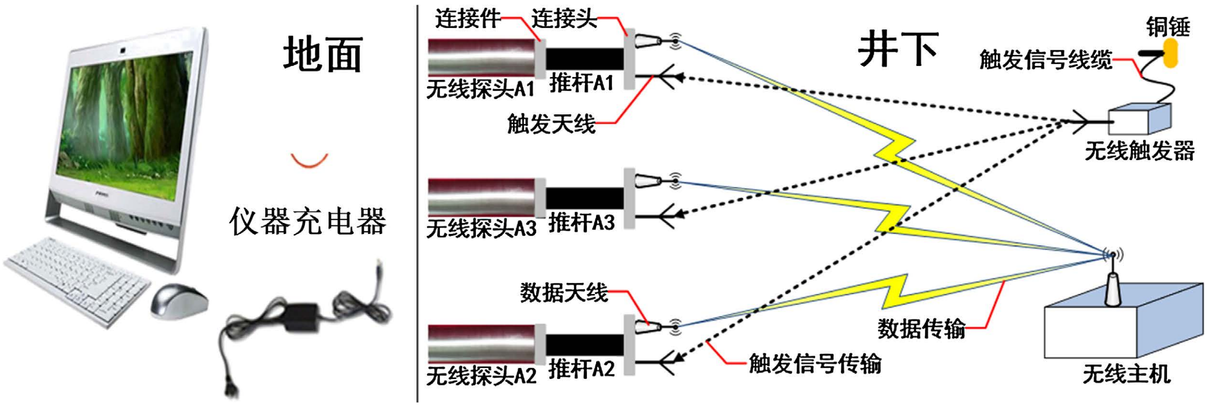


Figure 1. Schematic diagram of the high-precision borehole seismic prediction system for tunnel excavation
图 1. 巷道掘进中孔中地震高精度预报系统组成示意图

2.2. 预报系统各部件功能简介

(1) 无线主机：在数据采集之前，无线主机可对 3 个无线探头进行参数设置和命令，同时，在数据采集时实时接收 3 个无线探头发送的地震数据进行存储、预处理以及各种实时显示；数据采集开始时，通知 3 个无线探头(A1、A2、A3)执行提前并行数据采集；现场数据采集后，可把数据无线传输给计算机；**(2) 无线探头：**无线探头集成了 MEMS 三分量加速度模块、数据调理处理模块、模数转换模块、微控制器模块、WIFI 通讯模块、无线触发接收模块、电源模块(包括充电电路)、电池以及天线等。当现场数据采集时，3 个无线探头接收到无线主机的触发命令时，同时同步进行数据采集并把采集的地震数据传送给无线主机；**(3) 无线触发器：**通过触发信号线连接安装在震源铜锤上的蜂鸣器，把震源铜锤激发蜂鸣片产生的触发信号发送给无线主机。无线触发器有 3 种触发方式：先短后断触发、先断后短触发和外部信号触发；**(4) 震源铜锤：**采用 8 磅铜锤，激发产生地震波和触发信号为地震数据采集提供信号；**(5) 锤垫：**供震源铜锤激发时使用；**(6) 触发信号线：**连接蜂鸣器与无线触发器；**(7) 蜂鸣器：**产生触发信号。

2.3. 预报系统的特点

巷道掘进中孔中地震高精度预报系统具有独特和精细的设计风格，并应用最先进的电子技术，该系统普遍采用高集成度、高精度、低噪声、低功耗、小封装器件，在硬件、软件和设备组成上加强抗干扰措施，同时，极大限度地简化常规地震数据采集中所有的模拟电路和数字电路，这样，既可以减小硬件电路结构体积，又能避免因模拟器件的阀门效应而造成有效信号不可挽回性丢失。其特点如下：

