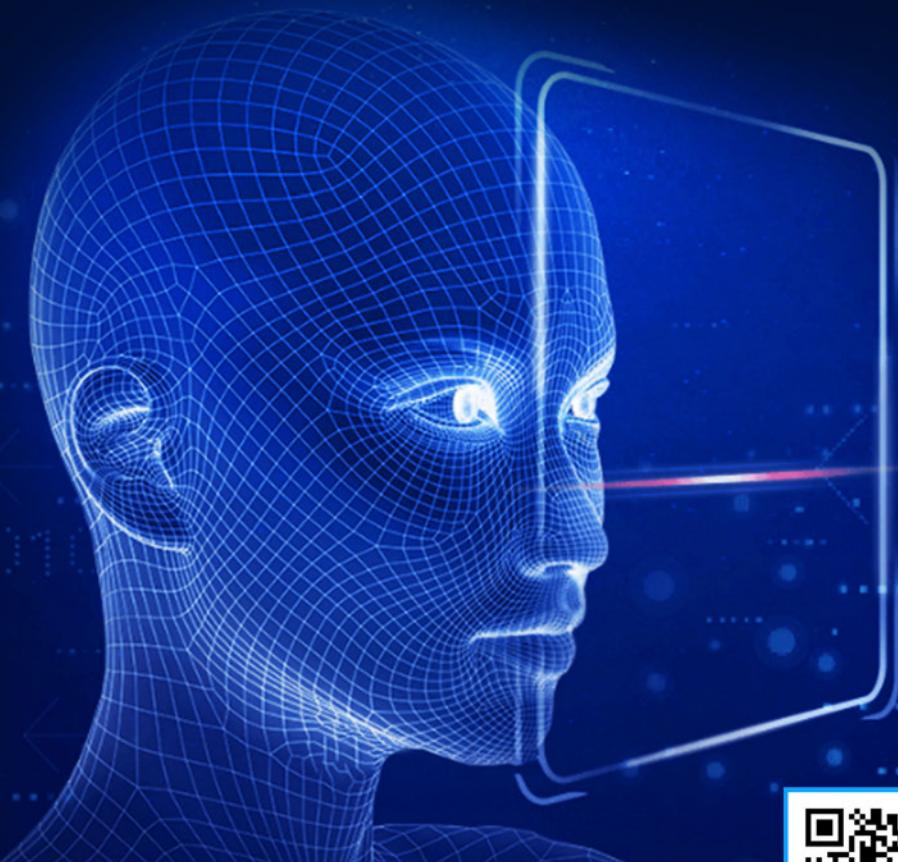




2018年02月刊

A I - F R O N T



关注落地技术，探寻AI应用场景



聚焦最新技术热点 沉淀最优实践经验

[北京站]2018

北京·国际会议中心

演讲：2018年4月20-22日 培训：2018年4月18-19日

精彩案例 先睹为快

《Netflix的工程文化：是什么在激励着我们？》

Speaker: Katharina Probst

Netflix 工程总监

《Apache Kafka的过去，现在，和未来》

Speaker: Jun Rao

Confluent 联合创始人

《人工智能系统中的安全风险》

Speaker: 李康

360网络安全北美研究院负责人，IoT安全研究院院长

《从C#看开放对编程语言发展的影响》

Speaker: Mads Torgersen

微软 C#编程语言Program Manager

《Lavas：PWA的探索与最佳实践》

Speaker: 彭星

百度 资深前端工程师

《浅谈前端交互的基础设施的建设》

Speaker: 程劭非（寒冬）

淘宝 高级技术专家

《深入Apache Spark流计算引擎：Structured Streaming》

Speaker: 朱诗雄

Databricks软件开发工程师, Apache Spark PMC和Committer

《AI大数据时代电商攻防：AI对抗AI》

Speaker: 苏志刚

京东安全 硅谷研究中心负责人

《QUIC在手机微博中的应用实践》

Speaker: 聂永

新浪微博 技术专家

《阿基米德微服务及治理平台》

Speaker: 张晋军

京东 基础架构部服务治理组负责人，架构师

8折 优惠报名中，立减1360元
团 购 享 受 更 多 优 惠

访问官网获取更多前沿技术趋势

2018.qconbeijing.com

如有任何问题，欢迎咨询

电话：15110019061，微信：qcon-0410



ArchSummit 全球架构师峰会

2018 · 深圳站

从2012年开始算起，InfoQ已经举办了9场

ArchSummit全球架构师峰会，有来自Microsoft、Google、Facebook、Twitter、LinkedIn、阿里巴巴、腾讯、百度等技术专家分享过他们的实践经验，至今累计已经为中国技术人奉上了近千场精彩演讲。

● 2017.07.07-08 深圳站

how to use sagas to maintain data consistency in a microservice architecture

--Chris Richardson, *POJOs in Action* 作者, 知名微服务专家

● 2017.12.08-11 北京站

《创新是人类的自信》

--王坚博士, 阿里巴巴集团技术委员会主席

● 2018.7.06-09 深圳站

限时**7折报名中**, 名额有限, 快快抢购。

7折报名中
名额有限, 快快抢购

华南地区架构领域最有影响力的会议，届时有哪些专题和演讲，敬请扫描右方二维码浏览官网。



卷首语：工程才是做好 AI 的钥匙

作者 文因互联 CEO 鲍捷

经常有人问我学好人工智能的秘诀。我会先问下对方对数据结构、代码设计、调试工具、代码版本维护的入门问题。如果这些都不过关，我的回答就是“工程”。

做好AI应用，不仅是要懂AI“算法”，更重要的是软件工程能力和系统能力。在实践中，Linux命令用得熟不熟，写程序是不是有良好的风格，版本控制是不是成为习惯，是不是掌握基本的网络服务构架，这些基本功比会用Keras/TensorFlow重要多了。有想法的人很多，具体工程去做的人就少了，应先从最底层的工程练起。没有具体的工程经验，就是清谈，是浪费时间。先过了系统运维关、数据库关、代码习惯关、基本软件工程关，才能谈得上落地一个AI的系统。

现实系统里行之有效的人工智能算法，都是很简单的。能不能发挥好的根本，都在于如何把这些简单的东西因地制宜综合运用。为1%的核心算法代码跑好，要99%的“工程”代码的支持。



比如对机器学习，无免费午餐定理告诉我们，一个算法如果在一类问题上特别有效，那一定有一些问题它比随机算法还差。一个现实中可用的机器学习系统，几乎一定是多种问题的混合问题。不会存在一种算法是一个现实问题的灵丹妙药。现实的问题的解决，一定是用一个良好的工程架构，让多种算法混合在一起解决问题。能拿捏这个架构设计的“度”，就是人工智能工程师最核心的能力。

又比如逻辑这个分支。概念上其实没有比逻辑更简单的语言了：与、非、存在量词。但是为了工程化这个简单的东西，就衍生出巨大的一门学科：知识工程、语义网、知识图谱。知识工程之所以难不在“知识”，而在“工程”。当关注“知识”的时候，总是可以映射最优秀的人的智能。但工程化的时候，必须适应群体无限的奇葩，和不可避免的各种成本的折衷。

AI应用落地，核心是工程问题，不是算法问题，更不是“哲学”问

题。一定要特别特别“土”，踏踏实实从朴素的运维、数据库、数据清洗做起，从实际的工程中逐步演化。如何按天迭代？如何构造联调系统？如何无标注数据启动？如何分离准确度和召回率要求？如何统一运用规则和统计？如何适应无明确衡量标准的开发？如何设计可演进的数据模式？如何提升数据可理解性？如何逐步提升规则系统的表达力？如何平衡黑箱和白箱模型的优缺点？如何在优雅架构和工期间取舍？等等，这些都是教科书上没有的答案。只有扎扎实实从工程出发，才能实事求是地发展出低成本的、有生命力的AI系统。

如果仅仅是因为某个东西时髦就去学，比如因为这两年AI火就去学AI，满口CNN、RNN、LSTM，却没有兴趣去理解这些东西背后的基本原理和应用范围，对工程也是无益的。比如只知道“卷积”这个词，却不理解不同的卷积核对于图像到底起什么作用；只知道深度网络，却连其他的神经网络一概不知；只知道word2vec分布式表示，却连TFIDF和LDA都没用过。这种赶时髦，对工程实践害处大于用处。

掌握分很多层次。会用包是一个层次，会改进是一个层次，发优秀论文再进一个层次。至于懂得方法的边界、工程上和其他方法融汇使用，就只有凤毛麟角的人了。到AI架构师的层次，又需要通透理解多种方法的前沿。这样的人，学校、研究院都培养不出来，都是通过工程逼出来、练出来、打出来的。光是懂算法不行，还必须通透理解实践的前沿；光是理解一个分支也不行，还必须通透理解几个分支。

没有银弹，没有奇迹。都是扎扎实实的工程，多年的细节的打磨才能解决一点小事。也从来没有一个所谓的伟大的想法，能跳过工程的考验而就成功的。工程才是做好AI的钥匙。

AI 前线

InfoQ 中文站 AI 月刊 2018 年 2 月

生态评论

8 MIT 重磅报告：一文看清 AI 商业化现状与未来

重磅访谈

30 多款重磅翻译产品落地之际，我们独家专访了搜狗语音负责人王砚峰

落地实践

39 基于深度学习技术的 AI 输入法引擎

52 文档扫描：深度神经网络在移动端的实践

推荐阅读

64 Netflix 实战指南：规模化时序数据存储

精选论文导读

73 深度解读：深度学习在 IoT 大数据和流分析中的应用

MIT 重磅报告： 一文看清 AI 商业化现状与未来

作者 山世光



《麻省理工大学斯隆管理评论》(MIT Sloan Management Review) 是由知名高校麻省理工大学斯隆管理学院出版发行，也是全球十大综合管理类期刊之一，世界顶级商学院专家进行研究的必备工具。以下为报告部分内容。

概述

企业的愿景和现实之间存在着巨大的鸿沟。报告显示，四分之三的管理者认为 AI 将会帮助公司进入新的商业领域，将近 85% 的受访者认为

AI 将会帮助公司获得或保持竞争优势。然而，目前仅有五分之一的企业已在产品或服务中采用 AI 相关技术。20 家企业中仅有一家已大规模采用 AI，而仅有不到 39% 的企业已将 AI 作为公司的发展战略。员工数超过 10 万人的大公司表示均有制定 AI 战略的计划，但实际上只有一半企业已经制定了 AI 发展战略。

我们的调查发现，那些已经理解和采用 AI 的公司——先锋企业，与落后的企业之间有着巨大的鸿沟。其中一个相当大的差异是数据获取的方式。AI 算法不是生来就是“智能”的，它们只有通过不断分析数据才会变得“聪明”。虽然大多数公司管理层对 AI 非常感兴趣，已经建立起强大的数据架构，但仍有很多公司缺乏数据分析的经验或数据获取的渠道。我们的报告揭示了人们对 AI 训练所必需的资源方面的一些误解。先锋企业不仅比落后企业在 AI 训练所需要的资源方面有更深的理解，而且更倾向于在领导决策和 AI 商业落地层面给予 AI 发展更多的支持。

关于该报告

为了解 AI 相关的挑战和机遇，MIT 斯隆管理学院和波士顿咨询公司合作，联合完成了一年一度的调查报告，受访者来自全球逾 3000 名企业和组织的管理者、经理和分析专家。

该调查报告于 2017 年春季进行，获取了来自全球 112 个国家，21 各行业，各种规模的企业和组织管理人员对于 AI 的看法。其中，超过三分之二的受访者来自美国之外的国家和地区，样本来源多样，包括《MIT 斯隆管理学院评论》的读者和其他的团体。

此外，我们还采访了来自不同行业和学术界的管理人员作为补充，以了解如今企业面对的实际问题，他们的观点丰富了对数据的理解。

在此报告中，我们使用了牛津字典对“artificial intelligence”的定义：“AI 是指计算机系统能够完成通常需要人类智能才能完成的任务，例如视觉感知、语音识别、决策、语言翻译。”

然而，随着 AI 的发展，人们对 AI 的理解和定义在不断发正变化。

AI 在实际工作中

AI 对管理和组织实践会产生影响。现在已经有很多企业和组织使用不同的人工智能模型，但灵活性仍是所有模型的核心。在一些管理者看来，对于大公司，完成应用人工智能所需的文化变革将是一个艰巨的任务。

我们的受访者比那些可能会因 AI 而失业的人对 AI 更为乐观，他们中的大多数管理者并不认为人工智能会在未来五年内导致他们的公司大规模裁员。相反，他们希望人工智能将代替人类完成一些无聊和让人不愉快的任务。

Airbus（空中客车公司）是欧洲一家民航飞机制造公司，总部设于法国布拉尼亚克。随着 Airbus 开始增加 A350 飞机（新产品）的产量时，该公司面临着资金方面的挑战。用 Matthew Evans（一家位于法国图卢兹的数字化转型公司的副总裁）的话来说：“我们的计划是以前所未有的速度提高飞机的生产率。要做到这一点，我们需要解决快速响应生产中断等常见故障问题。”

为此，Airbus 将目光转向了人工智能，以将过去的生产计划数据与 A350 程序的持续输入、模糊匹配，以及自主学习算法结合起来，识别生产问题的模型。在某些领域，该系统几乎可以实时地匹配之前采用的 70% 的生产中断解决方案。Evans 描述了 AI 是如何让整个 Airbus 生产线快速学习，并应对业务上的挑战：

“系统所做的事情实质上是查看问题描述，并理解所有的上下文信息，然后将其与问题本身的描述进行匹配，进而为用户提出建议。虽然对系统来说这可能是新的问题，但其实可能在一周前的生产线，或在生产线不同的班次或部分遇到过类似的问题。这使得我们能够将处理生产中断所需的时间缩短三分之一以上。”Evans 表示。

采用人工智能让 Airbus 能够更快速、更有效地解决业务问题（例如代替人力，对数百甚至数千个案例进行原因分析）。

正如 AI 提高了 Airbus 公司的业务处理速度和效率一样，其他应用了 AI 的组织也开拓出更新、更好的处理程序，如 BP、Infosys、Wells、法戈和平安保险等大公司已经在使用 AI 解决重要业务问题。然而，仍有其他的组织尚未开始采用 AI。

各行业对 AI 的高期望

各个行业、各种规模和不同地域的公司对 AI 均抱有很高的期望。虽然目前大多数高管还没有看到人工智能的实质性影响，但他们显然期望在未来五年可以看到。在所有的组织中，只有 14% 的受访者认为人工智能目前在其组织产品中有很大（非常大或巨大）的影响。但是，63% 受访者表示希望在未来五年内可以看到效果。

人工智能将对公司产品产生影响，各个行业整体上期待值一直很高。（见图 1）在技术、媒体和电信行业中，有 72% 的受访者预计，未来五年人工智能会产生较大的影响，比报告中目前认为 AI 会对企业产生较大影响的受访者数量高出 52%。然而，即使公共部门（对人工智能效应总体预期最低的行业）也有 41% 的受访者预计，五年内人工智能产生的巨大影响将比目前的水平提高 30%。不同规模和地区的组织均对 AI 持看涨的态度。

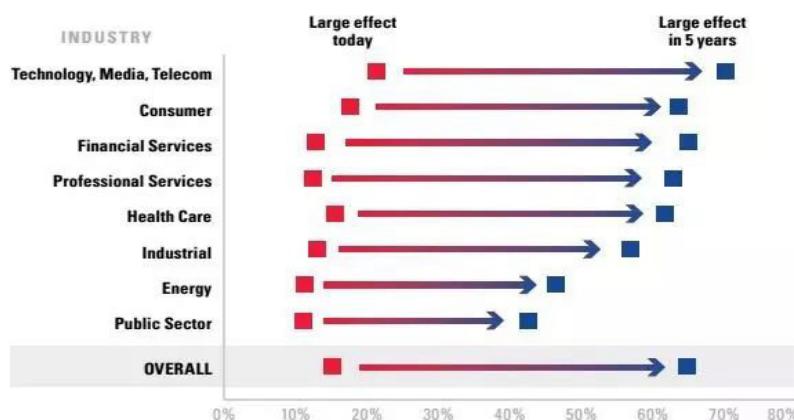


图 1：各行业对未来五年内 AI 将对企业产品产生影响的期望持续增长

在组织内部，受访者对人工智能将对流程产生巨大影响也抱有同样的高度期望。15%的受访者表示人工智能对当前流程有很大的影响，超过59%的受访者预计在五年内会出现较大的影响。如图 2大多数组织预计AI 将对信息技术、运营和制造、供应链管理，以及面向客户的活动产生巨大的影响（见图 3）。

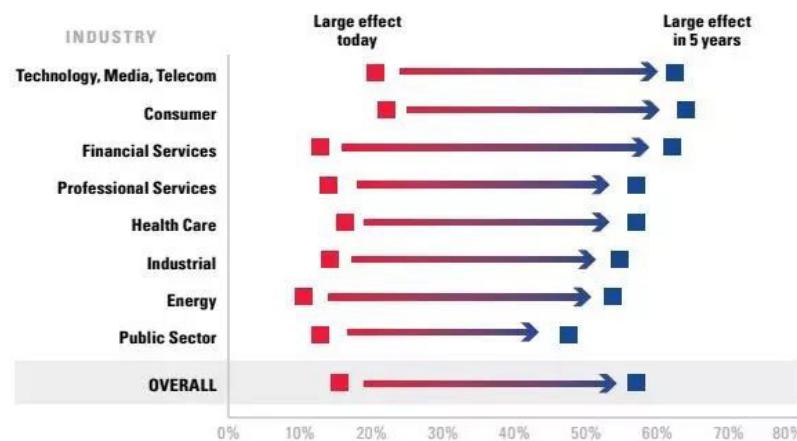


图 2：各行业对采用 AI 将对工作流程产生影响的预测



图 3：大多数企业预计 AI 将对 IT、运营和面对消费者的行业产生巨大影响

信息技术业：业务流程外包提供商是一个可以证明人工智能潜力的例子。Infosys 公司 CEO 兼董事总经理 Vishal Sikka 表示：“Infosys 在 IT 服务业举足轻重，这个行业在过去的 20 年左右发展迅猛。”许多被转移到低劳动力成本国家的工作是比较机械的工作：系统管理、IT 管理、商业运营、认证。随着 AI 技术的发展，我们的系统可以代替人类完成这些工作。虽然我们还处于完成工作的初始阶段，但是再过几年，系统将可以完成大部分，甚至全部此类工作。然而，AI 技术可以完成目前存在的、分工明确的任务，同样也可以创造不存在的、新的、具有突破性的工作。”

受影响最大的行业

运营和制造业：工业企业的高管预计，AI 将会对运营和制造业产生的影响将最大。例如，BP plc 通过人工智能提高人的技能，以改善现场操作能力。Upstream Technology 全球负责人 Ahmed Hashmi 表示：“我们设有一个 BP 钻井顾问的 AI “岗位”，它从钻井系统中提取数据，为工程师提供调整钻井最佳区域参数的建议，并提醒他们潜在的操作异常和风险。我们尝试自动分析失败的原因，并训练系统进行快速评估，并根据描述进行预测。

