

# 计算科学评论

环球科学

WWW.COMPUTER.ORG

2018年第2期

## 移动 嵌入式 深度学习

真实世界的计算是最大挑战

孩子的智能玩具安全吗?

P12

P80

合作机构



ISSN 1673-5153



9 771673 515122

# 云计算

人工智能

工控机

制造业

电子  
嵌入计算

传感器

互联网

3D 打印

绿色计算

图形图像

虚拟现实

市场 行业分析 职场 学习 科技新闻 竞赛  
**创业**

找工  
教育

# 互联网金融

芯片

## 大数据

纳米架构

软件架构

# 人机交互

多媒体

# 普适计算 MEMS

合作  
工作 研究综述  
申请  
进修

微信名：计算人 微信号：jisuanren



# Copyright

# 版权

## 主管单位 Authorite in Charge

中华人民共和国教育部 Ministry of Education of the People's Republic of China

## 主办单位 Sponso

中国大学出版社协会 China University Presses Association

## 出版单位 Publisher

《环球科学》杂志社有限公司 GLOBAL SCIENCE MAGAZINES Co.,Ltd

社址 Address: 北京市朝阳区秀水街1号建外外交公寓4-1-21 Office 4-1-21, Jianguomen Diplomatic Residence Compound, No. 1, Xiu Shui Street, Chaoyang District, Beijing, China. 邮编: 100600

联系电话: 010-85325810 / 85325871

## 社长 / 总编辑 Editor-in-chief

陈宗周 Chen Zongzhou

## 副社长 / 副总编辑 Deputy Editor-in-chief

刘芳 Liu Fang

## 执行出版人 Publisher

管心宇 Xinyu Guan  
张岚 Landy Zhang

## 资深编辑 Senior Editor

马法达 Falda Ma  
刘妍 Yan Liu

## 特约编辑 Contributing Editor

史彦诚 Yancheng Shi  
刘大明 Daming Liu  
高天羽 Tianyu Gao  
费麟 Yong Fei  
王璇 Xuan Wang

## 运营中心 OPERATING DEPARTMENT

运营机构 Publisher  
上海灵宸文化传媒有限公司

## 发行部 Circulation Department

发行总监 Circulation Director  
谢磊 Xie Lei 010 - 57439192

## 市场部 Marketing Department

市场总监 Marketing Director  
赵子豪 Zhao Zihao 010 - 85325810 - 807

## 广告部 Advertising Department

销售总监 Sales Director

范欢 FanHuan 010-85325871-802 010-85325981

## 读者服务部 Reader Service

杜君 Du Jun 010 - 57458982

## 印刷 北京博海升彩色印刷有限公司

如发现本刊缺页、装订错误和损坏等质量问题, 请在当月与本刊读者服务部联系调换(请将证书寄回)。

国际标准刊号: ISSN 1673-5153

国内统一刊号: CN11-5480/N

广告经营许可证号: 京朝工商广字第8144号

## 知识产权声明:

IEEE, IEEE Computer, IEEE中文网站的名称和标识, 属于位于美国纽约的电气电子工程师学会有限责任公司所有的商标, 仅通过授权使用。这些材料的一部分由IEEE Computer英文版翻译而来, 版权归IEEE所有, 并经IEEE授权翻译复制。

IEEE Computer杂志的中文版权归, 由美国电气电子工程师学会有限责任公司授予上海灵宸文化传媒有限公司, 并由本刊独家使用。

本刊发表的所有文章内容由作者负责, 并不代表上海灵宸文化传媒有限公司、美国电气电子工程师学会有限责任公司的立场。

本刊内容未经书面许可, 不得以任何形式转载或使用。

## 编辑团队

### 流程编辑

Carrie Clark

colark@computer.org

### 资深编辑

Chris Nelson

### 编辑

Lee Garber, Meghan O'Dell

Rebecca Torres, Bonnie Wylie

### 多媒体编辑

Rebecca Torres

### 设计与印刷

Carmen Flores-Garvey

Erica Hardison

### 封面设计

Matthew Cooper

### 资深广告经理

Debbie Sims

### 产品与服务总监

Evan Butterfield

### 会员总监

Eric Berkowitz

### 出版人

Robin Baldwin

## 主编

Sumi Helal

Lancaster University,  
sumi.helal@computer.org

## 副主编

Elisa Bertino

Purdue University,  
bertino@cs.purdue.edu

## 副主席, COMPUTING PRACTICES

Rohit Kapur

Synopsys, kapurfamily04@gmail.com

## 副主席, PERSPECTIVES

Jean-Marc Jézéquel

University of Rennes jean-marc.jezequel@irisa.fr

## 副主席, SPECIAL ISSUES

George K. Thiruvathukal

Loyola University Chicago,  
gkt@cs.luc.edu

## 2018 IEEE计算机协会主席

Hironori Katashara

Waseda University,  
kasahara@waseda.jp

## 行业编辑

### 大数据和数据分析

Naren Ramakrishnan

Virginia Tech

Ravi Kumar

Google

### 云计算

Schahram Dustdar

TU Wien

### 计算机结构

David H. Albonesi

Cornell University

Greg Byrd

North Carolina State University

Erik DeBenedictis

Sandia National Laboratories

### 信息物理系统

Oleg Sokolsky

University of Pennsylvania

### 数字健康

Christopher Nugent

Ulster University

## 顾问委员会

Doris L. Carver

Louisiana State University (EIC Emeritus)

Carl K. Chang

Iowa State University (EIC Emeritus)

Theresa-Marie Ryne

Consultant

Bill Schilit

Google

Savitha Srinivasan

IBM Almaden Research Center

Ron Vetter

University of North Carolina Wilmington (EIC Emeritus)

Alf Weaver

University of Virginia

# 领研网

专注科研招聘与学术分享

[www.linkresearcher.com](http://www.linkresearcher.com)



访问领研网获取招聘信息、一手科研资讯，动态追踪学者研究成果



领研网是《科学美国人》中文版《环球科学》旗下科研招聘与学术分享网站，  
服务百万学者，为高校、机构与科技企业搭建人才桥梁，  
助力学者传播优秀成果，提升学术生涯。

合作请致电: 010 - 85321181, 或邮件 [contact@linkresearcher.com](mailto:contact@linkresearcher.com)



科研求职者可扫描二维码  
注册，即可在站内信获得  
价值千元就业礼包



科研机构 / 科技企业可扫描二维码  
成功注册可免费发布职位信息

# 计算科学评论

06

## 导读

### 移动系统和嵌入式系统的 深度学习飞跃

移动和嵌入式设备越来越依赖深度神经网络来了解这个世界——哪怕是仅几年前，这么做就会消耗这些设备所有的系统资源。机器学习和嵌入式 / 移动系统的进一步整合将需要在高效学习算法领域产生新的突破，这些算法可以利用波动且有限的资源运行，从而形成跨越计算机体系结构、软件系统和人工智能的领域。

Nicholas D. Lane、Pete Warden，牛津大学、谷歌

2018年第2期

特刊

12

26

36

### 利用典型值加速深 度学习

撰文 Andreas Moshovos、  
Jorge Albericio、Patrick  
Judd、Alberto Delmás  
Lascorz、Sayeh Sharify、  
Zisis Poulos、Tayler  
Hetherington、Tor  
Aamodt、Natalie Enright  
Jerger

### 深度学习与物联网

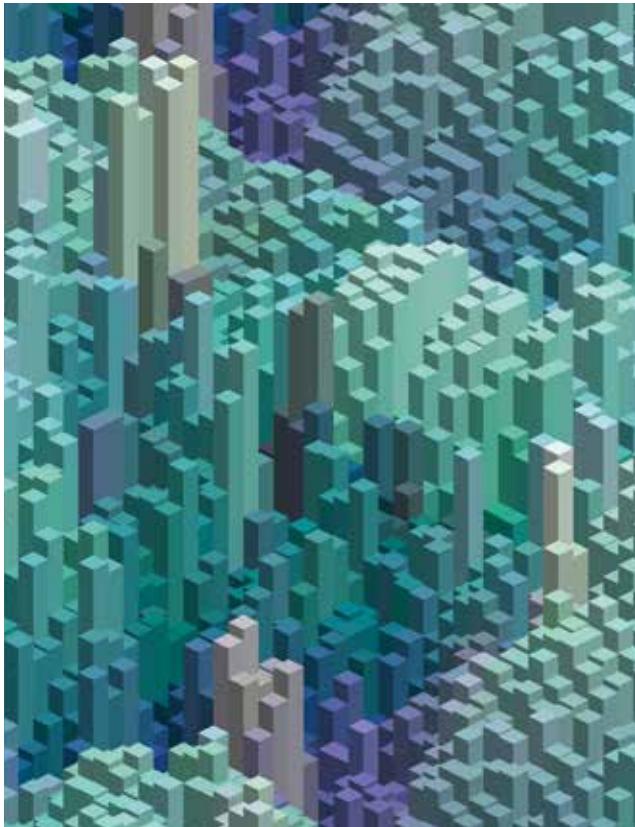
撰文 Shuochao Yao、  
Yiran Zhao、Aston  
Zhang、Shaohan Hu、  
Huajie Shao、Chao  
Zhang、Lu Su、Tarek  
Abdelzeher

### 使用混合边缘到 云深度学习进行 隐私的且可扩展 的个人数据分析

撰文 Seyed Ali Osia、  
Ali Shahin Shamsabadi、  
Ali Taheri、Hamid R.  
Rabiee、Hamed Haddadi

2018 年第 2 期

## 目录



### 教育

- 76 真实世界的计算是最大挑战

撰文 Marilyn Wolf

### 网络安全

- 78 孩子身边的网络威胁

撰文 Nir Kshetri、Jeffrey Voas

46

56

66

基于深度学习的人类行为识别在移动计算领域的使用

撰文 Thomas Plotz、Yu Guan

利用递归神经网络对资源受限的物联网设备进行基于呼吸的身份验证

撰文 Jagmohan Chauhan、Suranga Seneviratne、Yining Hu、Aruna Seneviratne、Aruna Seneviratne、Youngki Lee

找出小肠损伤：构建基于内镜成像的学习系统时遇到的挑战

撰文 JUNGMO AHN、HUYNH NGUYEN LOC、RAJESH KRISHNA BALAN、YOUNGKI LEE、JEONGGIL KO、



# 移动系统和嵌入式系统的深度学习飞跃

文 | Nicholas D. Lane, 牛津大学  
Pete Warden, 谷歌

移动和嵌入式设备越来越依赖深度神经网络来了解这个世界——哪怕是仅仅几年前，这么做就会消耗这些设备所有的系统资源。机器学习和嵌入式／移动系统的进一步整合将需要在高效学习算法领域产生新的突破，这些算法可以利用波动且有限的资源运行，从而形成跨越计算机体系结构、软件系统和人工智能的领域。

## 在

过去3年时间里，深度学习技术悄悄地改变了移动设备和嵌入式设备解读世界和进行交互的能力。<sup>1</sup>智能手机和智能手表在进行判别性任务（例如人脸识别、文字和物体识别）的时候，只需依赖低资源、高效率的深度学习模型。这些模型之前需要云级别的资源才能运行。经过最新发明的一系列技术，现在深度神经网络模型只需要移动设备，或移动-云混合设备即可运行，它带来的准确性和鲁棒性可以用于家庭、办公室、汽车和口袋里的设备。

这些进展正逐步消除之前因为计算平台限制产生的机器学习质量损耗，让和人类水平相仿甚至更出色的关键认知和感知能力（例如机器翻译、图片理解，以及语音合成）出现在计算能力受限的设备上。这一进展让嵌入式和移动系统在消费层面上的服务质量得到

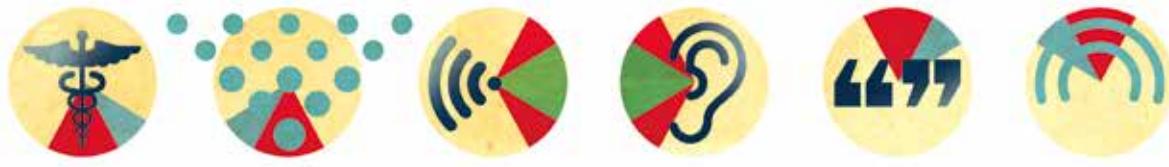
了提升。相关的例子包括微软的Seeing AI，这是一款可以看见东西并向视障患者准确描述物体的移动视觉应用程序。另一个例子是Babylon Health的Babylon，这是一款可以媲美医院非急诊咨询电话的医疗诊断和咨询数字助手。这两款移动应用程序每天都有数千人（Babylon）和数百万人（Seeing AI）使用，有效、基于移动平台的深度学习模型也应用在了多个不同的方面。

令人难以置信的是，这些巨大的变化只是一个开始。深度学习和机器学习，正从当今的主流分类和感知任务的完成者，扩展到对整个移动和嵌入式计算产生影响的角色。深度学习技术已经成为自动系统（从家庭机器人，到无人机，到汽车）控制算法的核心。更广泛地说，我们正亲眼见证即使非常成熟的系统和算法也被深度神经网络大幅增加性能的过程。例如，像B +树这样基本

的数据结构，<sup>2</sup>或是视频编解码器，网络协议，数据压缩和加密等领域。<sup>3</sup>

将机器学习无缝混合到移动/嵌入式系统的设计和操作中将持续获得动力。实际上，在这种转变过程中，深度学习在智能设备（如手机，手表和嵌入式传感器）的发展中将发挥比我们已经见证的更为关键的作用。正是通过这种学习算法的组合和对移动和嵌入式系统设计的重新思考 - 以及对使用深度神经网络以支持传感器系统的推理和推理需求的进一步研究投资 - 这将产生下一代的智能设备驱动未来雄心勃勃的传感器驱动应用。

出于这些原因，我们必须继续提升我们对如何最佳地减少、控制和塑造深层模型和算法的系统资源需求（包括能源，计算和内存）的知识。这是一个基础构建模块，如果没有进展，移动和嵌入式设备的资源限制将限制我们利用



深度学习和其他机器学习技术的能力。尽管在这类计算中出现深度学习越来越普遍，但将全新的深度学习创新转移到受限制的平台的过程仍然是一门“魔法”，需要许多的接受过高级训练的专家旷日持久的工作。这些深层网络由数

特定需求（例如，分别来自Google, Facebook, Qualcomm, ARM和Nvidia的TensorFlow, Caffe2, SNPE, Compute Library和TensorRT）已经开始得到解决，这一进程又进一步加快了。更重要的是，这个软件集合开始提

《物联网的深度学习》开始，Andreas Moshovos, Jorge Albericio, Patrick Judd, Alberto Delmás Lascorz, Sayeh Sharify, Zisis Poulos, Taylor Hetherington, Tor Aamodt和Natalie Enright Jerger提出了深度学习工作负载的硬件优化的创新理念，分析了现有深度模型呈现的核心低效率。这种类型的研究正在改变移动和嵌入式系统中出现的SoC的处理器架构设计。与之相辅相成的是，Shuochao Yao, Yiran Zhao, Zhang Aston, Shaohan Hu, Sha Hua, Zhang Chao, Lu Su和Tarek Abdelzaher在《深度学习与物联网》中详细介绍了专门用于处理嵌入式传感器（包括麦克风，加速度计，磁力计）捕获数据的深层网络的设计，以及一套降低其资源需求，以适应这些设备限制的、基于软件的方法。这两篇文章提供了以软件和硬件为基础的效率方法的示例，旨在将深度模型的系统资源需求降低到可接受的水平。

## 深度学习和机器学习，正从当今的主流分类和感知任务的完成者，扩展到对整个移动和嵌入式计算产生影响的角色。

百层互连节点组成，单个识别任务可能需要评估数以亿计的模型参数。这些模型的表示以及相关的推理算法很容易引入极端的资源开销。要完全解决当前学习算法与嵌入式/移动设备之间存在的障碍，可能需要彻底重新设计深层模型表示和算法；我们在这方面的最新进展虽然非常有前途，但仅代表向前推进的第一步。

解决机器学习效率的挑战将需要许多相互依赖的领域的进展，包括硬件、系统和学习算法本身。有希望的进展已经开始出现在这些传统上各不相同的领域。<sup>4-6</sup>随着大规模（通常是商业支持的）深度学习工具、图书馆，运行时系统和框架的成熟，受限设备的

供构建模块，通过简化关键步骤来直接支持该领域急需的基础研究，如：使得我们可以在Android设备上轻松测试和分析深层模型，针对特定处理器架构进行高度调整的矩阵运算库，甚至可以访问移动和嵌入式设备SoC上通常无法访问的非CPU处理器，如DSP和GPU。

## 本期内容

我们相信这个特刊为移动和嵌入式设备的深度学习研究提供了一个有代表性的快速概览。这六篇交叉的文章分析了算法效率、硬件专业化、传感器处理、活动识别和应用等核心问题。

从利用典型值加速深度学习

Seyed Ali Osia, Ali Shahin Shamsabadi, Ali Taheri, Hamid R. Rabiee和Hamed Haddadi在《使用混合边缘到云深度学习进行隐私的且可扩展的个人数据分析》中考虑了部署这项技术时对隐私的总体关注点。作者研究了如何在设备和云之间划分模型，这是一种常见的设计模式，可以为用户提



供保证,从系统捕获的数据中能推断什么、不能推断什么。

本期杂志的最后有三篇文章描述了从深度学习延伸出来的应用级别的进展。在《利用递归神经网络对资源受限的物联网设备进行基于呼吸的身份验证》一文中,Thomas Plötz和Yu Guan通过采用神经网络的原理和算法,概述了人类行为的模型以及移动、嵌入式设备的环境如何发生显著变化,并提供了具体实例。Jagmohan Chauhan, Suranga Seneviratne, Yining Hu, Archana Misra, Aruna Seneviratne和Youngki Lee在《基于呼吸的资源受限物联网设备认证中使用递归神经网络》中描述了一种用于移动用户认证的新应用,该认证依赖于深度学习方法;最后,在《找出小肠损伤:构建基于内镜成像的学习系统时遇到的挑战》中,Jungmo Ahn, Huynh Nguyen Loc, Rajesh Krishna Balan, Youngki Lee和JeongGil Ko概述了使用机器学习的新型医疗器械的潜力,并专门关注了由卷积神经网络实现的自动评估。

## 前方的路

对于各种受限类计算(如手机,手表,无人机,机器人和传感器等),集成和采用深度学习越来越普遍和普遍。推

## 关于作者

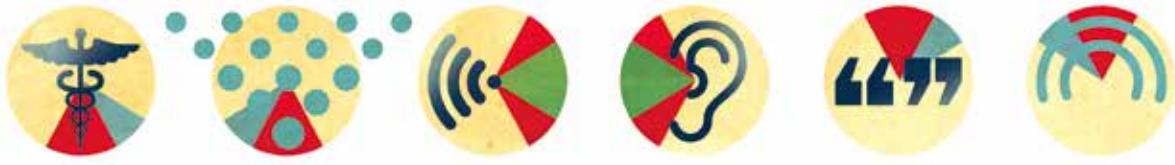
**NICHOLAS D. LANE**是牛津大学的副教授。他的研究兴趣包括移动和嵌入式系统限制下的高效深度学习,以及一些更广义的领域,包括机器学习和软件系统。Lane在达特摩斯学院获得博士学位。联系方式: nicholas.lane@cs.ox.ac.uk。

**PETE WARDEN**是TensorFlow Mobile部门的技术负责人,此前他是Jetpac公司的CTO。Jetpac因其优化的、可在移动和嵌入式设备上运行的深度学习技术而被谷歌收购。他是O'Reilly最近出版的电子图书《用TensorFlow搭建移动应用》(Building Mobile Applications with TensorFlow)的作者。联系方式: petewarden@google.com。

动这一革命性转变的是学术和工业研究人员,他们可以通过处理器架构、移动系统和学习算法的进步,将先前不相交的机器学习和嵌入式/移动计算的世界结合在一起。这个特刊中的论文只是这个新兴领域中正在进行的研究的一小部分。

在展望这个领域的未来时,我们认为,在短期内,随着深度学习向传感器推理的发展,设备理解和推理最复杂环境的能力将会持续提高。从仅使用推理的深度模式到直接在设备上进行培训和适应学习的模式的转变,将辅助实现这一进步。更深刻的是,尽管目前深度

学习的集成几乎完全局限于分类任务,未来将有更广泛的趋势,使得深度学习实现控制和决策任务。我们相信,以深度学习技术为主的学习算法可以取代(并增强)嵌入式和移动系统中的应用程序和系统组件逻辑,这种持久和长期的趋势将会发生。这将广泛地改变这些设备的内部功能,包括操作系统、无线和网络堆栈以及传感器处理管线。这样的改变将带来效率和功能的飞跃,嵌入式和移动设备将得以学习和动态地适应现实世界的情景和人类行为,而这是一种人类长期以来一直在寻求,但被证明难以鲁棒地实现的能力。■



## 参考文献

1. N.D. Lane et al., "Squeezing Deep Learning into Mobile and Embedded Devices," *IEEE Pervasive Computing*, vol. 16, no. 3, 2017. pp. 82–88.
2. T. Kraska et al., "The Case for Learned Index Structures," Arxiv.org, 11 Dec. 2017; <https://arxiv.org/abs/1712.01208>.
3. G. Toderici et al., "Variable Rate Image Compression with Recurrent Neural Networks," *Proc. Int'l Conf. Learning Representations (ICLR)*, 2016; <https://arxiv.org/abs/1511.06085>.
4. R. LiKamWa et al., "RedEye: Analog ConvNet Image Sensor Architecture for Continuous Mobile Vision," *Proc. 43rd ACM/IEEE Int'l Symp. Computer Architecture (ISCA 16)*, 2016, pp. 255–266.
5. S. Han et al., "Deep Compression: Compressing Deep Neural Networks, with Pruning, Trained Quantization and Huffman Coding," *Proc. Int'l Conf. Learning Representations (ICLR)*, 2016; <https://arxiv.org/pdf/1510.00149.pdf>.
6. G. Huang et al., "Densely Connected Convolutional Networks," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR 17)*, 2017, pp. 2261–2269.

**Call for Articles**

**IEEE Pervasive Computing** seeks accessible, useful papers on the latest peer-reviewed developments in pervasive, mobile, and ubiquitous computing. Topics include hardware technology, software infrastructure, real-world sensing and interaction, human-computer interaction, and systems considerations, including deployment, scalability, security, and privacy.

**Author guidelines:** [www.computer.org/mc/pervasive/author.htm](http://www.computer.org/mc/pervasive/author.htm)

**Further details:** [pervasive@computer.org](mailto:pervasive@computer.org) [www.computer.org/pervasive](http://www.computer.org/pervasive)

**IEEE Pervasive Computing**  
MOBILE AND UBIQUITOUS SYSTEMS



CCIR

## 知乎用户行为预测

# 700万用户，为谁点赞？

奖金：100000元

初赛：2018.5.21.-2018.7.7.

报名入口





# 利用典型值 加速深度学习

文 | Andreas Moshovos, 多伦多大学

Jorge Albericio, NVIDIA

Patrick Judd, 多伦多大学、NVIDIA

Alberto Delmás Lascorz, 多伦多大学

Sayeh Sharify, 多伦多大学

Zissis Poulos, 多伦多大学

Taylor Hetherington, 英属哥伦比亚大学

Tor Aamodt, 英属哥伦比亚大学

Natalie Enright Jerger, 多伦多大学

译 | 武昀, 浙江大学

为了提升支持深度学习创新所需的硬件计算能力, 识别出设计师可利用的深度学习特性十分关键。本文阐述了我们采用的策略, 概括了所发现的一些深度学习模型的价值特性, 以及能利用值特性减少计算量、片上和片外储存以及通信的硬件设计。

**深**度学习 (DL) 使计算设备能够“通过示例学习”, 从而解决传统计算无法胜任的任务。例

如, 如今利用深度学习, 计算设备能推断出图像或涂鸦描绘的对象。深度学习中最常用的形式是监督学习。通过对众多已知案例的初次培训, 监督学习能学会如何辨别物体。通过检验这些例子, 深度学习可以“学会”如何精确地辨

别一张从未见过图像描绘的是飞机还是茶壶, 只要之前检验的例子中包含足够的飞机和茶壶图像。

深度学习的核心组件已经存在了数十年, 但仅有少数几个合适案例。最近, 实际应用层出不穷, 定期示范操作的应用更是不计其数。是什么引发了这些“突然”的成功? 深度学习界已经能够收集或利用注释完善的示例, 并借助现代计算硬

件的出色计算能力，在机器学习的核心组件和连接方式上进一步创新。与我们的讨论最息息相关的是，在2010年后不久，商用计算机的处理能力以图形处理器的形式达到了一定高度，使得从前无法实现的深度学习应用得以发展。

尽管深度学习取得了巨大的成功，许多任务仍然无法实现，还有些需要进一步完善（例如，自动驾驶）。深度学习继续创新的一条清晰路径是利用更强大的计算能力以处理更多示例数据，构建更完善的组件和布局。正如以往，更强大的处理能力依旧是进一步创新的推动力，但未必一定能驱动创新。

我们的专长一直在于通用处理器的性能提升和能效增强技术。通用处理器如今是所有计算设备的核心，无论是服务器级机器、智能手机还是嵌入式设备。在过去四年左右，我们一直在探索深度学习应用的硬件层加速。目标是开发硬件层技术，将其纳入下一代硬件设备后，有望能使深度学习界探索更先进的应用。

本文回顾了深度学习硬件层加速的一般方法，重点介绍了我们开发的一些技术。这些技术的核心特征是以价值为基础。也就是说，它们通过利用应用的计算结构来开拓深度学习应用中数据流的特性，从而提升当前的性能和能效。本文中，我们主要讨论卷积神经网络的推断加速。

在解释基于价值的加速如何增强基于结构的方法，以及我们为什么选择该方向指导研究之前，我们首先回顾一下加速现在为何备受关注。

## 对加速的需求

在过去三十年左右，计算硬件性能大约每两年提升一倍。在1985年的台式计算机上需大约一个小时执行的任务，在1995年的计算机上只需不到一两分钟。摩尔定律推动了这种指数级的性能增长：半导体技术的进步使更多、更快的设备推动了计算机体系结构创新。

不幸的是，使用更多、更快的晶体管需要更大功率，但工作电压降低却抑制了计算硬件更新换代带来的总功率加速。然而，这些性能改进方法通常需要更多晶体管以实现性能优势，这就导致了不成比例的功耗成本。因此在21世纪初，处理器功耗和功耗密度超过了实际限制，性能定标显著减慢。单芯片多处理器出现，只要应用可以分解为线程，也就是基本可以同时执行的部件，就有望持续提高性能。图形处理器面向特定类别的工作负荷，这些工作负荷可以分解为数千个线程，每个线程大致以同等速度执行相同的代码。计算机图形学就是这种数据并行的工作负载。半导体技术潜在的定标趋势一直存在，当前的架构技术却正在接近极限，现在需要进一

步的创新以保持性能的提升。

由于现在功率是主要制约因素，提升性能需要减少每次操作的能耗——每次操作所需能量减少，就能用充足的硬件资源执行更多操作，同时保持处在功率范围内。硬件加速就是一种这样的方法。硬件加速器是部分或完全针对特定任务或任务种类专门设计的“处理器”。因此，我们来仔细看看深度学习，了解专门化如何提高每次操作的能量，从而提高性能。

## 硬件加速和深度学习

深度学习利用神经网络（NN）。图1a是前馈神经网络的示例，其中几个图层按顺序运行。在其他神经网络中，层和层之间存在反馈，每层接受几个输入数字，产生另一组输出数字。就图像分类而言，第一层的输入是图像。目前，决定卷积神经网络（CNN）执行时间的只有几种卷积层和更小范围上完全连接的图层。在讨论中，我们重点关注卷积神经网络和卷积层，因为完全连接的图层被当做卷积层的特例。

如图1b所示，卷积层（CVL）接受的输入是执行计算值的三维数组或激活（对层1来说，接受的是我们的外部输入——一张图像），并产生输出的激活三维数组。卷积层用几个过滤器来卷积输入激活，每个过滤器都存在预先定值

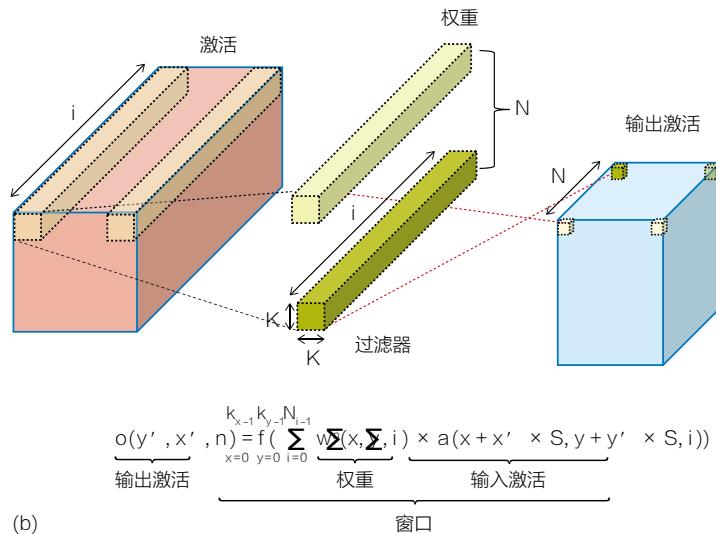
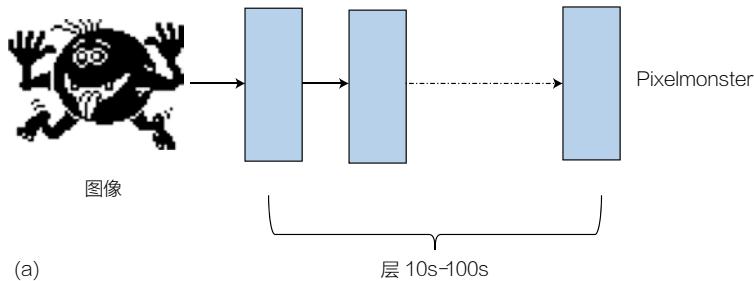


图1. (a) 前馈图像分类神经网络 (b) 卷积层

的三维数组或权重，这些权重值包含嵌入在神经网络中的“知识”，它们在训练时被计算出来，在推断过程中则为定值。对于图像分类而言，推断过程是通过网络确定描绘对象的过程。

典型的输入或输出激活数组包含几千个值，每层常有几百个过滤器，每个过滤器包含几百到几千的权重。每个输出激活计算等同于过滤器的点积，该

过滤器有和输入激活阵列等大的子阵。图1b显示了如何将输出激活 $o(x', y', n)$ 表示为输入激活 $a(x, y, i)$ 和权重 $w(x', y', i)$ 的函数。每个点积都包含了几百到几千个激活和权重，常量偏差通常在最后，结果会通过产生输出激活的非线性激活函数。多个激活函数与线性整流函数(ReLU)共存，ReLU函数常被用来进行图像分类，它将负激活值转

换为零，同时允许正激活通过。为了完全处理输入激活数组，过滤器使用步长 $S$ 扫描输入。每个过滤器参与的计算中所使用的输入激活子阵被称为窗口。由于偏差容易与激活函数一起执行，下文的讨论中不计偏差的增加。

## 加速深度学习的机会

点积可以执行为三重嵌套循环：

```
outa = 0
for xi = 0 to K
  for yi = 0 to K
    for ii = 0 to imax
      outa += a(x+xi, y+yi, ii) *
      w(xi, yi, ii)
```

通用处理器将这些循环执行为数十个机器代码指令，通常是一次简单计算或数据传送。执行每条指令需要进行一些操作，例如从存储器获取指令表示，进行解码以分析内容，读取并写入多个存储元件，例如寄存器或存储器。这种处理器十分灵活，能很好地执行任何代码。但如果只考虑执行点积，那么将会耗能巨大，从而导致性能成本增加。利用硬件专门化来执行点积可以大大减少这些开支。

## 基于计算结构的加速

专门化可以利用点积的计算结构。

图2是基于结构的加速器，其中包含16个激活和16个权重。将这些成对乘，然后使用加法器树减少16个权重，将结果累加到输出寄存器中。该硬件可以在多个周期计算一个输出激活，加速器可以用类似的多个单位来处理每个周期中更多的激活和权重。由于卷积层通常有多个过滤器，每个过滤器有一个独立的单元，所有单元重复使用16个相同激活。因为现代半导体技术中存储器访问比典型计算的能源成本高得多，所以重复使用数据十分可取。

寒武纪2号神经网络处理器就是这种基于结构的加速器。它利用卷积层中的激活重用，合理使用片上资源来平衡计算和通信需求。寒武纪2号包含256个处理单元，类似于图2，每个处理单元包含16个区，每个区包括16个单元，每个单元可以处理一个单独的过滤器。寒武纪2号每个周期总共计算4K个乘积和256个部分点积。不同应用可以且最好使用不同配置。

## 基于价值的加速

我们特地选取了可以补充基于结构的加速的技术：作为学术团体，我们认为在行业完善基于结构的方法后，如果能进一步研究可能有用方面，我们的贡献会更有意义。利用我们在通用处理器优化方面的经验，我们确定了以下三个原则：1) 尝试利用典型的执行行