(1) 采用专用模拟－数字抗干扰设计，实现强干扰中微弱有效信号的采集；
(2) 锤震或炮震探测施工方式可选；
(3) 采用高性能三分量加速度检波器，提高探测精度及距离；
(4) 同时对反射波、绕射波进行反演分析，保证分析成果的准确度；
(5) 波场分离采用 F-K 滤波，消除噪声提高信噪比；
(6) 三维观测三维成像；
(7) 数据通讯可选择为无线或有线传输方式；
(8) 具有远程升级功能；
(9) 低通、高通、带通、陷波四种数字 DSP 滤波方式；

(10) 锤震探测距离 80 m 以上，炮震探测距离 150 m 以上。

2.4. 预报系统的技术指标

巷道掘进中孔中地震高精度预报系统技术指标如表 1 所示。

Table 1. Key technical specifications of the high-precision borehole seismic prediction system for tunnel excavation
表 1. 巷道掘进中孔中地震高精度预报系统主要技术指标

性能指标	技术参数	性能指标	技术参数
无线主机处理器	200 MHz (FPGA)	存储容量(GB)	64
无线探头处理器	120 MHz (FPGA)	存储容量(GB)	8
采集精度	24 位	触发方式	模拟信号外触发
采集长度(dot)	2^k (k = 9、10…20)	模数转换率(KSPS)	128
放大增益(倍)	1、10、100、1000	动态范围(dB)	150
通讯协议	802.11 b/g 无线网络	通讯速率(Kbps)	1000
传感器	三分量加速度	灵敏度(mV/g)	1000
连续工作时间(小时)	≥8	工作温度(℃)	−10~+50

3. 预报系统的结构设计

3.1. 无线主机结构简介

巷道掘进中孔中地震高精度预报系统无线主机组成主要包括：主板、液晶显示、面板、锰酸锂电池组、无线路由器、WIFI 模块、2.4 G-WIFI 天线、外壳、密封圈、电池盒、托架等，如图 2 所示为无线主机外观示意图。

Figure 2. Appearance diagram of the wireless host in the high-precision borehole seismic prediction system for tunnel excavation
图 2. 巷道掘进中孔中地震高精度预报系统无线主机外观示意图

3.2. 无线探头结构设计

巷道掘进中孔中地震高精度预报系统中的无线探头主要包括：MEMS 三分量加速度传感模块、信号处理采集模块、2.4 GHz-WIFI 无线数据传输模块、1.0GHz 无线模拟接收模块、电源电路、锰酸锂柱状电池组、圆柱外壳、圆柱电池盒、推杆、连接部件、天线等。无线探头结构如图 3 所示，外观图如图 4 所示。