面向客户的业务：市值 1200 亿美元的中国第二大保险公司——中国平安保险股份有限公司（中国第二大保险公司），正在通过人工智能改善其保险和金融服务组合，为客户提供更好的服务。例如，平安现在可以在三分钟内提供在线贷款，这部分归功于一个内部开发的基于人脸识别功能的客户评分工具，它比人类的精准度更高。这个工具已经验证了 3 亿多人的面孔并用于各种用途，对平安的认知 AI 功能，包括语音和图像识别进行了补充。

采用 AI 带来的机会和风险

虽然高管对人工智能的期望值高涨，但同时也认识到其潜在风险。Sikka 对 AI 持乐观但又谨慎的态度：“从 1956 年起，纵观 AI 的历

史，我们会发现 AI 的发展经历过高峰，也经历过低谷，现在我们正处于一个 AI 发展火爆的时代，一切都似乎预示着 AI 处于快速发展的时期。

“超过 80% 的受访高管正瞄准这个高峰，把人工智能看作一个战略性的机遇。（见图 4）事实上，50% 的受访者只看到人工智能是一个机遇。而另一些人则看到了 AI 竞争加剧的潜力，以及将会带来的风险和收益。另外，有 40% 的管理者将人工智能视为战略性风险。而仅有 13% 受访者认为人工智能既不是机会，也不也是风险。



图 4：80% 以上组织认为 AI 是战略性机遇，将近 40% 将其视为战略性风险

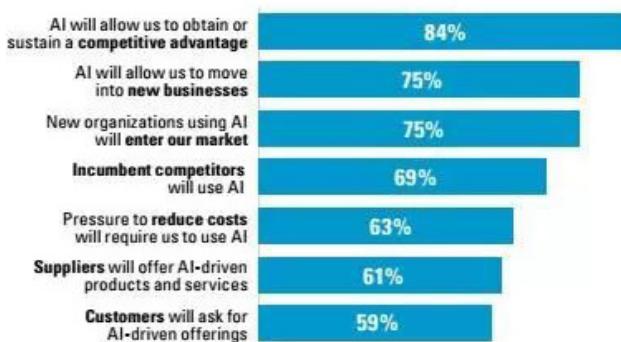
人们对 AI 商业化的高期望值和兴趣背后意味着什么呢？我们没有一个简单的解释。（见图 5）大多数受访者认为，人工智能将使组织受益，例如带来新业务或降低成本；84% 的人认为，AI 将让他们的组织获得或维持竞争优势。四分之三的管理者认为，人工智能将带领他们进入新的业务领域。

同时，高管们意识到，他们的组织不可能是 AI 的唯一受益者。受访者预计，新入局和已经进入的组织都同样有获益的可能性。四分之三的受访者预计，新的竞争者将通过 AI 进入市场，而 69% 的受访者预计，目前的竞争对手将会在他们的业务中采用人工智能。此外，他们意识到，他

们的商业生态系统中供应商和客户将越来越期望他们能够使用 AI 提供服务。

Reasons for adopting AI

Why is your organization interested in AI?



Percentage of respondents who somewhat or strongly agree with each statement

图 5：组织期望通过采用 AI 获得竞争优势，但竞争对手的加入使得竞争加剧

AI 采用和理解上的差异

尽管人们对 AI 抱有很高的期望，但商业化应用开发还处于初级阶段，即期望与行动之间存在着巨大的鸿沟。尽管五分之四的高管认同人工智能是他们的一次战略性机遇，但只有五分之一的组织已经在某些产品和流程中采用人工智能。仅有二十分之一的企业在其产品或流程中广泛地引入了人工智能。（见图 6）

采用 AI 的原因

组织采用 AI 情况的差异性是很惊人的，特别是在同一行业。例如，旗下有 110 名数据科学家的中国平安已经推出了约 30 个 CEO 发起的人工智能计划，部分原因是为了响应“技术是推动公司 2018 年快速增长关键动力的口号，“平安首席创新官 Jonathan Larsen 说道。然而，与保险行业的其他领域形成鲜明对比，其他大公司的人工智能计划仅限于”聊天机器人产品的试验。“这家大型保险公司的高管如此描述其公司的 AI 项目。

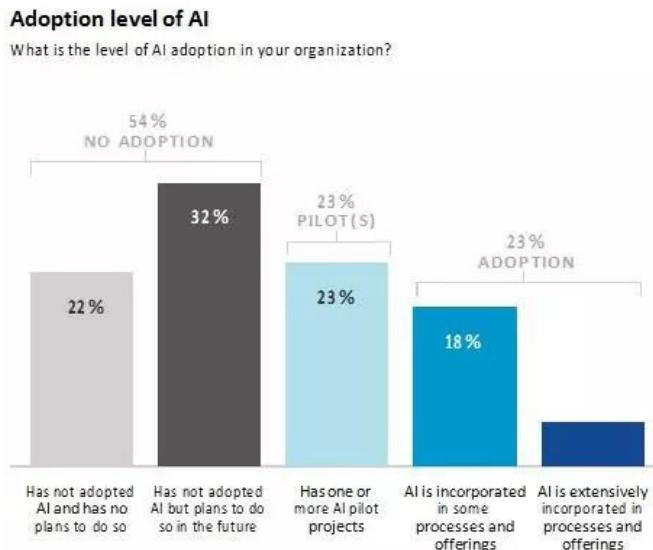


图 6：目前，仅有四分之一的企业采用了 AI 技术

另外，报告显示，企业对于 AI 的理解也是大相径庭。例如，16% 的受访者一致认为公司理解开发 AI 相关产品和服务产生的成本。然而，17% 的受访者表示其所在组织并不理解相关成本的产生。相似地，19% 的受访者认为所在企业理解训练 AI 所需的数据投入，16% 则不同意这一观点。

根据对 AI 的理解和采用程度，企业的成熟度可以分为四种类型：先锋、研究者、实验者和消极者。

- 先锋（19%）：了解并采用 AI 的组织。这些组织在将 AI 融入其组织产品和内部流程方面处于领先地位。
- 研究者（32%）：了解人工智能，但仅限于试验 AI 阶段的组织。这些组织对 AI 具有前瞻性的理解。
- 实验者（13%）：试点采用人工智能，但对其缺乏深入了解的组织。这些组织在实践中学习 AI。
- 消极者（36%）：没有采用或不了解 AI 的组织。

既然组织对 AI 的期望如此之高，那么是什么在阻碍企业采用 AI 呢？即使在一向具有整合新技术和管理数据理念的行业中，推广人工智能的障碍也很难克服。例如，在金融服务方面，瑞银集团（UBS）首席投资

官西蒙·斯迈尔斯（Simon Smiles）就这样说道：“大型金融机构在业务中更积极地利用技术（包括人工智能）和数据，为终端用户提供更好的客户体验的潜力是巨大的。但问题在于，这些传统机构是否真的能够抓住机遇。“抓住人工智能带来的机遇需要组织的承诺，并跨越许多伴随着人工智能而来的不可避免的挑战。

然而，导致这些差异的原因较少涉及技术限制，而更多的是商业。总体而言，受访者将竞争投资重点和不清晰的商业案例列为部署 AI 的更大的障碍，排在技术障碍之前。Airbus 的 Evans 指出了关键性的区别所在：“严格来说，我们不投资人工智能、自然语言处理和图像分析。相反地，我们投资是因为要解决具体的业务问题。“Airbus 采用人工智能，是因为它能解决业务问题；向人工智能投资比向其他方向投资更有意义。

瑞银集团的 Smiles 称，组织要面临的困难其实不一而同。对于大公司和金融科技创业公司，他说道：“它们之间存在着巨大的差异，前者的规模让他们足以开发比较大的平台，而后者虽然有更先进的模式，但是缺乏客户和相关数据来充分利用这个机会。“这样的差异导致不同组织人工智能采用率上的差异。

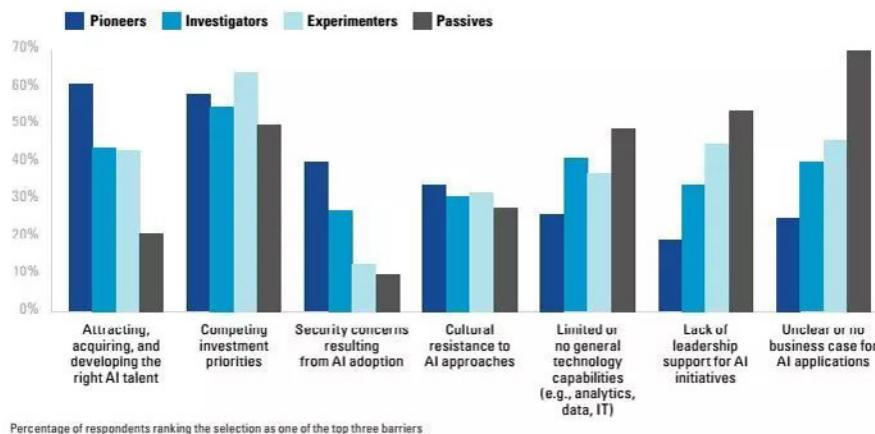
采用 AI 的障碍

这些组织分为不同的类型表明他们采用 AI 所面临的障碍不同，进而影响了 AI 的采用率。（见图 7）先锋组织已经克服了理解上问题：这些公司中有四分之三已经有了 AI 实践的商业案例。领导组织 AI 计划的高级管理人员面临的最大障碍，是如何挖掘 AI 人才，并获得优先投资，他们也更容易接受人工智能所带来的安全问题。相比之下，被动者型组织还没有认识到 AI 能为他们做什么，也并未确定符合他们投资标准的商业实践案例。缺乏 AI 计划上的领导，技术也是一个障碍，甚至许多人还未意识到他们在挖掘人工智能人才和专家上将面临的困难。

不同类型的组织在对 AI 的理解上业存在着巨大差异。

Barriers to AI adoption

What are the top three barriers to AI adoption in your organization?

**图 7：先锋企业采用 AI, 而消极者企业人不理解 AI**

商业潜力：人工智能可能会改变组织创造商业价值的方式。报告显示，先锋（91%）和研究者（90%）比实验者（32%）和消极者（23%）组织更能意识到 AI 对商业的影响。Airbus 的 Evans 表示：“我们只是在尝试解决飞机产品的服务问题。”

工作场所的影响：现如今，在工作场所把人类和机器的能力结合起来是需首要解决的问题。人工智能在很大程度上改变了日常的工作环境。先锋和研究者组织能够更好地意识到，工作场所中的机器将改变组织内的行为。麻省理工学院航空学副教授 Julie Shah 说道：“即使你可以开发一个针对某项任务（目前由人类完成）的系统，但除非流程中完全不需要人类工作，否则就会有新的问题产生，因为人类在协调工作，以及协助 AI 系统之间进行交流必不可少。这样的交流问题仍然是我们亟待解决的难题。

行业环境

企业是在行业规则和环境下运行的；实验者和消极者的受访者并未感受到 AI 将会对行业生态产生多大的影响。

数据、训练和算法需求

也许四种类型的组织最大的区别，在于它们对数据和 AI 算法之间独立性的理解存在的偏差。先锋对训练算法过程、AI 产品服务开发成本、训练算法所需数据的了解程度分别是消极者组织的 12 倍、10 倍和 8 倍。（见图 8）

Levels of AI understanding

To what extent do you agree with the following statements about your organization?

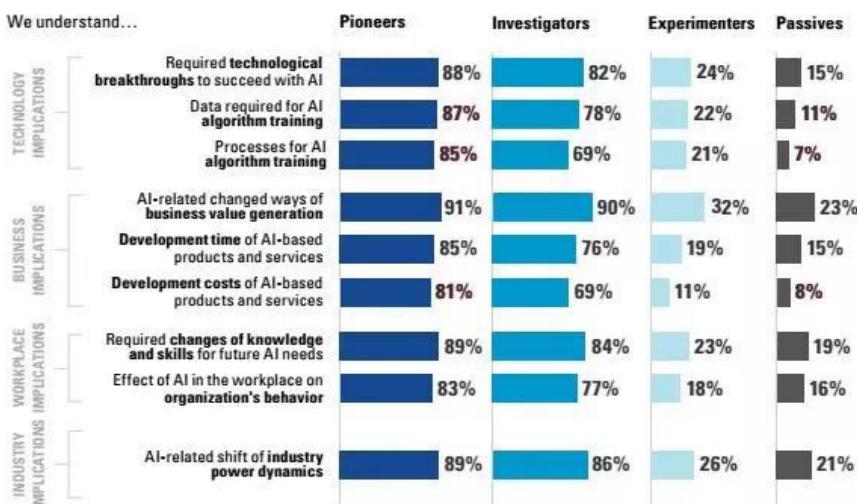


图 8：不同类型组织对 AI 相关技术和商业环境的理解程度不同

参与调查的大多数组织都对他们的数据进行 AI 算法训练的必要性了解不深，以解决类似于 Airbus 在应用 AI 的过程中所遇到的问题。不到一半的受访者表示，他们的组织理解训练算法的过程和算法的数据需求。

AI 产生业务价值，与 AI 算法的有效训练直接相关。许多现有的 AI 应用程序都是从一个或多个“裸”算法开始，只有经过训练（主要是公司特有的数据）才会变得智能化。成功的训练取决于完善的信息系统，可将相关培训数据汇总在一起。许多先锋组织已经拥有强大的数据和分析基础设施，同时对开发人工智能算法数据所需要的知识有广泛的理解。相比之下，研究者和实验者组织却因为他们几乎没有分析经验，空有一座“数据

“孤岛”而备受困扰。超过一半的先锋组织对数据和训练进行了大量投资，但其他类型的组织在这方面的投资却甚少。例如，只有四分之一的研究者组织在人工智能技术、训练人工智能算法所需的数据，以及训练过程方面进行了大量投资。

关于 AI 数据的误解

我们的研究表明，很多组织对数据有一些误解。其中一个误解是，无需足够的数据，仅靠复杂的 AI 算法就可以提供有价值的业务解决方案。微软的数据科学总监 Jacob Spoelstra 指出：

“我认为从人们对通过机器学习可以完成的事情的理解还是相当不成熟的。一个常见的误解是，一些企业并没有可以从中提取算法，以做出预测所需的历史数据。例如，他们请我们为他们建立一个预测性维护解决方案，但是我们发现有记录的故障很少。他们希望 AI 在没有学习数据的情况下能够预测什么时候会出现故障。” Jacob 如此说道。

没有任何一个算法可以克服缺乏数据的问题。这一点在所有企业希望 AI 能为他们的前沿业务带来进步提升时，显得尤为重要。

他们对于数据错误的认识不足：只有积极的结果对于训练 AI 来说是不够的。Citrine Informatics 是一个帮助加速产品开发的 AI 平台，使用相关研究机构提供的公开实验（成功实验）和为公开实验（包括失败的实验）数据。Citrine 的联合创始人兼首席科学家 Bryce Meredig 说道：“失败的数据几乎从未被公布过，但负面结果语料库对建立一个没有偏见的数据库至关重要。通过这种方法，Citrine 可以将研发时间缩短一半，以满足特定的应用需求。Gore-Tex 防水面料的开发商 W. L. Gore & Associates 公司也同样记录了成功和不成功的实验，这推动了他们的创新，了解不起作用的因素有助于帮助他们的下一步探索。

如果数据质量足够好，有时复杂的算法可以克服数据有限的障碍，但糟糕的数据只会导致算法瘫痪。数据收集和准备通常是开发 AI 的应用程序过程中最耗时的活动，比选择和调整模型耗时得多。正如 Airbus 的

Evans 所说：

“由于能够重复使用之前所建项目的资源，使得他们在成本降低的情况下工作效率更高，从而为数据湖增加更多的价值和更多的业务内容。”

先锋组织明白，他们的数据基础设施对于 AI 算法的价值。

此外，公司有时错误地认为，他们已经有权访问建立 AI 算法的数据。对于整个行业来说，数据所有权对管理者来说都是一个棘手的问题。一些数据是企业有的，他们似乎没什么理由共享出来。而其他数据源分散，为获得训练 AI 系统的更完整的数据，他们需要与其他多个组织进行整合，达成协议。在其他情况下，重要数据的所有权可能是不确定的或有争议的。理论上，靠 AI 获得商业价值是可能的，但在实践中却很难实现。

即使组织拥有所需的数据，多个系统之间分散也会阻碍 AI 算法的训练过程。富国银行公司风险模式执行副总裁 Agus Sudjianto 这样说道：

我们的工作很大一部分是处理非结构化数据（如文本挖掘），并分析大量事务数据，查看模型，致力于不断改进客户体验以及客户勘察、信贷审批和金融犯罪检测等方面的决策。在所有这些领域，应用 AI 都有很大的机会，但是在-一个非常庞大的组织中，数据往往是分散的。这是大公司要解决的核心问题——战略性地处理数据。

自建 vs 购买

使用合适的数据来训练人工智能算法的需求，对公司面临技术投资时决定自建还是购买系统有着很大的影响。AI 产生价值是一件比单纯地建立或购买 AI 复杂得多的事情。训练 AI 算法涉及多种技能，包括理解如何构建算法，如何收集和整合相关数据用于训练，以及如何监督算法的训练。“我们必须引进不同学科的人才。当然，我们需要机器学习和 AI 研究人员，”Sudjianto 说道，“能够领导 AI 项目的人才非常重要。”

先锋组织非常依赖于通过培训或聘用人才来提高工作人员的技能。对 AI 理解不深，缺乏经验的组织倾向于外包 AI 相关业务，但这样的模式

本身是有问题的。（见图 9）

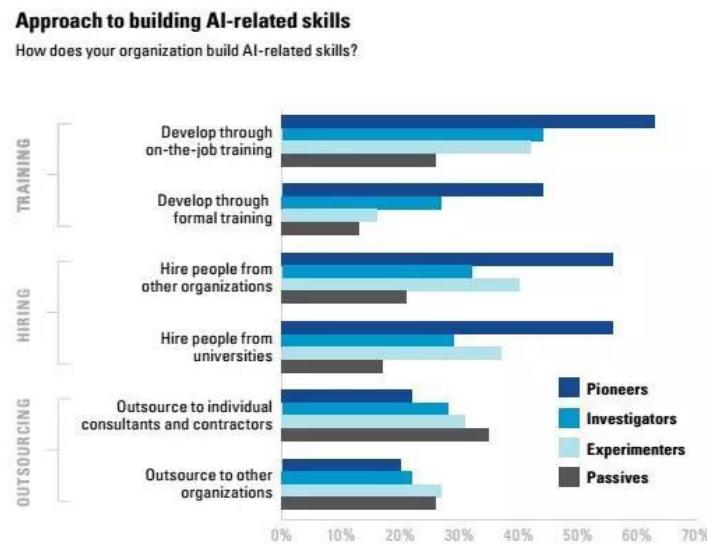


图 9：先锋组织通过训练和招聘获得 AI 相关技能，而消极者企业更多地依靠外包资源

一家大型制药公司的 CIO 认为，AI 服务商提供的产品和服务水平就像是“小孩子”一样。AI 技术供应商要求提供大量的学习数据，把 AI 训练成 17、18 岁智力的投入不敷出，他对此表示失望。