为，2) 不需要修改神经网络就能达到效益，3) 在尝试归纳前先深入研究专门化。原因如下：

**利用典型行为。**许多通用处理器性能技术都利用了典型的程序行为。以硬件缓存（一种关键的内存访问加速技术）为例，在当今的技术中，处理器执行计算速度比主存储器提供数据的速度快100倍。不幸的是，我们无法造出大而快的主存，但好在我们可以通过利用通用程序行为构建分级存储器体系，大多数情况下它能像大而快的存储器一样运行。这是唯一可行的手段，因为大多数组程序存在内存访问流局部性：它们倾向于及时获取相同或邻近的内存位置。因此，缓存（小而快的内存）可以使用如下策略来处理内存请求：保留一定最近访问的内存位置和附近内存位置的备份。程序不必表现出局部性，但大多数程序恰巧都有。

通过在通用处理器中复制这种经验，我们想问神经网络执行中是否存在能被硬件利用以提高性能的属性。我们希望对利用计算结构的方法进行补充，并以价值流为目标。在权重和激活之间，我们决定首先研究激活。我们认为虽然权重流中存在诸多机会，但由于权重事先可以得知，所以软件方法很可能提供许多潜在效益，或者至少应该归属于解决方案。激活是运行时得出的值，因此不宜采用静态分析。但随着我

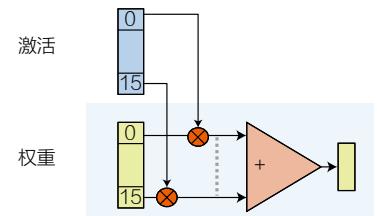


图2. 基于计算结构的加速器。

们基于激活的方法日益成熟，我们最近的确探索了利用两者属性的选项2-4。

**以即用网络为目标。**对于通用计算，需要更改软件的技术只取得了部分成功。软件开发已经十分艰难，需要大型开发团队多年开发的软件尤其如此。我们复制了这一经验，选择研究对即用神经网络有效的加速器。这并不是说共同设计的神经网络和硬件不值得提倡，共同设计反而可能会带来更多好处，但这需要时间才能成熟并取得成果。就算在执行前完成，这也可能会产生大量的开销。我们选择设计能够产生即时效益的硬件，同时激励神经网络设计中的相关进展，例如降低值精度。

**风险：广度与深度。**任何加速器设计都有风险。如果应用变动太大，无法在加速器上继续执行该如何？例如，由于视频解码算法发生了巨大变化，专门用于早期视频格式的加速器现在已经过时。又或者，如果一个应用程序混

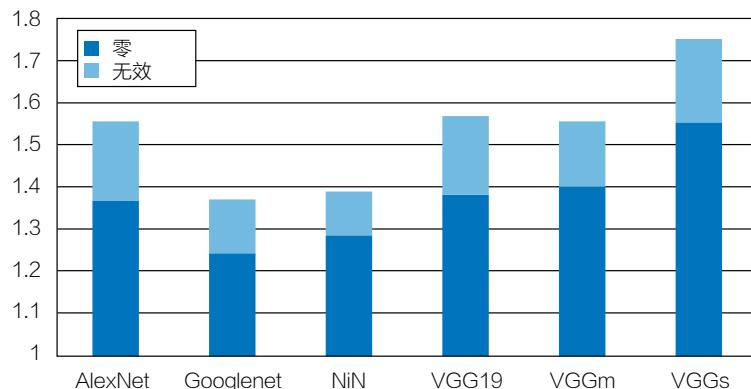


图3. 跳过无效激活时的性能改进。深蓝色：跳过零激活；浅蓝色：在保持精确度的同时进行阈值处理。

合使用其他技术，加速器无法从中获益呢？

理想的加速器应为：1) 足够专门化，能提供期望的性能水平，2) 足够通用，能支持更广泛的应用类别，3) 不会过时。不是所有目标都能实现。尽管我们需要广度，但单独深入研究每种感兴趣算法也有价值。这种探索最终可以为设计提供足够的通用性，同时也有益于已知预期使用寿命的专用应用和设备。因此，我们决定重点研究神经网络。由于现存的网络大多针对图像分类，我们主要研究这类神经网络。对这些神经网络的分析证实，卷积以及更小范围内完全连接的层决定了执行时间，因此我们着眼于这两个层面。最后，我们选择了第一个目标推理，部分是由于它是训练的基本架构，也是因为我们预计将来会有更多只需要进行推理的设备。

## 有趣的运行时间值特性

研究图像分类的价值流时，神经网络展现出了一些可以被用来加速的特性。

### 无效的激活

在所有研究的卷积神经网络中，许多乘法都是无效的，因为它们包含零值激活。只要激活输入值足够接近零，就可以避免更多的乘法运算。每个网络和每层的“足够接近”各不相同，我们开发了一种寻找层阈值的经验方法。这些无效的乘法意味着改善性能的机会，对于大规模数据并行的机器而言，利用它们是个挑战。为了获得任何性能，我们需要开发其他有用的计算，以取代无效计算。但仅仅检查激活是否无效花费的时间就与执行乘法相同，更糟的是，再次

激活需要另一次数据访问。幸运的是，第一层外的每个卷积神经网络层的输入都是前一层的输出，因此，我们可以在每层的输出处将有效的激活紧紧打包在内存中，以便下一层顺利处理，无需检查无效计算或执行额外的存储器访问。我们的Cnvlutin5正是按此设计，图3显示了它相对于塞武纪2号的性能提升。

为什么存在如此多的零激活或接近零激活？在图像分类和第一层的情况下，激活可以被认为是某个视觉特征（例如表示虹膜的圆）出现在某个位置的概率。除非图像充满了这样的圆圈，大多数这种激活将为零或接近零。虽然这过于简化，但它表明无效激活是神经网络的固有属性。

同样，有效推论机器跳过零激活，同时利用权重稀疏<sup>6</sup>，使用执行单次乘法累加的单位。SCNN也针对稀疏化的神经网络（需要额外步骤将权重转换为零的神经网络），跳过无效的权重和激活。<sup>7</sup>

### 精度可变性

我们很早就发现神经网络所需的精度每层都不同，其他人也注意到了这种特性。在这个过程中，我们开发了一种基于性能分析的方法，从而确定每层使用的精度，同时保持准确性。如表1所示，所需的精度从AlexNet一些层

的5位到VGG-19一些层的13位不等，这意味着精度通用的传统硬件执行了许多不必要的且耗能的计算。但我们能否构建避免这些计算的加速器，以提高能效和执行性能？具体而言，我们想问能否构建一个执行时间与所需精度成比例的加速器。和对所有激活使用固定精度（如16位）的设计相比，我们期望的加速器在执行层L时将会速度快 $16/PL$ 倍，其中PL是针对层选择的激活精度。我们的目标是仅从精度位减少也能获得更好的性能。例如，与通用16位精度相比，使用8位和7位精度的层，加速器会分别快2倍和2.3倍。现有的处理器以粗糙的间隔尺度（例如8位、16位或32位）利用精度可变性，性能优势远低于预期。我们的Stripes加速器使用位串行处理，同时利用数据并行性来提供所需的性能定标<sup>8</sup>。Stripes只提高卷积层的性能，Tartan将这些效益扩展到完全互连的层，尽管面积成本会随之增加<sup>3</sup>。Stripes和Tartan可以相应进行配置以服务于任何设备，上至高档服务器级别，下至嵌入式设备。对于较小规模的设备，变体Loom可以对权重和激活利用精度可变性，从而进一步提高性能<sup>4</sup>。上述加速器设计通过支持全范围的精度，激励在设计精度降低的神经网络中的任何进步，最终可能促使库尔巴里奥（Courbariaux）等人提出了二元模型。所有上述设计同样减少了存储器

表1. 激活精确概况

网络	以位为单位的激活精确 每层/对动态精确检测有效
AlexNet	9-8-5-5-7 / 5.4-7.4-4.2-4.4-5.8
NiN	8-8-8-9-7-8-8-9-9-8- 8-8 / 6.4-7.1-7.8-7.0-5.8-5.2-8.4-7.5-7.6-7.6 4.7-6.8
Goo- gLeNet	10-8-10-9-8-10-9-8-9- 10-7 / 6.2-5.8-6.8-6.3-5.3-6.7-6.3-5.0-5.5-7.9-4.8
VGG-M	7-7-7-8-7 / 5.3-5.1-5.8-3.4-4.8
VGG-S	7-8-9-7-9 / 5.3-5.1-5.0-5.4-4.0
VGG-19	12-12-12-11-12-10-11-11-13-12-13-13-13-13-13-13 / 9.1-7.7-10.0-9.0-11.1-8.8-9.7-8.3-11.6-10.4-12.2-11.7-11.5-11.5-10.4-5.9

和通信需求，因为它们只存储了所需位以表述存储器中的激活，这使得它们能够存储并处理更大的网络。Proteus扩展为现有的位并行执行器带来了这些优势，使内存占用空间和带宽平均降低了约40%<sup>10</sup>。它使用轻量级机制，将数据从便于数据存储和通信的表述法转换为便于数据处理的表述法。

**动态精度检测。**尽管性能分析让我们能确定每层精度，在执行时保持TOP-1（精确匹配）的精度，这些精度却是悲观的。分析找到了所有可能的图像以及该层所有激活所需的最差精度。然而，在实践中加速器将会处理：1) 在任何给定时间点的一个特定输入，2) 每个周期有限数量的输入激活，如256，并非层中的所有激活。当限制对同时处理的每组激活的注意时，我们可以进一步降低精度。动态Stripes是对Stripes

和Loom的精确、低成本的扩展，可以在运行时检测并利用精度变化<sup>11,12</sup>。如表1所示，在间隔尺度为256时动态检测这些精度，有效激活精度比那些通过分析检测到的更短。动态精确检测与权重精确检测相结合，也大大减少了片外流量以及片上存储和通信<sup>12</sup>。

## 重复计算

我们早期发现许多乘法恰好处理完全相同的值对，最有趣的情况是不同过滤器在相同坐标处恰巧具有完全相同的权重。在运行时，每个都会和相同的激活相乘，因此权重相同。为什么不同的过滤器在相同的坐标下等值？我们推测至少有两个原因：1) 过滤器容器是3D阵列，而过滤器正在寻找的特征并不是与该容器完全相符的形状，这会导致几个权重为零或接近零，或者不是所

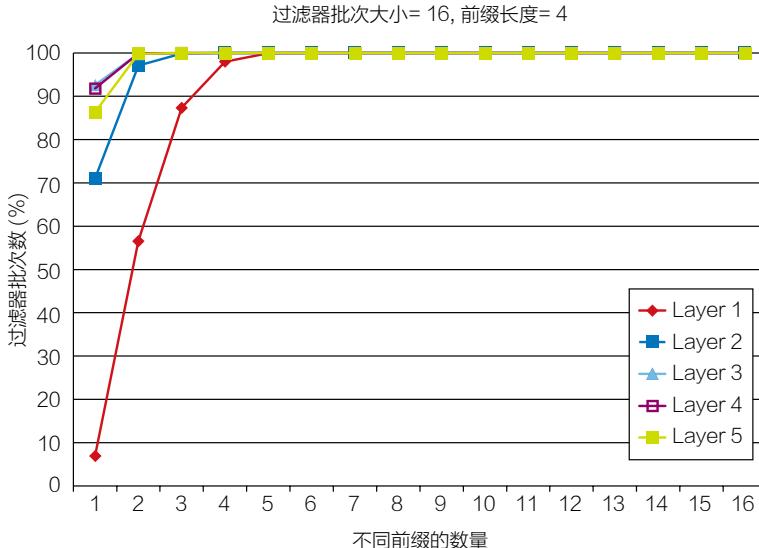


图4. 不同的4位前缀，用于出现在AlexNet卷积图层中16个不同过滤器的相同坐标的权重。

有特征都与所有潜在的客体类别相关。  
2)一些特征在一定程度上相似，这会使  
得一些权重相同或相似。一旦精确度下  
降，权重冗余就会增加到有趣的水平。

除了整体值之外，在限制对权重部  
分(如前缀)的注意时，还会有更多的  
冗余。图4展示了AlexNet中的一些冗  
余，这组测量结果分别来自不同过滤器  
的16个权重组，所有权重出现在相同的  
坐标上，图表显示了独特的4位前缀值  
的分布。尽管4位前缀有16种可能的组  
合，但在层2至层5中，至少70%的权重  
组仅包含单个前缀。层3到层5几乎所有  
的权重组都包含最多3个不同前缀。第1  
层的冗余度较低，只有7%的权重组包  
含单个前缀，但约88%的权重组只包

含3个不同的前缀值。这种冗余可能对  
于压缩存储器中权重表示以及减少所  
需的计算次数有用。

### 有效位“密度”

最后，激活在个体位明显偏向于零。  
如图5所示，具体而言，平均只有8%  
的激活位为1。我们将激活调整为每层  
所需的精度后，激活位值被用于乘法，  
图5测量了该值。我们在小学学习了如  
何用铅笔和纸做乘法：从乘数中取一  
个数字，将其乘以被乘数，重复下一个  
乘数位。由于数字是二进制，因此乘数  
位将为0或1，当为0时，不会改变最终  
结果。使用这种方法，我们在处理卷积  
神经网络时会有92%的时间乘以0位。

如图5所示，如果我们能够开发一种只  
处理有效位的加速器(也就是1位)，那  
么性能改进的潜力是12.5倍。图5还表  
明，即使我们可以用某种方式消除所有  
的零值激活，接近75%的激活位仍然为  
零，性能改善的潜力为4倍。即使用8位  
量化，这种情况虽然得到减缓，但仍然  
存在。

通过利用精度可变性，Stripes和  
动态Stripes可以消除一些无效计算，  
但还有一些零会被保留下。例如，在  
处理一对8位“0100 0000”和“0000  
0010”的激活时，Stripes即使用动态  
精确检测也只能处理6位。但是如果我  
们只处理每次激活的有效位，一步就足  
以处理两者。Bit-Pragmatic加速器或仅  
Pragmatic加速器就利用了这种卷积神  
经网络的特性<sup>13</sup>。

## BIT-PRAGMATIC 加速器

图6用一个简化的例子说明了  
Pragmatic加速器的核心基本概念。  
(a)部分是基于结构的加速器，使用  
16位定点表示法来处理两个激活(A0  
和A1)以及两个权重(W0和W1)。两个  
乘数 $16b \times 16b$ 产生 $32b$ 的乘积(A0×W0  
和A1×W1)，加法器将其减少到1个 $33b$   
值，输出寄存器累计结果。该加速器每  
周期处理两对激活和权重，处理16个激

活和权重对需要16个周期。在我们的例子中，每个激活值只包含一个二的幂， $A_0$ 为23， $A_1$ 为213。因此位并行加速器将处理 $15 + 15$ 个零激活位，不会影响最后输出。

(b) 部分是一个简化的Pragmatic加速器，只处理有效的激活位。激活现在不再以位置表示法来表示，而是用一系列2的幂来表示，因为每个只有一个2的幂，且 $A_0$ 和 $A_1$ 分别为(0011)和(1011)。如果 $A_0$ 是“0000 1100”，则它将被表示为(0100, 0011)。每个周期中，该单位将每次激活的2的幂乘以相应的权重。因为乘以2的幂相当于简单移位，乘法器被替换为移位器。由于2的乘积为32b时，每个周期都会被减少并累积起来，该单位的其余部分保持不变。该单位在一个周期内处理两个激活和权重乘积，因此与(a)部分的位平行单位速度一样快。但如果 $A_0$ 或 $A_1$ 包含多个无效位，则该单位需要成比例数量的周期来计算乘积。因此它的执行时间与有效位的数量成正比，这在某种程度上是我们想要的结果。但这种设计最多与位并行设计一样快，而且只有在所有激活只包含一个有效位时才能实现。最坏的情况是至少1个激活有16个有效位，这时它会慢16倍。

幸运的是，卷积层在整个窗口中表现出并行性和权重重用性，Pragmatic利用这两种特性以确保它总是至少与

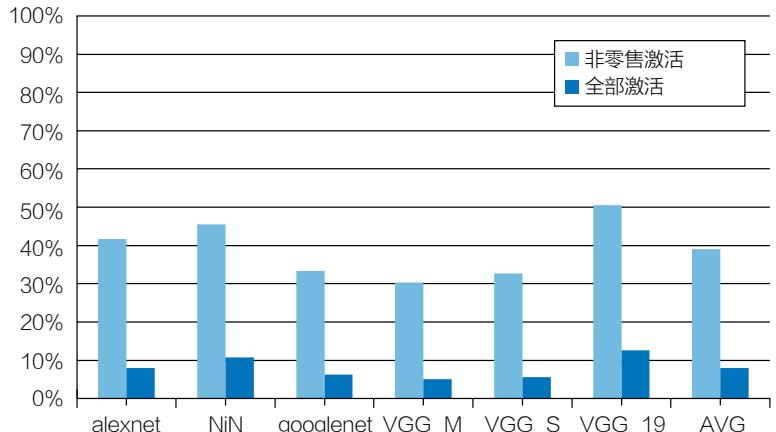


图5. 激活位是1的部分，根据使用频率加权的所有卷积层平均值。

位并行机器一样快，不需要从内存中读取更多的权重或激活位，后者需要更大内存和更昂贵的补充。(c)部分显示了Pragmatic的方法，(b)部分的单位已经被复制了16次，16个单元每个处理不同的激活对。但所有单位有相同的权重，这可以通过并行处理16个窗口来实现，每个单元一个窗口。位并行单元处理 $2 \times 16b$ 激活，每个周期总共处理32b激活输入，而Pragmatic单元处理32次激活，每次激活一个2的幂，相当于每个周期32位激活。Pragmatic每次激活用4位2的幂，但该转换在从存储中读取激活之后才完成。

在最坏的情况下，当至少一次激活的全部16位为1时，该单元需要16个周期来处理所有32次激活，产生32次激活和权重积，这与(a)部分位并行执行

器的处理能力相符。两个执行器按不同的顺序进行计算，但最后产生了相同的结果。当所有激活最多只有一个有效位时，Pragmatic单位只需要1个周期来完成传统单位16个周期完成的工作，因此速度快16倍。一般来说，如果每次激活的最大有效位是N，那么Pragmatic将快 $16 / N$ 倍。

## 使其具有实用性

然而，上述Pragmatic简单易懂的执行却并不实用。Pragmatic单位比其位平行大4倍左右，且性能改进不尽如人意。我们只能开发多种技术，这些技术结合在一起才能执行Pragmatic，我们的讨论中强调了其中三个技术。首先是两阶段转变。在简化的设计中，对于每个输出激活，我们同时处理16个权

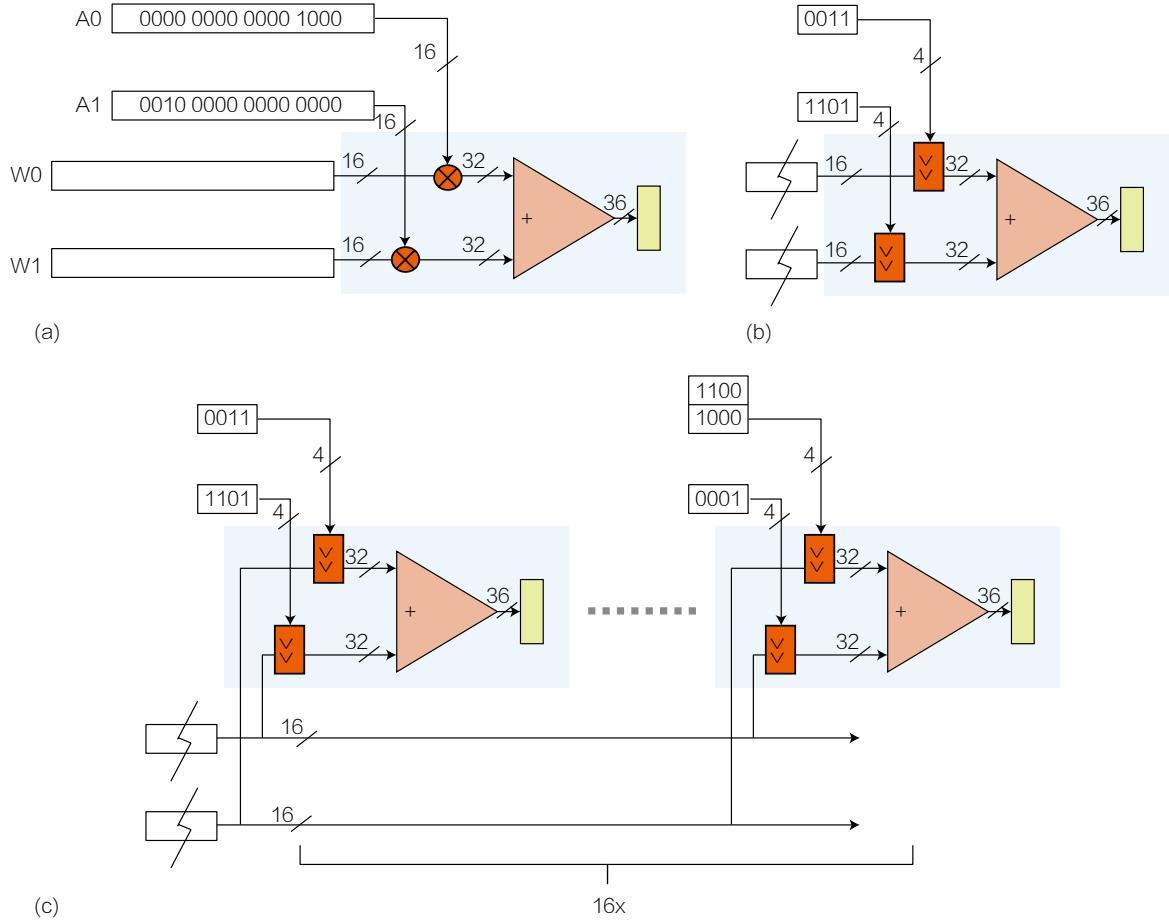


图 6. Pragmatic 的方法：例 (a) 基于结构的加速器。 (b) 连续处理 2 的幂。 (c) 超出基于结构的加速器的性能。

重和激活偏移对。由于我们用4位2的幂移位每个权重，在最坏的情况下，其中一个幂为0，另一个为15，这样每个移位器都需要接受16b权重并产生32b输出“乘积”，因此，加法树需要接受32b的乘积输入。虽然这种设计为我们提供了最大的灵活性，以消除无效激活

位，但成本也很高。两阶段转变放弃了这种灵活性，因此可以大幅度降低成本，从而提高性能，目的是将输入激活处理成子组。例如，我们可以四位组一次处理每个激活，而不是同时处理多个2的幂。在这种情况下，处理两个值为“0100 0000 0000 0000”和“0000

0000 0000 0010”的激活将在两个周期内完成，即使每个激活只包含单个有效位。在第一个周期中，我们将处理四个最低有效位0000和0010，在第二个周期处理四个最高有效位0100和0000。我们发现，实际上处理四个一组的位足以实现大部分性能，且处理不受

限制。Pragmatic在运行时动态选择每组的开端，例如，它会在一个周期内处理“0000 0000 0001 0000”和“0000 0000 0000 1000”。

第二种技术是激活通道的部分去耦技术。在简化设计中，Pragmatic进入下一个组之前会处理组中的所有激活。通过在权重输入中增加缓冲区，并将激活静态地放置到子组中，我们可以允许一些子组先于其他组运行。在实践中，如果只使用一个权重缓冲区，从而允许子组仅运行一个激活组，这种做法可以显著提高性能。不管怎样，这些缓冲区在执行完全连接的层时都支持充分利用。

最后，到目前为止，我们假设激活被表示为2的幂之和，但基本设计可以轻松处理幂的增加和减少。这是Booth编码的一种形式，常被用于减少高性能乘法器的延迟。例如，激活“0011 1100 0000 0000”可以表示为“(0010 0000 0000 0000 - 0000 0010 0000 0000)”或“(213-29)”。Pragmatic使用修改后的Booth编码形式，以避免与阶段转化连接的周期数的增加。

## 减少执行时间

图7显示了相较于类似于寒武纪2号的配置等效加速器，性能（与执行时间相反）如何得到提高。三个参数定义一个配置：每个过滤器的项，每个区的过滤器以及区数量。每个过滤器的项

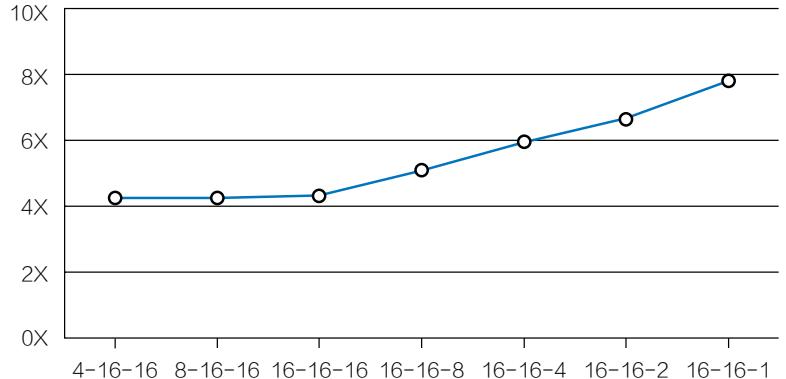


图7. 各种实用配置的性能改进。

是每个过滤器计算的激活和权重积的数量，每个区的过滤器数量是每个处理器区处理的过滤器的数量。 $x$ 轴以项-过滤器/区-项/过滤器格式显示配置。一个16-8-4配置有16个区，每个处理8个过滤器和4个权重，共处理512个周期。为匹配寒武纪2号16-16-16原始服务器级的配置设计，Pragmatic平均提升性能4.3倍。在处理项较少的过滤器时，Pragmatic在激活间的失衡降低。处理过滤器的项为1的配置时，性能将提升8倍，这可能更适合嵌入式设计。

## 总结

表2总结了一些我们的设计，并标明了它们与寒武纪2号等同的配置的相对性能、能效和面积。该表还记录了最近的加速器Tactical2，它结合了Pragmatic或动态Pragmatic的优点以

及轻量级零重量的跳跃前端，从而具有乘法优势。如图所示，Laconic配置使用了一半的权重存储器线，若使用相同数量的权重存储器线，速度可以提升30倍。Tactical的报告结果是针对AlexNet、Googlenet和ResNet-50的删减版本，此外，Loom的报告结果支持动态精确检测。

**深**度学习早期硬件加速的成功凭借的是计算结构和数据重用，例如，Y. Chen和Chen<sup>1,14</sup>。最近，许多深度学习硬件加速器利用了深度学习神经网络（DNN）展现的各种形式的信息无效，我们和其他人的研究就是很好的典范。我们发现在深度学习神经网络中，信息无效表现为无效的神经元<sup>6,15</sup>、激活<sup>6,5,15</sup>或权重<sup>16,15</sup>、过度精确（如沃登（Warden）和贾

表2. 相对寒武纪2号的基于价值的加速器特性<sup>1</sup>

加速器	配置	性能	功率	面积	频率	技术节点
DaDianNao	16-16-16	3.9 Tmul/sec	17.6 Watt	78mm <sup>2</sup>	980 Mhz	65nm
加速器	与寒武纪2号配置相比	相对性能	相对能耗	相对面积	值属性	
Cnvlutin <sup>5</sup>	16-16-16	1.6×	1.47×	1.05×	无效激活值	
Dynamic stripes <sup>11</sup>	16-16-16	2.6×	1.54×	1.35×	动态激活精确值	
Loom <sup>4</sup>	1-8-16	3.6×	2.9×	0.94×	动态激活+权重精确	
Pragmatic <sup>13</sup>	16-16-16	4.3×	1.71×	1.68×	无效激活位	
Tactical <sup>2</sup>	4-16-16	10.2×	2.4×	1.14×	零权重+无效激活位	
Laconic	1-8-16	16×	1.63×	2.39×	权重+激活的无效激活位	

“寒武纪2号配置”栏下注明了性能、能效和面积(与寒武纪2号同等配置相比)。寒武纪2号的配置被标为“区块-过滤器/区块-乘积/过滤器”。

德(Judd)<sup>17,8</sup>)、无效激活位<sup>13</sup>或者是过度配置。无论是静态、动态还是两者兼有,低效率是否被充分利用是一个悬而未决的问题。此外,随着深度学习神经网络的发展,哪种形式的低效率仍将持续存在也是问题。过去的成功表明,在我们探索如何能最好地提供所需的硬件性能改进,以支持深度学习创新时,确定硬件和/或软件可利用的深度学习神经网络属性非常重要。此外,从只考虑精确度以减少存储到考虑有效位密度以提高性能,技术革新的发展表明,事先预见未来的创新并非易事。因此,我们应该鼓励进一步的探索,即使是在那些今天看来似乎毫无益处或毫无关联的方向。

顺着这些路径,我们的加速器利用深度学习神经网络的一些价值特性,同时使用即用网络,从而完成了当前的部署。更重要的是,它们为深度学习神经网络设计人员提供了安全的创新途径,为微小的进步提供奖励,开拓了新的机会,并创造了新的激励机制。具体而言,如果得到部署,它们有可能在以下方面加速创新:1)极低精度的神经网络设计,关注三元或三元网络(Stripes, Loom, Pragmatic, Tactical),2)降低权重(Tactical)。它们能够实现全范围精度选择的实验,同时还为全精密网络提供出色的性能。它们有可能“激励”机器学习界进一步在这些方向进行投资,及时提供相符的奖励。最终,如

果极低精度和过度削减的网络取代,更高效的硬件平台可以安全地接管。新的机会也有可能出现,例如进一步减少1位的数量;采用其他量化方案,例如在具体问题具体分析的基础上,只使用2的幂;或者甚至是重新排列过滤器以减少有效位不平衡,所有这些都无需新开发的方案就能适用于所有网络。

感兴趣的读者可以访问第一作者的网页,阅读我们的最新发现。

## 参考文献

1. Y. Chen, T. Luo, S. Liu, S. Zhang, L. He, J. Wang, L. Li, T. Chen, Z. Xu, N. Sun, and O. Temam, “DaDianNao: A Machine-Learning Supercomputer,”

- in Microarchitecture(MICRO), 2014 47th Annual IEEE/ACM International Symposium on, Dec 2014, pp. 609–622.
2. A. D. Lascorz, P. Judd, D. M. Stuart, Z. Poulos, M. Mahmoud, S. Sharify, M. Nikolic, and A. Moshovos, “Bit-Tactical: Exploiting Ineffectual Computations in Convolutional Neural Networks: Which, Why, and How,” CoRR, vol. abs/1803.03688, 2018. [Online]. Available: <https://arxiv.org/abs/1803.03688>
  3. A. Delmás, S. Sharify, P. Judd, and A. Moshovos, “Tartan: Accelerating fully-connected and convolutional layers in deep learning networks by exploiting numerical precision variability,” CoRR, vol. abs/1707.09068, 2017. [Online]. Available: <https://arxiv.org/abs/1707.09068>
  4. S. Sharify, A. D. Lascorz, P. Judd, and A. Moshovos, “Loom: Exploiting weight and activation precisions to accelerate convolutional neural networks,” CoRR, vol. abs/1706.07853, 2017. [Online]. Available: <https://arxiv.org/abs/1706.07853>
  5. J. Albericio, P. Judd, T. Hetherington, T. Aamodt, N. Enright Jerger, and A. Moshovos, “Cnvlutin: Ineffectual-neuron-free deep neural network computing,” in 2016 IEEE/ACM International Conference on Computer Architecture (ISCA), 2016.
  6. S. Han, X. Liu, H. Mao, J. Pu, A. Pedram, M. A. Horowitz, and W. J. Dally, “EIE: Efficient Inference Engine on Compressed Deep Neural Network,” arXiv:1602.01528 [cs], Feb. 2016, arXiv: 1602.01528. [Online]. Available: <https://arxiv.org/abs/1602.01528>
  7. A. Parashar, M. Rhu, A. Mukkara, A. Puglisi, R. Venkatesan, B. Khailany, J. Emer, S. W. Keckler, and W. J. Dally, “Scnn: An accelerator for compressed-sparse convolutional neural networks,” in Proceedings of the 44th Annual International Symposium on Computer Architecture, ser. ISCA ’17. New York, NY, USA: ACM, 2017, pp. 27–40. [Online]. Available: <https://doi.acm.org/10.1145/3079856.3080254>
  8. P. Judd, J. Albericio, T. Hetherington, T. Aamodt, and A. Moshovos, “Stripes: Bit-serial Deep Neural Network Computing,” in Proceedings of the 49th Annual IEEE/ACM International Symposium on Microarchitecture, ser. MICRO-49, 2016.
  9. M. Courbariaux, Y. Bengio, and J.-P. David, “BinaryConnect: Training Deep Neural Networks with binary weights during propagations,” CoRR, vol. abs/1511.00363, Nov. 2015. [Online]. Available: <https://arxiv.org/abs/1511.00363>
  10. P. Judd, J. Albericio, T. Hetherington, T. M. Aamodt, N. Enright Jerger, and A. Moshovos, “Proteus: Exploiting numerical precision variability in deep neural networks,” in Proceedings of the 2016 International Conference on Supercomputing, ser. ICS ’16. New York, NY, USA: ACM, 2016, pp. 23:1–23:12. [Online]. Available: <http://doi.acm.org/10.1145/2925426.2926294>
  11. A. Delmás, P. Judd, S. Sharify, and A. Moshovos, “Dynamic stripes: Exploiting the dynamic precision requirements of activation values in neural networks,” CoRR, vol. abs/1706.00504, 2017. [Online]. Available: <http://arxiv.org/abs/1706.00504>
  12. A. Delmás, S. Sharify, P. Judd, M.