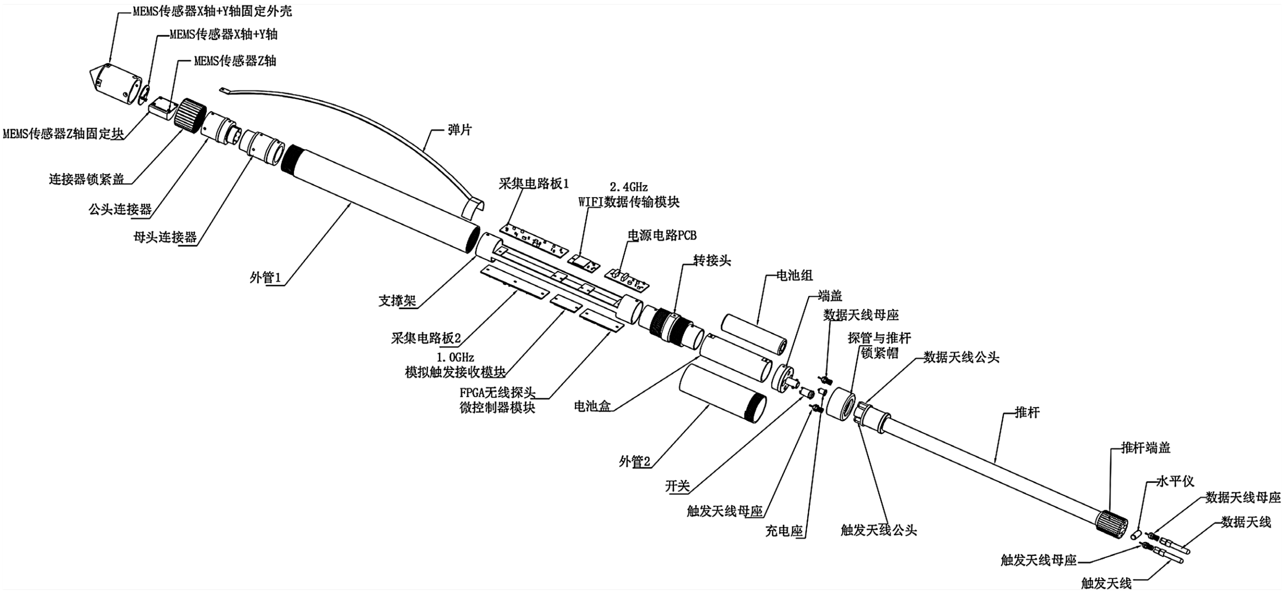


Figure 3. Structural diagram of the wireless probe in the high-precision borehole seismic prediction system for tunnel excavation
图 3. 巷道掘进中孔中地震高精度预报系统无线探头结构示意图

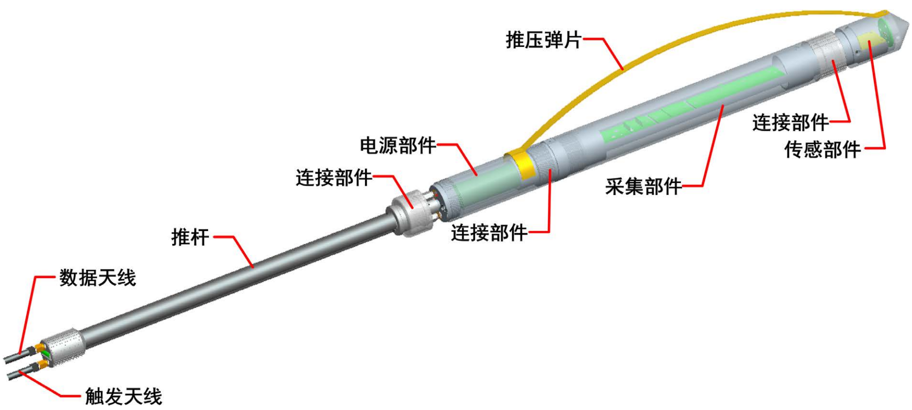


Figure 4. Appearance diagram of the wireless host in the high-precision borehole seismic prediction system for tunnel excavation
图 4. 巷道掘进中孔中地震高精度预报系统无线主机外观示意图

3.3. 预报系统硬件设计特点

巷道掘进中孔中地震高精度预报系统硬件布置主要是采集到有效的微弱信号。为了实现这个目标，本设计调研了目前市场上同类产品的使用状况，充分了解这些产品应用过程中存在的优缺点，在这基础上进行精心设计、反复试验、步步把关，克服了其它设备存在的缺陷，设计出优良的巷道掘进中孔中地震高精度预报系统。其硬件特点如下：

(1) 无线主机和无线探头分开设计，并利用无线联系，避免它们之间因有线联系而产生的互相干扰。

(2) 无线主机硬件结构简单，探测时置于巷道空旷处；无线探头包括 FPGA 微处理中央单元、高性能 MEMS 三分量加速度传感器、信号处理电路、模数转换电路、无线触发接收电路、无线 WIFI 通讯电路以及电源电路等，设计时把这些电路合理设计成模块板，各板布局得当、合理安排，使获取微弱信号能力最强，抗干扰能力最强。