为了获得 IT 管理类似的功能，很多公司选择把整个流程外包。当然，尽管这些工作外包出去，他们也需要自己人了解如何解决问题、处理数据，以及当机遇来临时能够有意识。

“五年之前，我们可以通过外包获得成本较低的人力去做此类工作，同时供应商可以自动处理这些工作，但往往是我们自己的系统使用我们的框架，但是用的是他们的技术。这样的方法显然不适用于公司的特定需求和核心业务。

微软研究室主任 Eric Horvitz 认为，“市面上已经有很多好用的 AI 算法和工具，包括 Google 的 TensorFlow，GitHub 和来自技术供应商的应用程序编程接口。但是，因为这是一个竞争激烈的领域，虽然外界提供的工具和服务越来越便利，但并不意味着企业不需要拥有自己的内部专家，对于每个组织而言，拥有自己的机器学习和 AI 技术还是非常重要的。”

隐私和管理

训练 AI 所需的数据和算法能达到一定的准确性和性能还不够，遵循隐私问题和相关法规也是一个需要提上议程的问题。然而，在我们的调查中，只有一半的受访者认为其所在的行业已经形成了数据隐私相关的规则。

具有强大的数据管理实践能力才能保障数据隐私。先锋（73%）比实验者（34%）和消极者（30%）组织更有可能有良好的数据管理实践。（见图 10）这个巨大的鸿沟是落后企业面临的另一个挑战。

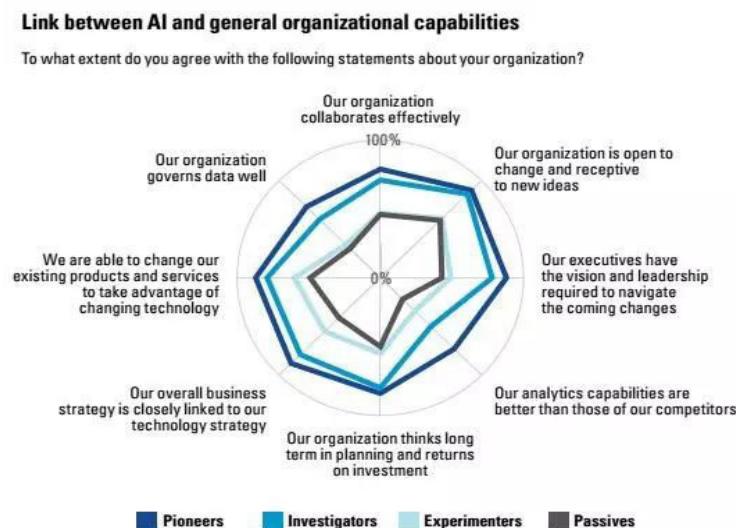


图 10：先锋组织将公司放在管理和领导维度之上

在监控较为严格的行业，例如保险行业中产生数据隐私问题的可能性较大，这些行业处于从基于风险池的模式向综合预测具体风险元素的风控方法转变。但有些元素在法律上是禁止使用的，例如，虽然性别和宗教因素可以用来预测一些风险，但在某些应用和司法管辖区，监管机构是不能接受这些信息被使用的。

其他金融市场的监管机构也有严格的透明度要求。正如富国银行的 Sudjianto 所说：“模型必须非常非常透明，并始终接受监管机构的审

查。我们不使用机器学习的原因在于，监管要求解决方案少一些“黑匣子”，以便监管机构监察。但是我们使用机器学习算法来评估模型的非线性结构、变量和功能，并作为传统模型表现的基准。

随着技术竞争也越激烈，企业和公共部门在 AI 计划、隐私保护和客户服务之间的规则越来越细化。一些金融服务提供商正在使用语音识别技术识别来电客户，以节省验证身份的时间。客户对此表示欢迎的部分原因是他们喜欢这项服务，并且相信公司不会滥用用户的数据。技术服务还提供人工智能服务，使用用户的语音数据，帮助呼叫中心运营商实时进行客户的情绪分析。然而，不太受欢迎的应用程序可能即将出现。几年后，中国安装的 1.7 亿台摄像机和美国 5000 万摄像机能够识别出人脸。事实上，据说上海已经应用这些图像数据源来惩罚街头流浪者。

技术之外：管理挑战

AI 需要的不仅是数据，组织在引入 AI 时也面临着许多管理方面的挑战。

不出所料，先锋组织的受访者对其所在公司的总体管理和领导力方面：愿景和领导力、开放性和变革能力、高瞻远瞩的思维、业务和技术战略之间的紧密结合，以及有效的合作方面评价更高。与其他技术驱动的转型一样，这些是公司保持良好经营状况必不可少的能力。

但是，公司在管理方面也面临一些具体的挑战：高管可能仍然需要

1. 更深入地了解更 AI；
2. 深化理解如何将业务与 AI 结合；
3. 以更广阔的视野看待业务竞争格局。

挑战 1：培养了解 AI 的直观思维

管理人员和其他管理人员至少需要对人工智能有基本的理解，这一观点得到了高管和学者的一致赞同。TIAA 公司企业数据管理总监 JD Elliott 补充说：“我不认为每个前线经理都需要了解神经网络深度学习

和浅层学习之间的差异。但是，对于依靠分析和数据，而不是直觉能够产生更好、更准确结果，我们需要有一个基本的认识，这是非常重要的。

“多伦多大学罗特曼管理学院市场营销学教授 Avi Goldfarb 指出：“我们会担心一个不成熟的管理者在看到一次预测之后就下结论认定这个模型好或不好。”麻省理工学院媒体实验室主任认为，“每个经理都必须对 AI 有一个直观的理解。”

管理者应该花一些时间来学习基础知识，比如可以从简单的在线课程或在线工具开始。了解程序如何从数据中学习，也许是他们理解人工智能如何让业务受益的最重要的方法。

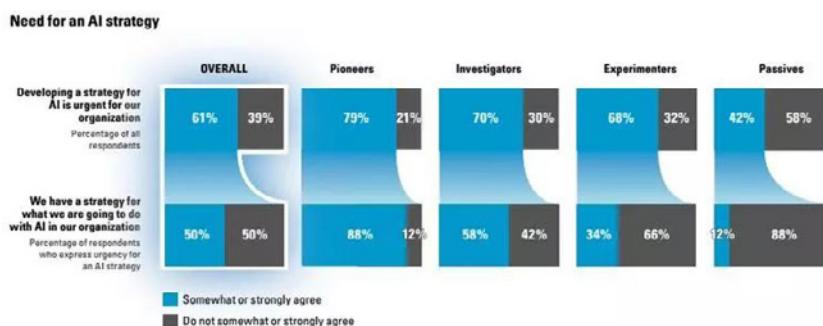


图 11：虽然大多数组织认为发展 AI 刻不容缓，但只有一半已采用 AI

挑战 2：组织 AI 部门

调查发现，这些公司为发展 AI 进行了很多探索。先锋组织选择的 AI 部门组成类型均匀分布在集中、分散和混合型。研究者和实验者组织也倾向于选择混合型的组织方法，但两种类型的企业中均仅有 30% 的 AI 部门有明确的职责。约有 70% 的消极者组织还未开始明确 AI 部门的职责，也许部分原因是不足 50% 的消极者组织认为，AI 在未来五年会对其工作流程和产品产生较大影响。

最后，混合型组织方式可能对于大多数企业来说意义更大，因为他们需要集中化和本地化的 AI 资源。以 TIAA 为例，其拥有一个高性能的数据分析中心和一些分散的团体。“整个组织的数据分析工作不全由数据

中心来完成，它为其他内部 AI 部署和分析团队提供专业知识、指导和方向。

而四种类型组织均将文化列在采用 AI 面临的障碍中相对靠后的位置，仅有一半的受访者表示公司理解 AI 需要的知识和技能方面的挑战。平安执行副总裁 Jessica Tan 表示，其公司面临的最大挑战是各部门之间的配合，以及建立集中和分散式的科技团队，他们需要三种人才：可以用不同方法工作的技术人员；了解特定商业领域的技术人员，以及有组织、咨询能力或项目经验的人。

接下来怎么做？

人工智能只是公司完成整体数字化转型的一个要素，还是探索人工智能需要另辟蹊径？一方面，AI 和其他数字技术一样存在许多相同的问题和挑战，公司可以通过多种方式建立数字和分析程序。但另一方面，AI 也具有鲜明的特点。

确保客户的信任。人工智能的功能类似于许多数字计划，它们依赖于客户数据，客户也信任公司会尊重和保护他们的个人数据。但是，确保人工智能值得信赖的方法与其他数据相关的数字计划有所不同。首先，管理者可能无法准确解释客户的个人数据会如何被用来生产某些 AI 产品，因为一些机器学习程序的内部运作是不透明的。其次，越来越多的人工智能系统能够模仿人类的代理人，在这种情况下，管理者有责任明确地与客户沟通，告知他们是在与机器还是与人类交流。第三，一些人工智能系统能够远程评估人类的情绪，识别细节。这种能力会产生新的信息管理问题，包括哪些员工可以访问这些信息，以及在什么情况下可以访问等。

进行一次 AI 健康检查。这与数字健康检查有一些相似之处，从支持基础架构的程序、技术、流程，以及快速响应故障进行检查。与许多数字计划一样，人工智能的成功取决于数据来源的访问权（内部或外部），以及对数据基础架构的投资。大公司可能拥有他们所需要的数据，但是如果这些数据是分散、孤立的，则会大大限制其战略的发展和进步。与其他

数字计划不同的是，人工智能健康检查包括对正确执行人工智能训练所需的技能进行评估，包括训练系统变得更聪明，直至部署后继续学习的全过程。

认识到不确定性。公司通常通过预估一个项目创造的价值和所需时间来确定其优先级，但是 AI 进行实验和学习可能会比其他数字计划花费更多的时间，成功和失败的不确定性更高。因此，管理者需要认识到这种不确定性。

基于场景需要。与数字相同，人工智能有可能改变企业创造价值的方式。AI 需要更激进的思维，因此，企业需要更加广泛地思考自己的业务，构建连贯的应用场景，并测试这些场景对计划的依赖性。这种基于场景的计划将提高系统识别有可能将触发影响业务的大事件的能力。

重视劳动力问题。人工智能会影响人们的工作和事业已成事实，也会造成社会的不安。因此，建立一个 AI 计划相关的工作项目是十分必要的，这个项目应包括 AI 相关的知识交流、教育和培训。另外，吸引和训练对 AI 感兴趣的人才，将商业和技术结合起来也变得非常重要。

AI 的未来之路

人工智能的采用可能会对工作，价值创造和竞争优势产生深远的影响。在未来，企业应该如何应对这些变化呢？

未来的工作

随着人工智能日益应用于知识相关工作，此前有众多预测称，AI 将使得工作场所发生重大转变。相反地，我们的报告显示，多数企业对这个问题持谨慎乐观的态度。例如，大多数受访者并不认为人工智能会在未来五年内导致其组织中的工作岗位减少。近七成的受访者表示，他们并不担心 AI 会取代他们的工作。相当一部分的受访者表示希望 AI 可以代替他们做一些无聊或让他们感到不愉快的任务。然而，受访者一致认同，AI 将迫使员工在未来五年内学习新的技能，并提高现有技能。（见图 12）

麻省理工大学斯隆管理学院 Schussel Family 教授 Erik Brynjolfsson 说道：“即使发展迅速，人工智能也不会很快取代大部分人类的工作。但几乎在每个行业中，使用人工智能的人都会替代不使用人工智能的人，而这种趋势只会加剧。”

AI's effect on the workforce

How do you expect AI will affect the workforce in the next five years?

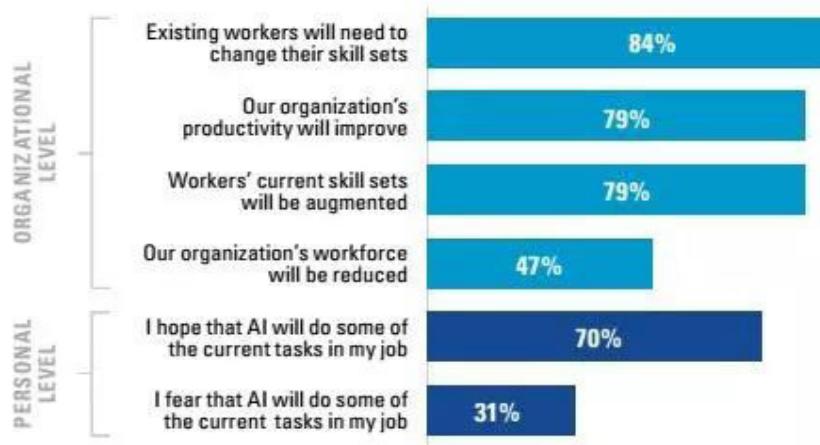


图 12：组织对 AI 未来五年内将对劳动力产生的影响持谨慎乐观的态度

价值创造发生变化

人工智能在哪些领域将会创造、摧毁或转移经济价值呢？

健康护理支出在美国经济总量中的占比达六分之一，平均约为经济合作与发展组织（OECD）成员国的十分之一。

AI 已经在改变医疗产值链：机器可以读取诊断图像，外科医生越来越依赖机器人，并且越来越多的实时医疗设备数据可以改善疾病预防和慢性疾病护理。

虽然人工智能可以在一个行业内创造价值，但是究竟哪个领域的产值将因此上升，哪些将下降还是一个未知数。当 IT 服务商、医疗技术公司、放射科医生网、医院、创业公司，甚至保险公司都开始利用人工智能

来降低诊断成本时，人工智能的影响可能会发生不均衡的状况。

因此，现在就下结论说哪种类型的组织可能从医疗保健 AI 中获益还为时过早。正如慕尼黑保险集团业务负责人马库斯·温特（Marcus Winter）所言：“在当今世界，随着大数据的普及，仅有少数几个独家数据集。大多数情况下，我们可以通过其他来源获得我们需要的信息。”换句话说，数据和 AI 算法的结合可以创造新的更有效的解决方法。例如，当诊断成像不可用时，更准确地分析血液或其他体液样本可能有助于诊断。这样，价值创造的变化其实很难预测。

保持竞争优势

许多公司的管理人员都专注于如何通过 AI 显著改善现有的流程和产品的性能。然而，仅仅改进产品并不能创造可持续性的竞争优势——当每个人的工作效率都提高到相同水平时，基准会相应地发生变化。要想通过 AI 获得竞争优势，企业必须明白如何将人类和计算机的优势结合起来，创造最大的竞争优势。而这并不容易：企业需要获得数据的访问特权，而这是目前很多公司所做不到的。他们必须学会如何让人和机器高效地共同工作，这是一个目前的先锋组织都不具备的能力。他们需要建立灵活的组织结构，而这意味着企业和员工需要经历一次文化大变革的洗礼。

多款重磅翻译产品落地之际，我们独家专访了搜狗语音负责人王砚峰

作者 刘海星



AI 前线导读

1月24日，搜狗在北京国贸举行了2018合作伙伴大会。会上，搜狗CEO王小川对搜狗的AI战略及布局进行了详细阐释，公布了搜狗在翻译领域的最新动作，同时发布两件重磅智能硬件新品“搜狗旅行翻译宝”和“搜狗速记翻译笔”。通过14年的积累，目前搜狗已成长为用户规模仅次于BAT的互联网公司，搜狗搜索为中国第二大搜索引擎，输入法为全球第一大中文输入法。作为人工智能的创新者，搜狗以“语言”处理为核心，目前已先后将自研AI技术落地于搜狗搜索与输入法等核心产品中。

搜狗如何解决人工智能技术落地

AI 前线：手机上面已经有翻译 APP 了，为什么还要单独做一个翻译机？

王砚峰：在做翻译机的时候，也经常会有行业的人问我们这个问题。但实际上如果你经常到美国、日本等国家，会发现其实这些国家的网络状况和国内是根本没法比的，大部分还是停留在 3G 的状态。因此在手机网络，或者通过移动 WiFi 网络访问语音或者翻译服务的时候，大概率会出现速度慢，没有返回的情况。因此我们认为这个阶段如果能有一个具有很强离线翻译功能的设备，帮助克服网络问题，那么是可以非常切实的解决用户这方面痛点的。

AI 前线：搜狗旅行翻译宝有什么独创的技术？

王砚峰：我们的独创性主要在于，一方面我们同时集成了语音和 OCR 拍照翻译的入口，针对旅游场景比如点菜这样旅游者们非常有感知的情景，做了针对性的优化。更重要的是这些功能，从语音识别，到 OCR，到翻译本身，都是离线完成的。这个在技术上是一个巨大的挑战。尤其是在翻译方面，我们使用了最新的 T2T 深度学习翻译框架。这个技术很多公司刚刚能够做到在线服务上使用，而我们就已经把它离线化了。从旅游场景的翻译效果上，我们离线的性能和在线性能基本接近，这是之前任何产品都不曾做到的能力。这个需要更大计算能力的支持，而目前市面上最优配置的手机上面都无法支撑这样的计算力需求。所以总结一句话，语音翻译和拍照翻译可能已经有一些手机 APP 具有能力，但是完全不能离线化。而之前市面上已经发布的专门的翻译设备，无论是从能力的全面性上还是从技术效果上，都不能够跟我们的翻译产品相比。

AI 前线：搜狗对翻译产品寄予什么样的期待？

王砚峰：2016 年在乌镇互联网大会上，我们在国内首发了中英语音



同传产品。几天前，我们又首次使用了英中同传系统，把 native 英文演讲者的话语直接翻译成中文。无论是翻译机，还是未来的同传，我们的期待都是产品能够在真正的刚需场景中，带给用户更实际的价值。而具体到翻译机产品，就是能够在旅游场景中，帮助出国的旅游者，更好地解决出国游的跨语言交流问题，带给他们旅游中的幸福感，减少他们的不便和焦虑。而且从这些年出国旅游人数增长的趋势来看，这本身也是一个足够大的市场，具有足够大的价值。

AI 前线：人工智能从技术到产品之间会有一个很长的距离，你们是怎么把技术应用到产品当中的？

王砚峰：第一我们有大量用户，第二我们在做 to C 的产品，我们在技术应用的过程当中，第一个思考方向就是怎么能够把技术用到我们已有的用户量比较大的产品上，并且能够让技术本身在这个产品上带来一个产品创新。因为虽然我们是技术基因，但是我们是产品导向的公司。几年之前我们做语音的时候，当时语音的识别准确率还没有那么高，我们就会想有没有可能在输入法上做出一个纯语音交互的产品来，当语音输入出错以后，能不能通过语音去进行编辑修改。当我们把语音识别做得更准，能够有条件来支持这样的产品的时候，我们也是在业内第一家推出这样的功能。包括我们现在做的对话也好，语音理解也好，也是在积极的想怎么能够在输入法或是搜索当中，更多地把这种技术加进去，提升用户在这个主路径上的产品体验。搜索，未来的方向是问答，就是怎么能够把问答跟对话的技术放到搜索这种核心上去；而输入法这块，我们目前提的是叫辅助对话，怎么能够在大量的日常聊天对话当中，输入法给你合适的候选，帮助你去进行输入，帮助你去进行表达，这都是我们的对话技术在当前产品上的应用。而我们优势就是我们有大量的用户，我们可能随随便便一个功能，每天有上亿次的请求量，像这种高频的用户场景跟技术结合起来，就能够更好地推动技术迭代，让技术真的在产品中落地，并且给用户带来价值。