## 关于作者

**ANDREAS MOSHOVOS**, 多伦多大学电子计算机工程系教授, 研究兴趣是构建高效、高性能的计算硬件。莫夏沃斯拥有威斯康星大学麦迪逊分校的计算机博士学位, 是电气与电子工程师协会(IEEE)的高级会员以及美国计算机协会(ACM)的会员。请通过moshovos@eecg.toronto.edu与他联络。

**JORGE ALBERICIO**, NVIDIA深度学习高级架构师, 拥有萨拉戈萨大学的系统工程和计算博士学位。2013年至2016年为多伦多大学的博士后研究员, 从事分支预测、近似计算和硬件加速器的研究。阿尔伯利乔是电气与电子工程师协会(IEEE)的成员。请通过jorge.albericio@gmail.com与他联络。

**PATRICK JUDD**, 多伦多大学四年级博士候选人, 以深度学习高级架构师身份加入NVIDIA。研究兴趣包括计算机体系结构、机器学习和近似计算, 主要研究深度神经网络硬件加速器的设计, 该技术利用近似法来提高性能和能效。贾德是IEEE的学生会员。通过patrick.judd@mail.utoronto.ca与他联络。

**Alberto Delmás Lascorz**, 多伦多大学三年级博士候选人。重点研究机器学习加速器的硬件设计, 研究兴趣包括计算机体系结构、深度学习和嵌入式可重构系统。达尔马斯·拉斯克兹此前在萨拉戈萨大学主修计算机工程, 是IEEE的学生会员。请通过a.delmaslascorz@mail.utoronto.ca与他联络。

**SAYEH SHARIFY**, 多伦多大学三年级博士候选人。研究兴趣包括计算机体系结构、机器学习、嵌入式系统以及

可重构计算, 为机器学习算法设计硬件加速器。舍里此前在谢里夫理工大学学习计算机工程, 是IEEE和ACM的学生会员。请通过sayeh@ece.utoronto.ca与她联络。

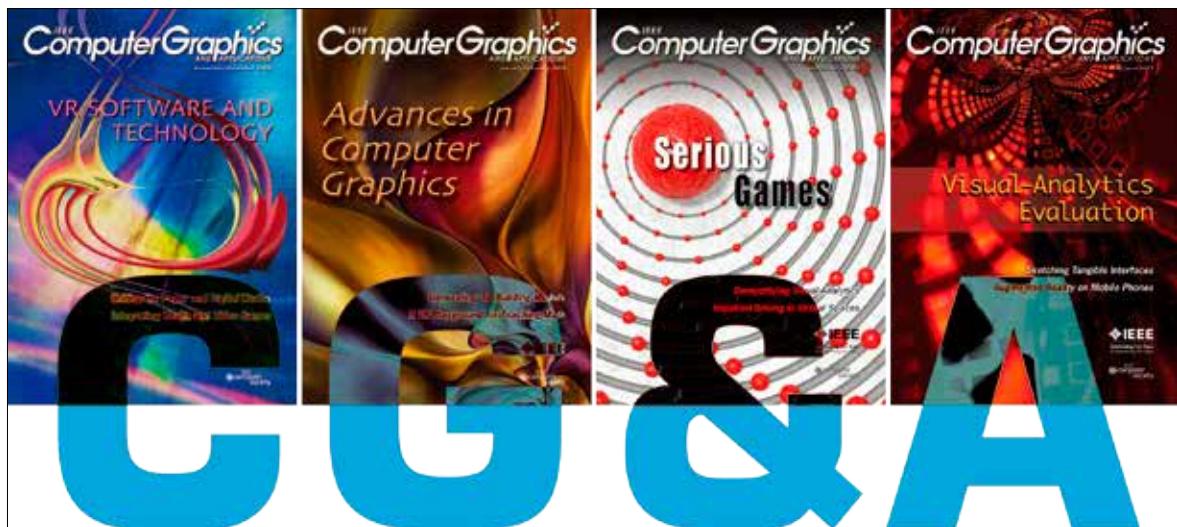
**ZISSLIS POULOS**, 多伦多大学电子计算机工程系的博士候选人。研究兴趣在于设计用于机器学习应用的高性能硬件, 开发网络扩散和社交图谱推理的近似方法。普洛斯拥有多伦多大学电气工程学的应用科学硕士学位(MASc), 是IEEE的学生会员。请通过zpolous@eecg.toronto.edu与他联系。

**TAYLER HETHERINGTON**, 英属哥伦比亚大学计算机工程系毕业班博士研究生, 目前在甲骨文实验室工作。研究兴趣包括计算机体系结构(特别是通用GPU)、机器学习加速器和系统软件。赫瑟林顿是IEEE的学生会员。请通过taylerh@ece.ubc.ca与他联系。

**TOR AAMODT**, 英属哥伦比亚大学电子计算机工程系教授。研究兴趣包括通用GPU和机器学习加速器的架构。阿莫特拥有多伦多大学的电子计算机工程博士学位, 是IEEE和ACM的会员。请通过aamodt@ece.ubc.ca与他联系。

**NATALIE ENRIGHT JERGER**, 多伦多大学电气与计算机工程的珀西·爱德华·哈特教授(表彰卓越研究水平和出色指导能力的教授职位)。研究兴趣包括计算机体系结构、近似计算、互连网络和机器学习的硬件加速。恩莱特·耶格尔拥有威斯康星大学麦迪逊分校电气工程博士学位, 是IEEE和ACM的高级会员。请通过enright@ece.utoronto.ca与她联系。

13. J. Albericio, A. Delma's, P. Judd, S. Sharify, G. O'Leary, R. Genov, and A. Moshovos, "Bit-pragmatic deep neural network computing," in Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture, ser. MICRO-50 '17. New York, NY, USA: ACM, 2017, pp. 382–394. [Online]. Available: <http://doi.acm.org/10.1145/3123939.3123982>
14. Chen, Yu-Hsin and Krishna, Tushar and Emer, Joel and Sze, Vivienne, "Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks," in IEEE International Solid-State Circuits Conference, ISSCC 2016, Digest of Technical Papers, 2016, pp. 262–263.
15. A. Parashar, M. Rhu, A. Mukkara, A. Puglielli, R. Venkatesan, B. Khailany, J. Emer, S. W. Keckler, and W. J. Dally, "Scnn: An accelerator for compressed-sparse convolutional neural networks," in Proceedings of the 44th Annual International Symposium on Computer Architecture, ser. ISCA '17. New York, NY, USA: ACM, 2017, pp. 27–40. [Online]. Available: <http://doi.acm.org/10.1145/3079856.3080254>
16. S. Zhang, Z. Du, L. Zhang, H. Lan, S. Liu, L. Li, Q. Guo, T. Chen, and Y. Chen, "Cambricon-x: An accelerator for sparse neural networks," in 49th Annual IEEE/ACM International Symposium on Microarchitecture, MICRO 2016, Taipei, Taiwan, October 15–19, 2016, 2016, pp. 1–12. [Online]. Available: <https://doi.org/10.1109/MICRO.2016.7783723>
17. P. Warden, "Low-precision matrix multiplication," <https://petewarden.com>, 2016.



《IEEE计算机图形及应用》(IEEE Computer Graphics and Applications, 简称CG&A)把计算机图形学领域的理论和实践联系在一起。《IEEE计算机图形及应用》提供了包括从某个特定算法到全系统实现在内的同行评议的深度报道。它为那些处于计算机图形技术前沿的人们提供了必不可少的资料。无论他们处于商界还是艺术界, 这本杂志都能让他们受益。

请点击: [www.computer.org/cga](http://www.computer.org/cga)



# 深度学习与物联网

文 | Shuochao Yao, Yiran Zhao, 伊利诺伊大学厄巴纳 - 香槟分校 (UIUC)

Aston Zhang, 亚马逊人工智能研究

Shaohan Hu, IBM Thomas J. Watson 研究所

Huajie Shao, Chao Zhang, 伊利诺伊大学厄巴纳 - 香槟分校 (UIUC)

Lu Su, 纽约州立大学布法罗分校

Tarek Abdelzehher, 伊利诺伊大学厄巴纳 - 香槟分校 (UIUC)

译 | 张彧, 浙江大学

如何将深度学习的优势带入新兴的嵌入式物联网设备领域? 围绕这一问题, 本文作者讨论了嵌入式和移动式深度学习模式中的几个核心挑战, 以及最近的解决方案, 展示了构建以有效高效和可靠的深度学习模式为动力的物联网应用的可行性。

**过** 去十年中, Twitter和Facebook等社交媒体平台已成为在全球范围内广泛的人际网络中进行社交和政治交流的关键工具。各种主题的信息在这些平台上以前所未有的规模交换, 为用户提供了一个开放的、与特定领域无关的交流场所, 同时带来了益处和挑战。

互联网络移动终端和嵌入式设备的普及为物联网 (IoT) 的发展带来了愿景, 也催生出一个充满各种感应设备的世界, 让我们的日常生活越来越因计算、传感和通信功能而变得丰

富多彩。这些功能将有望推动人类和物体之间互动的变革。

事实上, 目前已经有不少重大研究致力于在移动式和嵌入式设备和传感器上构建更加智能、更加用户友好型的应用程序。同时, 近年来, 深度学习的发展极大地改变了计算装置处理一些以人为中心的内容的方式, 这些内容包括图像、视频、语音和音频等。将深度神经网络应用到物联网装置中赋予新一代应用程序执行复杂的传感和识别任务的能力, 以支持人类和物理环境之间互动的新领域。本文围绕人类与深度学

习型物质世界之间新型互动方式讨论了四个关键性研究问题，包括：什么样的深度神经网络结构能够有效地处理和融合不同物联网应用的传感输入数据？如何减少资源消耗，让深度学习模式更有效地应用于资源受限的物联网装置中？如何对深度学习对物联网应用预测的准确度进行置信度量？最后，如何降低深度学习模式对标记数据的需求？

为了详细解释上述问题，首先需要注意，物联网应用装置通常依赖于各种传感器之间的合作，而这则需为多传感器数据融合建立新型神经网络结构。这些结构需要能够模拟各种传感输入数据之间复杂的交互作用，并有效地编码与所需识别和其他任务相关的感官输入特征。为了实现这一目的，我们回顾了一个名为DeepSense的通用深度学习框架，该框架为不同的物联网应用深度学习需求提供了统一且定制化的解决方案。它表明，深度神经网络拓扑的某些组合特别适合从传感器数据中学习。

其次，物联网装置的低端系统在计算能力、能量和内存资源等方面通常会受到限制。将深度神经网络应用于物联网装置中，受训后的深度神经网络模型对于资源的高度要求是一个主要障碍。尽管现有的神经网络压缩算法能够有效减少模型参数的数量，但是并非所有这种模型的矩阵表达都能有效地运用到商品物联网装置中。最近的研究记录

了一种有效的深度学习压缩算法，这种称为DeepIoT的算法能够直接压缩常用于深度神经网络的结构，在缩短执行时间，降低能量和内存的同时几乎不影响最终预测的准确度。

第三，可靠性保证对网络物理和物联网应用也十分重要。可靠性保证要求对和学习结果有关的不确定性进行良好校准的估计。我们将介绍RDppeSense，这是一种能对深度神经网络计算做出经过良好校准的不确定性估计的简单方法。这种方法通过改变目标函数来忠实地反映预测的正确性，从而实现精确和校准的估计。

最后，为了学习的目的而标记信息是十分耗时的，这要求让传感装置能够在没有任何样本、目标和概念的基本真值的条件下，识别目标和概念。因此，可以通过监督的和半监督的解决方法来帮助系统学习有一定标记的（大部分是未标记的）样本，并使结果与完全标记的数据取得相似效果。

我们将详细阐释这些核心问题以及现有的解决办法，以此来为构建一种有效、高效、可靠的物联网深度学习模型奠定基础。

## 传感器数据深度学习模型研究

实现学习型物联网系统的一项关

键性研究挑战在于深层神经网络结构的设计，该结构可以有效地估计来自噪声时间序列多传感器测量的兴趣输出。

尽管在物联网环境下，嵌入式和移动式运算任务种类繁多，但通常可以将它们分为两种常见的子类型：估算任务和分类任务，具体取决于它们各自产生的预测结果是连续的还是分类的。因此，具体问题变成了：是否有一个通用的神经网络架构存在，能够有效地从传感数据中学习估算和分类任务所需的模型结构。从理论上来说，这种通用的深度学习神经网络架构能够克服目前基于分析模型简化或者手工工程特性的缺陷。

在以估算为导向的问题上，例如追踪和定位，传感器输入根据所涉及现象的物理模型进行处理。传感器生成物理量的测量值，例如加速度和角速度。其他的物理量能够通过这些测量数据而衍生出来，例如位移可以通过加速度对于时间的二重积分测算出。然而对于商品传感的测量是很嘈杂的。测量的噪声是非线性的，并且可能与时间相关，很难进行建模。因此将信号和导致估计误差和偏离率的噪声分隔开成为了一项挑战。

在以分类为导向的问题，例如活动和环境识别的问题上，对从未加工的传感数据中产生的特征进行合理估算是一种典型的方法。这些手工特征会被送入分类器进行训练。设计出适当的手工

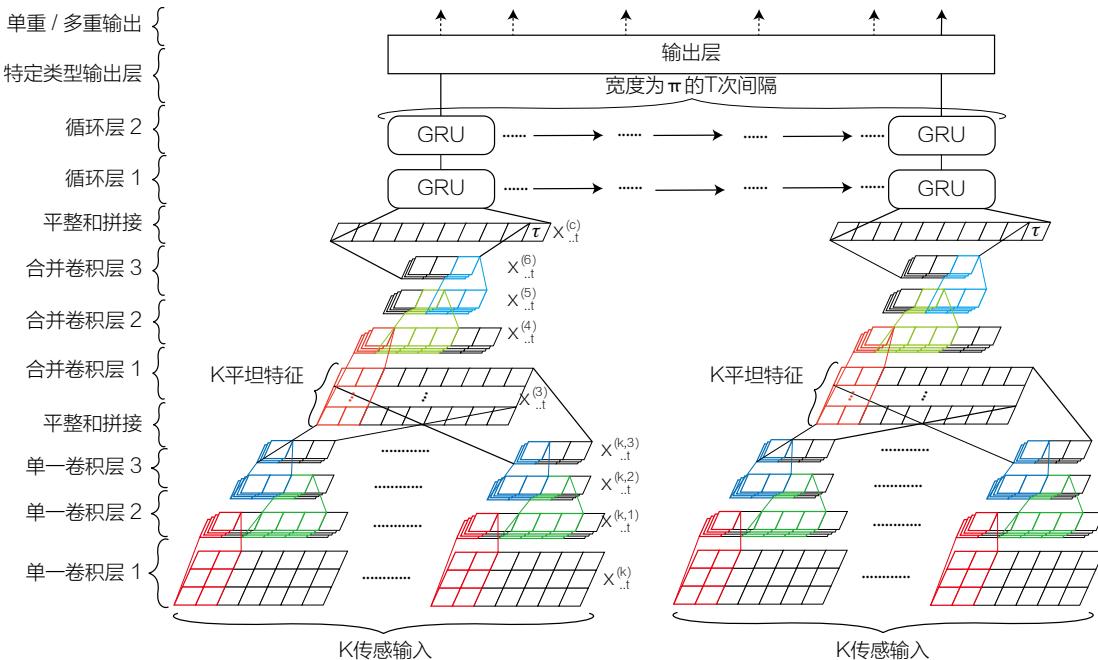


图 1. DeepSense 框架的主要结构。

特征需要花费大量时间，因为这需要大量的实验来归纳各种设置，例如不同的传感噪声模式和异构用户行为。

一个通用的深度学习框架能够有效的处理上述两种问题。它能自动地将学习后的神经网络适应复杂的关联噪声模式，同时集中提取最适合于手头任务的最强健信号特征。最近，DeepSense框架演示了一种具有通用解决办法可行性的案例。

从图1可以看出，DeepSense融合了卷积神经网络(convolutional neural networks, CNNs)和递归神经网络(recurrent neural networks, RNNs)。传感输入被对齐并划分成用于处理时间序列数据的时间间隔。对每一时间间隔，DeepSense首先将单个CNN应用于每个传感器，再对传感器数据流中的相关本地特征进行编码。接着，在各个输出端上应用(全局)CNN，模拟多个传

感器之间的相互作用以实现有效的传感器融合。然后，应用RNN来提取时间模式。最后，我们可以使用仿射变换，或者使用softmax输出，取决于我们需要对估计还是分类任务建模。

这一架构解决了学习多传感器融合任务以从时间序列数据进行估计或分类的一般问题。关于以估算为导向的问题，DeepSense学习了物理系统和噪声模型，从而能够直接从嘈杂的传感器数据中获得输出。而在以分类为导向的问题上，DeepSense则能够作为自动特征提取器编码本地信息、总体信息和时间信息。

作为一个统一模型，DeepSense可以轻松地针对特定的物联网应用程序进行定制。程序的设计者只需要决定传感器输入的数量，输入、输出维度，以及训练目标函数等问题。关于DeepSense的详细数学公式可以在相

关文章中找到。

在将DeepSense应用在两个具有代表性的传感任务的过程中，我们得到了鼓舞人心的结果，这两个任务分别是异构人体活动识别(heterogeneous human activity recognition, HHAR)和基于生物特征运动分析的用户识别(user identification with biometric motion analysis, UserID)。HHAR是基于运动传感器的活动识别任务，它能够对没有在训练集中出现过的新用户进行测试。与此相反，UserID则使用运动传感器从用户的活动，例如步行，骑自行车和爬楼梯等识别用户。

为了更好地理解不同架构成分的作用，DeepSense的多种变异模型将被引入使用，代替通用结构中的一些设计成分。单个门控循环单元(DS-singleGRU)通过使用覆盖更大维度的单层GRU以代替RNN中的双层堆叠式

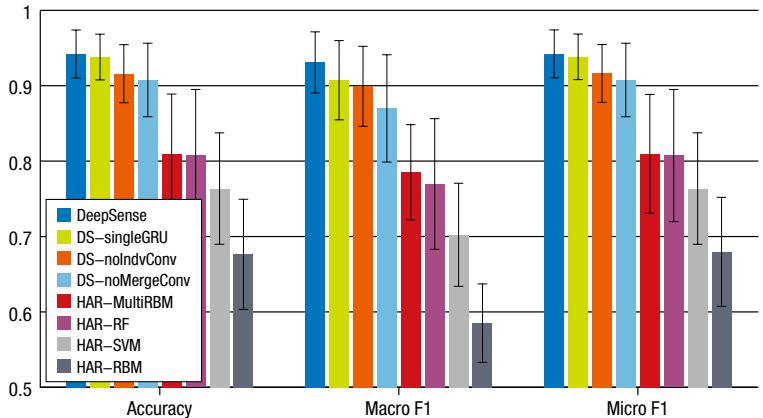


图2. 使用 DeepSense 框架进行 HHAR 的性能指标。

GRU结构,从而简化了RNN。非单一卷积成分( DS-noInConv)跳过单个传感器的卷积子网,在每个时间窗口中保留一个CNN,用于合并来自所有传感器的数据。最后,非合并卷积成分( DS-noMergeConv)跳过了合并传感数据的全局卷积子网,将每个单卷积子网的输出平坦化,并将其连接成单个矢量作为RNN的输入。

这些模型(连同整体DeepSense模型)与各种应用的定制设计或手工制作的基线进行了比较,其中包括HAR-RF, HAR-SVM, HRA-RBM, 用于活动识别的HRA-MultiRBM和用于用户识别的GaitID和IDNet。

图2和图3分别阐释了HHAR和UserID任务执行表现的准确度结果。基于DeepSense的算法(包括DeepSense及其三种变体)在性能上优于其他基线算法(在HHAR任务上优于其他算法至少为10%,在UserID任务上由于其他算法至少为20%)。证据表明,通用的深度学习架构优于为个人应用空间设计的手工解决方案。虽然目前的工作还无法对模型的普遍性进行完美的证明,但这些性能(如果属实)将是非常重要的。因为在物联网环境中应用深度学习的主要吸引之处,在于它避免了对每个应用定制理论推导和手工特性。我们还需要更多的研究来证实或驳斥早期的结果和依据,以进一步理解学习模型

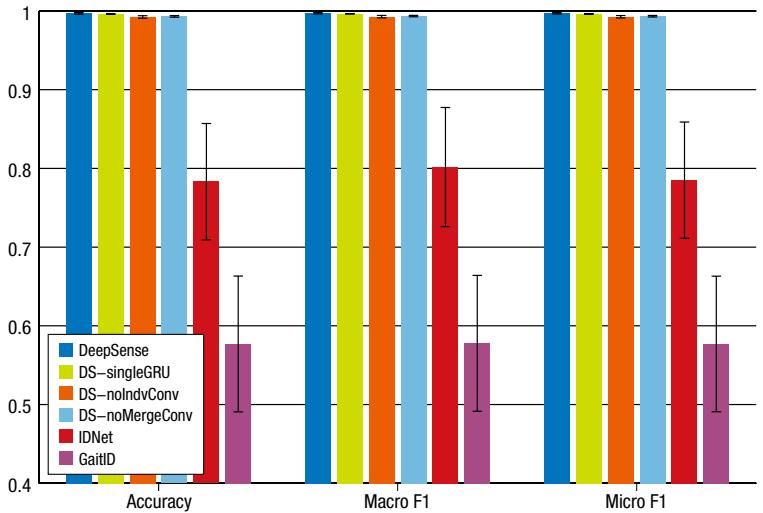


图3. 使用 DeepSense 框架进行 UserID 任务的性能指标。

和物联网系统交叉的普遍化限制。

## 压缩神经网络结构

物联网设备的资源限制也然是执行深度学习模式的重要障碍。因此,关键的问题在于是否有可能将深度神经网络(如前一节中所述)压缩到适合低端嵌入式设备的程度,从而实现与其环境的实时“智能”交互。那么是否有一种统一的方法可以压缩常用的深度学习结

构,包括完全连接神经网络、卷积神经网络、递归神经网络以及它们的组合?压缩的结果能在多大程度上减少实际使用中的能量、执行时间和内存需求?

图4展示了这种称为DeepIoT的压缩框架。DeepIoT从被广泛使用的一种称为“丢弃操作(dropout)”的深度学习正则化方法中,借用了隐藏元素的思想。退出操作给每个隐藏元素一个丢弃的概率。在操作过程中,隐藏元素可根据其丢弃概率进行修剪,由此生成“瘦

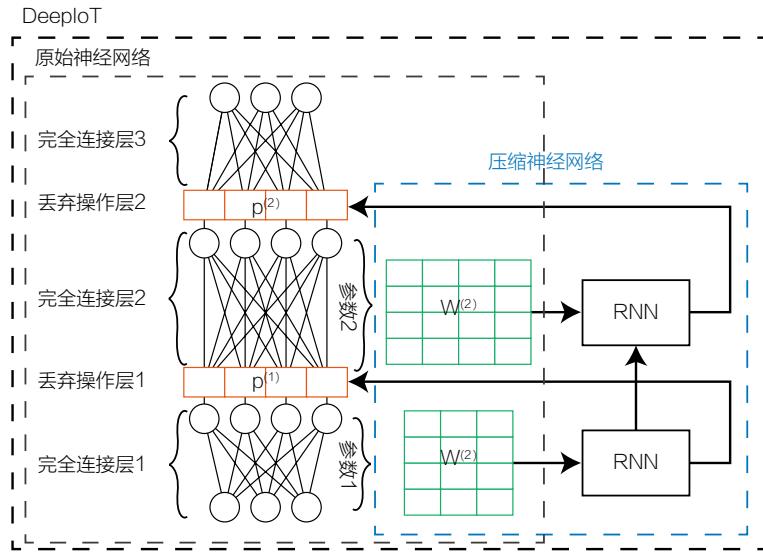


图4. 整体 DeepIoT 系统框架。图中橙色框表示丢弃操作 (dropout operation)，绿色框表示原始神经网络的参数。

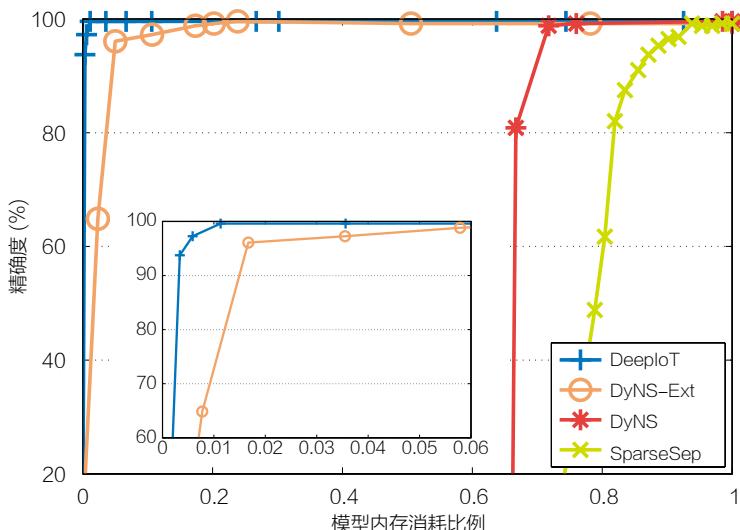


图5. 模型测试精度和内存消耗之间的参数权衡。

身”网络结构。这个过程中面临的挑战是如何设置最合适 的丢弃概率来生成最佳的“瘦身”网络结构，从而实现在最大限度地减少其资源消耗的同时，确保传感应用的准确性。因此，DeepIoT的一个重要目标是找到神经网络中每个隐藏元素的最优丢弃概率。

为了获得神经网络中节点的最优丢弃概率，DeepIoT开发了自己的网络参数。从模型压缩的角度来看，越冗余的元素应当有更高的丢弃概率，而DeepIoT的贡献在于利用新颖的压缩器神经网络来解决这个问题。它将每个层的模型参数作为输入信息，学习参数冗

余度，并相应地生成丢弃概率。压缩器神经网络与原始神经网络一起被优化，以迭代的方式压缩，来试图达到使原始物联网应用的损失函数最小化的目的。

评估结果表明，DeepIoT压缩算法能够在不损害预测精度的前提下，大幅度降低网络规模、执行时间和能耗。我们继续使用UserID作为正在运行的应用程序示例，并将压缩效果与几条基线的效果进行比较，包括DyNS、SparseSep和DyNS-Ext。DyNS是基于量级的网络修剪算法，它根据量级来修剪卷积内核的大小和完全连接的层数。SparseSep通过稀疏编码技术来简化完全连接的层数，并且通过矩阵分解来压缩卷积层。DyNS-Ext将DyNS中使用的基于量级的方法扩展到了循环层。就像DeepIoT一样，DyNS-Ext可以应用于所有常用的深度网络模块，包括完全连接层，卷积层和循环层。所有模型都使用32位浮点数而不进行量化，这些模型在英特尔的爱迪生平台上 (Intel Edison platform) 进行了实验。

图5说明了测试准确度与所得模型的内存消耗之间的详细参数权衡。我们使用了不同的压缩比率对原始的DeepSense神经网络进行了压缩，通过观察最终的测试准确度，最后得出DeepIoT实现了最佳的参数权衡。

图6阐释了执行时间和测试准确度之间的参数权衡。同样的，图7阐释了能

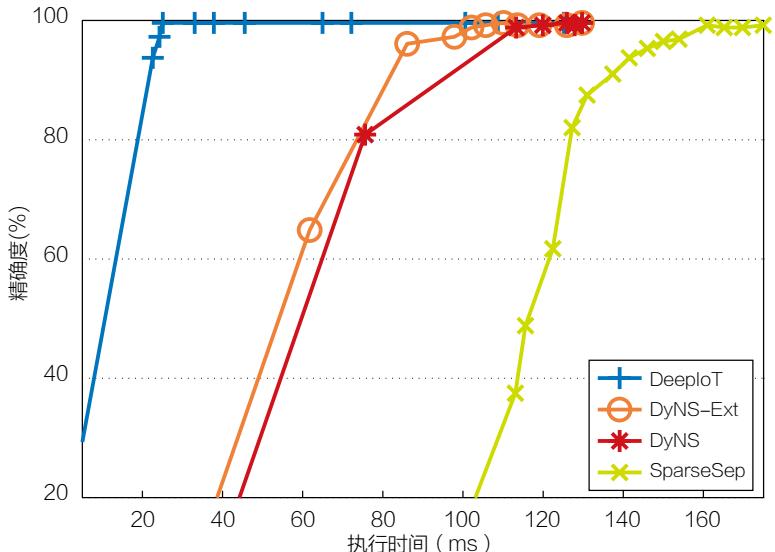


图 6. 测试精度和执行时间之间的参数权衡。

耗和测试准确度的参数权衡。在没有明显精度损失的情况下，DeepIoT 提供了最佳的减少执行时间（约为 80.8%）和降低能量消耗（约为 83.3%）方案。

压缩算法能在不影响准确性的情况下显著缩小网络。这表明，物联网应用的基础模型在本质上是低维的，因此它能使经过学习的神经网络结构得到显著简化。而这对于资源有限的硬件而言是个好消息，例如在上述用于评估的爱迪生平台上实现深度学习的可行性。

## 估算不确定性

下一个问题涉及深度学习模型的可靠性。特别是如何提供原则性的不确定性估计以忠实地反映模型预测的正确性？当深度学习用于支持需要量化可靠性保证的物联网应用程序时，原则性的不确定性估计是至关重要的。

近期的工作主要关注了下面两方面的问题：如何开发能够为深度学习模型的预测结果提供准确的不确定性评估方法？如何为不确定性估计问题开发资源节约型的解决方案，以在资源有限的物联网应用程序上运行？

在这一部分，我们将介绍一种多层感知器的不确定估计算法，叫做 RDeepSens，该方法简单、校准度良好并且效率高，能够在评估不确定性的同时，在理论上验证物联网应用的误差范围。

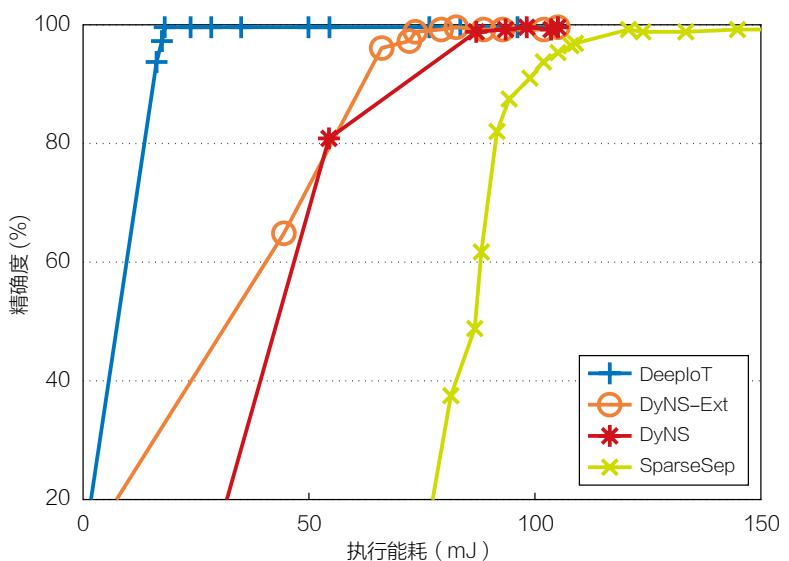


图 7. 测试精度和能耗之间的参数权衡。

计算任意完全连接的神经网络的不确定性只有两个步骤。首先，将丢弃操作插入到每个完全连接层中。第二，采用适当的评分规则作为损失函数，并在输出层使用分布估计方法，而不是点估计方法。

直观地说，丢弃操作将具有参数的传统（确定性的）神经网络转换为具有

随机变量的随机贝叶斯神经网络模型，将神经网络做为统计模型。适当的（基于损失函数的）评分规则将随之测量出概率预测的准确性。

损失函数对最终结果有很大的影响。以回归问题为例，使用均方误差作为损失函数往往低估不确定性关系。这是因为训练过程注重预测准确的平均

表1: NYCOMMUTE任务的平均绝对误差 (MAE) 和负对数似然 (NLL)

Deep learning algorithm	MAE	NLL
RDeepSense	5.64	7.7
SSP-1	8.15	4.86
SSP-3	7.90	4.67
SSP-5	7.51	4.84
SSP-10	7.03	4.81
MCDrop-3	5.69	19,995.6
MCDrop-5	5.64	1,335.73
MCDrop-10	5.61	640.35
MCDrop-20	5.61	640.35
Gaussian Process	11.84	7.46

值而不关心方差。同时，使用负对数似然作为损失函数往往高估不确定性。原因在于，在训练具有对数似然损失的神经网络的早期阶段，对平均值进行精确估计相对较为困难。增加估计方差值可以持续降低可能性极高的负对数似然损失。综上所述，预测的不确定往往是由于高估了真实不确定性的较大方差。

根据负对数似然函数和均方误差的加权和，RDeepSense将可调函数应用为损失函数。通过调整加权总和，均方误差的低估效应和负对数似然的高估效应能够达到平衡。因此，RDeepSense被证明可以产生良好校准的不确定估计。

在资源效率方面，由于RDeepSense在输出层使用了分布估计而不是点估计，它能够在单次运行中进行不确定性估计。基于采样和基于集成的方法要求k样本需要运行k次模型，与这些方法相比RDeepSense大大缩短了执行时间，降低了能耗。