(3) 无线主机和无线探头硬件设计是基于最先进的强大电子技术的基础上完成的，它采用分布式控

制、无线传输和集中处理方式，核心芯片选用美国 Altera 公司片上可编程 SOPC 技术，FPGA 控制人机对话、机机通讯以及相关的各种算法和控制，FPGA 控制信号的放大、去噪、实时采集和存储，接着把采集的地震数据传输给系统主机进行显示和存储，最后把有效的地震数据传给 PC 机由后台分析软件来处理。

4. 预报系统的软件设计

巷道掘进中孔中地震高精度预报系统分析软件基于所设计硬件作为观测系统，首先对三维波场中纵横波信号进行分离和共接收点信号编排，然后应用“F-K”二维波速滤波方法，提取保留掌子面前方的回波信号(负速度)，滤除巷道侧面及其它方向的干扰信号；接下来进行围岩速度扫描分析，确定围岩的速度分布；最后是在围岩波速的基础上，应用观测到的纵横波信号进行地质构造的偏移成像。该系统通过上述系列处理过程即可解决波场分离和速度分析问题，具有先进水平。

巷道掘进中孔中地震高精度预报系统包括嵌入式软件、初始化软件、信号调理和处理软件、模数转换控制软件、数据存取控制软件、触发信号处理软件、通讯控制软件、电源管理监控软件以及检测警示软件。

5. 预报系统的应用

巷道掘进中孔中地震高精度预报系统设计适用于探测巷道前方 150 米以内软弱带、破碎带或裂隙发育情况以及探测巷道前方断层、陷落柱、采空区或赋水情况等。预报系统应用多项创新技术，可以有效解决巷道在掘进过程中存在的安全隐患，在矿井安全生产预测预报中起着重要作用。自巷道掘进中孔中地震高精度预报系统上市以来，应用次数超过百例，预测预报准确率超过 90%，带来较高的社会效益和经济效益。

探测案例：

(1) 现场施工：本施工案例为在山西潞安集团某矿工作面巷道掘进前方进行探测，进一步了解工作面前方的地质情况。在布置中，左右两侧各布置 12 个间距 2 m 的锤击点，其中锤击点距掌子面最近 5 m 左右，锤击点距检波器 1 为 5 m 左右，锤击点距检波器 2 为 5 m 左右，检波器 1、检波器 2 在同一个平面上。

(2) 探测结果(如图 5 所示)：①在掘进头前方 24 m 附近有反射界面，推测为陷落柱局部裂隙较发育及岩层变化影响所致。实际情况在 20 m 处发生岩性变化，泥岩变砂岩；②在 60 m 附近为异常界面二，推测为陷落柱中部裂隙及岩层变化影响所致；③在 98~108 m 范围为异常带三，推测为陷落柱边界，煤岩层变化影响所致。以上结果在后期施工过程得到验证。