AI 前线：相当于也是在搜索引擎的基础上去开发更多应用？

王砚峰：就是语义理解的技术，就是根据你的上文，我来理解你的意图是什么，下面给你对应的答案，它有可能是涉及日常聊天的，有可能是涉及知识的。

AI 前线：您刚提到的这几项功能，现在的成熟度怎么样？

王砚峰：如果站在传统的角度来看，不管是搜索也好，还是输入法也好，还是很多其他产品也好，一定是相对成熟的了，但是如果我们放在一个新的 AI 大背景下，比如搜索在 AI 的形态下就是问答，而输入法在 AI 的形态下我们希望去替代人去打字，替代人去对话，距离成熟还有很长的路要走。

AI 前线：您曾经说过语音交互的三个刚需场景是车内、客厅和户外，搜狗在这三个场景都有哪些最新的布局和进展？

王砚峰：我们在车内呢，是做了一个智能副驾，智能副驾更多的是解决怎么能够通过更好的语音交互来完成导航，这个我们是作为一个 to C 的产品发布的，相当于它是搜狗地图的一个版本，更多的面向车载这样一个场景，把语音作为一个更重要的交互手段，而不像以前那样使用文字和搜索。To B 产品这块，我们已经跟一些厂商建立了合作，2018 年就可以看到搭载搜狗能力的一些产品出来。客厅场景，我们去年是发布了糖猫在家这样一个产品。户外这块，目前的核心就是面向翻译这个领域。

搜狗如何看目前智能语音行业

AI 前线：您怎么看待智能语音行业的市场规模？

王砚峰：我很难给出一个确切的数字，因为现在大家统计这件事的口径也不一样。如果只是针对 to B 市场，可能是千亿或者万亿的一个规模。但是如果后面把 to C 市场拿进来，也就是说是不是把智能语音看成是未来搜索的一个中控，如果把它看成一个中控，那么未来它包含的应该是个更大的市场规模。所以现在哪一家给出来的数据应该都不是一个很科学的数据。

AI 前线：目前智能语音技术竞争和产品竞争的格局如何？



王砚峰：技术竞争这一块，我

认为是相对比较充分的竞争，因为这一波人工智能技术的兴起更多的是靠三个东西，一个是深度学习，一个是大量的数据，另外一个就是计算设备的能力。而这三个模块本身来讲，技术和计算能力对于各家来讲都是开放的，深度学习最初是学术界提出来，然后再渗透到产业界，产业界再跟进，跟进以后加上自己的数据产生好的效果。现在反观学术界，已经不再做语音识别了，或者已经不再做这种相对偏工业级的语音识别了，因为他们没有数据。既然深度学习已经变成了主流，而大家对于深度学习的使用仍

然处在相对初期的一个状态，所以说在技术这块，并没有说谁家的技术就一定比谁家的技术有个很强的壁垒。语音未来会变成一个更加像空气和水这样的基础性的技术。

产品竞争，两个方向去看，第一个方向就是，如果是 toB 类的行业产品，那么这种竞争更多的是看你在这个行业当中生根的时间，你在行业建立起的行业壁垒。我们再来看偏和消费者领域结合的，不管是车内的，还是音箱这样的产品，最终它其实会变成一个集团式作战的一个整体的竞争，就是你只有语音技术是不够的，你要有内容，甚至你要有产品前端，这也是为什么现在人工智能公司都要去做硬件，很多像小米这样的硬件公司都要去做人工智能技术，都要去做自己的内容，其实是一个道理。

AI 前线：搜狗在智能语音方面有哪些优势和劣势？

王砚峰：我们的优势就是，第一，我们的用户量确实更大一些，然后

从语料的获得上，资源的切入上，我们肯定都会更有优势，同时我们又是一个有很多流量的一个平台，从搜索，到输入法，再到浏览器上的各种流量。有流量以后，做偏智能语音这方面的硬件的时候也会有很好的销售能力，比如像我们的糖猫手表，2017 年大概是突破百万的这样一个销量。但是在 to B 方面，因为我们本身不是做 to B 的公司，我们在这方面还需要积累。

AI 前线：做产品实际上需要大量的数据，搜狗现在的语音数据是什么量级？

王砚峰：我们语音数据已经标注的量级，就是在万这样一个量级，大几万，或者十万左右这样一个量级，然后每天能够新增大概是不到 30 万个小时，这样的一个规模。而现在行业主流的，大家的训练数据基本上是万这样一个量级，所以现在不是去解决数据量的问题，而是数据量怎么能够用起来的问题，这是第一点。第二点，去解决当很多场景下你没有数据，你怎么能够在这个场景下去做到一个更好的效果，就像刚才说的听写这个产品，我们之前没有上线，那么我们可能在对应的这个场景下效果就没有那么好。那我们怎么能够去解决更多的没有场景数据的问题，这个是未来大家面临的核心问题。

AI 前线：用什么方式去解决呢？

王砚峰：还是技术问题，因为技术做得不够好，导致现在太依赖数据，现在我们所说的这种人工智能就是大数据加上深度学习的技术，但是是一旦你缺少数据，就不会有好的效果。现在我们语音识别什么领域做得好？就是数据充分的领域，我们日常的对话，手机这种相对标准的场景，数据自然是最多的。但是一旦切换到一个新的场景，那么这个新的细分场景，数据就会变少，效果就会变差。但是如果技术足够的好，能够去弥补数据这块的问题，最终就能够去解决语音在全方位各场景落地的问题。

搜狗为什么不做智能音箱？

AI 前线：现在市面上比较有代表性的几个智能音箱，您是怎么看他

们的切入点和前景的？

王砚峰：首先大家都相信语音是下一代搜索的入口，通过语音，然后把语音变成一个完整的服务，把内容提供给你，这也是为什么大家都在投巨量的成本在里面的原因。但是不管怎么样，现阶段的产品都是不好使的，只能是定个闹钟，查询个天气，这种最简单的操作。大家定义的是一种未来的场景，因为我们现在的场景是在手机上，是通过搜索，是通过各家的 APP 来满足你的服务，而且现在挺好的。而你要去做一个更好的，更有科技感的服务，并且能够通过语音的入口来替代手机，这是一个未来的产品，不是一个现阶段的产品。那么未来的产品到底是什么样的，现在还是个问号，大方向是可以的，但是切入点到底是不是应该是音箱，包括前景怎么样，是否还是这几家公司存活到最后，我相信最终是有特别大的一个变数的。

AI 前线：现在聊天机器人也很火，你们有兴趣吗？

王砚峰：确实它更容易去博得一些眼球，因为相对比较有意思，也确实像小冰这样的产品，会处在相对显得比较明星的这样一个形象。但是呢，为什么会这么重视聊天机器人，我个人的看法，它可能更多的是“人工智能”这四个字本身的原罪。就是当你说人工智能的时候，你头脑当中的第一个印象并不是说这个机器能用怎样的计算能力完成一件机器该做的事，你第一个想到是这个机器像人。所以我一直更喜欢 Google 的吴军老师对于这件事的定义，他觉得我们这一波智能叫机器智能，不叫人工智能，就是让机器通过计算能力，通过大数据，通过机器特有的方式，让它变得更聪明，更能够预测你的行为，更能够帮助你去解决问题。但当我们把它定义成人工智能的时候，我们更希望机器表现的像人一样。最像人的是什么呢？就是聊天，就是这种情感类的东西，只有人是带情感的，机器是不带情感的，这也是为什么人们一提起人工智能，就觉得聊天，或者情感是人工智能里面更有趣的。

AI 前线：似乎人们对这种更像人的机器人天生有一种情节。

王砚峰：虽然我们看到平台上很多都是聊天数据，用户时不时与机器

人互动。但是从未来大方向上来讲，聊天机器人不是一个产品。我需要的产品到底是什么？是陪伴。而陪伴的话，只有聊天能够做陪伴吗？我们现在市面上所有的产品，都是说我来当你的秘书，同时呢，你还可以跟我对话聊天。但是，当这个秘书是帮你去打理你生活当中的各种事的时候，你是否还真的需要跟它聊天？当你真的无聊的时候你会去打游戏，会去和朋友吃个饭。只是现阶段用户处于新鲜感中，还没有见过这样的东西，所以想去尝试与机器人互动。

搜狗未来的人工智能之路怎么走

AI 前线：现在人工智能领域的企業竞争越来越激烈了，搜狗打算怎样应对？

王砚峰：分三点来看。第一点就是我们之所以能够被大家认可，第一就是持续在技术上投入，去占据技术的制高点，保持技术的一个阶段性的领先。我们需要坚持这样一个理念，人工智能的核心是技术，如果没有这样一个核心，是不能够支撑你各个产品和业务的。虽然像我之前说的，技术构不成一个绝对的壁垒，但是如果你的技术更好，有一年，半年的领先，那么你就会有个不错的窗口期，你可以在窗口期内产生出更好的产品。

第二点就是我们能够坚定地去跟我们现阶段的有用户规模的产品和场景去联动，能够在这里面去迭代人工智能的产品和技术，比如像我刚才说的在输入法当中，怎么能把输入法变成一个智能对话，或者哪怕是一个辅助的对话。如果真的拿下输入法这个场景，它将是中国最大的一个场景，每天用户所有的聊天都是通过输入法来进行的，所有的信息的产品也都是通过输入法来进行的，我们希望能够利用好输入法跟搜索这两大产品。

然后第三点，还是坚持产品导向。只有好的产品，才有长久的生命力。我们希望做出来的产品，并不是一个冷冰冰的产品放在那，仅仅局限在聊天就足够了，我们还是希望做出一个产品，用户每天都能够用，每天都能够带来价值的。

AI 前线：对现在的搜狗输入法有什么不够满意的地方？

王砚峰：输入法还没有做成一个特别聪明的输入法，从输入效率上来讲，它确实是比上一代输入法要强很多，但是它还没有聪明到让你输入特别快。现在输入法输的快慢，仍然取决于你的手速。我们能不能做特别好的预测，联想功能，能够让那些手速特别慢的用户，很少需要去敲拼音键，而直接通过联想去完成输入，我们仍然需要去努力。

还有就是，大家现在对于输入法的认知仍然是一个输入工具，那么我们究竟什么时候能够把输入法从一个工具变成一个服务，因为你确实是掌握大量的用户信息的一个入口，当你能够把工具变成服务的时候，输入法的商业价值就会有一个极大的发挥。

AI 前线：您最近还关注哪些大的技术方向？

王砚峰：我现在也会开始关注图像这一块，因为不管是做业务也好，还是做产品的过程当中，能够很明确的感受到，一个完整的产品，如果有语音能力的话，是不够的。比如糖猫在家这样的产品，我们是把它定义成一个家庭的陪伴机器人，基本上能够去满足你任何时候你想看看家里面什么样，想看看小孩在干吗，可以随时接入视频通话的这样一个功能，首先它是在满足用户刚需的一个功能，那么在这个功能之下，它能不能做的更智能，能不能做得更有趣，在更有趣和更智能这个方向上，两个方向延伸出来，就是语音的对话够不够好，另外一个图像，这家伙是长了一只眼睛的，它是盯着家里的情况，盯着小孩的情况，比如这个小孩有没有有意思的一个瞬间，它能捕捉到，拍下来发给你，小孩是不是摔倒了，捕捉到了以后，马上开始哇哇叫，你家孩子摔倒了。像这样的产品，只有语音是不够的，它需要好的图像能力，包括我们现在的车载产品，那么车载产品除了大家现在都谈的功能，自动驾驶，辅助驾驶功能还有好多在视觉上可以去做的事，这个全都是图像的领域，所以现在我们也在看图像这块怎么更好的嵌入进去。

基于深度学习技术的 AI 输入法引擎

作者 姚从磊



各位好，我是姚从磊，非常高兴能够有这样一个跟大家交流的机会。今天主要想为大家介绍一下手机输入法最核心的模块 – 输入法引擎的技术方案，为什么以及如何从传统 N-gram 引擎演化到深度神经网络引擎的。

主要的内容分为五个部分：

什么是输入法引擎；

基于传统 N-gram 语言模型的输入法引擎；

为什么要转向深度神经网络引擎；

深度神经网络输入法引擎的那些坑；

高级预测功能。

先用一张图介绍一下我们公司的情况。



作为一家面向全球用户提供 173 种语言输入法的公司，Kika 利用 AI 技术，为用户提供了一流的输入体验，也在全球获得了大量的用户。



这张图中列出了目前全球输入法市场上用户量较大的产品，背后的公司既包括 Kika、百度、搜狗、Go 以及触宝这样国内的公司，也包括 Google(产品为 GBoard)、微软 (Swiftkey) 等国外大公司。大家都在输入法引擎的核心技术上投入大量研发精力，期望为全球各国用户提供一流的输入体验。

什么是输入法引擎

输入法 (Input Method, 简称 IME) 是最常用的工具软件之一，也

常被称为 Keyboard、键盘等。对每种语言，输入法会提供一个字母布局 (Layout)，上面按照用户习惯将对应语言的基础字母放置在合适的位置，比如英文键盘的 QWERT、汉语键盘的九宫格等。用户输入文字其实就是按照顺序来敲击 Layout 上的字母，字母敲击序列称为键码序列；在用户敲击字母的过程中，键码序列以及之前用户输入的词会被传入 Layout 下层的「输入法引擎」，引擎会根据从大规模数据中训练得到的语言模型，来预测用户当前以及接下来可能输入的词 / 词序列，并将最可能输入的词 / 词序列在键盘的候选区上展示给用户，供用户选择。

例如，如果一位用户期望输入的完整文本内容为「What's the weather today？」，当前输入到了「weather」的第三个字母「a」，此时词序列「What's the 」和键码序列「W h a t ' s SPACE t h e SAPCE w e a」（SPACE 表示空格）作为输入传送至输入法引擎，引擎基于训练好的语言模型进行预测，并将最有可能的候选词「weather」、「weapon」等展示给用户，供用户选择。在这个 case 中，如果「weather」排在第一位，则可以认为引擎是合格的，可以打 60 分。如果仅输入到「weather」的第一个字母「w」，就可以将「weather」排在第一位，则可以打 70 分。如果在输入到「weather」的第一个字母「w」后，就可以直接预测用户接下来要输入的词序列为「weather today?」，那就会更好，可以认为是 90 分。

总的来讲，输入法引擎的功能可以细分为「纠错」、「补全」和「预测」三类。

- 所谓纠错，指的是在用户输入一个错误的词，比如「westher」，会自动建议改为「weather」；
- 所谓补全，指的是输入一个词的一部分即预测整体，比如「w」预测「weather」；
- 所谓「预测」，指的是用户没有输入任何字母时直接预测用户接下来会输入什么，比如输入「What's the 」，预测出用户会输入「weather today?」。

同时，在拉丁等语系的输入法中，会提供滑行输入的功能。

用户在键盘上快速滑行词的字母序列，即便滑行轨迹有所偏差（因为滑行速度很快，用户较难准确定位各个字母的位置），也可以准确预测用户所想输入的词。在滑行输入中，引擎的输入是滑行点的轨迹，输出是预测的词。在本文中，我们不会深入探讨滑行输入的引擎实现逻辑。

更进一步，随着用户越来越多地倾向于利用 Emoji、表情图片等非文字内容表达自己的情感，引擎也需要能够根据用户输入词 / 键码序列来预测 Emoji 或表情图片。而 Emoji 往往具有多义性（表情图片也类似），此类预测的复杂度会更高，我们已经利用基于深度学习的建模技术较好地解决了这一问题。本文不会深入探讨，有兴趣的小伙伴可以单独探讨。

本文主要讨论在手机的按键文本输入场景下，输入法引擎高效准确地预测的相关技术。

此类技术的演化可以分两个阶段：

- 1) N-gram 统计语言模型阶段；
- 2) 深度神经网络语言模型阶段。

前者主要基于大规模语料进行统计，获取一个词在 $N-1$ 个词组成的序列 (N-gram) 之后紧邻出现的条件概率；但由于手机内存和 CPU 的限制，仅能对 N 较小 ($N \leq 3$) 的 N-gram 进行计算，预测效果存在明显天花板。后者通过构建深度神经网络，利用大规模语料数据集进行训练，不仅可以突破 N-gram 中 N 的限制，且可利用词与词的语义关系，准确预测在训练语料中未出现的词序列，达到远超统计语言模型的预测效果。

基于传统 N-gram 语言模型的输入法引擎

在输入「What's the weather today？」的这个 case 中，当用户输入到「weather」的第一个字母「w」时，引擎要做的事情，就是根据前面输入的词序列「What's the」来预测下一个最可能的以「w」打头的单词，而其中最关键的就是如何预测下一个最可能的词。

假设输入词序列为 w_1, \dots, w_{N-1} , 预测下一个词的问题实际上变成了 $\text{argmax}_{\text{WNP}}(\text{WN} \mid w_1, \dots, w_{N-1})$, 这个简单的模型称为输入法引擎的语言模型。

根据条件概率计算公式, $P(\text{WN} \mid w_1, \dots, w_{N-1}) = P(w_1, \dots, w_{N-1}, \text{WN}) / P(w_1, \dots, w_{N-1})$, , 根据最大似然估计原则, 只有在语料数据规模足够大以至于具备统计意义时, 上述概率计算才会具有意义。

但事实上, 如果 N 值过大, 并不存在「足够大」的语料数据可以支撑所有概率值的计算; 并且, 由于 WN 实际上仅同 w_1, \dots, w_{N-1} 中的部分词相关, 上述计算会造成大量的计算资源浪费。

因此, 实际计算中, 一种方式是引入马尔科夫假设: 当前词出现的概率只与它前面有限的几个词有关, 来简化计算。如果当前词出现的概率只与它前面的 $N-1$ 个词相关, 我们就称得到的语言模型为 $N\text{-gram}$ 模型。常用的 $N\text{-gram}$ 模型有 Unigram ($N=1$), Bigram ($N=2$), Trigram ($N=3$)。显然, 随着 N 的增大, 语言模型的信息量会指数级增加。

为了得到有效的 $N\text{-gram}$ 语言模型, 一方面需要确保语料数据规模足够大且有统计意义, 另一方面也需要处理「数据稀疏」问题。所谓数据稀疏, 指的是词序列 w_1, \dots, w_N 并没有在语料数据中出现, 所以导致条件概率 $P(\text{WN} \mid w_1, \dots, w_{N-1})$ 为 0 的情况出现。这显然是不合理的, 如果数据规模继续扩大, 这些词序列可能就会出现。我们可以引入平滑技术来解决数据稀疏问题。平滑技术通过把在训练语料集中出现过的 $N\text{-gram}$ 概率适当减小, 而把未出现的概率适当增大, 使得全部的 $N\text{-gram}$ 概率之和为 1, 且全部的 $N\text{-gram}$ 概率都不为 0。经典的平滑算法有很多种, 个人推荐 Laplace 平滑和 Good-Turing 平滑技术。