我们评估了RDeepSense和

NYCommute任务中相关基线的不确定性估计的准确性。NYCommute根据出租车乘客的上车/下车时间和位置数据集预测纽约市的通勤时间。

我们还对比了RDeepSense与其他三种基线算法。这三种基线算法分别是MCDrop, SSP, 和高斯过程(GP)，所有基于深度学习的算法都使用具有500个隐藏维度的四层完全连接神经网络。MCDrop算法基于蒙特卡洛丢弃操作(Monte Carlo dropout)。MCDrop与RDeepSense相比，主要区别在于它没有通过适当的评分规则进行优化。MCDrop需要多次运行神经网络来生成用于不确定性估计的样本。我们使用MCDrop-k来表示具有k个样本的MCDrop。而SSP用适当的评分方法来训练神经网络，与RDeepSense相比，

它们的主要区别在于SSP使用集成方法而不是在每层中的应用丢弃操作。SSP需要为集成方法训练多级神经网络。我们使用SSP-k来表示具有k个单独神经网络集合的SSP。GP是基于高斯过程的

算法，用于解释由统计模型生成的不确定性估计的质量。在测试中，我们根据每个算法的预测均值和方差计算z%的置信区间。然后，测量属于这个置信区间的部分测试数据。对于经过良好校准的不确定性估计来说，落入置信区间的测试数据部分应该与z%相近。

比较结果如表1所示，MCDrop-k结果显示了低平均绝对误差(MAE)和高负对数似然(NLL)，而SSP-k结果显示高平均绝对误差(MAE)和低负对数似然(NLL)。MCDrop-k试图最小化均方误差，而SSP-k试图最小化负对数似然。因此，MCDrop-k更多地关注预测分布的均值，而SSP-k更多关注整体的可能性。RDeepSense结合两个目标函数，即均方误差和负对数似然性，在这两者之间寻找一个平衡点。

校准曲线如图8和图9所示。MCDrop-k和SSP-k不是低估就是高估了不确定性，均未能产生高质量的不确定性估计。但是RDeepSense得出了高质量的不确定性估计，并且相比于GP，表现大幅度提高。这些对比结果为精确估计深度学习模型输出的不确定性提供了一条途径。

## 最小化标记数据

对深度学习方法而言，需要大量标记数据是一个一般性缺点。为了从实证

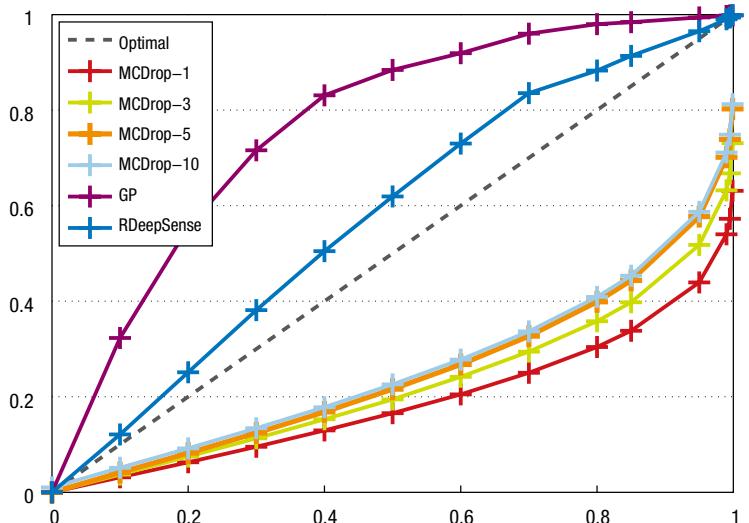


图 8. RDeepSense, GP 和 MCDrop-K 的校准曲线。

测量中更好地学习, 神经网络需要足够数量的标记示例来从中估计网络参数。由于参数的数量很大, 所需的标记示例数量也很大。对在物联网背景下使用深度学习而言, 这种对标记数据的需求成为了巨大的实际障碍, 因为在这些环境中完成标记并不容易。

最近, 生成对抗网络(GAN)作为一种前景良好的深度学习技术被提出, 用于无监督和半监督学习。GAN训练的策略是定义两个竞争网络之间的博弈。发生器网络将噪声源映射到输入空间, 鉴别器网络接收生成的样本或真实的数据样本, 并且必须将这两者区分开来。发生器网络通过训练来欺骗鉴别器网络。我们在测试中将输入概率空间定义为输入传感数据和分类标签的联合概率分布。GAN的训练策略利用未标记的数据来增加发生器和鉴别器网络的容量, 从而明确地提高分类器的识别能力。

评估显示, 称为SenseGAN的半监督策略大大降低了标记数据的要求。我们继续使用HHAR和DeepSense框架作为运行测试的应用程序示例, 并且将整个数据集的p%作为了标记数据。

如表2所示, 半监督训练可以利用90%的未标记数据, 在仅保留10%的标记数据的同时, 确保分类准确性。然而, 在物联网背景下, 探索使用更少量的标记和无标记数据进行训练的可能性仍

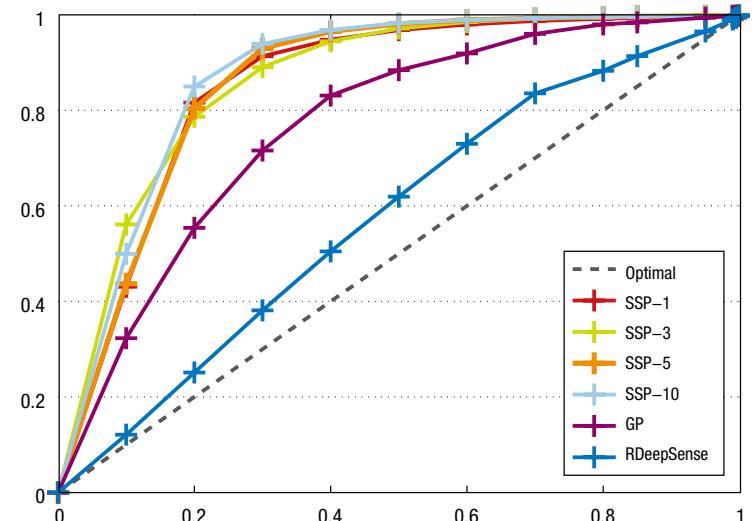


图 9. RDeepSense, GP 和 SSP-k 的校准曲线。

需要更广泛的研究。

我们在文章中介绍了主要的问题和最新的一些解决方案, 这些问题和解决方案为构建在深度学习技术的丰富下有效、高效和可靠的物联网系统的可行性提

出了相关建议。但还需更多的研究来进一步证实结果的适用性。是否可以针对音频信号, Wi-Fi信号和动作输入等主要的异构感觉输入构建统一的深度学习框架? 神经网络压缩对系统性能, 比如执行时间和能耗, 会带来何种影响? 除了多层感知器(MLPs)之外, 是否还能

表2：采用DEEPSENSE框架进行半监督式HHAR训练。

p%	10%	5%	3%	2%	1%
Sense-GAN	94.8%	92.5%	91.4%	90.4%	88.3%
DeepSense	92.0%	89.3%	85.3%	83.6%	79.1%

将不确定性测量扩展到其他深度学习模型？如何在高度动态的环境中，在无法收集大量数据样本的情况下进行学习？这些问题的解决还需更多相关研究。

## 致谢

本文报道的研究部分由美国国家科学基金会(NSF)下的CNS 16-18627和CNS 13-20209津贴资助，部分由陆军研究实验室(Army Research Laboratory)的合作协议W911NF-09-2-0053和W911NF-17-2-0196提供资助。本文中的观点和结论仅代表本文作者的个人意见，不能被解读为陆军研究实验室，美国国家科学基金会或美国政府的官方政策，无论是明示的还是暗示的。本文有版权标记，但是美国政府有权出于政府目的进行复制和传播。

## 参考文献

1. S. Yao et al., “DeepSense: A Unified Deep Learning Framework for Time-Series Mobile Sensing Data Processing,” Proc. 26th Int'l Conf. World Wide Web, 2017, pp. 351–360.
2. S. Yao et al., “DeepIoT: Compressing Deep Neural Network Structures for Sensing Systems with a Compressor-Critic Framework,” Proc. 15th ACM Conf. Embedded Network Sensor System, 2017; <https://arxiv.org/abs/1706.01215>.
3. S. Yao et al., “RDeepSense: Reliable Deep Mobile Computing Models with Uncertainty Estimations,” Proc. ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 1, no. 4, 2018, p. 173.
4. A. Stisen et al., “Smart Devices are Different: Assessing and Mitigating Mobile Sensing Heterogeneities for Activity Recognition,” Proc. 13th ACM Conf. Embedded Network Sensor System (SenSys 15), 2015, pp. 127–140.
5. V. Radu et al., “Towards Multimodal Deep Learning for Activity Recognition on Mobile Devices,” Proc. ACM Int'l Joint Conf. Pervasive and Ubiquitous Computing: Adjunct (UbiComp 16), 2016, pp. 185–188.
6. H. M. Thang et al., “Gait Identification Using Accelerometer on Mobile Phone,” Proc. Int'l Conf. Control, Automation and Information Sciences (ICCAIS 12), 2012, <https://doi.org/10.1109/ICCAIS.2012.6466615>.
7. M. Gadaleta and M. Rossi, “Idnet: Smartphone-Based Gait Recognition with Convolutional Neural Networks,” 2016; <https://arxiv.org/abs/1606.03238>.
8. Y. Guo, A. Yao, and Y. Chen, “Dynamic Network Surgery for Efficient DNNs,” Proc. 30th Int'l Conf. Neural Information Processing System (NIPS 16), 2016, pp. 1387–1395.
9. S. Bhattacharya and N.D. Lane, “Sparsification and Separation of Deep Learning Layers for Constrained Resource Inference on Wearables,” Proc. 14th ACM Conf. Embedded Network Sensor Systems

## 关于作者

**SHUOCHAO YAO**, 伊利诺伊大学厄巴纳-香槟分校 (UIUC) 计算机科学系博士研究生, 上海交通大学信息工程学士, 研究兴趣包括物联网 (IoT) 的深度学习、网络物理系统以及人群和社会感知。联系方式: syao9@illinois.edu。

**YIRAN ZHAO**, UIUC计算机科学系博士研究生, 获得上海交通大学信息工程学士学位, 研究兴趣包括网络物理系统, 机器学习和物联网应用。联系方式: zhao97@illinois.edu。

**ASTON ZHANG**, 亚马逊人工智能 (AI) 应用科学家, 获得UIUC计算机科学博士学位, 研究重点是深度学习。曾在谷歌, 微软研究院, 雅虎实验室, 瑞银集团和自营交易公司实习, 并曾在WWW, KDD, SIGIR和WSDM的计划委员会任职, 也是Apache MXNet / Gluon的深度学习教程的合作者和创始人。联系方式: lzhang74@illinois.edu

**SHAOHAN HU**, IBM Thomas J. Watson研究中心研究员, UIUC计算机科学博士。研究兴趣包括网络物理系统、移动普适计算、人群和社会感知、大数据分析、云计算和量子计算。联系方式: shaohan.hu@ibm.com

**HUAJIE SHAO**, UIUC计算机科学系的博士研究生, 获得浙江大学硕士学位, 研究兴趣包括社交网络数据分析、

应用机器学习、传感器网络和分布式数据中心。联系方式: hshao5@illinois.edu。

**CHAO ZHANG**, UIUC计算机科学系的博士研究生, 获得浙江大学硕士学位, 研究兴趣包括社交媒体分析、时空数据挖掘、文本挖掘、图形挖掘和城市计算。联系方式: czhang82@illinois.edu。

**LU SU**: 纽约州立大学布法罗分校计算机科学与工程系助理教授, 在UIUC获得计算机科学博士, 曾在IBM T. J. Watson研究中心和国家超级计算应用中心工作, 也是ACM和IEEE的成员, 曾获得美国国家科学基金会职业奖, 布法罗大学青年研究者奖, ICCPS 17最佳论文奖以及ICDCS 17最佳学生论文奖。研究兴趣包括一般领域下的移动和人群感知系统、物联网和网络物理系统。联系方式: lusu@buffalo.edu。

**TAREK ABDELZAHER**, UIUC计算机科学系的教授兼Willett Faculty学者, 密歇根大学实时系统服务质量专业博士, IEEE和ACM的成员, 在实时计算、分布式系统、传感器网络和控制方面撰写或合著了200多份参考出版物文献。研究兴趣包括理解和影响网络嵌入式、社交和软件系统的性能和时间特性, 以应对与外部物理环境日益复杂、分布和程度的交互。联系方式: zaher@illinois.edu。

- |  |   |  |
|--|---|--|
| (SenSys 16), 2016, pp. 176–189.  | pp. 1050–1059.  | 12. I. Goodfellow et al., “Generative Adversarial Nets,” 2014; <a href="https://arxiv.org/abs/1406.2661">https://arxiv.org/abs/1406.2661</a> . |
| 10. Y. Gal and Z. Ghahramani, “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning,” Proc. 33rd Int’l Conf. Machine Learning (ICML 16), 2016, | 11. B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles,” 2016; <a href="https://arxiv.org/abs/1612.01474">https://arxiv.org/abs/1612.01474</a> . |  |



# 使用混合边缘到云深度 学习进行隐私的且可扩 展的个人数据分析

文 | Seyed Ali Osia, 谢里夫理工大学

Ali Shahin Shamsabadi, 伦敦大学玛丽皇后学院

Ali Taheri, 谢里夫理工大学

Hamid R. Rabiee, 谢里夫理工大学

Hamed Haddadi, 伦敦帝国理工学院

译 | 周璐莹

尽管信息物理系统和物联网所催生的海量数据收集、整理和分析能力对用户和行业都大有裨益，但这一过程带来了许多挑战，包括隐私问题和可扩展性问题。本文作者提出了一个混合框架，通过以用户为中心的边缘设备与资源来弥补云在提供注重隐私、准确高效的分析上的不足。

**信**息物理系统和物联网（IoT）设备的迅猛发展和大量制造正改变着我们与外界的互动方式。如今，智能设备和环境传感器不断四处收集和传输用户数据，这些数据规模庞大，内容多样，被用于各种目的，包括安全监控、健康监测和城市规划等。大部分物联网设备都默认始终处于联机状态，并通过基于云的机器学习应用程序从收集的数据中洞悉一切。

企业云计算服务拥有按需给予、高性能、高效的计算能力，大幅降低了成本。除却这些优势，云计算也带来了某些挑战。未来十年，智能家居和智能汽车将从数百个传感器上传海量数据到云处理器，到那时，手机、宽带带宽和效率将成为主要瓶颈。这些基于云的模型也会对边缘设备施加重大能源限制。

基于云的系统导致的另一个重要威胁是隐私问题——用

户在分享个人信息，允许服务提供商收集、分析或变现这些信息时，有暴露自己敏感数据的风险。例如，大多数基于云的移动应用程序都是免费的，它们从用户的个人数据中收集信息，从而进行针对性的广告投放。这种做法涉及大量用户隐私，对用户资源有许多影响。<sup>1</sup> 基于云的机器学习算法可以提供有益的服务（例如，健康类的或基于图像的搜索应用），但是它们对过度收集数据的依赖可能会导致一些不为用户所知用户的结果（例如，通过面部识别进行有针对性的社交广告投放）。

最近，边缘计算作为应对这些挑战的解决方案被提出。它将处理能力部署在离终端用户更近的边缘节点上——类似于位于网络边缘的雾计算。通过这种方式，延迟敏感数据可以在边缘节点上得到分析，而云服务则被用于更多延迟容忍任务。但是，分析服务商或应用程序提供商可能并不乐意共享其宝贵的数据处理模型。我们不能总是假设本地处理（例如，在智能手机或计算机等边缘设备上部署深度学习模型）是可行的解决方案，即使用户不看重任务时间、内存和处理要求，即使任务在用户不活跃期间（例如，设备正在充电时）也可以执行，也是如此。

有人可能会说完全基于加密的算

法是理想的解决方案；然而，对于许多物联网应用，尤其是那些依赖机器学习模型或那些要求持续可用或持续在线的模块（如多媒体应用或无人驾驶汽车中的传感器）而言，其加密方法是非常复杂的。深度模型由非线性复杂函数构成，其加密方法更为复杂。这些模型很难用多项式函数来估算，而多项式函数是基于同态加密的方法的重要组成部分。<sup>2</sup>

一方面，将数据完全转移到云服务在现在或将来都可能会带来可扩展性风险和隐私风险；另一方面，依靠在用户端执行全部分析的技术有自己的资源限制（如存储和带宽限制，能源限制或计算成本），用户体验也受到损害。

本文提出了一种边缘到云的混合架构，其中数据处理是由隐私边缘数据处理单元和云服务协作完成的。通过这种方式，我们可以利用边缘预处理来解决隐私问题，而终端用户也能从高效的云处理中获益。该框架示意图如图 1 所示。

我们的研究重心是找到一个折中方案，平衡私人边缘节点上需要大量资源的本地分析与需要大量数据、侵犯了隐私的基于云的服务。在边缘节点上进行的处理量被降到了最低，这保护了隐私，而其余的处理都在云中进行。我们的主要目标是将特征提取

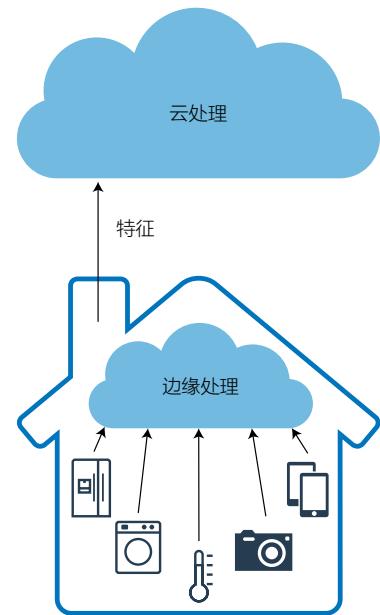


图 1. 保护隐私的机器学习的混合边缘到云框架。在私人边缘节点本地收集和处理用户数据，以此保护敏感信息。独立于敏感信息的数据表征被传输至云数据中心，用以运行复杂的推理。

阶段和推断阶段分离开来；前者发生在本地，后者发生在云端。通过这种方法，数据中的敏感信息可以在边缘节点的特征提取阶段被去除，同时传输到云端的数据比率也会有所降低。提取所得特征被传输到云服务器进行后处理，随后，用户从云端收到结果。

## 实际应用

计算机视觉、机器学习和云计算技术的进步为一大批多媒体物联网服务提供了新的机会。<sup>3</sup> 本文探讨了这些基于云的多媒体物联网应用在以下领域面临的隐私挑战。

> 图像处理。除了以图像为中心的社交媒体的普及，智能手机相机和传感器的质量也日益提高，由此，

各种图像分析应用程序应运而生，如场景标记、图像分类、人脸识别、人脸属性预测、年龄估计、性别分类和情绪检测等。

> 视频处理。无处不在的闭路电视摄像机表明了视频录制、视频索引和视频处理的重要性。出于各种目的，许多家庭和室外场所都配备了视频监控系统来捕捉视觉信息。

海量数据处理，通常这些数据处理通过机器学习算法来实现。思考一个分类问题，比如人脸识别。这一分类模型应该用一个大型的数据集来训练，该数据集由标有个人身份的脸部照片组成。经过训练，该模型可以标明该照片主人的身份。一般来说，机器学习问题可以是监督式的，无监督式的或半监督式的。在监督式问题中，训

在所有这些应用中，运营商可能会担心在宽带边缘或移动网络边缘产生的大量物联网数据的传输问题，而客户则担心其敏感信息的潜在泄露问题。在许多应用程序中，很大部分个人数据不需要被服务提供程序识别。<sup>5</sup> 在监控或分析应用程序中，个人身份已经是被收集的信息中最敏感的一部分了。例如，尽管应用了分类或光学字符识别技术，安装在停车场内的车牌识别相机不应该识别路过的行人的身份。换句话说，人们可能希望得到保护，免遭不想要的面部识别。同理，使用物联网设备语音提示的人可能不希望自己的身份通过语音会话被识别。健康分析也会产生隐私问题，用户可能不想透露自己的隐私信息。

## 在监控或分析应用程序中，个人身份已经是被收集的信息中最敏感的一部分了。

例如，护理设备中会安装智能相机来监护患者，此外自动驾驶车辆要安装许多相机来实现安全运行。

> 语音处理。在物联网领域，语音越来越多地被运用到人机交互之中。许多智能电视、电话、手表、烤箱和灯具都有声控功能。像 Google Home 和 Amazon Echo 这样的设备日益以智能助理的身份进入家庭。在未来几年里，语音识别系统将成为人们日常生活中不可或缺的一部分。

所有这些应用程序都需要复杂的

练习数据的真实标记是可获得的——其目标是预测测试数据的标记，类似于人脸识别的例子。

本文的关注点是监督式应用程序，尤其是分类程序。感兴趣的读者可以参考毕肖普（C.M. Bishop）的《模式识别和机器学习》（Pattern Recognition and Machine Learning）<sup>4</sup> 了解更多关于机器学习的知识。当真实标签不可得时，这一问题被称为无监督学习或聚类。当有少量标记数据和大量未标记数据时，半监督方法会用未标记数据来改善基于标记数据的监督式分类的结果。

这些隐私问题显示了能够解决隐私问题，并且在机器学习应用中有着悠久历史的通用框架的价值和重要性。训练数据的隐私性已在一些著作中得到解决——例如，查鲁·阿革瓦（Charu C. Aggarwal）和菲利普·余（Philip S. Yu）调查了涉及公共数据库隐私性的经典方法，例如随机法和 k-匿名法。<sup>6</sup> 此外，研究者为将差别隐私应用于学习模式做出了不少努力。<sup>7</sup> 例如，雷扎舒可利（Reza Shokri）和维塔利·施玛蒂科夫（Vitaly Shmatikov）<sup>8</sup> 以及马丁·阿巴迪（Martin Abadi）和他的同事们<sup>9</sup> 都试图让深度模型具有差别隐私。尽

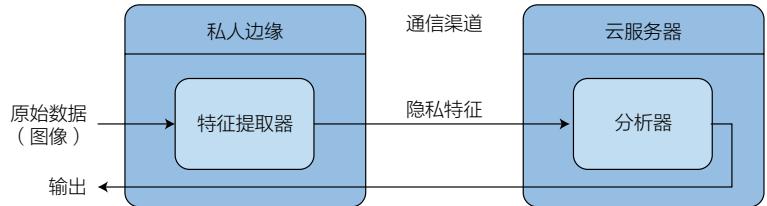
管如此，在测试阶段，用户数据的隐私受到关注较少，而这正是本文的主要关注点。

## 框架提取

假设我们想要通过云服务执行一项主要任务（如语音识别或图像分析）。由于本地处理能力有限或与商业考量冲突，我们可能会遇到一些限制。我们还希望保留用户的敏感信息（比如，一个人的身份可以通过他的声音，或者他在人行道上行走时闭路电视记录下来的镜头被识别）。因此，与云服务共享的数据应具有两个重要特性：推断这一主要任务是可能的，而敏感信息是不可能的。

在云中共享数据使对敏感信息进行进一步推断成为可能。基于边缘的原始数据预处理可以避免不想泄露的数据特征被泄露，但是由于客户端的各种限制，这样的任务需要将负担降至最低。为此，我们提出了一个通用混合架构，该架构包含两个主要模块：特征提取器和分析器。前者构建在私有边缘节点（如个人计算机或家庭机顶盒）上，而后者存储在云中。

这些模块及其交互方式如图 2 所示。客户端设备的数据在私有边缘节点上被收集，然后被发送到特征提取器。特征提取器获取输入数据，给数



**图 2.** 本文所提框架的各个模块。云服务器内的分析器可以获得一组删减后的由特征提取器提供的数据的隐私特征。

据应用函数，接着输出一组新的中间特征，随后这些特征将被转移到云中用以执行主要任务。分析器接收中间特征，推断出主要信息，并在需要时将结果返还给客户端。

在这个框架中，设计一个好的特征提取器至关重要。中间特征需要在保护敏感信息的同时，保留有关主要任务的必要信息。特征提取器在本地运行，因而它不应是一个复杂的例程。因此，设计这个模块是项重要且富有挑战性的任务。

作为用例，请思考某个图片标记云服务，其中个人身份可以通过实时视频流的图像被识别。在这一案例中，一个简单的特征提取器就可以检测人脸，并用阴影替换人脸。分析器接收到删改后的图像，然后执行图像标记程序（例如，给这一图像标注标签）。另一个常见的例子是语音识别，人们可能会担心个人身份因为自己的声音而被识别。一个简单的解决方案就是

改变特征提取器中语音的音高频率，从而达到匿名的效果。在上述两个案例中，特征提取器的设计过程都很简单，且不会影响分析器的结果；然而，情况并非总是如此。

在这些案例中，部分包含敏感信息的数据被去除，而其余数据则被视为中间特征。但是，当要去除的部分包含关乎主要任务的重要信息时，这种方法就不适用了。例如，遮挡某个脸部区域在去除敏感信息（身份）的同时，也去除了脸部属性（如情绪或性别）。因此，当主要任务是人脸属性预测等时，我们就不能使用这种方法。

当主要信息和敏感信息交织联结时，我们就遇到了一种复杂的情况。此时，虑及主要任务，我们应该设计一种用于去除敏感信息的特征提取器。在我们的框架中，我们提出了一种基于深度学习的方法，在设计过程中兼顾主要任务和敏感信息。假设

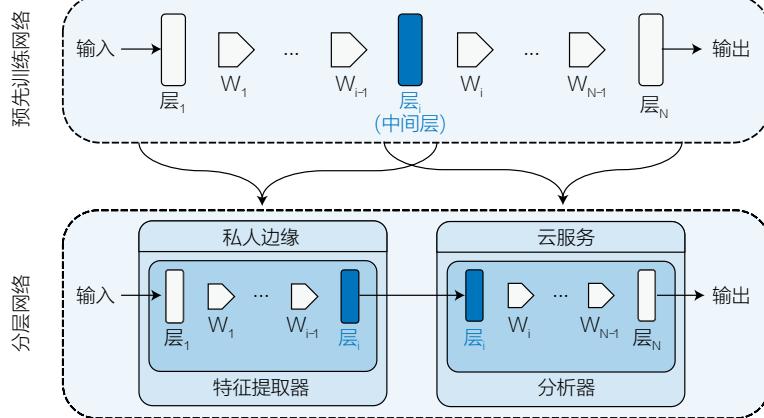


图 3. 分层机制。深度网络的基础层对应特征提取器，而模型的其他部分则被视作是分析器。

服务提供程序知道敏感信息的类型（如身份），则下列设想可以发生：服务提供程序把特征提取器转移到客户机上。这就确保能够同时兼顾主要任务和敏感信息。虽然服务提供程序不必分享分析器，但它必须为隐私保护定义一个验证方法。此过程定义了服务提供程序应当采用的隐私标准。

## 深度学习应用程序

深度神经网络（DNNs）在机器学习领域已是广为流行，尤其是在多媒体应用程序中。<sup>10</sup> 它们提供高精度的分类器，可从原始数据中提取出高级信息。深度网络由不同单层网络叠加组成。每层网络都是前一层的一个简单函数，代表了比前一层更复杂的概念。

初始层是原始输入数据，而最后一层给出推理结果。所有层加在一起形成一个复杂的函数，被应用于输入数据，然后得出一个感知推理。在训练阶段，通过对训练数据应用优化方法可以习得中间函数。模型训练完毕后，就可以对任意输入数据执行推理。

卷积神经网络（CNNs）和递归神经网络（RNNs）是两种最著名的使用于多媒体应用的结构。前者适用于图像和视频处理，后者主要用于序列化数据的处理（如文本和视频）。本文，我们重点研究卷积神经网络，它是最受欢迎的图像和视频处理结构。假设主要任务的相关推理是通过预先训练好的深度网络（如一个即用型多层网络）完成的。我们将通过如下方式解决如何在我们所提出的边缘到云的框

架中嵌入这一训练模型的问题。

## 层结构分离

在深度模型中，网络层级越高就越明确地针对主要任务，同时也丢失了其他包含我们所担心的敏感信息的不相关信息。基于这一观察，我们为预先训练好的深度网络设计了一个分层机制。

> 首先，选择一个中间层作为分割点。

> 接着，将中间层之前的层存储到边缘作为特征提取器。

> 最后，将中间层之后的层存储到云作为分析器。

在选择中间层时有一个权衡之法——从较高层选择能够更好地保护敏感信息的隐私性，但也会增加客户端的计算成本。我们在之前的研究中对不同层的隐私 - 复杂度权衡情况进行了详细分析，同时，也探讨了如何基于边缘设备资源和用户隐私限制选择合适的中间层。<sup>11</sup>

我们将边缘和云之间层结构的简单分割称为简单嵌入式，如图 3 所示。

## 孪生嵌入

在向服务器显示中间特征过程中，为了增加隐私性，我们可以用一个特

定的方法微调现有的用于执行主要任务的深度模型。微调是一个常见的训练深度模型的任务。我们从一个预先训练的深层模型开始，继续对其进行训练以达到某个预期目标。最后，得到一个可用于分层机制的更新版模型。

本文提出方法的主要创新点在于利用了基于所选中间层的孪生结构，<sup>12</sup> 以此来微调主要任务模型。孪生结构是一种常用的训练学习模型的方法。面部验证应用程序常用这种结构来确定两幅图像是否属于同一人。孪生网络的核心思想是使相似数据点的表征（同一个人的不同脸部图像）靠近彼此，而不同数据点的表征（不同人的脸部图像）远离彼此。

为了实现这个目标，我们的训练数据集应该由成对的数据点组成，每对数据点可以是相似，或不相似的。给一对数据点中的两个数据点添加同一个函数，并计算两个输出的距离。通过一个对比损失函数完成优化。在这一损失函数中，两个不同数据点的距离最大，两个相似数据点的距离最小。这种方法使特征提取器更加隐私化，保护用户免遭云上的推理攻击。我们将其称为为孪生嵌入。

## 孪生隐私性

如何将孪生结构与隐私性联系起来？假设我们的主要任务是依靠一个预

先训练过的深度模型，根据面部肖像进行性别识别。个人身份是敏感信息，不应由于使用中间数据（例如，被面部识别系统）而被泄露。在上述场景中，我们唯一关心的是脸部肖像的性别而非其身份。我们可以定义一个新的相似性标准来将这个事实做成模型，然后用对比损失函数来微调我们的模型。所有具有相同性别的身份都被视作是相似的，不仅性别识别模型更加稳固了，更多来自中间特征的身份信息也被清除了。用这种方法微调之后，男性表征彼此非常接近，女性表征亦是如此，但男女表征之间相差甚远。

图 4 为隐私保护微调结构图。通过定义合理的相似性标准，我们将此理念应用于任何应用。实验表明，使用孪生嵌入在保护隐私的同时保持了主要任务的准确性。

## 降维

所有基于云的服务都要面对一个重要问题，即通常过高的通信成本。我们通过降低中间特征的维度来解决这一问题。

从可视化到特征提取，降维被用于一系列数据和机器学习应用。数据的降维可以通过一些线性或非线性的，降低高维空间维度的变换实现。一个最常用的降维方法是主成分分析（PCA）。主成分分析采用的是线性变换。此外，

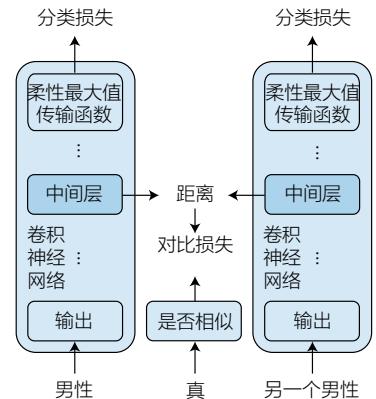
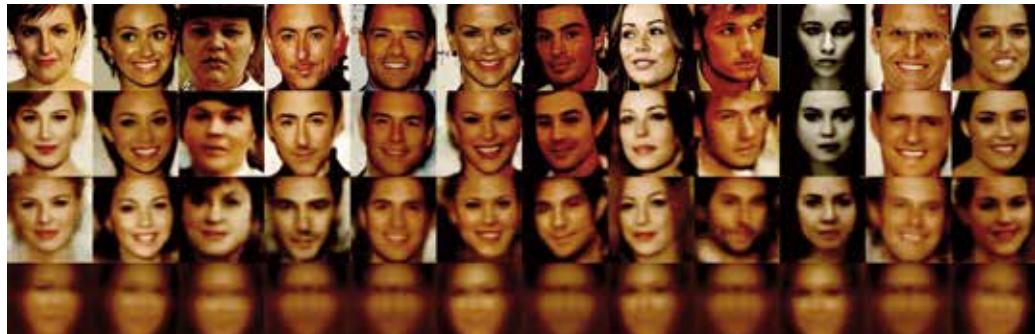


图 4. 对主要任务的孪生微调。通过两个完全相同的卷积神经网络分别提取两张男性面部图像的中间特征。它们应当紧挨彼此，因为它们被认为是相似的。

降维和重建过程可以通过矩阵乘法来实现。

孪生微调大大巩固了特征空间，在微调后的空间上进行主成分分析不会明显降低主要任务的准确率。在没有明显降低主要任务的准确率的基础上，对中间特征空间进行降维还带来了两大优势：极大地降低了边缘到云的通信成本；基于降维 - 重建过程的固有性质，大幅提高了隐私性。