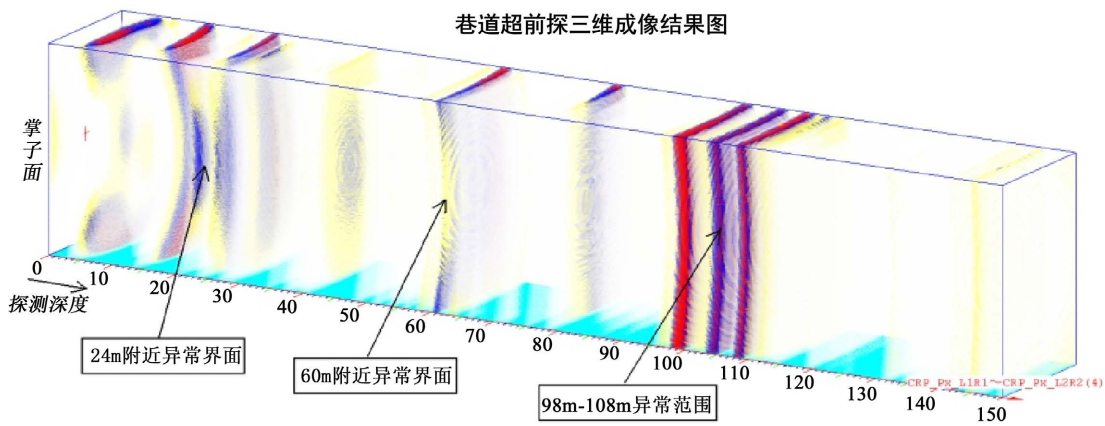


Figure 5. 3D prediction results diagram
图 5. 三维预测预报成果图

6. 结论

本文详细介绍了巷道掘进中孔中地震高精度预报系统，该系统的独特设计以及在应用过程中的验证，其有益效果在于：

(1) 采集部件与振动传感部件各自独立而集成一体，避免线缆连接易受到外部电磁波干扰，增强采集极弱信号的能力。

(2) 启动触发信号和数据均采用无线传输，没有线缆受损而影响施工的问题，数据采集完成后立即被数字化保存并进行无线传输，不受环境干扰。另外，无线主机与无线探头因采用无线传输而没有线缆连接，无线主机对数据采集产生不了干扰，简化施工复杂度，减少工作量，提高了采集极弱信号的能力和探测效率。

(3) 探测施工时，无线探头埋在钻孔中，不受外部环境影响，降低外部干扰，提高了采集数据分辨率，而且埋在孔中的传感器，与被测介质耦合得更好，更容易感知微弱的地震信号。

综上所述，该系统实现了在巷道掘进中的地震勘探，对高精度、高灵敏度以及准确率的需求，与目前同类产品相比具有更高的稳定性和探测准确性，具有广泛的应用前景。

参考文献

[1] Alimoradi, A., Moradzadeh, A., Naderi, R., Salehi, M.Z. and Etemadi, A. (2008) Prediction of Geological Hazardous Zones in Front of a Tunnel Face Using TSP-203 and Artificial Neural Networks. *Tunnelling and Underground Space Technology*, **23**, 711-717. https://doi.org/10.1016/j.tust.2008.01.001

[2] Dong, X., Zhu, J., Wang, Q., *et al*. (2023) A Technical Study of Advanced Geological Prediction of Tunnels under Complex Geological Conditions. *Academic Journal of Architecture and Geotechnical Engineering*, **5**, 52-57.

[3] 徐磊, 尹剑, 张建清, 等. TGS360Pro 技术三维正演及其与 TSP 技术对比试验研究[J]. 地球物理学进展, 2022, 37(3): 1321-1329.

[4] 谷江洪, 申康路. TST 地质超前预报技术在 N-J 水电工程中的应用[J]. 土工基础, 2018, 32(2): 224-228.

[5] 王健. TGP206 在隧道超前地质预报中的应用[J]. 石家庄铁路职业技术学院学报, 2022, 21(2): 44-47.

[6] 赵国军, 李俊杰, 江宗高, 等. AGI-T3 在输水隧洞超前地质预报中的应用[J]. 水利水电技术, 2018, 49(6): 164-170.

[7] 任云峰, 白如镜, 王晋宁. 基于 MSP 勘探法在煤矿掘进巷道地质构造超前探测应用[J]. 煤炭与化工, 2021, 44(10): 72-74.

[8] Fan, J., Yuan, Q., Chen, J., *et al*. (2024) Investigation of Surrounding Rock Stability During Proximal Coal Seams Mining Process and Feasibility of Ground Control Technology. *Process Safety and Environmental Protection*, **186**, 1447-1459.

[9] 李小雷. 矿井巷道震波超前探测技术在煤矿的应用研究[J]. 煤矿现代化, 2012(3): 34-35.

嵌入式技术与智能系统

主编：何立民　北京航空航天大学教授

主办：汉斯出版社　珠海吴谷电子科技有限公司

编辑：《嵌入式技术与智能系统》编委会

网址：https://www.hanspub.org/journal/etis

电子邮箱：etis@hanspub.org

出版：汉斯出版社