在利用 $N\text{-gram}$ 语言模型完成下一个词的预测后, 还需要根据用户的按键序列来对预测的结果进行调整, 可以利用编辑距离等衡量序列相似度的方法, 将按键序列同预测词的字母序列进行对比, 细节不再赘述。

利用 $N\text{-gram}$ 语言模型构建的输入法引擎, 在手机端运行时, 存在着如下问题:

不能充分利用词序列信息进行预测：受制于手机有限的 CPU 和内存资源，N-gram 中的 N 通常都不能太大，基本上 N 为 3 已经是极限。这意味着只能根据最近的 1 到 2 个词来进行预测，会丢失大量的关键信息；

不能准确预测语料数据集中未出现的单词序列。例如，如果在语料数据中出现过「go to work」，而没有出现过「go to school」。即使用户输入「A parents guide to go to s」，引擎也不能准确地将「school」排在候选区靠前的位置。

上述问题，利用深度神经网络技术可以很好地解决。

为什么要转向深度神经网络引擎

深度神经网络 (Deep Neural Networks, DNN) 是一种具备至少一个隐层的神经网络，通过调整神经元的连接方式以及网络的层数，可以提供任意复杂度的非线性模型建模能力。基于强大的非线性建模能力，深度神经网络已经在图像识别、语音识别、机器翻译等领域取得了突破性的进展，并正在自然语言处理、内容推荐等领域得到广泛的应用。

典型的深度神经网络技术有卷积神经网络 (Convolutional Neural Networks, CNN) 、递归神经网络 (Recurrent Neural Network, RNN) 、生成对抗网络 (Generative Adversarial Nets, GAN) 等，分别适用于不同的应用场景。其中，RNN (如图 1 所示) 特别适合序列到序列的预测场景。

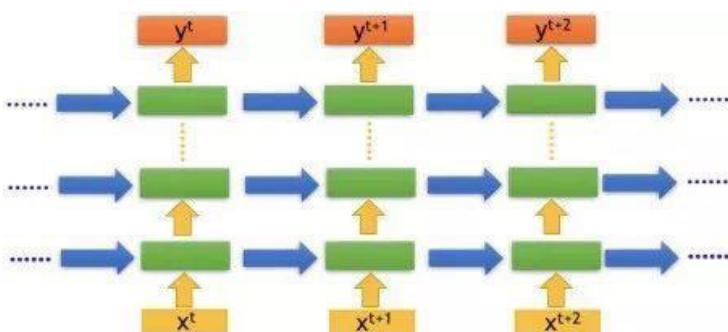


图1：RNN网络结构

传统的神经网络中层与层之间是全连接的，但层间的神经元是没有连接的（其实是假设各个数据之间是独立的），这种结构不善于处理序列化预测的问题。在输入法引擎的场景中，下一个词往往与前面的词序列是密切相关的。RNN 通过添加跨越时序的自连接隐藏层，对序列关系进行建模；也就是说，前一个状态隐藏层的反馈，不仅仅作为本状态的输出，而且进入下一状态隐层中作为输入，这样的网络可以打破独立假设，得以刻画序列相关性。

RNN 的优点是可以考虑足够长的输入词序列信息，每一个输入词状态的信息可以作为下一个状态的输入发挥作用，但这些信息不一定都是有用的，需要过滤以准确使用。为了实现这个目标，我们使用长短期记忆网络 (Long-Short Memory Networks, LSTM) 对数据进行建模，以实现更准确的预测。

LSTM (图 2) 是一种特殊的 RNN，能够有选择性地学习长期的依赖关系。LSTM 也具有 RNN 链结构，但具有不同的网络结构。在 LSTM 中，每个单元都有三个门（输入门，输出门和遗忘门）来控制哪部分信息应该被考虑进行预测。利用 LSTM，不仅可以考虑更长的输入序列，并且可以利用三种门的参数训练来自动学习筛选出真正对于预测有价值的输入词，而非同等对待整个序列中所有的词。

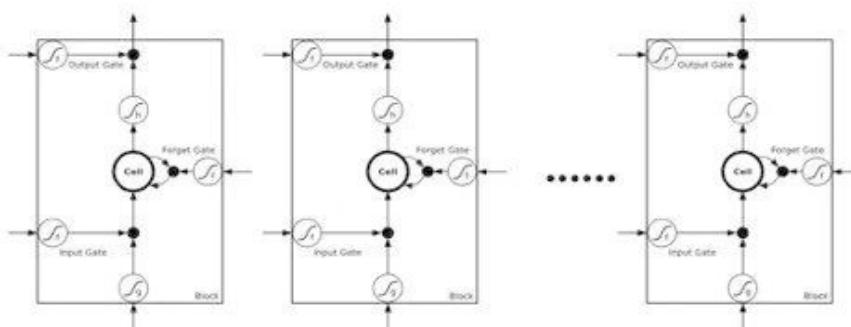


图 2 LSTM 网络结构

并且，可以在 LSTM 的网络结构中添加嵌入层 (Word Embedding Layer) 来将词与词的语义关系加入到训练和预测过程中。通过 Word

Embedding，虽然在语料数据中没有出现过「go to school」，但是因为「go to work」出现在语料库中，而通过 Word Embedding 可以发现「work」和「school」具有强烈的语义关联；这样，当用户输入「A parents guide to go to s」时，引擎会根据「work」和「school」的语义关联，以及 LSTM 中学习到的「parents」同「school」间存在的预测关系，而准确地向用户推荐「school」，而非「swimming」。

深度神经网络输入法引擎的那些坑

从理论上来讲，LSTM 可以完美解决 N-gram 语言模型的问题：不仅能够充分利用词序列信息进行预测，还可以准确预测语料数据集中未出现的单词序列。但是，在实际利用 LSTM 技术实现在手机上可以准确流畅运行的输入法引擎时，在云端和客户端都存在一些坑需要解决。

在云端，有两个问题需要重点解决：

充分利用词序列和键码序列信息。如前所述，在输入法引擎的预测过程中，LSTM 的输入包含词序列和键码序列两类不同的序列信息，需要设计一个完备的 LSTM 网络可以充分利用这两类信息。对此，我们经过若干实验，最终设计了图 3 所示的两阶段网络结构。在第一个阶段，词序列信息被充分利用，然后将最后一个词对应隐层的输出作为下一个阶段的输入，并和键码序列一起来进行计算，最后通过 Softmax 计算来生成最终结果。同时，在两阶段间加入「Start Flag」，以区隔词序列和键码序列。

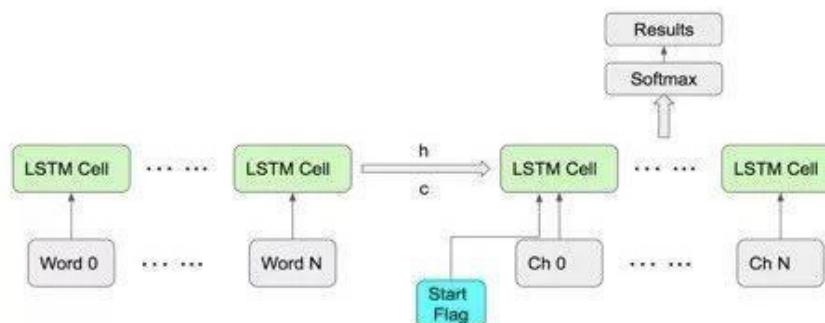


图3 词序列 / 键码序列混合

高质量训练数据的生成：在训练 LSTM 语言模型时，训练语料的质量和覆盖度是关键因素。从质量角度来讲，必须确保其中没有乱码、其他语言以及过短的句子等数据。从覆盖度角度来讲，一方面要确保训练语料的规模，使其能够覆盖语言中的大部分词汇，并足以支持语言模型的统计有效性，一般来讲训练语料的量级应该在千万或者亿；覆盖度的另一个角度为文本类型，需要确保训练语料中文本类型（比如新闻、聊天、搜索等）的分布同目标应用场景一致，对于手机输入法来讲，日常聊天类型的数据应该占足够大的比例；覆盖度的第三个维度为时间维度，需要确保训练语料可以覆盖对应国家 / 语言固定时间周期（通常为年）中各个时间段的数据，尤其是大型节日的数据。

在客户端，性能和内存是必须解决的关键问题。一个优秀的输入法引擎，在手机端运行时，需要始终稳定地保持低内存占用，确保在 Android Oreo (Go edition) 系统上也可以稳定运行，且保持良好的性能（每次按键响应时间小于 60ms）。而 LSTM 原始模型通常较大（例如美式英语的模型超过 1G），在手机端运行时响应时间也远超 1s，需进行大幅优化。可以利用稀疏表示与学习的技术，来压缩图 3 LSTM 网络中的 word/ch embedding 矩阵及输出端 softmax 向量矩阵，同时基于 Kmeans 聚类对模型参数进行自适应量化学习，最终可以将超过 1G 的模型量化压缩到小于 5M。性能优化则意味着需要控制手机端的计算量，需要在保证效果的前提下优化模型结构，减少不必要的层数和神经元；同时，可基于 TensorFlow Lite（而非 TensorFlow Mobile）进行手机端计算模块的开发，大幅提升性能和内存占用，唯一的成本是需要自己实现一些必备的 operators。我们采用该方案可以将运行时内存占用控制在 25M 以内，且响应时间保持在 20ms。图 4 是 TensorFlow Mobile 和 TensorFlow Lite 在相同 benchmark 上的对比数据。

基于以上的云端建模和客户端预测技术，我们完成了基于深度神经网络（LSTM）的输入法引擎方案的整体部署，并在大量语言上，同基于 N-gram 的语言模型进行了对比测试。在对比测试中，我们关注的关键指

标为输入效率 (Input Efficiency) :

输入效率 = # 输入的文字长度 / # 完成文字输入所需的按键次数

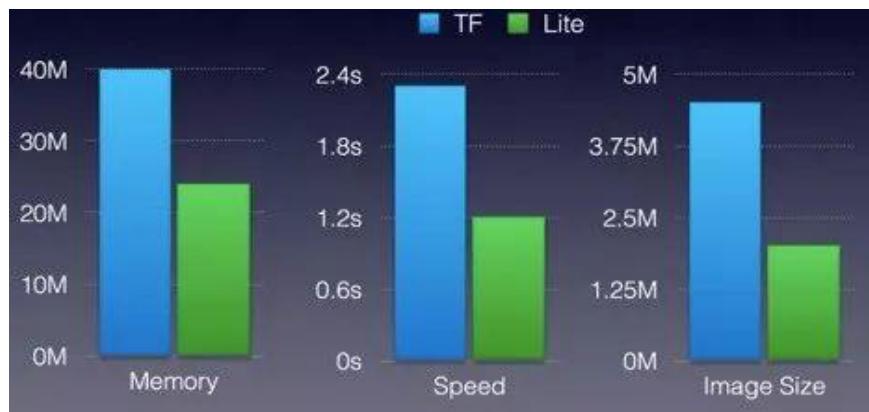


图 4 TensorFlow Mobile 和 TensorFlow Lite 在相同 Benchmark 上的对比数据

我们期望输入效率越高越好；同时，我们也会关注每种语言对应的线上用户的回删率，在此不再赘述。

下图是在一些语言上 LSTM 引擎相对 N-gram 引擎输入效率的提升幅度。

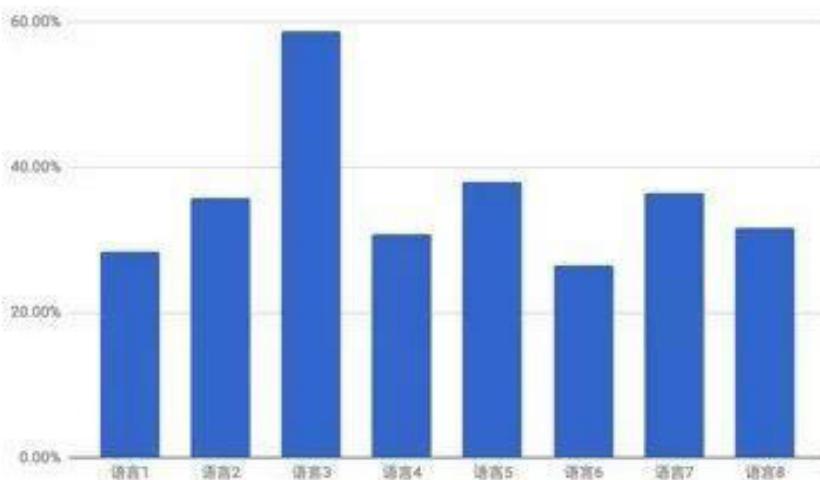


图 5 LSTM 引擎相对 N-gram 引擎输入效率的提升幅度

高级预测功能

高级预测功能在第一部分提到过，输入法引擎也需要能够准确预测用

户可能输入的 Emoji 或表情图片，而这些内容往往具有多义性，因此此类预测的复杂度会更高。

同时，对于 Emoji 来讲，用户往往会创造出一些有趣的 Emoji 组合，如何自动挖掘出这样的 Emoji 组合，并将其整合进 LSTM 的模型框架中，也是一个很有趣的问题。

另一方面，从文字输入效率的角度看，如果每次预测能够准确预测不只一个词，而是词组，对用户的体验会是一个很大的提升。如何发现有意义的尽可能长的词组，并且整合进 LSTM 的模型框架中，也会是一个很有挑战的工作。

Q/A 环节

Q1：请问姚老师，这里 softmax 输出是？

A1：因为这个网络的目的是预测下一个词，所以 softmax 的输出是预测词的 id 和概率值。在实际产品中，我们会选择 Top 3 的预测词，按照概率值从高到底显示在候选区。

Q2：输入法本质上就是根据用户前面的输入预测他接下来要输入什么。训练数据集和预测模型是在云端完成然后定期更新到手机端的吗？还是完全在手机上端完成的？

A2：我们的方案是包含两个部分。对每种语言，我们会在云端迭代训练一个新 general language model，在新 model 效果得到离线评测验证后，下发到手机端。并且，在手机端，会根据每个用户的个人输入历史，来训练 personalized model（这个 model 的训练频率会更高）。在实际预测时，会将这两个 model 的 inference 结果 merge 起来得到最终结果。在手机端的训练，需要尤其注意训练的时机，不能在用户手机负载高的时候执行训练。

Q3：TF Lite 对手机有要求吗？

A3：TF Lite 对手机没有要求。但是 TF Lite 为了性能的考虑，砍

掉了很多的 operators，我们在实现的过程中实现了自己模型 inference 必需的 operators.

Q4: 接上 Q1 的问题，针对英文情况，假如词表为 1w，softmax 层 1w 个节点，怎么优化 softmax 层呢？

A4: softmax 层的压缩，本质上就是 softmax 向量矩阵的压缩，其原理就是将巨大的向量矩阵转换为少量的过完备基向量组合，而过完备基向量可以自动学习获得。

Q5: 请问：每训练一次模型 LSTM 需要多少个 cell 这是由什么决定的？

A5: 每训练一次模型 LSTM 需要多少个 cell，决定因素大体有两类：1) 我们可以接受的模型的复杂度，这直接决定了最终量化压缩后的模型的大小；2) 我们期望达到的效果。最终的决定主要是在这两类之间进行平衡。当然，也同语言本身的复杂度有关，比如德语同英语对比，会更加复杂，因此 cell 的数量多一些会更好。如果不考虑这个限制，我们可以通过云端 service 的方式来进行 inference.

Q6: 未来输入法会支持语音输入吗？

A6: 我们正在开发 Kika 的语音识别和语义理解引擎，目前在英文上的语音识别水平接近 Google 的水平，所以会逐步上线 Kika 的语音输入功能。同时，我们基于 Kika 的一系列语音技术，已经在 CES 2018 发布了 KikaGO 车载语音解决方案，获得了很多好评和 CES 的四项大奖，并正在准备产品的正式发布。我们的全语音解决方案除了为车载场景下提供服务外，还会在场景上做出更多的尝试。

Q7: ”可以在 LSTM 的网络结构中添加嵌入层（Word Embedding Layer）来将词与词的语义关系加入到训练和预测过程中” 能具体解释下，在实际的数据预处理中，是如何添加的？简单拼接还是啥？LSTM cell 的输入又是啥？

A7: Word Embedding Layer 的作用是将高维的词空间映射到低维的向量空间，确保在低维空间上语义相似的词的向量距离比较小。Word

Embedding Layer 的输出作为 LSTM 的输入。TensorFlow 本身自带 Word Embedding Layer，其实就是一个简单的 lookup table。但如果所处理的问题领域不是 general domain，建议用所在 domain 的语料数据利用 Word2vector 来训练得到对应的 domain specific word embedding，用来替换掉 TensorFlow 自带的 Word Embedding Layer。

Q8：输入法的研究应该是很有挑战的领域，尤其是人类语言太多了。能谈谈这方面的发展趋势吗？

A8：「趋势」都是大拿才可以谈的事情，我只谈一些自己的浅薄想法。输入法本质为了解决人类通过机器（手机、电脑、智能家居等）和网络达成的人与人的沟通问题。而这样的沟通问题，最重要的是能够达到或超越人与人在现实世界中利用语音、表情和肢体动作面对面沟通的效果，能够全面、准确、快速地达到意图、信息和情感的沟通。

因此，我们认为输入法的下一步发展一定是围绕着「全」、「准」、「快」三个方面进行的。「全」是指能够提供文字、语音、表情、多媒体内容等的沟通方式，「准」指的是使得接收方能够准确接收到表达方的意图、信息和情感，不会产生误解，「快」指的是表达方产生到接收方接收的时间足够短。

而「全」、「准」、「快」这三方面的用户体验，通过 AI 的技术，都可以大幅提升。

Q9：我还想提问这样的 AI 输入法和百度输入法怎么竞争。谢谢。

A9：产品之间的竞争，本质就是如何为用户创造价值。Kika 输入法的定位，就是解决全世界用户在人与人沟通之间产生的问题，就是为用户提供极致的「全」、「准」、「快」的沟通方式，为用户创造更多的价值。从市场划分上来看，两款产品也不在一个维度上竞争——国内的输入法巨头像搜狗、百度主要解决的是中文，而 kika 则重点为中文之外的其他语种的用户创造价值。

文档扫描：深度神经网络在移动端的实践

作者 有道技术团队



背景篇

首先介绍一下什么是文档扫描功能。文档扫描功能希望能在用户拍摄的照片中，识别出文档所在的区域，进行拉伸（比例还原），识别出其中的文字，最终得到一张干净的图片或是一篇带有格式的文字版笔记。实现这个功能需要以下这些步骤：