如下为对中间特征进行主成分分析的过程。服务提供程序分别在特征提取器的末尾和分析器的开头添加主成分分析投影和重建。提取得到的中间特征将是一个低维向量，只需少量通讯成本就能被转移到云。在使用这两种方法的过程中，我们引入了高级



**图 5.** 通过可视化手段比较不同的隐私保护深度模型。所有模型的性别分类准确率都很相似，即 93%。第一行展示的是原始图像，剩下三行展示了从中间表征重构而得的图像。在所有重构的图像里，个人性别的识别结果都与原始图像的识别结果一致。此外，从简单嵌入到高级嵌入，个人的身份特征被慢慢移除，说明高级嵌入在隐私保护方面效果最佳。

嵌入的概念。其中李生微调被用作预处理，而主成分分析投影则被添加在中间特征上。

## 初步评估

我们针对人脸图像做了大量实验，将性别分类问题作为主要任务，将个人身份作为要保护的敏感信息。我们评估了每一种嵌入方法中中间特征有关性别和身份的信息量。我们使用了一种直观的可视化技术，展示了从中间数据表征重建原始图像的可能性有多高。我们对自己以前的研究方法进行了更严格的分析，提出了一个隐私措施，用以正式量化此框架保护敏感信息的能力。<sup>11</sup>

为了比较不同的深度嵌入方法，我们采用了拉斯马斯·罗特 (Rasmus

Rothe) 和他的同事们提出的性别分类模型。<sup>13</sup> 这个模型是一个 16 层的卷积神经网络，它采用了常用的 VGG-16 结构。罗特和他的同事建立了一个大型人脸数据集，其中包括从互联网电影数据库 (IMDB) 和维基百科收集而来的年龄和性别属性。他们建立的模型在维基百科图像上的准确率高达 93%。为了公平比较起见，我们的实验也是在这个数据集上进行的。

我们选择第五个卷积层作为我们的中间特征，以此构建特征提取器。与基于本地设备的解决方案相比，我们的方法平均可将内存使用量降低 50%，并将智能手机负载降至 20% 以下，这证明了我们提出的混合方案具有高效性。

简单的嵌入只需分层即可。李生嵌入则需要微调预先训练的模型，然

后再进行分层。高级嵌入的步骤同上，另外再加一个主成分分析处理。我们将中间特征的维度缩小到了<sup>8</sup>。我们分析了性别分类（主要任务）的准确性和身份（敏感信息）的隐私性之间的权衡。令人惊讶的是，这些模型的平均准确率几乎与性别分类的准确率持平（93%）。因此，他们在完成主要任务方面表现相似。于是，这次比较研究唯一的关键问题在于，这些模型通过身份保护能力保留更多隐私的能力如何。

我们通过可视化技术比较了这些方法的隐私保护能力。可视化试图回答一个关键问题：如果只使用深度网络的中间层，则原始输入图像最多能被识别什么程度？阿列克谢·多索维斯基 (Alexey Dosovitskiy) 和托马斯·布洛克斯 (Thomas Brox) 训练了一个

解码器来回答这个问题——他们把中间层作为解码器的输入，并将生成的图像作为期望输出。<sup>15</sup> 我们用他们的方法比较了不同深度模型的结果（它虽然算不上卓越性能的严格证明，但非常直观）。

图 5 显示了用不同的方法从中间特征还原原始图像的结果。图 5 表明，在简单嵌入和孪生嵌入中，所有图像的性别与原始图像的性别一致。由于性别分类的准确性，尽管我们更难将其与重构图像区分开来，高级嵌入的图像性别同样与原始图像一致。简单嵌入几乎可以复原原始图像。因此，仅对深度网络进行分层不能确保获得可接受的隐私保护性能。由于面部的固有特征（例如骨架），在使身份失真方面，孪生嵌入比简单嵌入效果更好。高级嵌入给出的结果是最好的，因为解码器不可训练，而且从中间图像推导不出任何信息，包括个人的身份。这种方法的一个优点在于，与其他方法相比，其通信成本可以忽略不计，因为我们只需要将八个实数上传到云。我们之前的研究对此作了更详细的分析。<sup>11</sup>



前，我们的框架是为预先训练的机器学习推理而设计的。我们当下的研究目标是设计一个作为业务的机器学习

的框架，从而推广我们的方法<sup>16</sup>。在这个框架中，用户可以在隐私得到保护的情况下共享他们的数据，从而在云服务器中训练新的学习模型。我们的框架的另一个潜在发展方向是为其他类型的神经网络（如递归神经网络）提供支持。这对处理时序数据和序列化数据大有帮助。■

## 参考文献

- N. Vallina-Rodriguez et al., “Breaking for Commercials: Characterizing Mobile Advertising,” Proc. 2012 Internet Measurement Conference (ICM 12), 2012, pp. 343–356.
- R. Gilad-Bachrach et al., “Cryptonets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy,” Proc. 33rd Int’l Conf. Machine Learning (ICML 16), 2016, pp. 201–210.
- I.F. Akyildiz, T. Melodia, and K.R. Chowdhury, “A Survey on Wireless Multimedia Sensor Networks,” Int’l J. Computer and Telecommunications Networking, vol. 51, no. 4, 2007, pp. 921–960.
- C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
- A. Chaudhry et al., “Personal Data: Thinking Inside the Box,” Proc. Fifth Decennial Aarhus Conf. Critical Alternatives (AA 15), 2015, pp. 29–32.
- C.C. Aggarwal and P.S. Yu, “A General Survey of Privacy-Preserving Data Mining Models and Algorithms,” Privacy-Preserving Data Mining, 2008, pp. 11–52.
- C. Dwork, “Differential Privacy: A Survey of Results,” Int’l Conf. Theory and Applications of Models of Computation (TAMC 08), 2008, pp. 1–19.
- R. Shokri and V. Shmatikov, “Privacy-Preserving Deep Learning,” Proc. 22nd ACM SIGSAC Conf. Computer and Communications Security (CCS 15), 2015, pp. 1310–1321.
- M. Abadi et al., “Deep Learning with Differential Privacy,” Proc. 2016 ACM SIGSAC Conf. Computer and Communications Security (CCS 16), 2016, pp. 308–318.
- I. Goodfellow, Y. Bengio, and A.

## 关于作者

Courville, Deep Learning, The MIT Press, 2016.

11. S.A. Osia et al., “A Hybrid Deep Learning Architecture for Privacy-Preserving Mobile Analytics,” preprint, 2017; <https://arxiv.org/abs/1703.02952>.

12. S. Chopra, R. Hadsell, and Y. LeCun, “Learning a Similarity Metric Discriminatively, with Application to Face Verification,” IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR 05), 2005, doi: 10.1109/CVPR.2005.202.

13. R. Rothe, R. Timofte, and L. Van Gool, “DEX: Deep EXpectation of Apparent Age from a Single Image,” 2015 IEEE Int'l Conf. Computer Vision Workshop (ICCVW 15), 2015; doi: 10.1109/ICCVW.2015.41.

14. K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” preprint, 2014; <https://arxiv.org/abs/1409.1556>.

15. A. Dosovitskiy and T. Brox, “Inverting Visual Representations with Convolutional Networks,”

**Seyed Ali Osia**是谢里夫理工大学计算机工程学院人工智能方向的在读博士。他的研究领域包括统计机器学习、深度学习、隐私和计算机视觉。他的联系方式: osia@ce.sharif.edu。

**Ali Shahin Shamsabadi**是伦敦大学玛丽皇后学院智能遥感中心深度学习和隐私方向的在读博士。他的研究领域包括在深度学习, 以及分散式和集中式学习中的数据隐私保护。沙姆沙巴蒂获得了谢里夫理工大学的电气工程学(数字化方向)理科硕士学位。他的联系方式: a.shahinshamsabadi@qmul.ac.uk。

**Ali Taheri**是谢里夫理工大学计算机工程学院的硕士生。他的研究领域包括深度学习和隐私。塔赫里获得了谢里夫理工大学人工智能方面的理科硕士学位。他的联系方式: ataheri@ce.sharif.edu。

**Hamid R. Rabiee**是一位计算机工程教授, 是谢里夫理工大学通信技术研究所(AICT)、数字媒体实验室(DML)和移动增值服务实验室(MVASL)的主任。目前, 他正处于学术休假期, 在伦敦帝国理工大学担任访问教授。拉比的研究领域包括统计机器学习、贝叶斯统计学、多媒体系统中的数据分析和应用程序复杂网络、社会网络、云和物联网数据隐私、生物信息学和大脑网络。他获得了普渡大学电子与计算机工程方面的博士学位。拉比还是美国电气和电子工程师协会的高级会员, 拥有三项专利。他的联系方式: rabiee@sharif.edu。

**Hamed Haddadi**是帝国理工大学戴森设计工程学院(Dyson School of Design Engineering)的副教授, 以及数据科学研究所(Data Science Institute)的学术成员。他的研究领域包括用户中心体系、物联网、应用机器学习和数据安全及隐私。哈达迪享受设计和构建能够更好地利用我们的数字足迹, 同时又尊重用户隐私的体系的过程。他获得了伦敦大学学院电子工程方面的博士学位。他的联系方式: h.haddadi@imperial.ac.uk。

preprint, 2015; <https://arxiv.org/abs/1506.02753>.

16. S. Servia-Rodriguez et al., “Personal

Model Training under Privacy Constraints,” preprint, 2017; <https://arxiv.org/abs/1703.00380>.

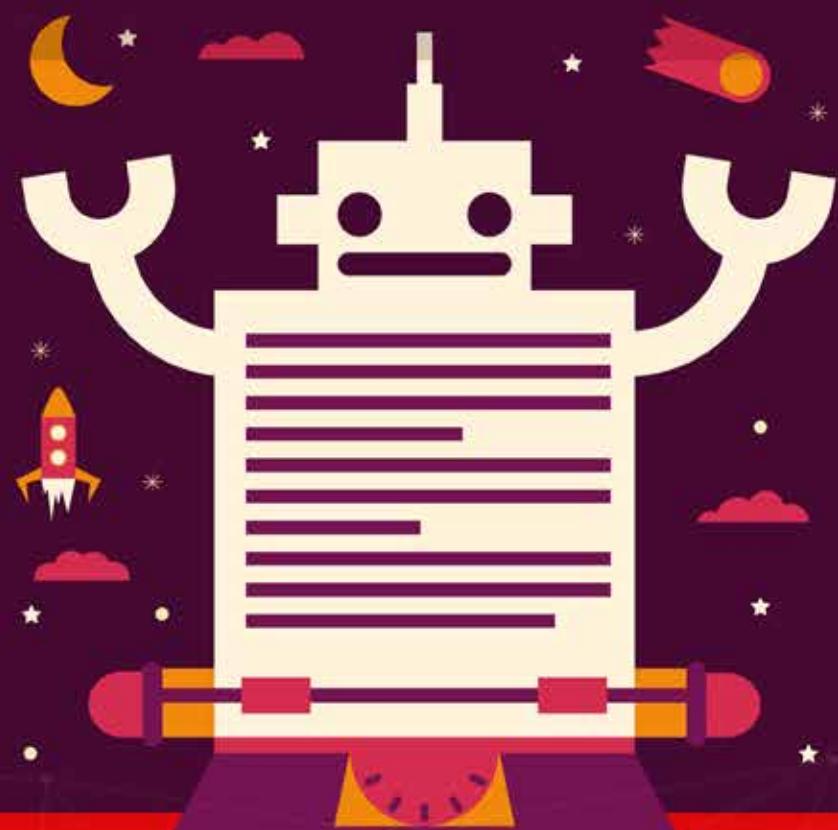
SMP EUPT 2018

# SMP & 今日头条

猜猜今日头条的文章作者

是人、写作机器人、翻译机器人还是摘要机器人？

奖金: ￥29000



SMP

X

 TOUTIAO AI LAB  
今日头条人工智能实验室

报名入口





# 基于深度学习的人类行为识别在移动计算领域的使用

文 | Thomas Plotz, 乔治亚理工大学 (Georgia Tech)、  
Yu Guan, 英国纽卡斯尔大学 (Newcastle University, UK)  
译 | 黄美桃

通过利用深度学习所取得的进步, 具有挑战性的模式识别问题已经在计算机视觉、语音识别、自然语言处理等方面得到解决。移动计算也采用了这些强有力的数据建模方法, 在该领域的核心应用域取得惊人的成功, 其中包括通过机器学习持续改变人类行动识别 (HAR) 技术。

**深**度学习技术以其他学习技术所没有的方式彻底革新了机器学习及其应用程序。<sup>1</sup>这一转变基于三项主要突破, 这些突破均在2000年代中期出现。首先, 开发了非常有效的建模技术, 尤其是预训练技术, 这些技术变得非常流行, 使研究人员能够引导更为复杂的模型架构, 并产生了可训练的分层数据表达法。然后, 开发了大规模并行计算架构, 例如(组合许多)图形处理器(GPU);这些架构已经成为主流并且得到广泛应用, 此外, 通过基本上消除计算约束和限制, 从而大大降低了训练复杂模型的阻碍。最

后, 对于许多应用领域而言, 注释样本数据的大量数据库越来越适用, 消除了学习复杂且深层次模型的另一个障碍。这些突破令深度学习研究激增, 在应用领域取得了令人印象深刻的进展, 这些应用领域的研究艰难并且具有挑战性, 其中包括计算机视觉、语音识别和自然语言处理等。

移动计算旨在将计算融入日常环境中, 使其始终为所有人提供服务。移动计算的关键组成部分, 特别是情景感知, 是对用户正在做什么进行自动评估;这就是HAR(HAR), 并且通过分析移动平台收集的传感器数据来实现。通过移动计算

实现的HAR有值得注意的限制和要求：移动平台增加了大量的资源限制；往往需要近实时的推断；部分需要处理的传感器数据具有挑战性，其中包括噪音、模糊不明确、数据缺失等；通常，只有非常有限的标签数据可用于训练HAR模型；而地面实况标志往往质量混杂甚至模糊不清，这使得训练和验证变得复杂。过去，已经开发了大量方法以利用和扩展传统机器学习方法以解决在这些硬性限制下HAR使用移动平台的问题（例如，参见Andreas Bulling及其同事的教程<sup>2</sup>）。多年来，已经开发出的各种系统在真实世界的应用场景（例如医疗、运动或一般的人机交互）中使用强大的HAR，这让人印象深刻。

然而，HAR在移动计算方面存在大量挑战，这在一定程度上限制了使用传统机器学习技术可能取得的进步。随着深度学习方法不断取得成功和普及，如今在移动计算场景中利用这些技术的要求也越来越高。从最初探索深度学习技术如何帮助克服为惯性测量装置数据找到适当的特征表达这一问题开始，<sup>3</sup> 利用深度卷积神经网络开发端到端的HAR系统；<sup>4</sup> 或实际序列模型，<sup>5, 6</sup> 到最近探索的迁移学习；<sup>7</sup> 深度的长短时记忆（LSTM）<sup>8</sup> 网络的集合以及资源受限的优化<sup>9</sup>，在移动计算场景下基于深度学习的HAR已经取得了很大进展。我们现在开始观察这些方法在苛刻且非常具

有挑战性的领域里的全部潜力。

尽管已经取得了令人瞩目的进展，而且深度学习模型现在具有竞争力，如果不是优于用于移动计算的基于传统机器学习模型的HAR，需要解决许多问题以实现真正稳健可靠的系统进一步发展，即自动分析人类行为。在本文中，我们概述了下一个基于深度学习的HAR的边界线路图。在深入探讨问题空间和迄今为止提出的解决方案基础上，我们提供了对现有障碍的洞察，旨在进一步改进并讨论下一步可能采用的措施以应对这些挑战。

## 移动计算中的HAR

HAR相当于自动识别一个人做出某些行为，因此其有效地回答了人们在做什么以及何时做的问题。虽然这个问题领域有所扩展，例如评估一个人做某些行为的效果这在技能评估领域很常见，<sup>10</sup> 但本文的重点在于前者，即更为传统地解释行为识别问题。HAR系统的基础是观察传感器器捕获的行为，通常侧重于记录运动数据，例如通过利用身体佩戴的惯性测量装置（包括加速计、陀螺仪），这是智能手机或智能手表的标准配置。

所有传感通道都记录暂时数据，因此序列数据流和自动识别方法需要解决双重问题：在数据流内定位可能与

当前行为识别问题相关的连续部分（分割），并且通过自动分派分类标签（通常来自固定的、限定的词典）对提取的片段进行分类。这个双重问题是“鸡或蛋”问题，因为需要有关的行为信息才能确定行为何时进行，但分类要求事先在传感器数据流中定位。

对于这种双重评估任务而言，最大的问题是分类阶段将无法恢复任何在初始分割步骤中被错误忽略的分段。这个问题是许多研究人员绕开双重问题的原因，取而代之的是使用基于滑动窗口的处理管道，其颇具启发性。小分析窗口沿着连续数据流移动，提取传感器读数的连续部分。然后对提取数据的窗口进行单独分析，如果分析窗口的长度配置适当（即通过使用事先设定的阈值），则对于（准）周期性/重复性行为（例如步行或爬楼梯）已经实现了很好的结果。传统的机器学习方法预处理分析窗口所覆盖的传感器数据，<sup>11</sup> 提取特征，<sup>12</sup> 并采用基于概率的分类后台将行为标签分配给各个分析窗口。<sup>2</sup> 许多HAR系统实现了这类基于滑动窗口的分析流水线变体。

## HAR的深度学习革命

深度学习技术承诺克服许多典型问题：更传统的机器学习技术存在颇具挑战的模式识别任务。首先，最重要的

是，通过自动学习（分层）数据消除手动指定适当特征表达的需求，并将其整合到包罗万象的分类模型中，这对许多人来说是最有吸引力的。更深层的潜力在于深度神经网络的强大建模能力，其允许人们学习极其复杂的决策函数，这在解决具有挑战性的分析问题时非常重要。

这些承诺对于移动计算领域中的HAR研究人员也非常具有吸引力，因此该社区采用了深度学习技术，并取得了巨大成功，因为这些方法在许多具有挑战性的任务中优于传统的机器学习技术。由于上文所提及的大量特定的挑战，在HAR社区采用和扩展深度学习方法并非一朝一夕的事。因此，在该领域内采用深度学习技术遵循相当有组织的轨迹，下文将进行叙述。

## HAR 的特征学习

深度学习方法在引入HAR领域之初，首先使用移动计算，希望找到更具区分性和特别可概括的特征表达。<sup>3</sup> 相对较少的系统研究已经解决了特征设计问题，几乎所有以前的工作都采用进行启发式选择的一般性措施。这些特征或在时域中计算，在传感器器数据的符号表达上计算，又或是以光谱为基础。这些特征的主要注意事项是，必须针对每个应用程序域手动对其进行优化，这使得系统设计变得相当繁琐并且往往容

易出错，从而很容易出现未达最佳标准的行为识别系统。这一发现产生了第一个基于深度学习的特征提取方法设计：“最直接的特征设计方法是调查要分析的数据性质，并开发一个明确捕捉其核心特征的表达。”<sup>3</sup> 与计算机视觉或自动语音识别等其他领域不同的是，基于移动计算的HAR问题，不存在包罗万象的模型，以提供由专家驱动设计的通用特征表达。然而，深度学习方法已经确定有可能通过自动发现这种传感器数据的通用特征表达以克服该缺点。

基于深度学习的特征学习利用了自动编码网络，旨在学习输入数据的低维表达，从而最大限度地减少重建原始数据时产生错误。前馈神经网络由输入层和输出层以及奇数隐藏层组成，其中每层使用非线性激活函数完全连接到相邻层。网络最内层有较低的维度，这是故意设定的障碍，其迫使网络学习输入数据的紧密重构误差达到最小化，然后将其用作该域的通用特征表达。这些模型在自下而上的过程中得到充分的训练，编码器中的每一对相邻层均视为受限玻尔兹曼机（RBM），这是一个完全连接的二分图形模型。训练随机二进制隐藏单元集合旨在有效地让其充当低级别特征检测器。通过将RBM中的特征检测器的激活概率视为下一个RBM的输入数据，为每对后续层训练一个RBM。一旦对堆叠在一起的RBM进行

训练，将揭开生成模型以获得最终的、完全初始化的自动编码器网络，用于随后的特征学习。

在HAR场景中，首次尝试在移动计算中利用深度学习方法的潜力，结果是为运动传感器数据提供了可概括的、丰富的特征表达。这已经非常成功地在很多应用场景中得以利用。更重要的是，这一发展旨在为端到端识别系统启动了广泛采用和进一步发展复杂的、以深度学习为基础的建模方法。

## 用于时间序列分析的卷积神经网络

最初的深度学习着重于提取传感器数据里丰富和可概括的特征表达。在2010年代初期，深度学习方法开始在许多研究和应用领域得以广泛推动。事实上，移动计算领域的研究人员开始致力于端到端识别器，将深度学习方法中前景很好的表达-学习方面与其卓越的识别能力相结合，这是由有深度、分层次的分析结构所促进的。

深卷积神经网络（CNN）已经研究了数十年，且得到了令人印象深刻的图像处理结果，其中分析了单一图像（即二维输入数据）。13CNN成功的关键在于采用卷积滤波器层级结构，从原始传感器数据中连续提取越来越复杂的特征表达。行为识别研究人员分析时间序列，即序列数据流，通过使用众所周知

的滑动窗口程序提取原则上的二维数据块（每个窗口样本数乘以传感器通道数）以应用CNN，从而允许以与上述解释的（静态）图像数据相同的方式使用CNN。该诀窍是一个重大突破，为按时间序列进行分析的端对端识别器做好

对方法参数化的过程中必须非常小心。不适当的窗口长度不可避免地导致得不到最佳的识别结果，因为CNN的过滤器将捕获不相关的或仅仅是部分应该进行分析的信息。

严格地说，序列数据的分析应采用

进行优化给出具体建议。<sup>6</sup>目前，最有效且最成功的建模变体是CNN和LSTM模型的组合，其将CNN强大的特征学习功能与LSTM的序列建模功能整合在一起。<sup>5</sup>图1显示了当前在移动计算领域内，HAR社区所使用的最相关模型架构。

## 移动计算在HAR情景下使用深度学习技术的最大挑战可能是没有大规模标记训练数据集。

充分准备，该分析不仅仅是通过特征学习合并强大的识别后台。现存许多基于CNN行为识别系统的例子，所有这些例子都对具有挑战性的任务能达到出色的识别性能<sup>4,5</sup>。

### HAR的序列建模

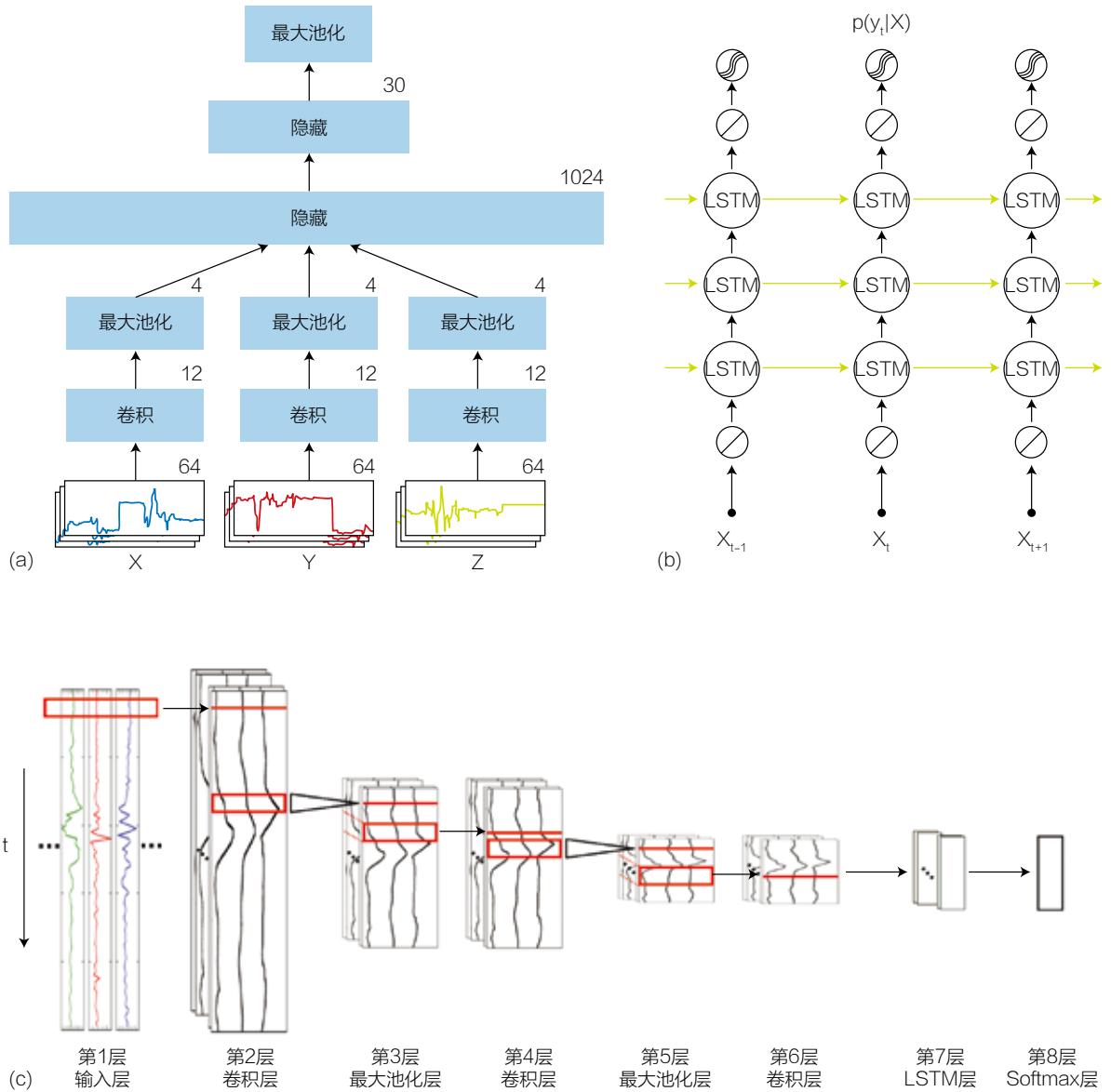
如上所述，CNN现在广泛应用于序列数据分析。然而，CNN只能用于这种数据，其通过使用滑动窗口预处理步骤这一“窍门”来挖掘连续的样本数据，因此“假装”时间序列数据实际上是静态数据，可以使用CNN进行分析。当然，时间顺序保存在分析窗口内，因此CNN实际上可以高效地处理有序的数据。然而，正如常规机器学习中滑动窗口程序的警告（如上所述），在

序列模型。一般而言，这是真的，特别是对于深度学习方法而言，由于其复杂性大大增加（与更传统的机器学习技术相比），出错的可能性也越来越大。因此，HAR研究人员现在采用更有序的深度学习模型，其中突出的例子包括通用循环神经网络，以及更为重要的LSTM模型。LSTM模型特别具有吸引力，因为其专业的内部结构使用的内存包含忘记功能，根据训练过程，可以非常有效且有选择性地将聚焦集与识别过程相关的实际传感数据流部分。因此，这些模型不仅可以整合表达的学习和分类，还可以有效地进行分割，这对于HAR而言是最重要的。不同建模变体的功能已经得到有效分析，因此针对何时使用特定的建模变体以及哪些（超）参数应该

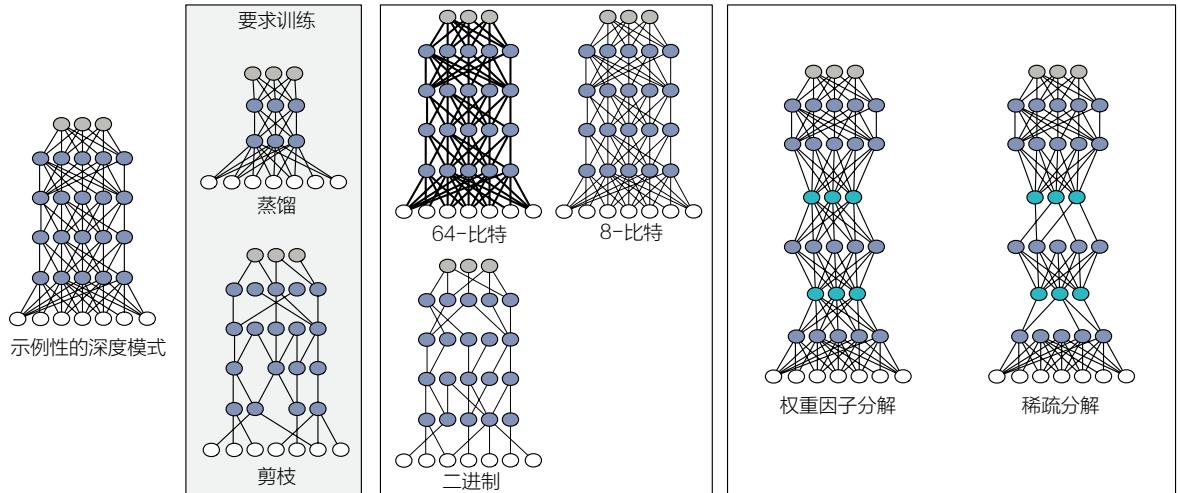
### 追踪小型训练集

移动计算在HAR情景下使用深度学习技术的最大挑战可能是没有大规模标记训练数据集。这与深度学习方法在其他领域成功应用形成了鲜明对比，例如计算机视觉拥有大量的注释图像的数据库，从而创建了非常有深度的模型。虽然记录几乎无限量的未标记数据是简单直接的，但大多数当代深度学习技术要求带注释的培训数据，因其遵循监督式的学习范例。大规模的标记数据集的确是严重的限制，如果不小心处理，可能很快导致过拟合，因为复杂的深度神经网络及其数百万参数将简单地存储小型训练集。

与其他领域不同的是，在HAR情景中，简单地集中精力收集和注释更大的数据集不容易。其原因在于，不容易观察到移动用户实际情况的注释，并且要求用户提供自己的标签这一举措只能在一定程度上扩大规模<sup>14</sup>。相反，研究人员专注于开发更有效利用现有数据集的技术。这些技术有两个主要的方



**图 1.** 移动计算中用于 HAR (HAR) 的深度学习模型架构。(a) 用于基于帧的 HAR 的卷积神经网络 (CNN) (经 M. Zeng 等人许可使用, 《利用移动传感器实现人体行为识别的卷积神经网络》 Proc. Int. Conf. Mobile Computing, Applications and Services (MobiCASE 14), 2014, pp. 1-18); (b) 用于 HAR 的深度的长短时记忆 (LSTM) 网络 (经 N.Y. Hammerla, S. Halloran, and T. Ploetz 允许使用, 《使用可穿戴设备识别人类行为的深度、卷积和递归模型》 Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI 16), 2016); (c) CNN 和 LSTM 网络的组合 (经 F.J.O. Morales and D. Roggen 允许使用, 《跨越移动行为识别域, 传感器模式和位置的深度卷积特征传输》 Proc. Int'l Symp. Wearable Computers, (ISWC 16), 2016)。



**图 2.** 在移动平台上通过模型手术、压缩和缩小，处理基于深度学习的 HAR 资源模型手术的限制。(经 N.D. Lane 等人允许使用，《把深度学习融入移动和嵌入式设备》，IEEE Pervasive Computing, vol. 16, no. 3, 2017, pp. 82–88)。

向：第一，使用迁移学习技术，让在特定情景训练的模型可以从其他基本相似的领域受益；<sup>7</sup>第二，采用分类器集合，这些是不同训练模型的集合，当其相结合时，由于在训练期间捕获更多的数据变化而增加建模能力<sup>8</sup>。迄今为止，迁移学习仅在特征提取级别进行了探索，只能对分类准确性适度改进。然而，一般的证据表明，原则上可以在域之间交换特征表达，并可适度地改进性能。集合学习方法应对这样的设想，即可用于深度学习模型的监督训练的小型标签数据集包含冗余和可能错误的数据，这些数据对训练过程具有负面影响并会产生未达最佳标准的模型。通过随机移除部分传感器数据，以及结合在相同训

练数据上训练的原始深度学习模型变体（例如：通过组合不同的损失函数），结果表明，LSTM 的集合更有效地利用小的、有潜在问题的训练数据集，从而令到总体上过拟合较少以及识别性能更好。

### 处理资源限制

移动计算至少将复杂模型分析任务的推理阶段推到了资源受限的平台上，例如智能手机、智能手表甚至是智能纺织品。这些环境的特点是存在限制，包括板载存储、计算能力以及最重要的是因为需要维持“持续”运作而依赖于电池运行时间。这种苛刻的计算环境与在云端基础架构中得到培训的深

度学习方法形成鲜明对比，后者在存储和（大规模并行化）计算能力方面都可获得近乎无限的资源。

移动计算研究人员和从业者很久以前就开始研究深度学习方法的概念验证是否有利于HAR分析任务。自那时以来，大量的工作致力于使深度学习模型可用于移动平台上的（接近）实时推断。<sup>9</sup>这些工作大部分针对自动优化，因此复杂模型适用于板载内存小的平台，并避免不必要的计算。这种优化使移动计算场景中的近实时推断（模型评估）成为可能。图2概述了常用的模型优化技术。

减少深度神经网络的内存占用（通常有数百万个参数），很大程度上与模

型手术相对应，这删除了模型中被认为不必要的部分或者至少对整个识别过程起最少作用的部分。这里的优化标准是减少层数、节点和权重，而不过多地损害整体识别性能。该文献中已经描述了许多技术，旨在对模型的某些部分进行“蒸馏”（压缩）或消除（修剪）。<sup>9</sup>此外，研究人员已成功地尝试降低模型参数的内部表达精度（例如，从16位降到8位，这大大减少了内存占用），而有趣的是，如果正确完成，这对识别功能没有过度负面影响。当在推断过程中评估修剪过的模型时，较小的模型大大地减少了运行时间，并且已经表明非常复杂的深度神经网络现在可以用于移动平台的推断，例如Aualcomm的SnapDragon，这是一套常见的SoC解决方案，广泛应用于现代智能手机和智能手表等可穿戴设备。

## 挑战和新边界

HAR在移动计算领域的深度学习方法已经走过了很长的路，现在代表了一系列具有挑战性的分析任务中最有前途的分析方法。整个领域在推进最先进的技术方面非常具有活力和创造性，因此可以在识别功能方面进行改进，并且这些强大的模型也可用于下一代移动计算平台。要做到这点，我们需要应对一系列重大挑战并开发新方法。我们

已经列出了线路图，以解决移动计算中关于HAR最迫切的问题，下面将描述这些问题。

## 利用未标记数据

大多数当代深度学习方法和应用程序都基于监督式训练程序；即那些从样本数据中导出并由此与真实情况注释相关联的模型。在诸如计算机视觉或自动语音识别等领域（其中有很多占据大量内存的带注释数据集可用），复杂深度神经网络的真正潜力可以通过令人印象深刻的识别能力得以展现。

在移动计算领域，记录样本数据非常简单直接，因为这些设备“持续开启”并不断记录传感器数据。相反，如果要收集类似规模的实际情况注释，那么则是非常具有挑战性的。这个问题的主要原因实质上是由于实用性，因为其通常不能为了标记移动计算用户的行为而对其进行（持续）观察，或者反复要求他们提供自己实际情况的注释。

克服这种稀疏注释问题的好方法是从本质上跨越无监督或半监督学习技术的边界。在移动计算中，基于深度学习的HAR的起源是从未标记的数据中进行特征学习，其中基于RBM的自动编码器网络生成可概括的数据表达。<sup>3</sup>随着可以使用非常大的数据集（如英国生物银行联盟<sup>15</sup>收集的2000人加速计数

据集）和新型建模技术（如LSTM自动编码器），更有效地利用未标记数据有很大的潜力可获得重大进展。

半监督学习将标记和未标记的训练数据结合，以派生出更有效的识别系统。事实证明，迁移学习可能是推导模型的可行方法，这种模型在不同任务甚至领域中有一定程度的推广。<sup>7</sup>随着可以使用大数据集，通过迁移学习的方法，（当前）应该有可能大幅度推动适度的改进。另一个推动可能来自基于集成识别概念的更多实质性投资，其中不同训练的分类器集体协作解决复杂的识别任务。合并LSTM网络出现满意的效果，为我们证明该方法可行提供了证据<sup>8</sup>。

## 个性化

HAR最具挑战性和理想化的其中一方面是个性化。个性化使技术能够实现不依赖于大量用户特定的注释示例数据建立用户特定的模型，并且在引导HAR系统时不会给用户本身带来不必要的负担。个性化的关键在于调整通用模型，从而使我们能够专注于（例如特征表达）有针对性的转换或整体模型架构优化，以更好捕捉特定用户的特质。

复杂的深度学习模型主要针对泛化，因此其包含大量参数和复杂的模型架构。未来的个性化研究应侧重于分析未标记的、特定用户的数据，例如通过分析独立于用户的模型所实际涵盖的内

容。这种比较分析加上对建模过程本身的深入理解(见下文)将使个性化方案成为可能,从而可以提高个人用户的识别性能,而不会给用户造成不必要的负担。

## 模型优化

模型优化是移动计算中基于深度学习的HAR更广泛研究议程中的一个关键项目。如前所述,其基本原理是减少复杂模型的内存占用量,并在推断过程中加快处理速度——所有这些都必须在不损害实际识别能力的情况下完成。现在已有令人印象深刻的结果,但需要进一步努力才能使移动平台能够访问最复杂(且功能最强大)的深层模型架构。

研究人员已经开始致力于研究专用编译器,将模型转换为高度优化、可能是特定硬件的表达。这与开发专用深度学习硬件的工作是一致的,例如Google的TensorFlow处理器。然而,对于移动计算而言,这些工作应该朝着相反的方向发展;也就是说,研究人员不应构建最佳的硬件,而应着眼于优化现有(和未来)的硬件平台模型。这种模型优化工作也将支持前面所提到的个性化努力。

## 可理解性问题

我们日常生活中越来越多的功能依赖于自动分析,使得使用这些功能

的人们看不见此类功能。这一点非常重要,特别是对于基于移动计算的自动化而言,可靠性是至关重要的。尽管深度学习方法产生了令人印象深刻的识别系统,但是对于这些复杂模型而言,本质上类似于“黑盒子”并不能解释如何做出决定的事实,尤其是在从业人员的社区当中,人们的敌意日益增加。不可理解性非常普遍,不仅影响深度学习,还影响所有基于机器学习的方法。然而,随着深度学习模式变得越来越复杂,这个问题变得越来越相关,因为用户在不可解释的、任务关键的自动化决策过程中,变得越来越不适。

开发可理解的建模技术需要大量研究工作。理想情况是,应该弄清楚模型究竟是如何捕获数据以及如何做出决策。除了回应用户的需求外,可理解的基于深度学习的HAR还具有实际改进模型识别能力的潜力,并进一步优化模型以整合到资源受限的移动平台。

## 有效分割时间序列数据

实质上,行为识别对应时间序列分析问题。因此,需要解决两个基本问题:分割和分类。到目前为止,大多数分析方法,包括基于深度学习的分析方法,通过采用如上所述的滑动窗口预处理步骤的变体或多或少避免了分割问题。在最近工作中,研究人员为摆脱固定长度的滑动窗口方法,通过使用例如

随机长度框架,甚至将其放置在输入数据流上,以实现在集合方法中更为有效的小批量学习过程。<sup>8</sup>

正如Nils Hammerla及其同事所建议的那样,逐个样本处理的替代方案可能并不总是一种选择,因为此类方法需要更复杂的模型架构(如LSTM模型,这些模型对于某些资源受限的应用程序领域并不总是可行),或者在某些情境下因为太具有挑战性而无法进行训练。替代方法可以恢复到传统的两阶段方案,包括明确的分割步骤及其后的分类。最新建模技术(包括注意力模型)是有吸引力的,因为其可以自动将分类工作集中于传感器数据流的相关部分,这与分割类似。



如其他基于机器学习的模式识别领域一样,深度学习已经彻底改变了移动计算中的HAR领域。在具有挑战性的分析任务中,令人印象深刻的识别能力现在已成为现实,这要归功于多年的研究和根本性突破,这有效地令我们能够利用深度神经网络架构中极妙的学习表达和分类功能。现在,该领域的最新技术由用于有效特征学习的(多个)卷积层组合和包含诸如LSTM单元的序列建模节点的(多个)层级组成。此外,研究人员高效地解决了严重资源受限操作的基本问题,因为这在移动计算领域很常

见。通过有效的模型手术和模型压缩，现在可以在移动设备上运行基于深度学习的识别行为，这种移动设备具有有限的内存和计算能力，可以有效地实现接近实时的推理，这是许多新型移动计算交互方案所要求的。

移动计算为HAR而设的下一代深度学习方法必须解决若干重大挑战，若未能解决，将限制提高识别能力方面的进展，并阻碍这些技术在该领域的更广泛应用。处理嘈杂的、往往不确定的传感器数据是最紧迫的挑战；缺失大规模的注释训练数据集；在不依赖不切实际的大量专业培训数据或漫长的适应程序的情况下，对个性化识别系统的需求；增加模型优化，以便在严重资源受限的平台上整合复杂模型；对开放式识别计划的需求；以及对可理解、可解释的模型推断的需求。

随着这些令人兴奋的发展，以及在移动计算中将深度学习应用于HAR的许多重要挑战，我们期待着从社区中获得进一步发展，以便应对新边界。深度学习停留在那里；并且由于建模、数据挖掘以及模型优化方面的进一步改进，开创全新一代的方法和应用程序成为可能。总体而言，我们期待看到这些属于未来的进步能得以实现。■

## 参考文献

- Y. LeCun, Y. Bengio, and G. Hinton,

## 关于作者

**Thomas Ploetz**是乔治亚理工学院交互计算学院副教授。他的研究方向包括主要应用在医疗领域的行为分析计算、开发应用于传感器和一般时间序列数据分析的机器学习技术。Ploetz从比勒菲尔德大学获得计算机科学博士学位。联系邮箱: thomas.ploetz@gatech.edu。

**Yu Guan**是纽卡斯尔大学计算科学院数据科学讲师。他的研究方向包括机器学习及其各种应用，如行为分析，可穿戴设备，普适计算，计算机视觉和生物统计学。Guan在华威大学获得计算机科学博士学位。联系邮箱: yu.guan@newcastle.ac.uk。

- “Deep Learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.
- A. Bulling, U. Blanke, and B. Schiele, “A Tutorial on Human Activity Recognition Using Body-worn Inertial Sensors,” *ACM Computing Surveys*, vol. 46, no. 3, Jan. 2014, pp. xx-xx.
  - T. Ploetz, N. Hammerla, and P. Olivier, “Feature Learning for Activity Recognition in Ubiquitous Computing,” *Proc. 22nd Int'l Joint Conf. Artificial Intelligence (IJCAI 11)*, 2011, pp. 1729–1734.
  - M. Zeng et al., “Convolutional Neural Networks for Human Activity Recognition using Mobile Sensors,” *Proc. Int. Conf. Mobile Computing, Applications and Services (MobiCASE 14)*, 2014, pp. 1–18.
  - [F. Ordóñez and D. Roggen, “Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition,” *Sensors*, vol. 16, no. 1, 2016, p. 115.
  - N.Y. Hammerla, S. Halloran, and T. Ploetz, “Deep, Convolutional, and Recurrent Models for Human Activity Recognition Using Wearables,” *Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI 16)*, 2016, xx.
  - F.J.O. Morales and D. Roggen, “Deep Convolutional Feature Transfer across Mobile Activity Recognition Domains, Sensor Modalities and Locations,” *Proc. Int'l Symp. Wearable Computers, (ISWC 16)*, 2016.
  - Y. Guan and T. Ploetz, “Ensembles of Deep LSTM Learners for Activity Recognition Using Wearables,” *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. (IMWUT 17)*, 2017.

9. N.D. Lane et al., "Squeezing Deep Learning into Mobile and Embedded Devices," *IEEE Pervasive Computing*, vol. 16, no. 3, 2017, pp. 82–88.
10. A. Khan et al., "Beyond Activity Recognition," Proc. Int'l Joint Conf. Pervasive Ubiquitous Computers (UbiComp 15), 2015, pp. 1155–1166.
11. D. Figo et al., "Preprocessing Techniques for Context Recognition from Accelerometer Data," *Personal and Ubiquitous Computing*, vol. 14, no. 7, 2010, pp. 645–662.
12. N. Hammerla et al., "On Preserving Statistical Characteristics of Accelerometry Data Using Their Empirical Cumulative Distribution," Proc. Int. Symp. Wearable Computers (ISWC 13), 2013.
13. Y. LeCun, S. Chopra, and R. Hadsell, "A Tutorial on Energy-Based Learning," *Predicting Structured Data*, 2006.
14. T. Miu, P. Missier, and T. Ploetz, "Bootstrapping Personalised Human Activity Recognition Models Using Online Active Learning," Proc. IUCC, 2015.
15. A. Doherty et al. , "Large Scale Population Assessment of Physical Activity Using Wrist Worn Accelerometers: The UK Biobank Study," *PLoS One*, vol. 12, no. 2, 2017, pp. e0169649–14.



SUBMIT  
TODAY

## IEEE TRANSACTIONS ON BIG DATA

► SUBSCRIBE  
AND SUBMIT

For more information on paper submission, featured articles, calls for papers, and subscription links visit:

[www.computer.org/tbd](http://www.computer.org/tbd)



*TBD* is financially cosponsored by IEEE Computer Society, IEEE Communications Society, IEEE Computational Intelligence Society, IEEE Sensors Council, IEEE Consumer Electronics Society, IEEE Signal Processing Society, IEEE Systems, Man & Cybernetics Society, IEEE Systems Council, IEEE Vehicular Technology Society

*TBD* is technically cosponsored by IEEE Control Systems Society, IEEE Photonics Society, IEEE Engineering in Medicine & Biology Society, IEEE Power & Energy Society, and IEEE Biometrics Council



IEEE  
computer society

myCS

Read your subscriptions through the myCS publications portal at

<http://mycs.computer.org>



# 利用递归神经网络对资源受限的物联网设备进行基于呼吸的身份验证

文 | Jagmohan Chauhan, 阿尔托大学  
Suranga Seneviratne, 悉尼大学  
Yining Hu, Data61 研究所和新南威尔士大学  
Aruna Seneviratne, 新加坡管理大学  
Aruna Seneviratne, 新南威尔士大学  
Youngki Lee, 新加坡管理大学

译 | 詹冰冰, 浙江大学

递归神经网络 (RNNs) 在音频和语音处理应用中显示出了可喜的效果。物联网 (IoT) 设备的日益普及为 RNN 的推理应用 (如基于声学的身份验证和智能家居的语音命令) 提供了强有力的支持。然而, 关于这些资源受限设备的推论, 其可行性和性能在很大程度上仍尚未得到探索。作者将传统的机器学习模型与 RNN 深度学习模型进行了比较, 以进一步研究基于呼吸声学的端到端认证系统。

**嵌**

入在智能手机、可穿戴设备和其他物联网 (IoT) 设备中的传感器越来越多地被用于支持对用户活动/周围环境的细粒度监控以及更丰富

的网络物理交互形式 (通过手势或自然语言界面) 之中。说明性情景包括: 监测用户的步数以估计每日消耗的卡路里, 跟踪饮食行为以获取食物的摄入量, 以及使用配备麦克风的

设备进行基于语音的家庭自动化控制等等。由于这些个人和边缘设备能够学习并存储私人特定信息，并且越来越多地以非传统形式出现，因此开发安全可用的用户认证技术十分紧迫。

在最近的一项研究中，我们引入了呼吸打印(Breath Print)系统这一概念，该系统利用用户呼吸的声学特征(由普通麦克风捕获)来支持在移动和物联网设备上无处不在的用户身份验证。<sup>1</sup>在本文中，我们对递归神经网络(RNN)深度学习模型能否有效地应用于Breath Print的资源受限设备进行了研究。最初的Breath Print是基于云计算的机器学习模型，该模型建立在传统的高斯混合模型(GMM)的基础上，具有人工处理的特性。

这项工作有两个主要目标。首先，与传统方法相比，如支持向量机(SVMs)、高斯混合模型(GMMs)和隐马尔可夫模型(HMMs)等，RNN最先进的语音识别和说话人识别方法具有明显优势，特别是在嘈杂的环境中，<sup>2</sup>因此，使用RNN模型对Breath Print的性能进行评估，对提高其性能具有重要意义。其次，对于不引人注意且无处不在的身份验证应用程序，Breath Print需要在资源受限的设备上实现，并且必须能够在没有云访问的情况下对用户进行身份验证。因此，掌握Breath Print的精确度与资源之间的折衷方案非常

重要。

为此，在三个具有代表性的硬件平台上，我们对基于RNN的Breath Print端到端身份验证系统进行了性能评估，这三个平台分别为：移动(智能手机)、可穿戴(智能手表)和物联网(Raspberry Pi<sup>3</sup>)。据我们所知，这是第一项基于RNN声学深度学习模型的研究，它展示了这一模型在有限资源占用设备上实现的可能性与性能。

我们取得了以下重要成果。

- > 基于浅层模型(SVM)和长短期记忆模型(LSTM, RNN的一种变体)，我们得出了端到端认证系统的性能评估结果，在三种典型的物联网设备上使用呼吸声学原理：智能手机、智能手表和迷你电脑(Raspberry Pi)。

- > 研究表明，相比于先前报告中的卷积神经网络(CNN)模型，RNN声学分类模型尺寸更小，重量也更轻，因此可以用于物联网设备中。具体来说，一个未压缩的RNN模型只有1.1兆字节(用于相关的呼吸类型)，并且可以在智能手机和智能手表上运行，其延迟时间分别约为100到200毫秒和700到1000毫秒。

- > 我们展示了线性量化的模型压缩技术如何在确保精度的前提下，

将RNN模型的内存占用减少5倍(大约150到50KB)。

## 相关工作

深度神经网络(DNN)在训练方面取得了一定的突破，性能也得到了大幅提升(深层网络性能已超越了传统机器学习模型，甚至人类专家)，受此启发，我们近期的各种工作都集中于尝试在资源受限设备上执行深度学习模型(如更高的内存需求或过度的计算延迟)。表1总结了一些值得注意的措施和所用的技术。广义而言，这些措施采用以下三种方法中的一种或多种：将神经网络处理转移到GPU上，使其在矢量计算中效率更高；降低加载完全连接层的时间和内存要求；以及更快地执行卷积层任务。

在早期的研究工作中，尼古拉斯·莱恩(Nicholas D. Lane)和他的同事们对在移动、可穿戴和物联网设备上的深度学习模型(CNN和DNN)进行了一些研究，涉及模型的性能特征、资源需求以及执行瓶颈等方面，以支持基于音频和视觉的应用程序。<sup>7</sup>结果表明，虽然在这些设备上，应用小型的深度学习模型没有任何问题，但在资源紧张的情况下，更复杂的CNN模型(如Alex Net)却无法很好地完成工作。为了解决这一问题，索瑞·巴塔查里亚(Sourav Bhattacharya)和他的同事们提出了SparseSep模型，

表1.资源受限设备上的深度学习(DL)

名称	深度学习(DL)类型	构建	应用	技术
Spars-eSep <sup>3</sup>	卷积神经网络(CNN) 深度神经网络(DNN)	多层	图像分类， 说话人识别和场景分析	稀疏化与分离
DeepEar <sup>2</sup>	深度神经网络(DNN)	5层	音频传感	NA
DeepEye <sup>4</sup>	卷积神经网络(CNN)	多层	连续视觉	交织、缓存和压缩
DeepMon <sup>5</sup>	卷积神经网络(CNN)	16层	连续视觉	GPU卸载，缓存和分解
MobiRNN <sup>6</sup>	递归神经网络(RNN)	2层	行为识别	GPU卸载

该模型主要关注于寻找一个完全连通层的稀疏表示，并使用独立的过滤器来处理卷积核。<sup>3</sup>这些技术减少了执行深度学习模型所需的参数和卷积操作的数量，从而大大降低了资源受限设备的计算复杂度和空间复杂度。

有几项研究致力于优化音频和图像传感应用的深度网络。DeepEar是一款基于DNNs的智能手机音频传感应用。<sup>2</sup>它是在智能手机的数字信号处理器中实现的，在日常能源消耗中只增加了6%的额外开销。DeepEye将CNNs部署到可穿戴设备上，用于连续的视觉应用。<sup>4</sup>通过交叉存取来编排执行大计算的卷积层和大内存的完全连接层，从而避开了资源瓶颈。DeepEye还利用缓存来更快地加载完全连接层，并通过奇异值分解(SVD)的分层分解法来实现对完全连接层的压缩。而DeepMon则侧重于降低卷积层的处理延迟(通过多种优化技术)，以实现连续的视觉应用。<sup>5</sup>它结合了缓存(利用连续图像之间的相似性)和模型分解(将卷积层分解成更小的层)技术来减少计算开销。DeepMon还将卷积层的任务分配到移动GPUs上以进行更快的处理。最后，Mobi RNN应用GPU卸载，在智能手机

上更快地执行RNNs，以支持活动识别任务。<sup>6</sup>

从表1可以看出，迄今为止的大部分工作都集中在CNNs和DNNs上。例如，甚至用CNN和DNNs来进行音频分析和说话人识别。一般来说，CNNs擅长利用定义在空间数据(比如图像)上的特性，而RNNs更适合于识别和使用在数据流(比如音频或文本)中定义的时间特性。与CNN的模型相比，RNN模型处理的数据并不复杂，且无需使用卷积过滤器，对计算能力和内存的要求也较低。因此，相对而言，CNN模型更复杂。

根据假设：每个人的呼吸模式都是独一无二的，Breath Print提出了一项技术，利用在移动和物联网设备上的呼吸声学特性来实现用户认证。由于该技术仅要求用户进行几次呼吸动作，因此可行性很强。在信念的驱动下，我们对RNNs产生了兴趣，我们相信，这种独特性可以通过用户呼吸模式的时间变化表现出来，而RNNs更有能力识别和利用这种时间特性。然而，只有在本地(设备)上以最小的延迟执行用户识别时，才能发挥Breath Print的全部潜能。因此，在这项工作中，我们研究的核心问题是：“在资源受限的设备上，

Breath Print能否通过使用RNN模型来实现？”

图1显示了一些真实的应用程序，在这些应用程序中，Breath Print可作为入口点或连续的身份验证机制。作为入口点认证，主要的用例是解锁智能设备，如智能手机和可穿戴设备。解锁设备不仅需要访问设备，还需要获得个性化的用户体验。例如，在工业环境中，设备上通常会有一层叫做增强现实的智能玻璃，覆盖着一些特定产品的信息，根据工作人员的个人资料，这种玻璃可以对内容和指令进行一定的调整。其他用例还包括在智能空间中操作物联网设备，如利用传感器启用的门或咖啡机。如果设备进入睡眠模式或处于非活动模式，用户则需要重新进行身份验证。此外，Breath Print还可应用于智能呼吸器，根据工作人员的个人偏好来调节工作环境中的湿度水平，或可根据呼吸模式对用户进行持续认证。连续身份验证方案中的重新认证时间可以根据所需的安全性级别进行配置。除非重新认证失败，否则持续认证无需用户干预。

这项工作在许多方面与Breath Print有着显著差异。首先，关注点不同。Breath Print为用户身份验证引入了一

种新的行为模式。而这项工作的重点则是评估浅层学习模型和深度学习模型的性能，以验证Breath Print在资源受限的设备上是否可用于实时身份验证。其次，本工作处理的是用户识别，而不是Breath Print的用户验证。在用户识别中，身份验证系统的任务是在一组封闭的用户中预测将试图访问的用户。而在用户验证中，身份验证系统检测试图访问系统的用户X是否为用户X本人。



图1. Breath Print 的应用。

## 实验设置

为了评估RNN驱动下呼吸认证的可行性，我们利用了先前工作中收集的呼吸声学数据集。<sup>1</sup>该数据集收集了对10位用户共进行的3组声学样本，其中包括三种呼吸类型——深呼吸、正常呼吸和嗅探（两次快速吸气）。对于每种呼吸类型，数据集分别在第一天（1组）、第4天（2组）和第7或第8天（3组）收集了30、30和10个样本。本文只关注两种呼吸类型（深呼吸和嗅探），因为之前的调查显示，与正常呼吸相比，这两种呼吸类型在身份验证应用中表现得更好。有关数据集的详细信息，请参阅原文。<sup>1</sup>

对于每个用户，我们选择了前50个样本进行模型培训和调优。由于训练深度学习模型需要更大的样本容量，因此我们采用了两种常用的数据增强技术

表2.设备的硬件配置

设备	操作系统	中央处理器	中央处理器	内存
Nexus 5 smartphone	Android 6.0	2.26千兆赫兹 双四核心	Adreno 330	2 GB
Pixel smartphone	Android 7.0	2.15 , 1.6 千兆赫兹 双四核心	Adreno 530	4 GB
LG G Watch R	Android Wear 2.0	1.2 千兆赫兹 双四核心	Adreno 305	512 MB
Raspberry Pi 3	Android Things 5.0	1.2 千兆赫兹 双四核心	VideoCore IV	1 GB

来增加数据样本量。此外，我们使用了频率框架<sup>8</sup>和幅值定标的组合。<sup>9</sup>在时间轴和幅度轴上，通过从均匀分布中选择两个不同的值来对每个样本进行10次缩放；~U (0.8,0.8)。总体而言，我们获得了十一倍的训练样例数量（每个参与者550个训练样本），包括原始样本及其扩充版本。剩下的10份原始样本（来自2、3组）保持不变以供测试。

我们进行了一项实验评估，使用了三种不同类型的四件设备（表2中列

出）——两部智能手机（移动），一只智能手表（可穿戴式）和一台迷你电脑（物联网）。所有这些设备均为安卓操作系统的不同变体，是当下流行的商用移动、可穿戴和嵌入式平台的典型代表。

## 方法论

我们的总体目标是用户认证，或者决定样本是否属于N个预先注册的可能

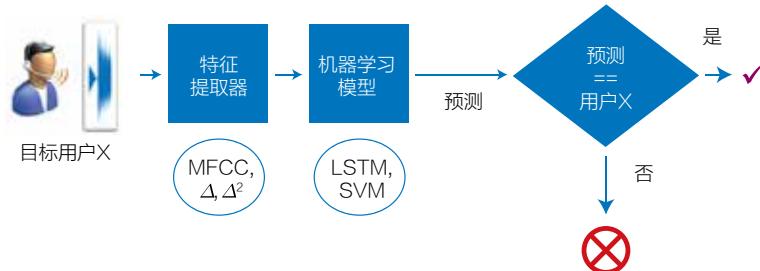


图2. 用户认证。MFCC: 梅尔频率倒普系数; LSTM: 长短期储存器; SVM: 支持向量机。

我们使用了扩充创建的样本进行训练。更具体地说，我们从这些样本中创建了重叠窗口，并将它们随机打乱——80%的窗口作为训练集，其余的(20%)作为验证集来调整超参数。我们将来自2组中10个音频样本的其余部分创建的重叠窗口称为帧内集合，将来自3组中10个音频样本创建的重叠窗口称为帧间集合。

用户之一。这实际上是一个封闭的用户识别问题，可以将其映射为多项分类问题（其中一个类代表一个用户）。图2显示了我们的系统。首先，目标用户X在设备（例如，解锁设备）的麦克风上进行一次呼吸，并从中输出特征。然后将特征输入分类器，得到预测输出。如果预测的用户与目标用户相同，则验证成功。

我们的方法包括对一些原始声学信号的预处理，以便将其输入到RNN模型中。对于使用浅层分类器的比较基准，我们还训练了SVM模型。

## 特征提取

我们将每个音频文件分割为<sup>10</sup>毫秒的非重叠帧，并用汉明窗平滑处理。在每一帧中，利用Java语音工具包 (Java Speech Toolkit; <https://github.com/sikoried/jstk>) 计算96个梅尔频率倒普系数 (Mel-frequency cepstral coefficients, MFCC) 特征 (32个MFCC, 32个Delta MFCC, 32

个Double Delta MFCC)。然后使用窗口来组合这些帧，以便在帧之间保留时间信息。对于嗅探和深呼吸，我们分别尝试了长度为{20,25,30,35}和{200,250,300,350}的窗口大小。在选择合适的窗口大小时需要考虑两个因素。每个窗口必须足够大，以保留每次呼吸的重要部分。然而，单次呼吸的持续时间在用户之间有显著差异；因此，如果选择一个较大的窗口尺寸，则可能会在呼吸持续时间相对较短的用户中遗漏一些测试样本。为了平衡这些考虑因素，我们最后选择了长度为{20,25,30}的嗅探窗口和{200,250}的深呼吸窗口。为了进一步扩充训练数据集，我们为给定的呼吸样本创建了重叠窗口，同时选择了三个重叠窗口尺寸，分别为(90%, 70%和50%)。对于每个窗口大小和重叠值对，我们训练了分类器，如下所述。

## 训练和测试数据集

## 模型

我们使用了RNN架构，类似于尼尔斯 Y. 哈默拉 (Nils Y. Hammerla) 和他同事所描述的一种体系结构。此架构如图3所示。我们运用了两个LSTM层，其中每个LSTM单元的隐藏单元大小为128，同时，通过张量流来实现模型，并以大小为32的批处理对网络进行了500次迭代训练。作为基线分类器，我们还使用LIBSVM ([www.csie.ntu.edu.tw/cjlin/ LIBSVM](http://www.csie.ntu.edu.tw/cjlin/)) 训练了一个具有线性内核的多类SVM分类器，并利用LIBSVM-AndroidJNI (<https://github.com/cnbuff410/libsvm-androidjni>) 在安卓设备上测试LIBSVM。但需要注意的是，尽管具有非线性内核的SVM模型可能会比线性内核提供更好的结果，但经发现，由于支持向量数量众多，具有非线性内核的SVM模型无法加载到任何设备上。在非线性核的支持向量模型中，所包含的支持向量的数量高于线

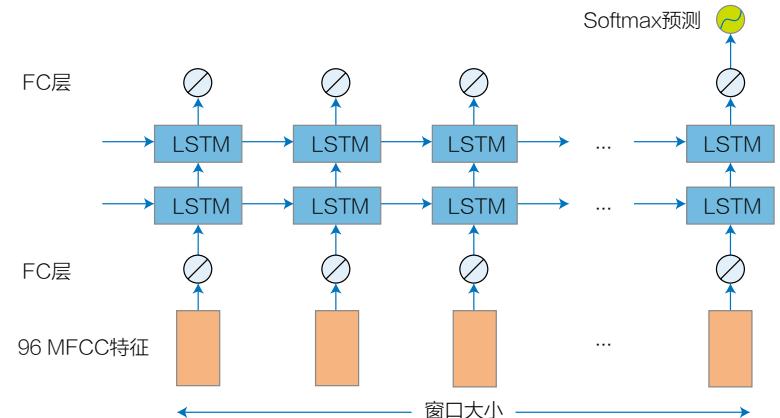


图3. 递归神经网络 (RNN) 架构。FC: 完全连接。

性核所需的支持向量。由于可扩展性问题, GMM不适合解决多类分类问题。而 SVM更适合多类分类。鉴于非线性内核和GMM存在的问题, 我们以线性内核作为基线来对SVM进行评估。

## 模型选择

我们用提前终止来选择最优的模型, 因为经观察, 一些迭代之后, 模型的精度达到了最大值, 并且在验证集上保持在几乎相同的区域, 而组内集和组间集的准确率则有轻微的下降趋势。

为了选择精确度图中合适的时点, 我们首先将20点移动平均应用于验证集精度图, 然后选择精度不会提高5%的点, 以用于接下来的四个连续点。精确度图表显示了深度学习模型在不同迭代中所能达到的精度。我们从经验上确定了移动平均窗和改进阈值。确定了时点后, 我们选择了11个模型(之前的5个模型和接下来的5个模型, 以及包括在时点的一个模型)。为了显示性能结果, 我们在三个数据集中选择了平均准确率最高的模型。在精度最高的窗口和重叠配置中, 嗅探为30窗口大小, 90%重叠, 深度呼吸为250窗口大小, 90%重叠。对于SVM, 我们选择了具有最佳交叉验证精度的模型。

## 模型降阶

为了实证研究, 计算开销与延

迟精度之间的权衡, 我们使用张量流([www.tensorflow.org/performance/quantization](http://www.tensorflow.org/performance/quantization))提供的内置256级量化函数对选定的RNN模型进行压缩。此函数为每个图层提取最小值和最大值, 然后将每个浮点值压缩为一个8位整数。请注意, 此选项将以压缩格式保存空间, 通常在安卓应用程序中使用。我们执行了原始模型和量化模型实验。

## 结果

我们以四个性能指标进行实验评估。

1. 精确度: 正确用户识别的百分比。
2. 特征提取时间: 从音频文件中提取MFCC特征的所需时间。
3. 模型加载时间: 将机器学习模型加载到内存中的所需时间。
4. 推断时间: 完成特征提取并将学习模型加载到内存中后, 预测用户标签的所需时间。

## LSTM 性能

报告显示了在分类过程中不同阶段的执行时间的平均值。在四种设备上, 嗅探的特征提取时间在40到60毫秒之间, 深度呼吸的特征提取时间在126到484毫秒之间。正如预期的那样, Pixel智能手机(拥有最高内存的最强大的平台)和智能手表的特征提取时间分别是最短和最长的。图4绘制了其他三个指标的结果。请注意, y轴上的时间刻度以对数刻度表示。

我们进行了以下观察:

- > 模型加载时间随着设备的RAM呈线性下降。对于嗅探而言, 将LSTM模型加载到Pixel智能手机(4 GB RAM)上需要大约100毫秒, 而在Nexus 5智能手机(2 GB RAM)上加载同一模型则需要大约两倍的时间(200毫秒), 是Raspberry Pi(1GB RAM)所用时长的四倍(400毫秒), 以及智能手表(512 MB RAM)的七倍(700毫秒)。在深度呼吸模式下的实验