识别文档区域

将文档从背景中找出来，确定文档的四个角；

拉伸文档区域，还原宽高比

根据文档四个角的坐标，根据透视原理，计算出文档原始宽高比，并

将文档区域拉伸还原成矩形。这是所有步骤中唯一具有解析算法的步骤；

色彩增强

根据文档的类型，选择不同的色彩增强方法，将文档图片的色彩变得干净清洁；

布局识别

理解文档图片的布局，找出文档的文字部分；

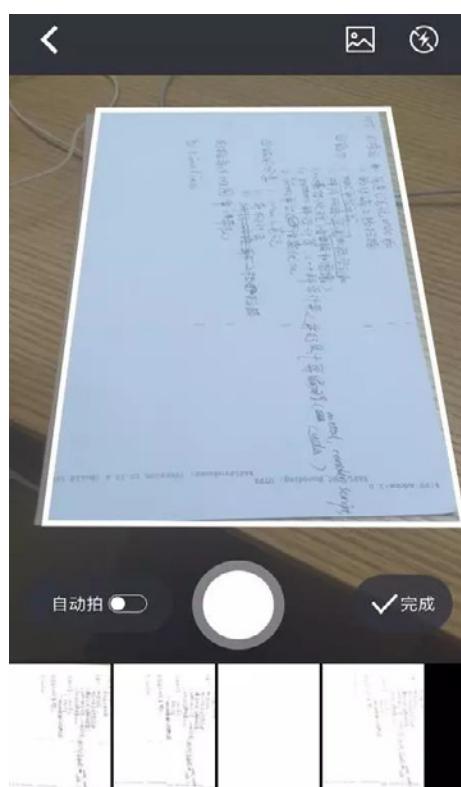
OCR

将图片形式的“文字”识别成可编码的文字；

生成笔记

根据文档图片的布局，从 OCR 的结果中生成带有格式的笔记。

在上述这些步骤中，“拉伸文档区域”和“生成笔记”是有解析算法或明确规则的，不需要机器学习处理。剩下的步骤中都含有机器学习算法。其中“文档区域识别”和“OCR”这两个步骤我们是采用深度神经网

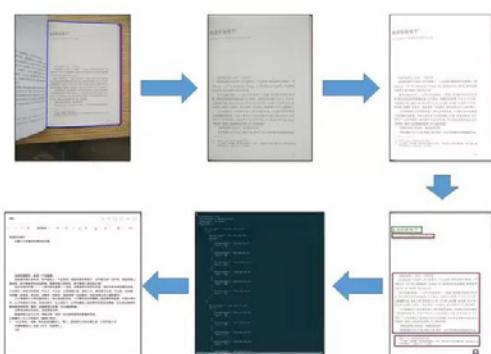


络算法来完成的。

之所以在这两个步骤选择深度神经网络算法，是考虑到其他算法很难满足我们的需求：

- 场景复杂，浅层学习很难很好的学习推广

同时，深度神经网络的一些难点在



这两个步骤中相对不那么困难

- 属于深度神经网络算法所擅长的图像和时序领域；
- 能够获取到大量的数据。能够对这些数据进行明确的标注。

接下来的内容中，我们将展开讲讲“文档区域识别”步骤中的神经网络算法。

算法篇

文档区域识别中使用的神经网络算法主要是全卷积网络（FCN）[1]。在介绍 FCN 前，首先简单介绍一下 FCN 的基础，卷积神经网络（这里假设读者对人工神经网络有最基本的了解）。

卷积神经网络 (CNN, Convolutional Neural Networks)

卷积神经网络 (CNN) 早在 1962 年就被提出 [2]，而目前最广泛应用的结构大概是 LeCun 在 1998 年提出的 [3]。CNN 和普通神经网络一样，由输入、输出层和若干隐层组成。CNN 的每一层并不是一维的，而是有（长，宽，通道数）三个维度，例如输入层为一张 rgb 图片，则其输入层三个维度分别是（图片高度，图片宽度，3）。

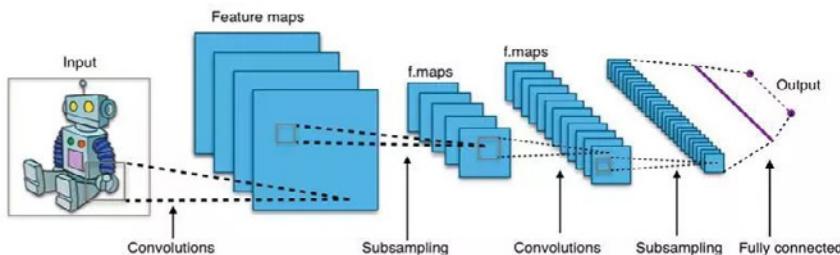
与普通神经网络相比，CNN 有如下特点：

- 第 n 层的某个节点并不和第 $n-1$ 层的所有节点相关，只和它空间位置附近的（ $n-1$ 层）节点相关；
- 同一层中，所有节点共享权值；
- 每隔若干层会有一个池化（pool）层，其功能是按比例缩小这一层的长和宽（通常是减半）。常用的 pool 方法有局部极大值（Max）和局部均值（Mean）两种。

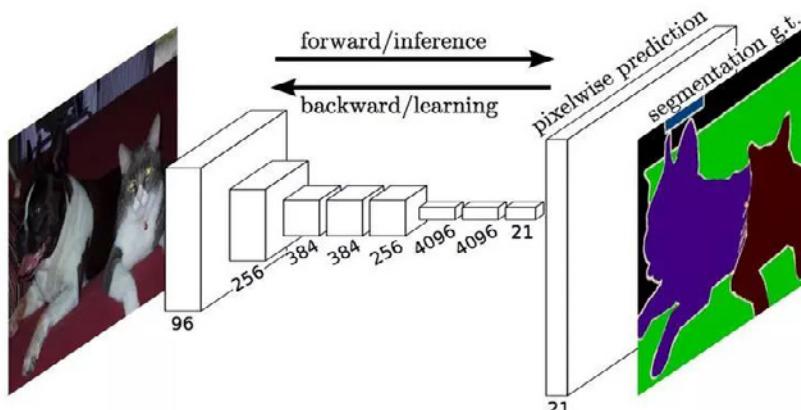
通过加入若干 pool 层，CNN 中隐层的长和宽不断缩小。当长宽缩小到一定程度（通常是个位数）的时候，CNN 在顶部连接上一个传统的全连接（Fully connected）神经网络，整个网络结构就搭建完成了。

CNN 之所以能够有效，在于它利用了图像中的一些约束。特点 1 对

应着图像的局域相关性（图像上右上角某点跟远处左下角某点关系不大）；特点 2 对应着图像的平移不变性（图像右上角的形状，移动到左下角仍然是那个形状）；特点 3 对应着图像的放缩不变性（图像缩放后，信息丢失的很少）。这些约束的加入，就好比物理中“动量守恒定理”这类发现。守恒定理能让物体的运动可预测，而约束的加入能让识别过程变得可控，对训练数据的需求降低，更不容易出现过拟合。



全卷积网络 (FCN, Fully Convolutional Networks)



全卷积网络 (FCN) 是 CNN 基础上发展起来的算法。与 CNN 不同，FCN 要解决这样的问题：图像的识别目标不是图像级的标签，而是像素级的标签。例如：

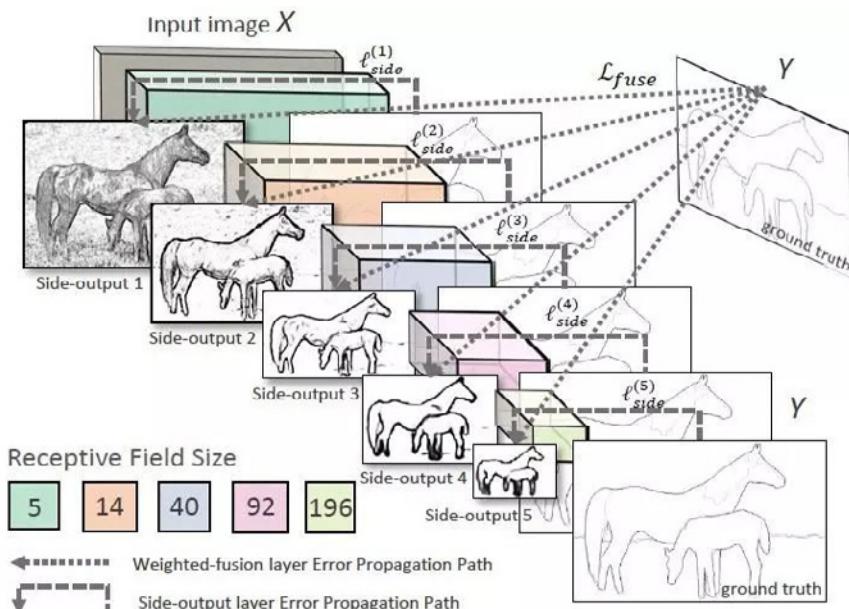
- 图像分割 需要将图像根据语义分割成若干类别，其中每一个像素都对应着一个分类结果；
- 边缘检测 需要将图像中的边缘部分和非边缘部分分隔开来，其中每一个像素都对应着“边缘”或“非边缘”。(我们面对的就属于这类问题)

- 视频分割 将图像分割用在连续的视频图像中。

在 CNN 中，pool 层让隐层的长宽缩小，而 FCN 面对的是完整长宽的标签，如何处理这对矛盾呢？

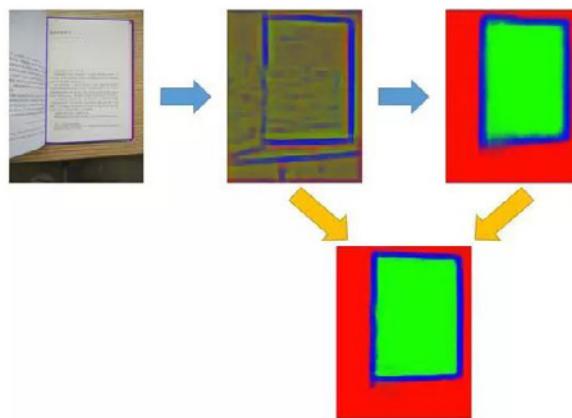
一个办法是不使用 pool 层，让每一个隐层的长宽都等于完整的长宽。这样做的缺点是，一来计算量相当大，尤其是当运算进行到 CNN 的较高层，通道数达到几百上千的时候；二来不使用 pool 层，卷积就始终是在局域进行，这样识别的结果没有利用到全局信息。

另一个办法是转置卷积（convolution transpose），可以理解为反向操作的 pool 层，或者上采样层，将隐层通过插值放缩回原来的长宽。这正是 FCN 采用的办法。当然，由于 CNN 的最后一个隐层的长宽很小，基本上只有全局信息，如果只对该隐层进行上采样，则局部细节就都丢失了。为此，FCN 会对 CNN 中间的几个隐层进行同样的上采样，由于中间层放缩的程度较低，保留了较多的局部细节，因而上采样的结果也会包含较多的局域信息。最后，将几个上采样的结果综合起来作为输出，这样就能比较好的平衡全局和局域信息。



整个 FCN 的结构如上图所示。FCN 去掉了 CNN 在顶部连接的全连接

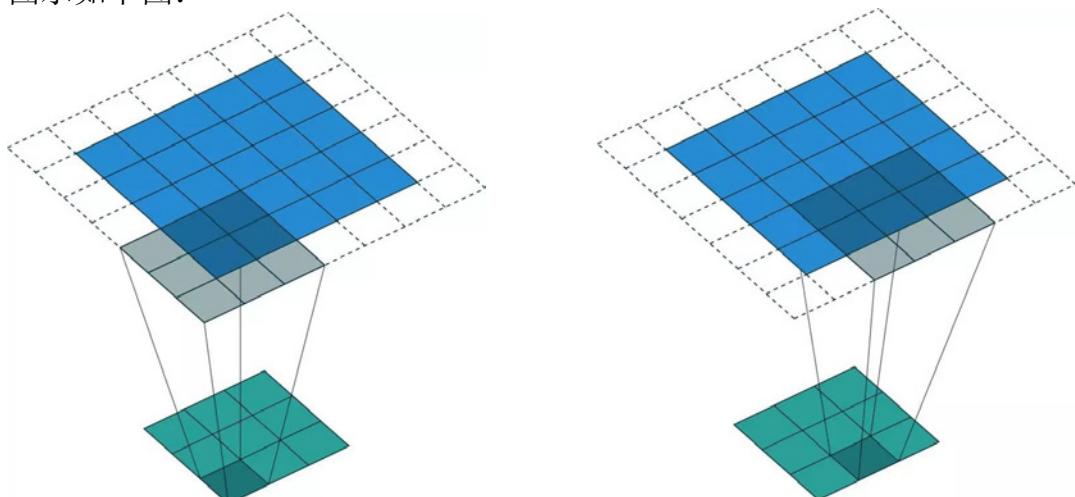
层，在每个转置卷积层之前都有一个分类器，将分类器的输出上采样（转置卷积），然后相加。



上图是我们实验中真实产生的上采样结果。可以看到，层级较低的隐层保留了很多图片细节，而层级较高的隐层对全局分布理解的比较好。将二者综合起来，得到了既包含全局信息，又没有丢失局部信息的结果。

转置卷积 (convolution transpose)

上文中出现的“转置卷积”是怎样实现的呢？顾名思义，转置卷积也是一种卷积操作，只不过是将 CNN 中的卷积操作的 Input 和 Output 的大小反转了过来。https://github.com/vdumoulin/conv_arithmetic 中提供了一系列转置卷积的图示，不过我个人认为更符合原意的转置卷积的图示如下图：



与 conv_arithmetic 提供的图示对比，可以看出上图只是卷积示意图的上下翻转。在实际运算中，Input 层的某个节点数值会（以卷积核为权重）加权相加到与该节点相关的每一个 Output 层节点上。

从维度上来看，如果记卷积核的高、宽为 H 和 W，Input 层的 channel 数为 C，Output 层的 channel 数为 0，那么一次正向卷积的输入节点数为 $H \times W \times C$ ，输出节点数为 0；而一次转置卷积运算的输入节点数为 C，输出节点数为 $H \times W \times 0$ 。

改进的 cross entropy 损失函数

在边缘识别问题中，每一个像素都对应着“边缘 - 非边缘”中的某一类。于是，我们可以认为每一个像素都是一个训练样本。这会带来一个问题：通常图片中的边缘要远少于非边缘，于是两类样本的数量悬殊。在模式识别问题中，类别不平衡会造成很多不可控的结果，是要极力避免的。

通常面对这种情况，我们会采用对少样本类别进行重复采样（过采样），或是基于原样本的空间分布产生人工数据。然而在本问题中，由于同一张图中包含很多样本，这两种常用的方法都不能进行。该怎么解决样本数量悬殊问题呢？

2015 年 ICCV 上的一篇论文 [4] 提出了名为 HED 的边缘识别模型，试着用改变损失函数（Loss Function）的定义来解决这个问题。我们的算法中也采用了这种方法。

首先我们概述一下 CNN 常用的 cross entropy 损失函数。在二分类问题里，cross entropy 的定义如下：

$$l = - \sum_{k=0}^n (Q_k \log p_k + (1 - Q_k) \log(1 - p_k))$$

这里 l 为损失值，n 为样本数，k 表示第几个样本，Q 表示标签值，取值为 0 或者 1，p 为分类器计算出来的“该样本属于类别 1”的概率，在 0 到 1 之间。

这个函数虽然看起来复杂，但如果对它取指数 ($L=\exp(-1)$)，会发现这是全部样本均预测正确的概率。比如样本集的标签值分别为 (1, 1, 0, 1, 1, 0, ...), 则：

$$L = p_0 \cdot p_1 \cdot (1 - p_2) \cdot p_3 \cdot p_4 \cdot (1 - p_5) \cdot \dots$$

这里 L 是似然函数，也就是全部样本均预测正确的概率。

HED 使用了加权的 cross entropy 函数。例如，当标签 0 对应的样本极少时，加权 cross entropy 函数定义为：

$$l = -\sum_{k=0}^n (Q_k \log p_k + W(1 - Q_k) \log(1 - p_k))$$

这里 W 为权重，需要大于 1。不妨设 $W = 2$ ，此时考虑似然函数：

$$L = p_0 \cdot p_1 \cdot (1 - p_2) \cdot (1 - p_2) \cdot p_3 \cdot p_4 \cdot (1 - p_5) \cdot (1 - p_5) \cdot \dots$$

可见类别为 0 的样本在似然函数中重复出现了，比重因此而增加。通过这种办法，我们虽然不能实际将少样本类别的样本数目扩大，却通过修改损失函数达到了基本等价的效果。

数据篇

文档区域识别中用到的神经网络算法就介绍到这里了，接下来聊一聊我们为训练这个神经网络所构建的数据集。

数据筛选

为了训练神经网络模型，我们标注了样本容量为五万左右的数据集。然而这些数据集中存在大量的坏数据，需要对数据进行进一步筛选。

五万左右的数据集，只凭人工来进行筛选成本太高了。好在根据网络的自由度等一些经验判断，我们的网络对数据集的大小要求尚没有那么高，数据集还算比较富足，可以允许一部分好的数据被错筛掉。

基于这一前提，我们人工标注了一个小训练集（500 张），训练了一

个 SVM 分类器来自动筛选数据。这个分类器只能判断图片中是否含有完整的文档，且分类效果并不特别强。不过，我们有选择性的强调了分类器分类的准确率，而对其召回率要求不高。换而言之，这个分类器可以接受把含有文档的图片错分成了不含文档的图片，但不能接受把不含文档的图片分进了含有文档的图片这一类中。

依靠这个分类器，我们将五万左右的数据集筛选得到了一个九千左右的较小数据集。再加上人工筛选，最终剩下容量为八千左右的，质量有保证的数据集。

实现篇

在模型训练中，我们使用 tensorflow 框架 [5] 进行模型训练。我们的最终目标是在移动端（手机端）实现文档区域识别功能，而移动端与桌面端存在着一些区别：

- 移动端的运算能力全方位的弱于桌面端；
- 带宽和功耗端限制，决定了移动端的显卡尤其弱于桌面端的独显；
- 移动端有 ios 和 Android 两个阵营，它们对密集运算的优化 API 各不相同，代码很难通用；
- 移动端对文件体积敏感。

这些区别使得我们不能直接将模型移植到移动端，而需要对它们做一些优化，保证其运行效率。优化的思路大致有两种：

- 选择合适的神经网络框架，尽可能用上芯片的加速技术；
- 压缩模型，在不损失精度的前提下减小模型的计算开销和文件体积。

神经网络框架的选择

目前比较流行的神经网络框架包括 tensorflow, caffe[6], mxnet[7] 等，它们大多数都有相应的移动端框架。所以直接使用这些移

移动端框架是最方便的选择。例如我们使用 tensorflow 框架进行模型训练，那么直接使用移动端 tensorflow 框架，就能省去模型转换的麻烦。

有的时候，我们可能不需要一个大而全的神经网络框架，或者对运行效率要求特别高。此时我们可以考虑一个底层一些的框架，在此基础上实现自己的需求。这方面的例子有 Eigen[8]，一个常用的矩阵运算库；NNPACK[9]，效率很高的神经网络底层库，等等。如果代码中已经集成了 OpenCV[10]，也可以考虑用其中的运算 API。

如果对运行效率要求很高，也可以考虑使用移动端的异构计算框架，将除 CPU 以外的 GPU、DSP 的运算能力也加入进来。这方面可以考虑的框架有 ios 端的 metal[11]，跨平台的 OpenGL[12] 和 Vulkan[13]，Android 端的 renderscript[14]。



模型压缩

模型压缩最简单的方法就是去调节网络模型中各个可调的超参数，这里的超参数的例子有：网络总层数、每一层的 channel 数、每一个卷积的 kernel 宽度 等等。在一开始训练的时候，我们会选择有一定冗余的超参数去训练，确保不会因为某个超参数太小而成为网络效果的瓶颈。在模型压缩的时候，则可以把这些冗余“挤掉”，即在不明显降低识别准确

率的前提下，逐步尝试调小某个超参数。在调节的过程中，我们发现网络总层数对识别效果的影响较大；相对而言，每一层的 channel 数的减小对识别效果的影响不大。

除了简单的调节超参数外，还有一些特别为移动端设计的模型结构，采用这些模型结构能显著的压缩模型。这方面的例子有 SVD Network[15]，SqueezeNet[16]，Mobilenets[17] 等，这里就不细说了。

最终效果

经过神经网络框架定制、模型压缩后，我们的模型大小被压缩到 1M 左右，在性能主流的手机（iphone 6，小米 4 或配置更好的手机）上能达到 100ms 以内识别一张图片的速度，且识别精度基本没有受到影响。应该说移植是很成功的。

总结

在两三年之前，神经网络算法在大家的眼里只适用于运算能力极强的服务器，似乎跟手机没有什么关联。然而在近两三年，出现了一些新的趋势：一是随着神经网络算法的成熟，一部分学者将研究兴趣放在了压缩神经网络的计算开销上，神经网络模型可以得到压缩；二是手机芯片的运算能力飞速发展，尤其是 GPU，DSP 运算能力的发展。伴随这一降一升，手机也能够得着神经网络的运算需求了。

“基于神经网络的文档扫描”功能得以实现，实在是踩在了无数前人的肩膀上完成的。从这个角度来说，我们这一代的研发人员是幸运的，能够实现一些我们过去不敢想象的东西，未来还能实现更多我们今天不能想象的东西。

参考文献

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3431-3440).

- Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1), 106-154.
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.
- Xie, S., & Tu, Z. (2015). Holistically-nested edge detection. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1395-1403).
- <https://www.tensorflow.org/>
- <http://caffe.berkeleyvision.org/>
- <http://mxnet.io/>
- http://eigen.tuxfamily.org/index.php?title>Main_Page
- <https://github.com/Maratyszcza/NNPACK>
- <http://opencv.org/>
- <https://developer.apple.com/metal/>
- <https://www.opengl.org/>
- <https://www.khronos.org/vulkan/>
- <https://developer.android.com/guide/topics/renderscript/compute.html>
- Denton, E. L., Zaremba, W., Bruna, J., LeCun, Y., & Fergus, R. (2014). Exploiting linear structure within convolutional networks for efficient evaluation. In *Advances in Neural Information Processing Systems* (pp. 1269-1277).
- Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. *arXiv preprint arXiv:1602.07360*.
- Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Netflix 实战指南：规模化时序数据存储



引言

因特网互联设备的发展，提供了大量易于访问的时序数据越来越多的公司有兴趣去挖掘这类数据，意图从中获取一些有意义的洞悉，并据此做出决策技术的最新进展提高了时序数据的收集存储和分析效率，激发了人们对如何处理此类数据的考量然而，大多数现有时序数据体系结构的处理能力，可能无法跟上时序数据的爆发性增长

作为一家根植于数据的公司，Netflix 已习惯于面对这样的挑战，多年来一直在推进应对此类增长的解决方案该系列博客文章分为两部分发表，我们将分享 Netflix 在改进时序数据存储架构上的做法，如何很好

地应对数据规模的成倍增长

时序数据: 会员视频观看记录

每天, Netflix 的全部会员会观看合计超过 1.4 亿小时的视频内容观看一个视频时, 每位会员会生成多个数据点, 存储为视频观看记录 Netflix 会分析这些视频观看数据, 实时准确地记录观看书签, 并为会员提供个性化推荐具体实现可参考如下帖子:

- 我们是如何知道会员观看视频的具体位置的?
- 如何帮助会员在 Netflix 上发现值得继续观看的视频?

视频观看的历史数据将会在以下三个维度上取得增长:

- 随时间的推进, 每位会员会生成更多需要存储的视频观看数据
- 随会员数量的增长, 需要存储更多会员的视频观看数据
- 随会员每月观看视频时间的增加, 需要为每位会员存储更多的视频观看数据

Netflix 经过近十年的发展, 全球用户数已经超过一亿, 视频观看历史数据也在大规模增长这篇博客帖子将聚焦于其中的一个重大挑战, 就是我们的团队是如何解决视频观看历史数据的规模化存储的

基本架构的初始设计

最初, Netflix 的云原生存储架构使用了 Cassandra 存储观看历史数据团队是出于如下方面的考虑:

Cassandra 对时序数据的建模提供了很好的支持, 支持一行中的列数动态可变

在观看历史数据上, 读操作和写操作的数量比大约为 1:9 因为 Cassandra 提供了非常高效的写操作, 特别适用于此类写密集的工作负载

从 CAP 定理方面考虑, 相对于可用性而言, 团队更侧重于实现最终一致性 Cassandra 支持可调整的一致性, 有助于实现 CAP 上的权衡

在最初的架构中, 使用 Cassandra 存储所有会员的观看历史记录其中, 每位会员的观看记录存储为一行, 使用 CustomerId 标识这种水平分

区设计支持数据存储随会员数量的增长而有效扩展，并支持简单并高效地读取会员的完整观看历史数据这一读取操作是历史数据存储上最频繁发生操作然而，随着会员数量的持续增长，尤其是每位会员观看的视频流越来越多，存储的数据行数和整体数据量也日益膨胀随着时间的推移，这将导致存储和操作的成本增大而且对于观看了大量视频的会员而言，性能会严重降低

下图展示了最初使用的数据模型中的读操作和写操作流。

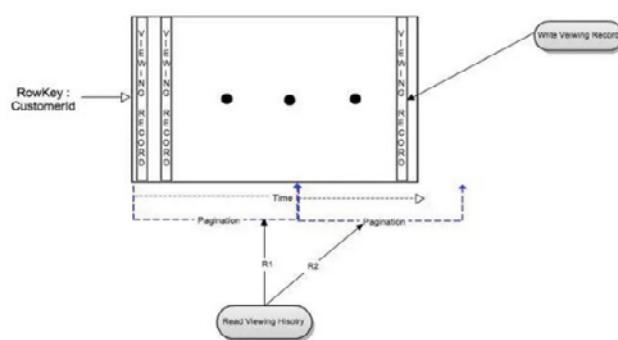


图 1: 单表数据模型

写操作流

当一位会员开始播放视频时，一条观看记录会以一个新列的方式插入当会员暂停或停止观看视频流时，观看记录会做更新在 Cassandra 中，对单一列值的写操作是快速和高效的

读操作流

为检索一位会员的所有观看记录，需要读取整行记录如果每位会员的观看记录数量不大，这时读操作是高效的如果一位会员观看了大量的视频，那么他的观看记录数量将会增加，即记录的列数增加读取一个具有大量列的数据行，会对 Cassandra 造成了额外压力，进而对读操作延迟产生负面影响

要读取一段时间内的会员数据，需要做一次时间范围查询这同样会导致上面介绍的性能不一致问题因为查询性能依赖于给定时间范围内的观看

记录数量。

如果要查看的历史数据规模很大，需要做分页才能进行整行读操作分页对 Cassandra 更好，因为查询不需要等待所有数据都就绪，就能返回给用户分页也避免了客户超时问题但是，随着观看记录的增长，分页增加了读取整行的整体延迟

延迟的原因

下面介绍一些 Cassandra 的内部机制，进而理解为什么我们最初简单的设计会产生性能下降随着数据的增长，SSTable 的数量也随之增加因为只有最近的数据是维护在内存中的，因此在很多情况下，检索观看历史记录时需要同时读取内存表和 SSTable 这对于读取延迟具有负面影响同样，随着数据的增长，合并（Compaction）操作将占用更多的 IO 和时间此外，随着一行记录越来越宽，读修复（Read repair）和全列修复（Full column repair）也会变慢

缓存层

Cassandra 可以很好地对观看数据执行写操作，但是需要改进读操作上的延迟为优化读操作延迟，我们考虑以增加写路径上的工作为代价，在 Cassandra 存储前增加了一个内存中的分片缓存层（即 EVCache）缓存实现为一种基本的键 - 值存储，键是 CustomerId，值是观看历史数据的二进制压缩表示每次 Cassandra 的写操作，将额外生成一次缓存查找操作一旦缓存命中，直接给出缓存中的已有值对于观看历史记录的读操作，首先使用缓存提供的服务一旦缓存没有命中，再从 Cassandra 读取条目，压缩后插入到缓存中

在添加了缓存层后，多年来 Cassandra 单表存储方法一直工作很好在 Cassandra 集群上，基于 CustomerId 的分区提供了很好的扩展到 2012 年，查看历史记录的 Cassandra 集群成为了 Netflix 的最大专用 Cassandra 集群之一为进一步实现存储的规模化，团队需要实现集群的规模翻番这意味着，团队需要冒险进入 Netflix 在使用 Cassandra 上尚未

涉足的领域同时，Netflix 业务也在持续快速增长，其中包括国际会员的增长，以及企业即将推出的自制节目业务

重新设计：实时存储和压缩存储

很明显，为适应未来五年中企业的发展，团队需要尝试多种不同的方法去实现存储的规模化。团队分析了数据的特征和使用模式，重新定义了观看历史存储团队给出了两个主要目标：

- 更小的存储空间；
- 考虑每位会员观看视频的增长情况，提供一致的读写性能。

团队将每位会员的观看历史数据划分为两个数据集：

- 实时 / 近期观看历史记录(LiveVH)：一小部分频繁更新的近期观看记录 LiveVH 数据以非压缩形式存储，详细设计随后介绍；
- 压缩 / 归档观看历史记录(CompressedVH)：大部分很少更新的历史观看记录该部分数据将做压缩，以降低存储空间。压缩观看历史作为一列，按键值存储在一行中。

为提供更好的性能，LiveVH 和 CompressedVH 存储在不同的数据库表中，并做了不同的优化。考虑到 LiveVH 更新频繁，并且涉及的观看记录数量不大，因此可对 LiveVH 做频繁的 Compaction 操作并且为了降低 SSTable 数量和数据规模，可以设置很小的 gc_grace_seconds 为改进数据的一致性，也可以频繁执行读修复和全列族修复 (full column family repair)。而对于 CompressedVH，由于该部分数据很少做更新操作，因此为了降低 SSTable 的数量，偶尔手工做完全 Compaction 即可在偶尔执行的更新操作中，会检查数据一致性，因此也不必再做读修复以及全列族修复。

写操作流

对于新的观看记录，使用同上的方法写入到 LiveVH

读操作流

为有效地利用新设计的优点，团队更新了观看历史 API，提供了读取

近期数据和读取全部数据的选项

读取近期观看历史：在大多数情况下，近期观看历史仅需从 LiveVH 读取这限制了数据的规模，进而给出了更低的延迟

读取完整观看历史：实现为对 LiveVH 和 CompressVH 的并行读操作
考虑到数据是压缩的，并且 CompressedVH 具有更少的列，因此读取操作涉及更少的数据，这显著地加速了读操作

CompressedVH 更新流

在从 LiveVH 读取观看历史记录时，如果记录数量超过了一个预设的阈值，那么最近观看记录将由后台任务打包（roll up）压缩并存储在 CompressedVH 中打包数据存储在一个行标识为 CustomerId 的新行中
新打包的数据在写入后会给出一个版本，用于读操作检查数据的一致性只有验证了新版本的一致性后，才会删除旧版本的打包数据出于简化的考虑，在打包中没有考虑加锁，由 Cassandra 负责处理非常罕见的重复写问题（即以最后写入的数据为准）

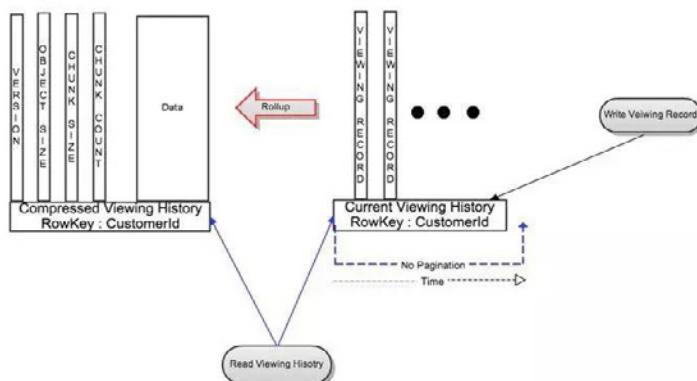


图 2: 实时数据和压缩数据的操作模型

如图 2 所示，CompressedVH 的打包行中还存储了元数据信息，其中包括最新版本信息对象规模和分块信息，细节稍后介绍记录中具有一个版本列，指向最新版本的打包数据这样，读取 CustomerId 总是会返回最新打包的数据为降低存储的压力，我们使用一个列存储打包数据为最小化具有频繁观看模式的会员的打包频率，LiveVH 中仅存储最近几天的观看历

史记录打包后，其余的记录在打包期间会与 CompressedVH 中的记录归并
通过分块实现自动扩展

通常情况是，对于大部分的会员而言，全部的观看历史记录可存储在一行压缩数据中，这时读操作流会给出相当不错的性能。罕见情况是，对于一小部分具有大量观看历史的会员，由于最初架构中的同一问题，从一行中读取 CompressedVH 的性能会逐渐降低。对于这类罕见情况，我们需要对读写延迟设置一个上限，以避免对通常情况下的读写延迟产生负面影响。

为解决这个问题，如果数据规模大于一个预先设定的阈值，我们会将打包的压缩数据切分为多个分块，并存储在不同的 Cassandra 节点中。即使某一会员的观看记录非常大，对分块做并行读写也会将读写延迟控制在设定的上限内。

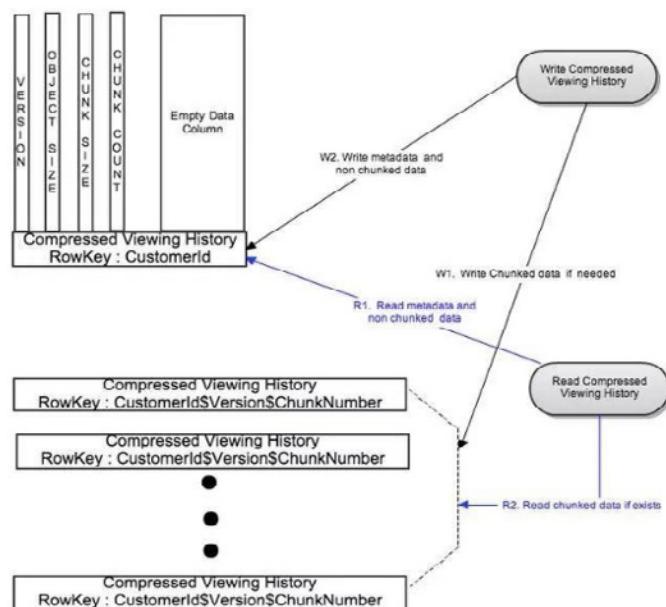


图 3: 通过数据分块实现自动扩展

写操作流

如图 3 所示，打包压缩数据基于一个预先设定的分块大小切分为多个分块。各个分块使用标识 `CustomerId$Version$ChunkNumber` 并行

写入到不同的行中在成功写入分块数据后，元数据会写入一个标识为 CustomerId 的单独行中对非常大的打包观看数据，这一做法将写延迟限制为两次写操作这时，元数据行实现为一个不具有数据列的行这种实现支持对元数据的快速读操作

为加快对通常情况（即经压缩的观看数据规模小于预定的阈值）的处理，我们将元数据与观看数据合并为一行，消除查找元数据的开销，如图 2 所示

读操作流

在读取时，首先会使用行标识 CustomerId 读取元数据行对于通常情况，分块数是 1，元数据行中包括了打包压缩观看数据的最新版本对于罕见情况，存在多个压缩观看数据的分块我们使用元数据信息（例如版本和分块数）对不同分块生成不同的行标识，并行读取所有的分块这将读延迟限制为两次读操作

改进缓存层

为了支持对大型条目的分块，我们还改进了内存中的缓存层对于存在大量观看历史的会员，整个压缩的观看历史可能无法置于单个 EVCache 条目中因此，我们采用类似于对 CompressedVH 模型的做法，将每个大型缓存条目分割为多个分块，并将元数据存储在首个分块中

结果

在引入了并行读写数据压缩和数据模型改进后，团队达成了如下目标：

1. 通过数据压缩，实现了占用更少的存储空间；
2. 通过分块和并行读写，给出了一致的读写性能；
3. 对于通常情况，延迟限制为一次读写对于罕见情况，延迟限制为两次读写。

团队实现了数据规模缩减约 6 倍，Cassandra 维护时间降低约 13

倍，平均读延迟降低约 5 倍，平均写时间降低约 1.5 倍更为重要的是，团队实现了一种可扩展的架构和存储空间，可适应 Netflix 观看数据的快速增长。

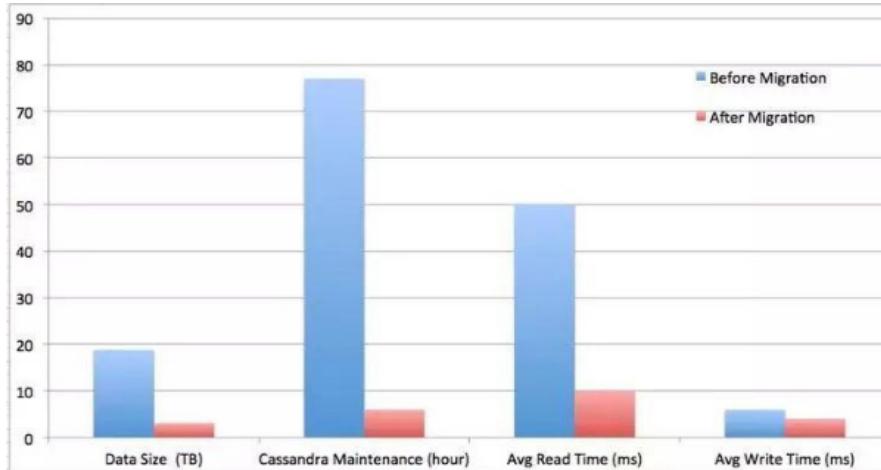


图 4: 运行结果

在该博客系列文章的第二部分中，我们将介绍存储规模化中的一些最新挑战这些挑战推动了会员观看历史数据存储架构的下一轮更新如果读者对解决类似问题感兴趣，可加入到我们的团队中