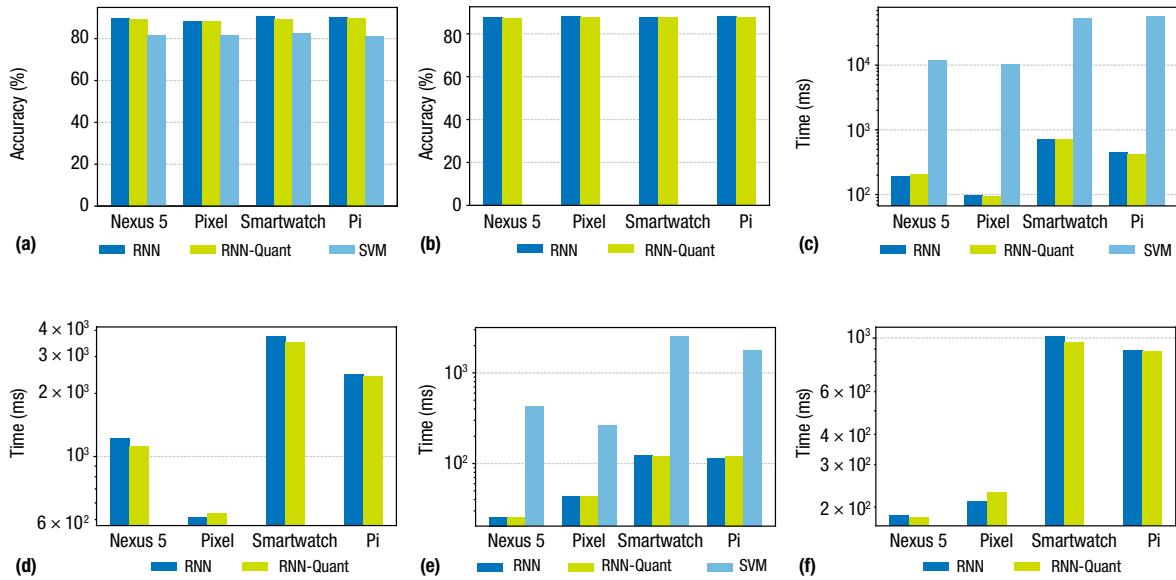


图4. 不同呼吸类型的度量指标: a) 精确度 - 嗅探, b) 模型载入时间 - 嗅探, c) 模型载入时间 - 深度, d) 推断时间 - 嗅探, e) 推断时间 - 深度, f) RNN-Quant。

观察结果也是如此。

> 推断时间取决于设备的处理能力。对于嗅探的平均推断时间(使用LSTM模型), Nexus 5为23毫秒, Pixel 40毫秒, 智能手表和Raspberry Pi均为100毫秒。相比之下, 对于深度呼吸的平均推断时间, Nexus 5约为180毫秒, Pixel 200毫秒, Raspberry Pi 900毫秒, 智能手表1,000毫秒。

> 在集合间的两种呼吸类型中, 系统的精确度都达到了90%。我们还观察到LSTM模型的精确度略高于SVM。这表明, 尽管训练数据量相当小, 但深度学习模型至少可

以和备选SVM方法一样有效。值得注意的是, 当嗅探和深度呼吸应用于集合间数据时, 用户识别的精确度分别下降到75%和70%。这个结果与我们之前的结果一致, 表明需要更大的训练语料库来适应呼吸模式更大范围的情境相关变化。

### 量化效益

图4a和4b中的结果也表明, 使用量化模型不会导致分类精度的任何损失(与原始LSTM模型相比)。然而, 量化降低了模型的大小, 带来了一些执行优势。嗅探的量化模型大小为175KB

(未压缩模型为1.1MB), 深度呼吸为264KB(未压缩模型为1.1MB)。因此, 量化模型可以显著地减少内存占用(4到6倍), 从而使其加载速度更快。

## LSTM与SVM

与普遍观点相反的是, 浅层SVM模型确实提供了与LSTM模型相当的精度, 但其加载到内存中却需要更长的时间(50到100秒)。此外, 用于预测用户标签的执行时间也比LSTM模型在所有设备上的相应时间长5到20倍。我们使用SVM测试嗅探的最佳模型是280MB。相比之下, LSTM模型的大小在未

压缩时为2MB，量化时仅为几百KB。

SVM模型的大尺寸是由于支持多类分类所需的大量支持向量造成的，但这些在二元类分类中是不存在的。实际上，SVM深度呼吸模型的大小为2GB或更高，因此无法在任何设备上加载。我们的研究结果表明，LSTM的深度学习模型比基于SVM的浅层分类器更轻、更耐用（尤其是量化的LSTM模型）。

实验表明，基于呼吸声学的RNN用户身份验证方法不仅耐用，而且轻便，足以在各种资源受限的嵌入式设备上有效地执行。一个适当量化的LSTM深度学习模型可以以高于90%的准确率对用户进行身份验证，并使用大小适中的模型（几百KB）。由此产生的用户身份验证延迟不仅适用于具有代表性的智能手机（小于或等于200毫秒），也适用于资源高度受限的智能手表（小于或等于1秒）。请注意，这些性能数字是通过使用纯CPU计算来实现的，还需通过其他研究人员提出的GPU卸载方法来进行一定的改进。

# 我

们的研究表明，对于许多由传感器驱动的普遍的应用程序来说，特别是当应用程序利用底层传感器数据的时间特性时，RNNs为CNNs提供了一种令人信服的轻量级替代方案。■

## 参考文献

- J. Chauhan et al., “BreathPrint: Breathing Acoustics-Based User Authentication,” Proc. 15th Ann. Int’l Conf. Mobile Systems, Applications, and Services (MobiSys 17), 2017, pp. 278–291.
- N.D. Lane, P. Georgiev, and L. Qendro, “DeepEar: Robust Smartphone Audio Sensing in Unconstrained Acoustic Environments Using Deep Learning,” Proc. 2015 ACM Int’l Joint Conf. Pervasive and Ubiquitous Computing (UbiComp 15), 2015, pp. 283–294.
- S. Bhattacharya and N.D. Lane, “Sparsification and Separation of Deep Learning Layers for Constrained Resource Inference on Wearables,” Proc. 14th ACM Conf. Embedded Network Sensor Systems (SenSys 16), 2016, pp. 176–189.
- A. Mathur et al., “DeepEye: Resource Efficient Local Execution of Multiple Deep Vision Models Using Wearable Commodity Hardware,” Proc. 15th Ann. Int’l Conf. Mobile Systems, Applications, and Services (MobiSys 17), 2017, pp. 29–41.
- L.N. Huynh, Y. Lee, and R.K. Balan, “DeepMon: Mobile GPU-Based Deep Learning Framework for Continuous Vision Applications,” Proc. 15th Ann. Int’l Conf. Mobile Systems, Applications, and Services (MobiSys 17), 2017, pp. 68–81.
- Q. Cao, N. Balasubramanian, and A. Balasubramanian, “MobiRNN: Efficient Recurrent Neural Network Execution on Mobile GPU,” Proc. 1st Int’l Workshop Deep Learning for Mobile Systems and Applications (EMDL 17), 2017, pp. 1–6.
- N.D. Lane et al., “An Early Resource Characterization of Deep Learning on Wearables, Smartphones and Internet-of-Things Devices,” Proc. 2015 Int’l Workshop Internet of Things towards Applications (IoT-App 15), 2015, pp. 7–12.
- M.T. Islam, B. Islam, and S. Nirjon, “SoundSifter: Mitigating Overhearing of Continuous Listening Devices,” Proc. 15th Ann. Int’l Conf. Mobile Systems, Applications, and Services (MobiSys 17), 2017, pp. 29–41.

## 关于作者

**JAGMOHAN CHAUHAN**是阿尔托大学的一名访问研究员，在新南威尔士大学(UNSW)获得了电气工程和电信学博士学位。他的研究兴趣包括移动和物联网系统的性能和安全性以及深度学习等。联系方式: jagmohan.chauhan@aalto.fi。

**SURANGA SENEVIRATNE**是悉尼大学信息技术学院的一名安全讲师，在新南威尔士大学(UNSW)获得了电气工程和电信学博士学位。他的研究兴趣包括移动系统中的隐私和安全，安全领域的人工智能应用以及行为生物识别等。联系方式: suranga.seneviratne@sydney.edu.au。

**YINING HU**是CSIRO Data61研究院的一名研究助理，主要从事可穿戴设备相关的研究领域。她在悉尼大学和哈尔滨工业大学(联合项目)获得了电气工程学士学位，目前正在攻读新南威尔士大学的电气工程和电信学博士学位。她的研究兴趣包括在间歇性网络连接的区域设计基于区块链的微支付系统。联系方式: yining.hu@data61.csiro.au。

**ARCHAN MISRA**是新加坡管理大学信息系统学院的教授和副院长，在马里兰大学帕克分校获得了电气和计算机工程博士学位，并于2005年至2007年期间担任IEEE计算机协会计算机通信技术委员会(TCCC)主席。他的研究兴趣包括新式可穿戴设备、物联网传感和分析技术，以及利用这些技术在零售、智能制造和智能家居等领域构建可感知环境的移动系统等。他在无线网络、移动和普适计算以及城市感知等领域进行了广泛的研究。联系方式: archanm@smu.edu.sg。

**ARUNA SENEVIRATNE**是新南威尔士大学的一名教授，并担任“玛哈那-科恩”(Mahana-korn)电信主席，在巴斯大学获得了电气工程博士学位。他的研究兴趣包括移动技术、网络和通信以及计算机系统安全等。联系方式: aruna.seneviratne@data61.csiro.au。

**YOUNGKI LEE**是新加坡管理大学的一名助理教授，在韩国科学技术院(KAIST)获得了计算机科学博士学位。他的研究兴趣包括建立基础移动和传感器平台，以便人们始终能活动并高度丰富地了解人类行为和环境，以及在诸如日常福利和儿童保育等各个领域构建创新的沉浸式移动应用程序等。联系方式: youngkilee@smu.edu.sg。

9. D.T. Nguyen et al., “SwallowNet: Recurrent Neural Network Detects and Characterizes Eating Patterns,” IEEE Int'l Conf. Pervasive Computing and Communications Workshops (Per Com Workshops), 2017, pp. 401–406.

10. N. Y. Hammerla, S. Halloran, and

T. Plotz, “Deep, Convolutional, and Recurrent Models for Human Activity Recognition Using Wearables,” Proc. 25th Int'l Joint Conf. Artificial Intelligence (IJCAI 16), 2016, pp. 1533–1540.

微信名: 计算人

微信号: jisuanren



# SMP ETST 2018

奖金: ￥52000

追寻文章的前身今世

报名入口





# 找出小肠损伤： 构建基于内镜成像的 学习系统时遇到的挑战

文 | JUNGMO AHN, 韩国亚洲大学 (Ajou University)

HUYNH NGUYEN LOC, 新加坡管理大学 (Singapore Management University)

RAJESH KRISHNA BALAN, 新加坡管理大学

YOUNGKI LEE, 新加坡管理大学

JEONGGIL KO, 韩国亚洲大学

译 | 吴霜, 四川大学

胶囊内镜能鉴定出病人小肠的受损区域, 但其输出的图像的质量通常不高, 也时常发生漏检, 造成误诊, 或使得病人不得不重复检查。本文的作者提出将深度学习模型应用于胶囊内镜, 实时自动处理内镜捕获的图像, 并鉴定出损伤, 胶囊由此可对某一特定区域进一步拍照、调整焦距, 或提高图像质量。同时, 作者也描述了为实现一个可用的自动胶囊内镜系统须面临的技术挑战。

## 胶

囊内镜是诊断小肠损伤的一种有效手段。有线内镜可以从嘴部和肛门进入人体, 对胃部和大肠进行检查, 但难以检查小肠, 因为小肠的位置和长度比较特殊。而胶囊内镜解决了上述问题。病人吞咽下一个

微小的胶囊类似装置, 装置上带有摄像头(如 Given Imaging 的 PillCam SB 系列, 见 [www.givenimaging.com/en-int/Innovative-Solutions/Capsule-Endoscopy/Pages/default.aspx](http://www.givenimaging.com/en-int/Innovative-Solutions/Capsule-Endoscopy/Pages/default.aspx); 或 IntroMedic 的 MicroCam, 见 [www.intromedic.com](http://www.intromedic.com))。

[com/item/item\\_010100.asp](http://com/item/item_010100.asp)) 和无线信号器, 可以在沿小肠运动的同时周期性地拍摄小肠照片, 并将图像传输至人体外的储存装置中。内科医生随后可手动分析收集到的图像, 以鉴别肠道损伤。这些损伤可能是包括消化道出血、小肠肿瘤和克罗恩病 (Crohn's disease) 等多种疾病的征状。

然而, 最先进的胶囊内镜技术仍有两个严重的短板。首先, 内镜拍摄的照片可能并不清楚, 很多时候甚至会漏拍一些损伤。理想情况下, 胶囊在通过可能的受损区域时应该多拍一些照片, 在通过没有问题的区域时少拍些照片, 但由于胶囊体积微小, 且电池容量有限, 其默认的拍照模式, 仅能设置为在固定的时间间隔下拍照 (如每秒2-4帧), 而非智能化采样。其次, 医生只能手动去查阅胶囊拍摄下的这个数量庞大的照片集 (每位病人大约会拍摄50000张照片), 然后做出相应的诊断。这是一个劳动密集型操作, 诊断每个病人大约须耗费医生4个小时, 成本高昂。

为了应对上述两个挑战, 我们提出了一个基于自动反馈机制的胶囊内镜系统, 能够识别小肠损伤, 进而进行适应性采样, 同时具有智能的图像审阅界面。前者使得胶囊内镜在可能的损伤区域采集到质量更高的图片, 可显著提高诊断的准确率; 后者可帮助医生快速找到损伤部位的照片, 医生不再需要手动

翻阅采集到的所有图片。

这一系统的关键是使用深度学习算法, 如卷积神经网络 (**convolutional neural network**) 和生成式对抗网络 (**generative adversarial network**),

地列出自动反馈的胶囊内镜系统的设计方案——特别是我们使用从3位病人体内拍摄到的13.3万张图片组成的数据训练出的损伤识别初级模型, 并列出了目前仍然面临的挑战, 以及为实现这

[ **这一系统的关键是使用深度学习算法, 以便准确地将采集到的图片实时划分  
为“正常”或“可能损伤”两类。** ]

以便准确地将采集到的图片实时划分为“正常”或“(可能的)损伤”两类, 同时根据划分结果操控病人体内的胶囊内镜。我们首先使用了已被诊断为各种小肠相关疾病的病人的肠镜照片, 训练深度学习模型, 然后将模型预装在外部的计算平台上 (通常病人会在肠镜过程中将其佩戴在腰部), 再对图片进行实时分类。当模型将图片识别为“损伤” (或极可能为损伤) 时, 系统向胶囊装置传送的控制信号将暂时提高其成像质量和帧率。如果系统没有检测到更多的损伤信号, 胶囊会将采样率回调至正常的2-4帧/秒, 以节约电能。此外, 系统还会将分类结果标记在采集到的图像上, 并按照分类后的时间线组织图像, 帮助医生快速找到感兴趣的照片。

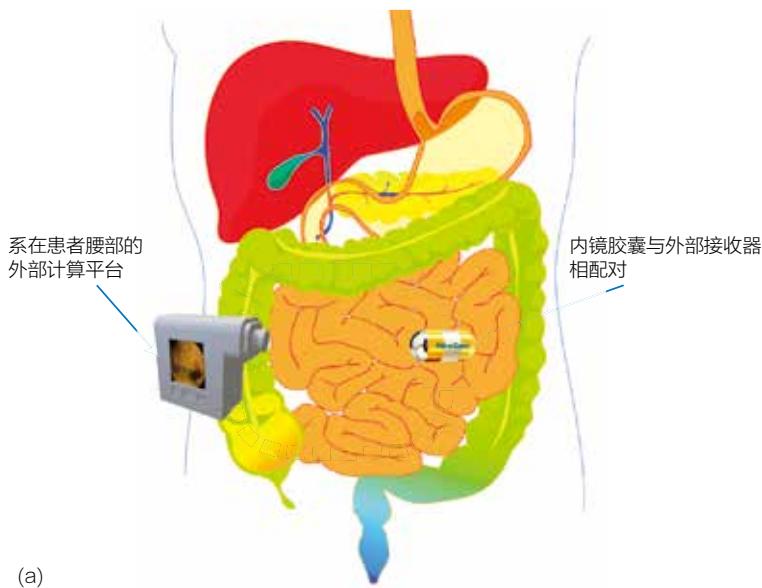
在这篇文章剩余的部分, 我们详尽

一系统而进行的研究的方向。

## 胶囊内镜

胶囊内镜系统包括一个微小的照相机、LED灯、收发器以及电池, 用以捕捉消化道内的图像。如图1a中的蓝色箭头所示, 胶囊会将采集到的图片传送到病人系在腰部的嵌入式设备中。在目前的胶囊内镜检测中, 系统在胶囊被排出病人体内前仅会将图片储存起来; 在图片收集完成后, 医生审阅拍摄到的图像, 从中鉴别出小肠损伤。

尽管胶囊内镜被认为是一种探查小肠损伤的有效手段, 但目前的系统具有局限。首先, 胶囊内镜与结肠镜或胃镜不同, 在后两种检测中, 医生可完全自主地观察器官的特定位置, 而胶囊内



(a)

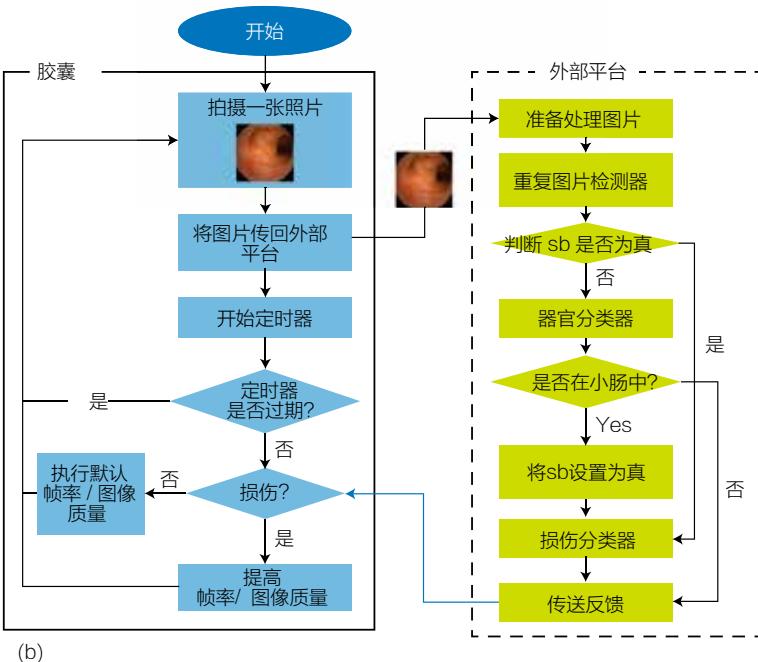


图1. 基于自动反馈的胶囊内镜。(a) 蓝色箭头指示内镜照片从胶囊传输至外部计算平台, 黑色箭头指示平台将基于图片分类结果的反馈送回胶囊。(b) 反馈过程的流程图; “sb”标记用于指示胶囊是否进入了小肠肠道。

镜只能以特定的帧率拍照, 无法适应性地关注某些特定区域。因此, 在许多情况下, 为了给某一部位拍到足够的照片, 病人需要进行多轮耗时长久且花费不菲的检查。其次, 医生需要花费数个小时, 手动审阅全部拍摄到的照片, 以找出病变。

### 基于自动反馈的胶囊内镜

为克服上述挑战, 我们提出, 将基于深度学习的分类反馈环路添加到胶囊内镜系统中(如图1a的黑色箭头所示)。我们希望, 利用胶囊传回至外部嵌入式装置的图片, 将肠道损伤进行实时分类, 并且根据分类结果, 随时调整胶囊拍照的帧率和图片的分辨率。这可以为医生提供足够的肠道损伤照片, 通过一次肠镜检查就能做出准确的诊断。此外, 由于胶囊肠镜检查通常需要8-10小时才能完成, 系统必须确保, 在频繁调整帧率图图像分辨率的同时, 胶囊不会因电池耗尽而关闭。

图1展示了我们的反馈过程的流程图。外部计算平台在收到图片后, 会运行一个特意以低延迟输出结果的、基于CNN的深度学习模型。我们的内镜图像数据库中的数据具有高相似性, 有差别的都是些微小的特征(比如肠道表面红色的出血点或疤痕), 这使得数据库变得独特, 因此, 使用最常用的(也就是ImageNet训练的)、基于

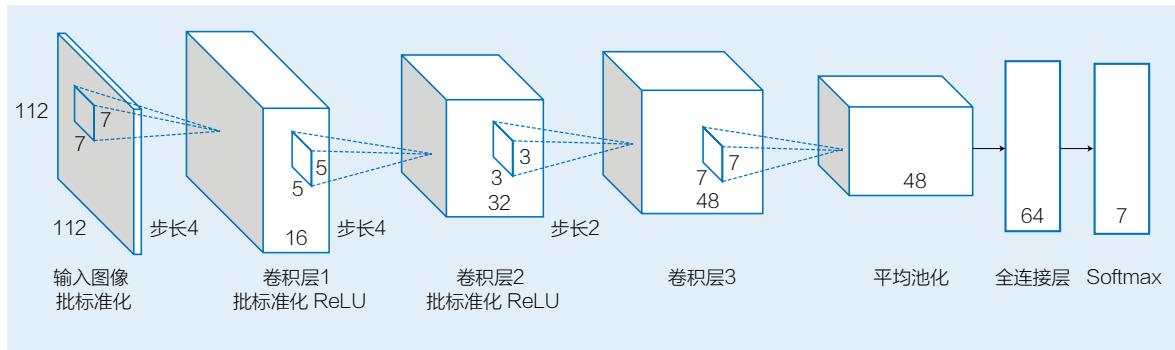


图2. 定制化的3层卷积神经网络，用于损伤分类。我们构建这一网络的目的是实现高准确性与低计算性延迟。

CNN的图像分类模型来对这个数据库进行分类变得困难——这些模型专门用来识别像猫或者狗这样的大型物体。为了克服这个局限，我们在一个训练数据集上从头训练这个网络直至其收敛，根据经验设计模型并挑选超参数，并使用一个验证数据集持续优化这个模型。

图2展示了我们的CNN，它包括3个卷积层。头两层分别进行批标准化（batch normalization）和修正线性单元（rectified linear unit，ReLU）激活。修正线性单元是一种常用的非线性激活函数。使神经网络探查到非线性特征；批标准化将输入的数据标准化，以避免梯度消失<sup>1</sup>。第三个卷积层进行全局平均池化（global average pooling），最后我们添加了一个softmax层进行分类。表1呈现了网络中每一层的细节信息。

在运行这个模型后，嵌入式计算平台会输出一个分类结果，判断一张图片是否可能指示了肠道损伤。尽管外部硬件可以有多种规格，这种嵌入式计算平台通常是一个资源受限的设备，如带有ARM Cortex-A级的处理器（[www.arm.com/products/processors/](http://www.arm.com/products/processors/)

表1. 带有超参数细节的内镜损伤检测神经网络结构

类型/步长	过滤器形状	输入形状
卷积层1/s4	$7 \times 7 \times 3 \times 16$	$112 \times 112 \times 3$
	批标准化及ReLU激活	
卷积层2/s2	$5 \times 5 \times 16 \times 32$	$28 \times 28 \times 16$
	批标准化及ReLU激活	
卷积层3/s2	$3 \times 3 \times 32 \times 48$	$13 \times 13 \times 32$
全局平均池化/s1	Pool $7 \times 7$	$7 \times 7 \times 48$
完全连接	64	$1 \times 48$
Softmax	分类器	$1 \times 64$

cortex-a）。不过，将嵌入式GPU（如英伟达 Jetson-series 处理器）集成，对于运行深度学习模型也是一个可选项。接下来的几个分区的内容会详细介绍我们提出的系统的构成。

### 重复图片检测器

一般来说，胶囊内镜会拍摄不少重复的图像。例如，设备在绘制曲线，或遇到肠道内容物时，可能不会移动，或会移动得非常缓慢，因此可能对同一位置拍摄多张照片。定量来说，之前的研究提示，删除胶囊内镜中的重复图片可以将数据集减小约68%。

剔除外部计算设备中的重复图片，对于我们的设计同样重要，这主要有两个原因。首先，剔除重复图片可将占用的资源最小化。如果不用对重复图片进行分类，我们可减少识别图像时的延迟，这样就能实时将反馈传回胶囊装置。<sup>3</sup>其次，医生必须在审阅整个图像数据集后，才能确认损伤是否存在，因此，去除重复图片可节约医生的时间，他们可以为更多病人提供高质量的医疗服务。为达到这一目的，此前的一些工作评估了相机运动算法，或设计了相似性计算算法，如SIFT 和 SURF。<sup>4,5</sup>我们的系统同时希望通过运用一系列图

片中的像素级相关性，鉴别出相似性极高的图片。

### 设计低延迟的损伤分类器

我们设计基于CNN的分类器的主要目标是低延迟，并能实时将反馈传送

络层 (layer decomposition) 和缓存对基于深度学习的分类器进行加速，以改善延迟。如DeepX8应用AlexNet以500毫秒的延迟分类图像，而Deep-Mon9借助移动GPU进一步将延迟减小到260毫秒。不过，之前的系统在支持实时高

不然只有运用有限的信息作出诊断。

我们的反馈系统试图克服这一局限。我们操控胶囊执行两种不同任务：首先，胶囊可以以一个更高的帧率拍摄图片，并可根据情景调整图片数量；其次，我们可使胶囊拍摄质量更高的图片。由于电池带来的局限，胶囊只能以 $320 \times 320$ 像素的低分辨率拍摄图片。尽管在内镜检测全程均采集高质量图片会给电池带来负担，在某些地方短时拍摄高分辨率的图片，可能辅助医生在首次内镜检查时就做出更优的诊断。我们在之前的评估中进行实验，区分正常图片和腐蚀图片（大约占整个数据集的40%），我们可以以86%的准确率检查出腐蚀图片，以94.4%的准确率分出正常图片。

不同病变的肠镜图片具有微小但可见的区别（如红色或皮疹状表面），因此，设计一个基于深度学习的病变鉴别器或分类器具备其可行性。

给胶囊装置。目前模型的参数设置非常实际，目的是探查到模糊的消化道出血、克罗恩氏病，监控消化道息肉，以及探查小肠肿瘤。我们注意到，以上不同病变的肠镜图片具有微小但可见的区别（如红色或皮疹状表面），因此，设计一个基于深度学习的病变鉴别器或分类器具备其可行性。

为提高深度学习模型的准确度，使用特定领域的数据集 (**domain-specific dataset**) 重复训练及精确调整 (**fine-tuning**) 模型是一种常用方法（如VGG-16和ResNet-1517）。然而，此前的特定领域模型并没有针对延迟和内存使用进行优化，但这对我们的场景是关键性要求。一些最近提出的系统通过应用多种优化途径，如分解神经网

帧率分析上都具有局限性。经过我们优化的推理模型含有32000个参数，运行470万个乘积累加运算式。值得提出的是，AlexNet包含6100万个参数，运行7.21亿个乘积累加运算式，我们的模型相对更精简。在延迟性上，我们的原型模型可在嵌入Jetson Tegra K1的GPU上，在约1.14毫秒内分类一张图片。

### 提高损伤检出率

医生没有办法手动操纵胶囊进行细节性观察，或对特定位置进行观察，这是胶囊内镜的一大局限。因此，拍摄到的图像可能是模糊的，图像上也可能仅有损伤的某一部分。在这种情况下，医生只能安排病人再进行一次内镜检测，以期得到一个更优视窗观察损伤，

### 利用器官分类将电池消耗最小化

我们还计划向系统中加入一个器官分类器，用以鉴定胶囊所处的位置（目前这还没有实现）。这对节约电池能源非常重要，因为胶囊在经吞咽后，至少会花费一个小时通过胃部，这大约是胶囊在人体内可正常工作的总时长的八分之一。在胶囊到达小肠前内关闭图片拍摄功能（或至少降低帧率）可以显著提高胶囊寿命。如图1b中所示，器官分类是一个须首先进行的筛选步骤。一旦系统认定胶囊已经到达小肠，系统可设置

一个标记，指示外部平台，其接受到的图片将不再需要进行器官分类的过滤。之前的工作显示，使用一个CNN的变体<sup>12</sup>可能实现这样的分类，我们计划采取相似的路径。

## 技术挑战

尽管我们目前做出了上述努力，要设计出一个基于全自动反馈的胶囊系统，我们还需克服许多技术或非技术性的挑战。

### 系统层面的挑战

一些挑战与搭建整个系统有关。

**反馈过程的低延迟。**反馈过程如要自动地控制胶囊，需要经过多步的信息传输（包括图片传输）与计算（如图像分类）。因此，整个过程会遭遇多重延迟。若反馈延迟过长，胶囊在接收到控制信号前就可能运动至一个新的位置，因此低延迟的反馈是系统的一个重要设计要求。一般来说，商用胶囊内镜产品的帧率为2帧/秒，这意味着，胶囊本应每500毫秒对一个新的位置拍摄一张新的照片。<sup>11</sup>我们的计算显示，胶囊的运动速度是0.56毫米/秒，而小肠全长约8米，所以胶囊会在约4.5小时后排出小肠。因此，外部反馈给胶囊的指令要在胶囊明显离开可疑区域前到达胶囊。



**图3.** 内镜图片的预过滤。(a) 在小肠中拍摄到的泡泡状小肠液; (b)重复的图片。理想情况下，考虑到资源的有限性，以上两种图片在分类前都应被过滤掉。

MiroCam可以6M/秒的速度传输图片，对于一个100K大小的标准图片（如一张320×320像素大小的原始拜耳模式图片）来说，传输过程需要140毫秒。这一无法避免的延迟使得系统更具挑战。以上数字都是胶囊内镜系统最小的计时限制，并且我们假定了胶囊并不会在小肠内来回移动。

**图片的预过滤。**为了实现实时的图片分类，我们必须减少须用深度学习模型处理的图片的数量。系统预先过滤掉肯定不会显示小肠损伤的图片（根据拍摄位置或图片质量判断），以及与之前处理过的图片相似的重复图片。图3a显示了系统可根据图像质量过滤图片。内镜还会采集到一些泡泡状的肠液图像，这些东西即便用一个经过了良好训练的模型来处理，也对探测目标损伤

无所裨益。图3b显示了胶囊采集到的重复图像，也须被过滤掉。

**低能耗。**病人吞咽下的胶囊须能对整个小肠进行检测。胶囊会先经过喉咙和胃部，然后到达小肠，这段时间叫做胃内滞留时间（Gastric transit time，GTT）。不同病人的GTT不同，少则30分钟，多则数小时。胶囊在整个GTT都处于激活状态，不断采集图像，其内嵌的LED灯也会一直开启，并与外部的嵌入式装置不断通讯。而系统需要保证的是，胶囊应在通过小肠的全部时间（SBTT）内保持激活。SBTT的平均值与标准偏差分别是4.1小时和2.2小时。

<sup>13</sup> 最近的一项研究显示，寿命为12小时的胶囊完成检查率比寿命为8小时的胶囊高出9%。<sup>14</sup> 这种耗时冗长的检查，使得我们提出的适应性图片采集算法必

若反馈延迟过长，胶囊在接收到控制信号前就可能运动至一个新的位置，因此低延迟的反馈是系统的一个重要设计要求。

须具有足够的能效性。

**胶囊位置。**鉴定胶囊的位置，对于确保胶囊在SBTT期间保持激活非常重要。为达到这一目标，我们提出应用一种简单的图片处理技术来判断胶囊从胃部到十二指肠（小肠的起始部位）的位移。此外，尽管这与检测损伤的关系不大，了解胶囊的位置依然有助于医生在检测到损伤后的治疗过程。内镜拍到的照片非常相似，仅用照片无法准确地判断出胶囊位置，因此研究者需要新的解决方案（如使用惯性传感器）。

### 设计学习模型的挑战

其余挑战属于设计分类模型的范畴。

**数据集的不对称性。**深度学习模型的训练原本就需要一个大容量的标记数据集，并且，理想情况下，训练数据集不能对某一特定的类别有过分的偏倚。然而，要找到一个平衡了正常小肠图片和病变小肠图片的数据集非常困难。例如，一个从6位小肠相关疾病患者身上取

样的胶囊内镜数据集中，有超过133000张图片，其中与病变损伤有关的图片只有100张。考虑到内镜需要对整个小肠进行监控和拍照，才能找到几个与病变有关的区域，这种情况是无法避免的。更棘手的是，这100张图片中包含了6种不同的疾病（出血、糜烂、溃疡、肿瘤等），每一个病变类别的样本数就更少了。这种具有偏倚的训练数据，使得我们难以设计出一个准确的针对胶囊内镜图片的深度训练模型。

**收集标记过的真实数据。**众所周知，医院存积了大量的病人数据。然而许多研究者都面临一个问题，即这些数据缺乏真实性标记。比如，对胶囊内镜照片而言，如果病人肠道的某个部位有损伤，其所有内镜图片都会被标记为“有损伤”。这并不意味着所有的图片中都出现了肠道损伤。对每一张图片进行再标记需要大量的劳动力，并且价格不菲，每个图片集至少要耗时四个小时。不过，谷歌<sup>15</sup>之前的工作表明，医学图片的分类标记，可以在专家的指导下由计算机完成。他们的系统根据54

位眼科医生标记的约130000张照片，检测糖尿病相关的视网膜病变，运用Inception-v3网络，其灵敏度达96%。

**同时达到低延迟与高分类准确性。**为了对胶囊实现实时的反馈控制，我们必须在外部平台上对图片进行低延迟的分类。然而，系统的假阴性和假阳性率都不能高，前两者会分别影响系统的临床可靠性和胶囊的能效性。这两点通常互相矛盾，难以平衡，因此，系统必须经过良好的调试，以同时达到以上两个要求。

### 研究指导

我们列出了未来的一些有趣的研究方向。

### 人工数据生成

临床数据集中阳性和阴性样本的严重不平衡，导致我们难以建立起一个准确的模型。

不幸的是，若要收集更多数据，成本极高，且耗时不菲，有时甚至是不可能的。好消息是，像自编码器及GAN这样的新型无监督学习工具及技术可以帮助研究者生成“虚假”但有意义的样本。值得一提的是，这些虚假图片的结构与真实图片类似，但其与真实图片又有足够的区别，可以帮助我们建立起一个准确的预测模型，且不容易导致过拟

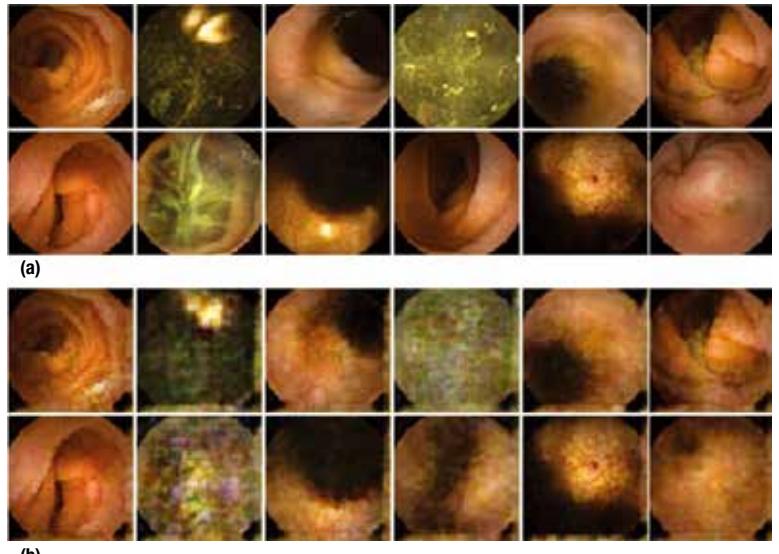
合。然而，一个关键的局限在于，虚假生成的数据的内容并不总是有意义的，比如，有时这些图片上会缺失代表肠道损伤的核心特征。

目前我们使用VAE-GAN模型<sup>16</sup>，这个模型带有一个编码器、一个解码器，以及一个鉴别器。编码器和解码器承担图像生成工作，鉴别器用于判断哪些图片是真实的。图4a展示了原始的图片，图4b展示了GAN利用从真实图片中提取出的特征生成的图片，与真实图片具有类似的视觉特征。

我们相信，在用更多的损伤图片进行训练后，GAN可能生成更多质量更高的“损伤类似”图像样本。必须注意的是，这些图片仍须经过医生的验证后才能使用。总的来说，我们相信，对于生成真实人工临床图片的研究，有助于许多医学学习/分类系统的发展。

### 准确的胶囊位置

正如此前提到的，辨明胶囊在小肠中的位置，也是目前仍须面对的一个挑战。然而，电池能源、装置的重量以及信号干扰等因素都使再安装一个额外的传感器变得困难。应用胶囊拍摄到的图像来定位胶囊，是一种更好的解决方法。不幸的是，由于缺少标记了小肠位置的图像数据，且在小肠不同的位置拍摄的图片差别并不大。如要建立一个准确的模型，就需要技术和临床人员合



**图 4.** 运用 GAN 人工生成的内镜图片: (a) 原始图片和 (b) 人工生成的假图片。人工生成的假图同样可以用于克服数据集里阴性和阳性结果不平衡带来的问题。

作，收集到大量标记过的图片数据，并对神经网络模型进行校验。

### 零假阴性分类器

即便我们的基线准确水平本就较高，系统仍不能将任一个病症图片分类为正常图片（即制造假阴性），这是至关重要的。假阳性图片可以在分类完成后被手动去除，但任意一个假阴性分类都可能导致临床上的严重后果。不幸的是，要训练处一个低假阴性率，或零假阴性率的模型困难重重——在我们的数据集是高度不对称的时候，尤其如此。此外，在系统做出假阴性推断时，我们还缺少一个被良好定义的代价函数

来对神经网络进行负向调节。像分类交叉熵代价函数这种常用的代价函数，会同时负向调节假阳性和假阴性结果。因此，对于合适的代价函数的研究，以及生成各种可用的肠道损伤图片的研究，都可以帮助研究者设计出有效的神经网络，以建立自动化的反馈胶囊内镜系统。

在这篇文章中，我们描述了一个初步的工作，即设计一个增强的胶囊内镜系统，以应对目前最先进的胶囊内镜系统仍面临的两大挑战：图片质量低或漏检，以及图片审阅耗时费力。我们特别提出了一个低延迟、高准确度、基于深度学习的自动化反馈机制，可以实时鉴

系统的假阴性和假阳性率都不能高，前两者会分别影响系统的临床可靠性和胶囊的能效性。

别出可能是肠道损伤的图片，并随时增加图片的采集速率和图片分辨率。我们同时对多种系统层面及深度学习相关的挑战进行的评议，例如实时反馈、能量的限制，以及图片质量问题。唯有解决了以上问题，这个自动化的反馈胶囊内镜系统才是可用的。最后，我们正在设计一个智能的图片审阅界面：应用我们的图片分类器，尽可能减少医生必须审阅的图片的数量。

## 致谢

这一工作得到了大韩民国科学部IRTC项目的支持(IITP-2018-2016-0-00309-002)，同时也得到了DGIST研究与发展项目(CPS国际中心)“寻找设计自动化临床事件监测装置中的未来可穿戴设备时的未尽需求”(Identifying Unmet Requirements for Future Wearable Devices in Designing Autonomous Clinical Event Detection Applications)项目的支持。

## 参考文献

1. Jiuxiang Gu et al., “Recent Advances in Convolutional Neural Networks,” *Pattern Recognition*, vol. 77, May 2018, pp. 354–377.
2. D.K. Lakovidis and A. Koulaouzidis, “Software for Enhanced Video Capsule Endoscopy: Challenges for Essential Progress,” *Nature Reviews Gastroenterology & Hepatology*, vol. 12, no. 3, 2015, pp. 172–186.
3. B. Kang and H. Choo, “A Deep-Learning-Based Emergency Alert System,” *ICT Express*, vol. 2, no. 2, 2016, pp. 67–70.
4. H.-G. Lee et al., “Reducing Redundancy in Wireless Capsule Endoscopy Videos,” *Computers in Biology and Medicine*, vol. 43, no. 6, 2013, pp. 670–682.
5. H. Liu et al., “Wireless Capsule Endoscopy Video Reduction Based on Camera Motion Estimation,” *J. Digital Imaging*, vol. 26, no. 2, 2013, pp. 287–301.
6. K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” preprint, arXiv:1409.1556, 2014.
7. K. He et al., “Deep Residual Learning for Image Recognition,” *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition (CVPR 16)*, 2016, pp. 770–778.
8. N.D. Lane et al., “DeepX: A Software Accelerator for Low-Power Deep Learning Inference on Mobile Devices,” *Proc. 15th Int'l Conf. Information Processing in Sensor Networks (IPSN 16)*, 2016, article no. 23.
9. L.N. Huynh, Y. Lee, and R.J. Balan, “DeepMon: Mobile GPU-Based Deep Learning Framework for Continuous Vision Applications,” *Proc. 15th Ann. Int'l Conf. Mobile Systems, Applications, and Services (MobiSys 17)*, 2017, pp. 82–95.
10. NVIDIA Corp., NVIDIA Tegra K1: A New Era in Mobile Computing, v1.0, white paper, 2014; [www.nvidia.com](http://www.nvidia.com)

## 关于作者

[/content/PDF/tegra\\_white\\_papers/Tegra\\_K1\\_whitepaper\\_v1.0.pdf](/content/PDF/tegra_white_papers/Tegra_K1_whitepaper_v1.0.pdf)

11. Z. Liao, C. Xu, and Z.-S. Li, "Completion Rate and Diagnostic Yield of Small-Bowel Capsule Endoscopy: 1 vs. 2 Frames per Second," *Endoscopy*, vol. 42, no. 5, 2010, pp. 360–364.

12. Y. Zou et al., "Classifying Digestive Organs in Wireless Capsule Endoscopy Images Based on Deep Convolutional Neural Network," *Proc. 2015 IEEE Int'l Conf. Digital Signal Processing (DSP 15)*, 2015, pp. 1274–1278.

13. G. Ou et al., "Effect of Longer Battery Life on Small Bowel Capsule Endoscopy," *World J. Gastroenterology*, vol. 21, no. 9, 2015, pp. 2677–2682.

14. M. Rahman et al., "Comparison of the Diagnostic Yield and Outcomes between Standard 8 H Capsule Endoscopy and the New 12 H Capsule Endoscopy for Investigating Small Bowel Pathology," *World J. Gastroenterology*, vol. 21, no. 18, 2015, pp. 5542–5547.

15. V. Gulshan et al., "Development and Validation of a Deep Learning

**JUNGMO AHN**是韩国亚洲大学 (Ajou University) 信息技术学院软件与计算机工程系博士生, 同时也是亚洲嵌入式智能系统工程实验室的成员。他的研究兴趣包括嵌入式学习系统, 移动计算以及健康应用。他的联系方式是ajm100@ajou.ac.kr。

**HUYNH NGUYEN LOC**是新加坡管理大学信息系统学院博士生。他目前的研究集中在移动计算, 并行编程以及深度学习应用方面。他的联系方式: nlhuynh.2014@phdis.smu.edu.sg。

**RAJESH KRISHNA BALAN**是新加坡管理大学信息系统学教授、Live 实验室城市生活方式创新平台负责人。他的研究兴趣包括移动系统、能源管理和可用性。巴兰在卡耐基梅隆大学获得计算机科学博士学位, 他的联系方式是: rajesh@smu.edu.sg。

**YOUNGKI LEE**是新加坡管理大学信息系统学助理教授。他的研究专注于移动和传感器平台, 以建立随时可用、高度丰富的针对人类行为的感知。他同时与各领域专家合作, 建立创新性的生命沉浸式移动应用, 以服务于人类日常健康, 儿童关护以及广告。他在韩国科学技术院获得计算机科学博士学位, 他的联系方式是: youngkilee@smu.edu.sg。

**JEONGGIL KO**是韩国亚洲大学信息技术学院软件与计算机工程系、医学院生物医学信息学系助理教授, 他同时也是亚洲嵌入式智能系统工程实验室的负责人。他的研究兴趣在于运用物联网环境智能及信息物理系统研发基于web及云端的感知系统。他在约翰斯·霍普金斯大学获得计算机科学博士学位, 他是IEEE高级会员, 也是这篇文章的通讯作者。他的联系方式是: gko@ajou.ac.kr

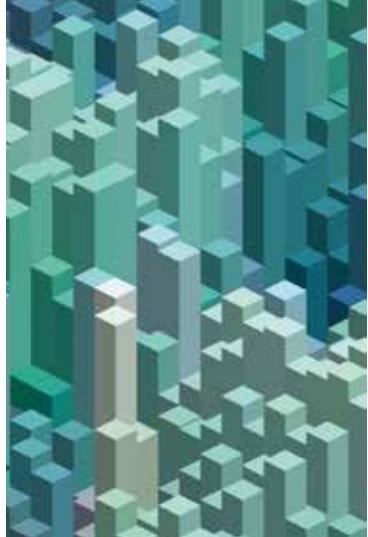
- |  |   |
|--|---|
| Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs," <i>JAMA</i> , vol. 316, no. 22, 2016, pp. 2402–2410. | Similarity Metric," <i>Proc. 33rd Int'l Conf. Machine Learning (ICML 16)</i> , 2016, pp. 1558–1566. |
| 16. A.B.L. Larsen et al., "Autoencoding beyond Pixels Using a Learned  |   |



# 真实世界的计算 是最大挑战

文 | Marilyn Wolf, 佐治亚理工学院  
译 | 韩晶晶

网络 - 实体系统 (*Cyber-physical system*) 不再是计算领域的边缘地带——它们现在是计算机科学家和工程师关注的核心问题。



# 在

计算科学中, 网络 - 实体

系统 (CPS) 通常被认为既小又简单。然而事实恰恰相反。这些系统依赖着庞大而复杂的硬件和软件。CPS不仅面临着信息技术 (IT) 系统的所有难题, 而且还要应对独特而艰巨的挑战。

福特汽车在 2016 年国际消费电子展上宣称, 他们的 F-150 皮卡——多年来一直最受美国人欢迎的汽车, 搭载的软件已经有超过 1.5 亿行代码。无论按何种标准, 这都是一个大型的软件项目。这种规模的软件系统面临着诸多难题。

它们一般是用多个不同来源的软件构建的, 而且往往要使用多种计算机语言。

版本控制和配置管理都必须在代码库上执行。软件也必须经过测试。

现在想象一下, 所有这些软件都被包裹在 4000 磅的金属和塑料中。网络 - 实体计算技术的实体一面提高了风险。失败的代价不仅是生产力的损失, 还有实物损坏, 甚至生命。

与实体设备联系紧密的计算系统必须满足一般 IT 系统无需关注的约束条件。时间是网络 - 实体系统运作的关键。未能满足时间限制和截止日期都可

能导致系统崩溃。在 CPS 中, 定时是最重要的功能特征。

网络 - 实体系统也面临着漫长生命周期的挑战, 与之相比, 一些传统 IT 系统那相当长的寿命也相形见绌。汽车通常要用上 10~20 年, 而许多飞机飞行了半个世纪或更长时间。越来越多地使用物联网传感器的道路和建筑物可能要持续使用数个世纪。规模横跨大陆的电网也必须要连续运行。

在网络 - 实体系统中, 安全也是一个上升到了全新层面的重要问题。网络 - 实体系统中为数众多不够安全的计

算设备不仅威胁信息安全，而且还威胁到实体设备的安全。迪米特里奥斯·塞尔帕诺斯 (Dimitrios Serpanos) 和我最近指出 (“Scanning The Issue,” Proc. IEEE, vol.106, no.1, Jan.2018, pp.7–8; <https://doi.org/10.1109/JPROC.2017.2777799>)，在网络 - 实体系统中，我们不能认为防范意外事故和防范人为破坏是相互独立，完全不同的：容易遭破坏的系统，防范意外的能力也会降低；要考虑防范意外事故，也意味着一些传统的计算机安全方法不能用于网络 - 实体系统。

这些挑战意味着什么？我们应该怎样应对？我们是否需要以有别于 IT 系统的发展方式来对待网络 - 实体计算技术？

我们需要做的第一件事就是认真对待 CPS 专业协会和大学经常把网络 - 实体系统视为次要的课题。而我们应该把网络 - 实体计算推向专业领域关注的最前沿。

所有计算机科学家和工程师都应该了解网络 - 实体计算的一些基本原则，就像计算机科学家理应知道算法和操作系统的基本原理一样。实时和低功耗计算是 CPS 的基础，也是其他主要计算应用的基础。多媒体和虚拟现实依赖于嵌入式计算和网络 - 实体计算建立起的基本概念。而数据中心必须满足功耗限制，这也是嵌入式计算开创的一个领域。

CPS 从业者需要吸取 IT 和科学计算的成功经验，但也不是总要遵循这一原则。CPS 设计师还需要了解这些知

识存在哪些不足或缺陷。在很多情况下，我们知道如何设计计算系统，使之适应 CPS 的需求；但在某些情况下，特别是在防范事故和防范人为破坏方面，我们还有更多的工作要做。

为了应对这两类安全需求的挑战，我们必须扩展传统的工程方法。实体设备的设计师习惯于制造适用于不变的环境的机器。不幸的是，计算机系统面临着一系列不断变化的安全挑战和潜在的攻击。即使是没有直接连接到互联网的网络 - 实体系统也容易受到来自这些媒介的攻击——许多研究都表明网络 - 实体设施可能遭受间接攻击。我们需要开发可升级的网络 - 实体系统来应对新的网络威胁。我们也需要针对各种攻击都有固有抵抗能力的网络 - 实体结构。

网络 - 实体系统 (Cyber-physical system) 不再是计算领域的边缘地带——它们现在是计算机科学家和工程师关注的核心问题。现在，还想回到计算机存放在数据中心与世隔绝的日子已经太迟了。把努力和技能结合起来，我们可以战胜这些挑战。■

**Marilyn Wolf** 是佐治亚理工学院嵌入式计算系统 Rhesa “Ray” S. Farmer, Jr. 杰出讲席教授和佐治亚研究联盟杰出学者。她是 IEEE 会员、ACM 会员、IEEE 计算机学会的金核心成员以及 ASEE Frederick E. Terman 奖获得者。请通过 [marilyn.wolf@gatech.edu](mailto:marilyn.wolf@gatech.edu) 与她联系。



## Call for Articles

**IEEE Software** seeks practical, readable articles that will appeal to experts and nonexperts alike. The magazine aims to deliver reliable, useful, leading-edge information to software developers, engineers, and managers to help them stay on top of rapid technology change. Topics include requirements, design, construction, tools, project management, process improvement, maintenance, testing, education and training, quality, standards, and more.

Author guidelines:  
[www.computer.org/software/author](http://www.computer.org/software/author)  
Further details: [software@computer.org](mailto:software@computer.org)  
[www.computer.org/software](http://www.computer.org/software)





# 孩子身边的网络威胁

文 | Nir Kshetri, 北卡罗来纳大学格林波若分校

Jeffrey Voas, IEEE 会员

译 | 韩晶晶

连接互联网的玩具为窃取个人数据提供了一条常被忽视的途径，尤其是那些最容易泄露的数据。政府和私营企业采取适当措施可以将风险降到最低，但监控智能玩具使用的责任归根结底应有家长承担。



## 根

据瞻博网络研究公司  
(Juniper Research) 的

数据，2017年全球智能玩  
具市场规模为50亿美元。智能玩具运用  
了传感器、摄像头、麦克风、数据存储、  
语音识别和GPS等技术。这些技术使  
玩具变得更加有趣和吸引人，但同时也  
为网络攻击提供了更多媒介。

2017年初，有人发现智能毛绒玩  
具 CloudPets 存在一个安全漏洞，这种  
玩具能通过蓝牙连接智能手机，可以让  
父母和孩子远距离相互发送语音消息。  
利用这个漏洞，黑客可以访问存储在云  
中的儿童个人信息、照片和录音。据一

位安全研究人员称，超过 820000 个用  
户账户被破解，其中包括 220 万份语音  
录音。曾有一次，有人利用窃取到的信  
息勒索金钱。

两年前，香港玩具制造商伟易达遭  
遇了更严重的网络数据泄露。通过该公  
司网站上的一个漏洞，黑客获得了美国、  
加拿大，欧洲，拉丁美洲，澳大利亚和  
新西兰的 630 多万儿童账户中的照片  
和聊天记录。

## 松懈的防范

部分原因是技术预算有限，许多智

能玩具制造商对安全和隐私的保护都  
相当脆弱。

例如，伟易达使用过时的协议保护  
玩具用户的数据。该公司的标准做法是  
用散列(hash) 算法处理密码，也就是  
把它们转换为不同的数字字符集，来保  
护数据库。据报道，伟易达使用的散列  
算法是 MD5，该算法的开发者在 2012  
年 6 月公开宣布，由于软件限制和算法  
最初发布以来计算能力的急剧提升，它  
已经过时了。

有时，即便遭受了重大攻击，玩具  
制造商也不会认真努力改善安全和隐  
私。例如，在大规模数据泄露之后，伟易

达也没采取什么措施来解决安全问题。相反它修改了用户协议的条款和条件，将未来任何数据泄漏的责任转移给家长，这让安全专家很是愤慨。<sup>4</sup>

## 盗用儿童身份

智能玩具制造商持有的信息可能比信用卡数据更敏感，更有价值。在伟易达的案例中，被盗数据包括父母名字、电子邮件地址、密码、用于验证帐户信息的提示问题及答案、IP 地址、邮寄地址和下载历史记录，以及有关孩子的信息，如姓名、性别和生日。<sup>5</sup>

一项研究发现，儿童身份遭到盗用的可能性比成年人高 51 倍。网络犯罪分子特别需要儿童的数据，因为在申请信用卡之前，儿童不太可能发现他们的身份被人盗用了，而那可能要到几十年后。联邦贸易委员会(FTC)官员在出席众议院筹款委员会的社会安全小组的听证会时指出：“儿童的社会安全号码具有独特的价值，因为它们没有信用记录，可以与任何姓名和出生日期配对。”<sup>12</sup>

此外，各类敏感数据，如 GPS 位置信息、图片或视频中的外貌特征、儿童的兴趣爱好(可用来获取他们的信任)都存在遭到滥用的风险。<sup>6</sup>

## 儿童保护工作

政府和消费者监督机构已开始采取措施保护儿童，以防他们遭受与玩具相关的数据窃取和隐私侵犯，并重点强

调某些玩具会带来风险。

### 监管措施

在某些国家或地区，监管部门已制定法规，以保护儿童免受智能玩具造成的有害影响。

在美国，FTC 于 2017 年 6 月扩大儿童在线隐私保护法案(COPPA)的适用范围，把针对儿童的智能玩具和其他连接互联网的设备包括进去。扩展后的法案明确规定这些设备属于“网站或在线服务”的受保护类别。<sup>13</sup>

在德国，带有隐藏摄像头或麦克风的无线设备是非法的。美国创世纪玩具公司(Genesis Toys)制造的 My Friend Cayla 娃娃通过无线方式连接到互联网，可回答问题。2017 年 2 月，德国联邦网络局在获悉黑客能通过这种娃娃听到儿童的谈话并利用其不安全的蓝牙连接窃取儿童个人数据后，将这种玩具分类为“间谍装置”。该机构用一个拆开的娃娃向家长展示了其内置的麦克风，并禁止它以后在德国销售。

欧盟的“通用数据保护条例”(GDPR)旨在全面加强和统一所有成员国的消费者数据保护，一些人希望该法案能解决与智能玩具相关的隐私和数据安全问题。GDPR 于 2018 年 5 月 25 日开始实施。

### 加强公众防患意识

政府机构和私人监督组织正在设法提高公众对智能玩具安全性和隐私问题的风险意识。

在 2017 年 7 月的公共服务公告中，

美国联邦调查局(FBI)的互联网犯罪投诉中心警告消费者，遭到破解的智能玩具可能会导致敏感信息泄露，包括儿童的姓名、学校、喜好和地理位置等，有可能导致身份被盗用。<sup>19</sup>

挪威消费者委员会审查了包括 My Friend Cayla 在内的各种联网玩具，并提出了四个关键问题：

- > 缺乏安全保障——任何人都可以轻松控制玩具；
- > 非法或不合理的服务条款——用户必须同意在没有通知的情况下更改条款，将用户个人数据用于定向广告，以及与第三方共享信息；
- > 侵犯隐私权——例如，儿童告诉 My Friend Cayla 娃娃的任何内容都会被记录并传输给制造商的技术合作伙伴 Nuance Communications；
- > 隐蔽的针对儿童的营销——例如，预编程的语言会推广某些特定的产品，如迪士尼电影。

英国消费者监督组织 Which? 同样测试了七款智能玩具，发现 CloudPets、Furby Connect、i-Que 智能机器人和 Toy-Fi Teddy 都存在安全漏洞。该组织敦促零售商停止销售这些存在安全缺陷的玩具。

这些组织和其他公共利益团体正在共同努力，以促成积极的政策变化。例如，在 2016 年，超过 18 个隐私团体向 FTC 和欧盟提出有关智能玩具的投诉。这些尝试最终会取得多大的成功仍

然是一个悬而未决的问题，因为消费者有时会无脑地抢着购买“智能”设备。

**经**验证明，许多智能玩具制造商完全忽视了安全和隐私问题，或是只在这方面花了一点点心思。因此他们的产品可能比其他物联网设备更容易受到网络攻击，这为黑客提供了一个经常被人忽视

可能促使玩具制造商提高其产品的安全性。■

## 参考文献

1. “Smart Toys: Market Summary 2017,” Juniper Research; [www.juniperresearch.com/resources/infographics/smart-toys-market-summary-2017](http://www.juniperresearch.com/resources/infographics/smart-toys-market-summary-2017).
2. S. Larson, “Stuffed Toys Leak Millions of Voice Recordings from Kids and Parents,” CNN Tech, 27 Feb. 2017; [money.cnn.com/2017/02/27/technology/cloudpets-data-leak-voices-photos/index.html](http://money.cnn.com/2017/02/27/technology/cloudpets-data-leak-voices-photos/index.html).
3. T. Hunt, “Data from Connected CloudPets Teddy Bears Leaked and Ransom, Exposing Kids’ Voice Messages,” blog, 20 Dec. 2017, [www.troyhunt.com/data-from-connected-cloudpets-teddy-bears-leaked-and-ransom-exposing-kids-voice-messages](http://www.troyhunt.com/data-from-connected-cloudpets-teddy-bears-leaked-and-ransom-exposing-kids-voice-messages).
4. L. Kelion, “Parents Urged to Boycott VTech Toys after Hack,” BBC News, 10 Feb. 2016; [www.bbc.com/news/technology-35532644](http://www.bbc.com/news/technology-35532644).
5. L. Eadicicco, “Everything to Know about a Massive Hack Targeting Children’s Toys,” Time, 1 Dec. 2015; <http://time.com/4130704/vtech-hack-childrens-toys>.
6. J. Kestenbaum, “The FTC and FBI Are Shining the Spotlight on Your Kid’s Smart Toys,” The Hill, 8 Aug. 2017; [thehill.com/blogs/pundits-blog/technology/345119-the-ftc-and-fbi-put-the-spotlight-on-your-kids-smart-toys](http://thehill.com/blogs/pundits-blog/technology/345119-the-ftc-and-fbi-put-the-spotlight-on-your-kids-smart-toys).
7. “10 Tips to Protect Your Kids’ Toys from Hackers This Holiday Season,” Vanderbilt Univ. News, 14 Dec. 2017; [https://news.vanderbilt.edu/2017/12/14/10-tips-to-protect-your-kids-toys-from-hackers-this-holiday-season](http://news.vanderbilt.edu/2017/12/14/10-tips-to-protect-your-kids-toys-from-hackers-this-holiday-season).
8. J. Keane, “VTech’s Hacked Toys: How Not to Rebuild Your Reputation after a Cyber Attack,” Paste Mag., 15 Feb. 2016; [www.pastemagazine.com/articles/2016/02/vtechs-hacked-toys-how-not-to-rebuild-your-reputation.html](http://www.pastemagazine.com/articles/2016/02/vtechs-hacked-toys-how-not-to-rebuild-your-reputation.html).
9. Z. Whittacker, “MD5 password scrambler ‘no longer safe,’” ZDNet, 7 June 2012; [www.zdnet.com/article/md5-password-scrambler-no-longer-safe](http://www.zdnet.com/article/md5-password-scrambler-no-longer-safe).
10. H. Kuchler, “Toymaker VTech Hit by Cyber Attack,” Financial Times, 29 Nov. 2015; [www.ft.com/content](http://www.ft.com/content)

## 黑客通过玩具公司的一个漏洞，获得了多个国家的 630 多万儿童账户中的照片和聊天记录。

的切入点。此外，联网玩具的安全漏洞使儿童面临身份遭人盗用的危险，这种盗窃行为，还有网络犯罪分子对用户家庭的监视可能已经持续了多年，都未被发现。

制造商可能缺乏增强智能玩具安全性的能力、资源和动力。解决这个问题的监管工作刚刚起步，各个政府机构和消费者监督组织正试图填补这一空白。不过，现在监控智能玩具使用和保护儿童个人数据的责任主要在父母身上。对于大多数父母而言，要充分理解智能玩具的安全和隐私风险，可能是不切实际的。但家长应知道一个通用的规则，就是要当心那些具有录音功能、可以连接互联网或要求提供个人资料的玩具。对那些“令人毛骨悚然”的玩具娃娃和其他可疑的智能玩具，家长应要求退货和换货，或干脆拒绝购买，这才

- /2bcf9ee6-9701-11e5-95c7-d47aa298f769.
11. R. Power, "Child Identity Theft; A Lot of Questions Need to Be Answered, but the Most Important One Is 'Has It Happened to Your Child?,'" blog, Carnegie Mellon Univ. CyLab, 1 Apr. 2011; [www.cyblog.cylab.cmu.edu/2011/03/child-identity-theft.html](http://www.cyblog.cylab.cmu.edu/2011/03/child-identity-theft.html).
12. Federal Trade Commission, "FTC Testifies on Children's Identity Theft," press release, 1 Sept. 2011; [www.ftc.gov/news-events/press-releases/2011/09/ftc-testifies-childrens-identity-theft](http://www.ftc.gov/news-events/press-releases/2011/09/ftc-testifies-childrens-identity-theft).
13. S.A. Reiter, "FBI and FTC on Privacy Risks Stemming from 'Smart' Toys," Lexology, 27 July 2017; [www.lexology.com/library/detail.aspx?g=78ff6c12-ed11-45d9-8fb2-48a1f8595f9c](http://www.lexology.com/library/detail.aspx?g=78ff6c12-ed11-45d9-8fb2-48a1f8595f9c).
14. S. Fogel, "Germany Bans Creepy Doll over Privacy Concerns," Engadget, 17 Feb. 2017; [www.engadget.com/2017/02/17/germany-bans-my-friend-cayla-doll](http://www.engadget.com/2017/02/17/germany-bans-my-friend-cayla-doll).
15. S. Bernando, "The Latest Hack Is Targeting Your Kids' Smart Toys," blog, Experian, 6 Dec. 2017; [www.experian.com/blogs/ask-experian/the-latest-hack-is-targeting-your-kids-smart-toys](http://www.experian.com/blogs/ask-experian/the-latest-hack-is-targeting-your-kids-smart-toys).
16. A. Petroff, "Germany Tells Parents to Destroy Microphone in 'Illegal' Doll," CNN Tech, 17 Feb. 2017; [money.cnn.com/2017/02/17/technology/germany-doll-my-friend-cayla/index.html](http://money.cnn.com/2017/02/17/technology/germany-doll-my-friend-cayla/index.html).
17. E. Silfversten, "A Smart Toy Could Have Personal Details for Life, Not Just for Christmas," blog, RAND Corp., 21 Dec. 2017; [www.rand.org/blog/2017/12/a-smart-toy-could-have-personal-details-for-life-not.html](http://www.rand.org/blog/2017/12/a-smart-toy-could-have-personal-details-for-life-not.html).
18. FBI Internet Crime Complaint Center, "Consumer Notice: Internet-Connected Toys Could Present Privacy and Contact Concerns for Children," alert no. I-071717(Revised)-PSA, 17 July 2017; [www.ic3.gov/media/2017/170717.aspx](http://www.ic3.gov/media/2017/170717.aspx).
19. A. Newcomb, "FBI Warns Parents of Privacy Risks with Internet-Connected Toys," NBC News, 18 July 2017; [www.nbcnews.com/tech/security/fbi-warns-parentsprivacy-risks-internet-connected-toys-n784126](http://www.nbcnews.com/tech/security/fbi-warns-parentsprivacy-risks-internet-connected-toys-n784126).
20. BEUC, "Consumer Organisations across the EU Take Action against Flawed Internet-Connected Toys," press release, 12 June 2016; [www.beuc.eu/publications/consumer-organisations-across-eu-take-action-against-flawed-internet-connected-toys/html](http://www.beuc.eu/publications/consumer-organisations-across-eu-take-action-against-flawed-internet-connected-toys/html).
21. R. Smithers, "Strangers Can Talk to Your Child through 'Connected' Toys, Investigation Finds," The Guardian, 14 Nov. 2017; [www.theguardian.com/technology/2017/nov/14/retailers-urged-to-withdraw-toys-that-allow-hackers-to-talk-to-children](http://www.theguardian.com/technology/2017/nov/14/retailers-urged-to-withdraw-toys-that-allow-hackers-to-talk-to-children).
22. "Connected Toys Have 'Worrying' Security Issues," BBC News, 14 Nov. 2017; [www.bbc.com/news/technology-41976031](http://www.bbc.com/news/technology-41976031).

**Nir Kshetri** 里北卡罗来纳大学格林波若分校布萊恩商学院经济学教授。请通过 [nbkshetr@uncg.edu](mailto:nbkshetr@uncg.edu) 与他联系。

**Jeffrey Voas** 是 Digital 的共同创始人、《计算机科学评论》Cybertrust 专栏编辑和 IEEE 会员。请通过 [j.voas@ieee.org](mailto:j.voas@ieee.org) 与他联系。



# 保持联系。

无论你在哪里，都能紧随IEEE计算机协会的脚步。

**在Twitter、Facebook、Linkedin和YouTube上关注我们。**



@ComputerSociety, @ComputingNow



[facebook.com/IEEEComputerSociety](https://facebook.com/IEEEComputerSociety)  
[facebook.com/ComputingNow](https://facebook.com/ComputingNow)



IEEE Computer Society, Computing Now



[youtube.com/ieeecomputersociety](https://youtube.com/ieeecomputersociety)