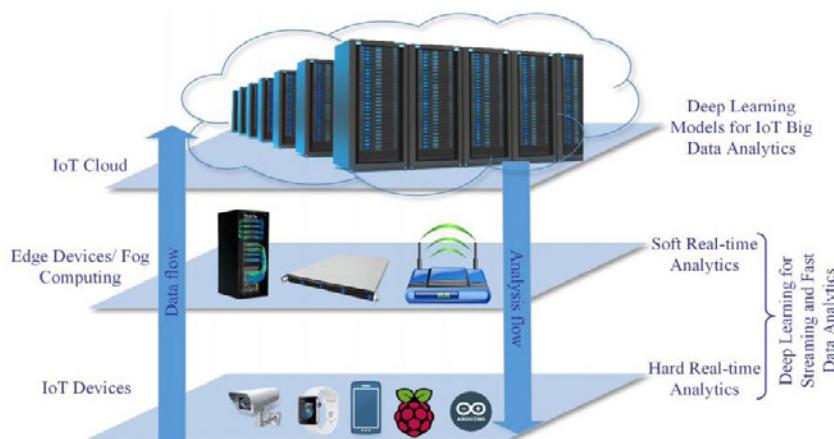
深度学习在 IoT 大数据和流分析中的应用



在物联网时代，大量的感知器每天都在收集并产生着涉及各个领域的数据。由于商业和生活质量提升方面的诉求，应用物联网（IoT）技术对大数据流进行分析是十分有价值的研究方向。这篇论文对于使用深度学习来改进IoT领域的数据分析和学习方法进行了详细的综述。从机器学习视角，作者将处理IoT数据的方法分为IoT大数据分析和IoT流数据分析。论文对目前不同的深度学习方法进行了总结，并详细讨论了使用深度学习方法对IoT数据进行分析的优势，以及未来面临的挑战。

论文贡献

为了更好的在IoT领域内应用深度学习方法，作者分析了IoT数据的关



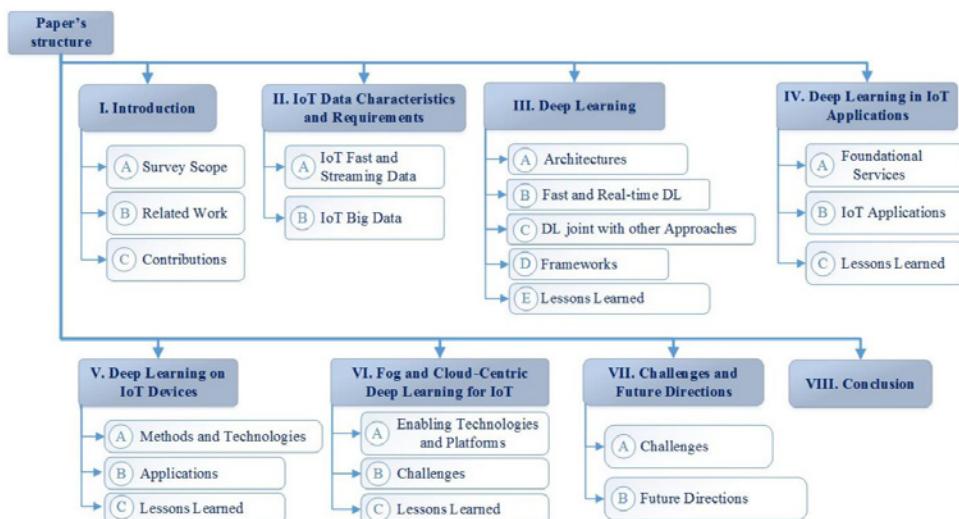
键特征和主要问题。

作者对于目前最先进的深度学习方法及其在物联网领域对于大数据和流数据的应用进行了详细的总结。

作者对于目前应用了深度学习方法的大量IoT应用进行了介绍，并且对不同类型的深度神经网络在各种IoT领域的应用进行了概括和对比。

强调了深度学习与物联网应用成功结合所面临的挑战和未来的研究方向。

论文结构



物联网数据特征及分析要求

IoT快速流数据

目前流数据分析都是基于数据并行计算或增量处理的框架，尽管这些技术减少了从流数据分析框架返回响应的时间延迟，对于IoT应用的严格时间要求，它们并不是最佳方案。IoT需要在数据源附近的平台（甚至是IoT设备自身）上进行快速流数据分析，以达到实时或近实时性的要求，传统的流数据分析方法则面临着计算、存储以及数据源能量方面的局限和挑战。

IoT大数据

IoT大数据具有“6V”特点：

1. 容量 (Volume)： 数据量是将数据集视为大数据、或传统的大规模/超大数据的一个决定性因素，使用物联网设备产生的数据量比以前要多得多，明显符合这一特点。
2. 速度 (Velocity)： 物联网大数据产生和处理速率要足够高，以支持实时大数据的可用性。鉴于这种高数据率，也证明了需要先进的工具和技术分析才能有效地运作。
3. 多样性 (Variety)： 一般来说，大数据有不同的形式和类型。这可能包括结构化的、半结构化的和非结构化的数据。各种各样的数据类型可以通过物联网产生，如文本、音频、视频、传感器数据等等。
4. 真实性 (Veracity)： 真实性是指质量，一致性，和数据的可信性，有真实性的数据才能进行准确的分析。这一点对于物联网来说尤其重要，特别是那些群体感知数据。
5. 易变性 (Variability)： 这个属性是指数据流的速率不同。由于物联网应用的性质，不同的数据生成组件可能会有不一致的数据流。此外，在特定时间，一个数据源的数据加载速率可能不

同。例如，利用物联网传感器的停车服务应用在高峰期的数据加载会达到峰值。

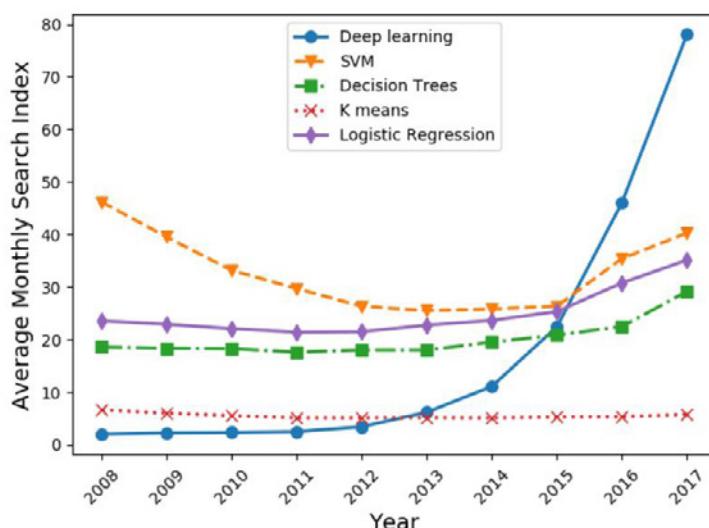
6. 价值 (Value)： 价值是指大数据转化成为有用的信息和内容，为组织带来竞争优势。数据的价值的高度不仅仅取决于对数据的处理过程或服务，还取决于对待数据的方式。

数据流处理的主要障碍是缺少能部署在系统边缘，甚至是IoT设备上的框架或算法。当采用深度学习方法时，也要折衷考虑运行在系统边缘的网络的深度和性能。

深度学习

与其他传统机器学习方法相比，深度学习结构在近几年受到越来越广泛的关注。

Google Trend显示近几年对深度学习的关注呈上升趋势。

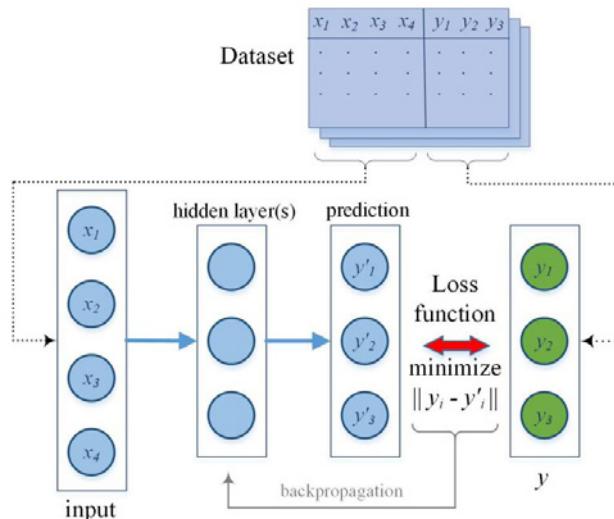


结构

- 1) 卷积神经网络 (Convolutional Neural Networks, CNN)

CNN的核心结构是卷积层，有一系列可学习的参数，称作滤波器。训练过程中，滤波器在全图按照卷积顺序进行移动，计算输入和滤波器的乘

积，得到该滤波器的特征图。CNN的另一个结构是池化层，将输入划分成不重叠的区域，然后用每个区域的最大值作为输出。CNN的最后一个结构是ReLU激活函数层，既可以缩短训练时间，也能避免影响网络的泛化能力。



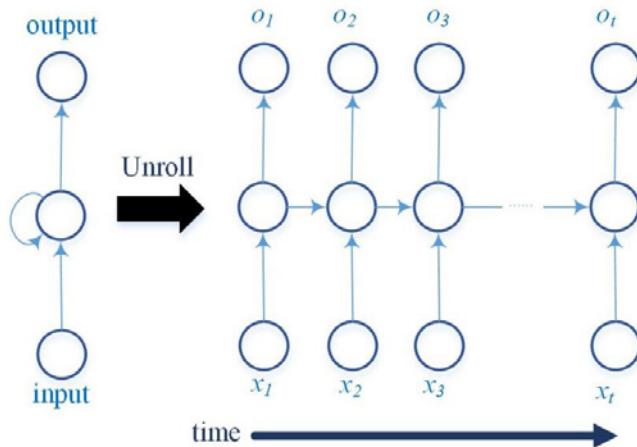
深度学习整体训练机制

Model	Category	Learning model	Typical input data	Characteristics
AE	Generative	Unsupervised	Various	<ul style="list-style-type: none"> Suitable for feature extraction, dimensionality reduction Same number of input and output units The output reconstructs input data Works with unlabeled data
RNN	Discriminative	Supervised	Serial, time-series	<ul style="list-style-type: none"> Processes sequences of data through internal memory Useful in IoT applications with time-dependent data
RBM	Generative	Unsupervised, Supervised	Various	<ul style="list-style-type: none"> Suitable for feature extraction, dimensionality reduction, and classification Expensive training procedure
DBN	Generative	Unsupervised, Supervised	Various	<ul style="list-style-type: none"> Suitable for hierarchical features discovery Greedy training of the network layer by layer
LSTM	Discriminative	Supervised	Serial, time-series, long time dependent data	<ul style="list-style-type: none"> Good performance with data of long time lag Access to memory cell is protected by gates
CNN	Discriminative	Supervised	2-D (image, sound, etc.)	<ul style="list-style-type: none"> Convolution layers take biggest part of computations Less connection compared to DNNs. Needs a large training dataset for visual tasks.
VAE	Generative	Semi-supervised	Various	<ul style="list-style-type: none"> A class of Auto-encoders Suitable for scarcity of labeled data
GAN	Hybrid	Semi-supervised	Various	<ul style="list-style-type: none"> Suitable for noisy data Composed of two networks: one generator and one discriminator
Ladder Net	Hybrid	Semi-supervised	Various	<ul style="list-style-type: none"> Suitable for noisy data Composed of three networks: two encoders and one decoder

深度学习模型总结

CNN和DNN的主要区别在于CNN具有局部相连、权值共享的特性，因此在视觉任务中具有独特的优越性，并且降低了网络的复杂性。

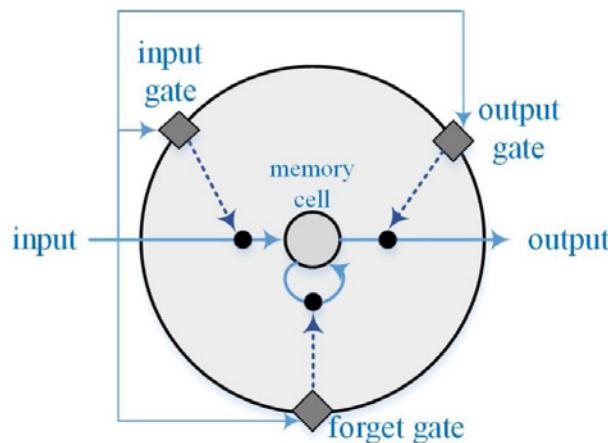
2) 循环神经网络 (Recurrent Neural Networks, RNN)



循环神经网络结构图

RNN主要适用于输入为序列（例如语音和文本）或时间序列的数据（传感器数据）。RNN的输入既包括当前样例，也包括之前观察的样例。也就是说，时间为 $t-1$ 时RNN的输出会影响时间为 t 的输出。RNN的每个神经元都有一个反馈环，将当前的输出作为下一步的输入。该结构可以解释为RNN的每个神经元都有一个内部存储，保留了用之前输入进行计算得到的信息。

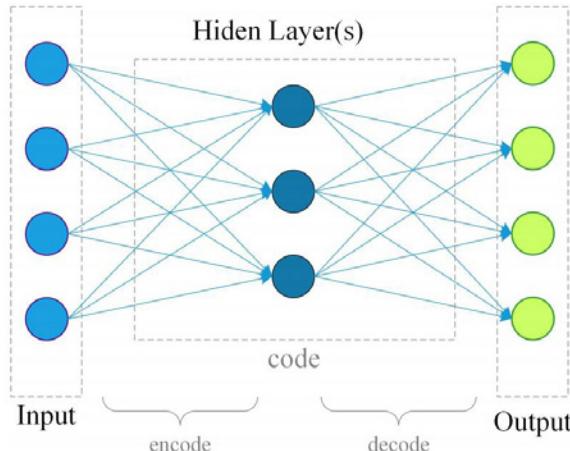
3) 长短时记忆 (Long Short Term Memory, LSTM)



LSTM记忆单元结构

LSTM是RNN的一种扩展。LSTM中，每个神经元除了有反馈环这一储存信息的机制，还有用于控制神经元信息通过的“遗忘门”、“输入层门”及“输出层门”，防止不相关的信息造成的扰动。

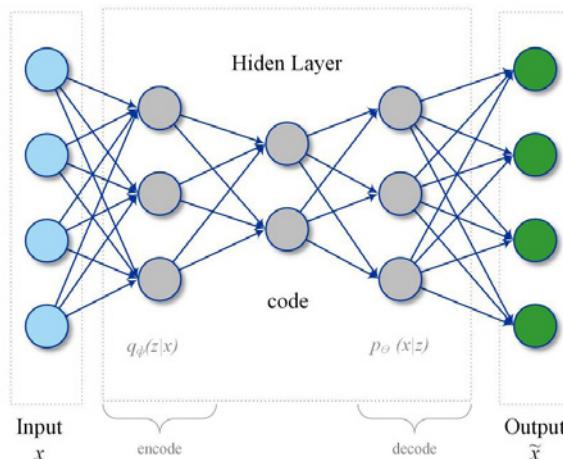
4) 自动编码器 (Autoencoders, AE)



自编码器网络结构

AE的输入层和输出层由一个或多个隐层相连接，其输入和输出神经元数量相同。该网络的目标是通过用最简单的方式将输入变换到输出，以重建输入信息。

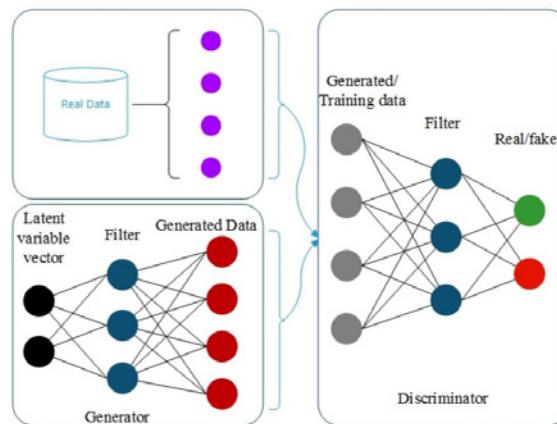
5) 变分自动编码器 (Variational Autoencoders, VAE)



变分自动编码器结构

VAE对数据结构的假设并不强，是较为流行的生成模型框架。它很适用于IoT解决方案，因为IoT数据呈现的多样性，以及标记数据的缺失。模型由两个子网络组成：一个生成样例，一个进行假设推理。

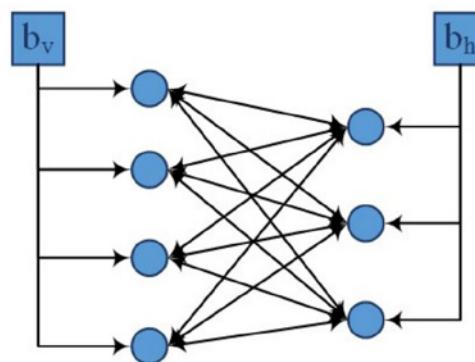
6) 生成对抗网络 (Generative Adversarial Networks, GAN)



生成对抗网络概念图

GAN由两个神经网络组成，一个生成网络，一个判别网络，共同工作来产生合成的、高质量数据。生成器根据数据在训练数据集中的分布生成新数据，判别器学习判别真实数据和生成器生成的假数据。GAN的目标函数是基于极大极小博弈的，一个网络要最大化目标函数，而另一个要最小化目标函数。

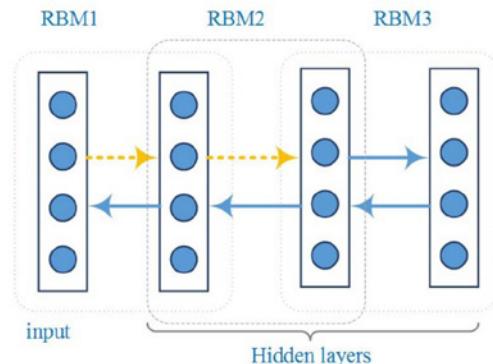
7) 受限玻尔兹曼机 (Restricted Boltzmann Machine, RBM)



受限玻尔兹曼机结构

RBM是一种随机神经网络，由两层组成，一层是包含输入的可见层，一层是含有隐变量的隐藏层。RBM中的限制是指同一层的任意两个神经元互不相连。除此之外，偏置单元与所有的可见层和隐藏层单元都相连。

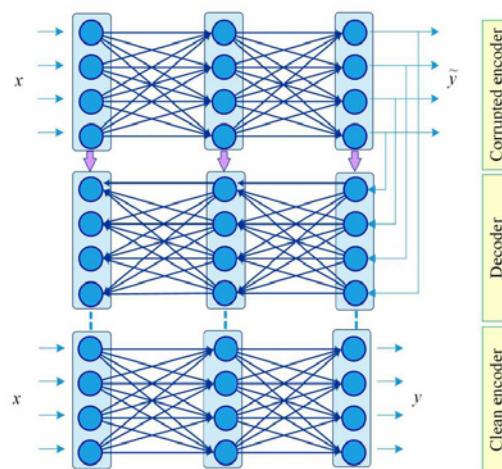
8) 深度信念网络 (Deep Belief Network, DBN)



深度信念网络结构图。虚线表示特征提取通道，实现表示生成通道

DBN是一种生成神经网络，由一个可见层可几个隐层组成。可以提取训练数据的多层表示，并且对输入数据进行重构。DBN的训练过程是逐层训练，将每一层视作一个RBM，在前一层的基础上进行训练。这样的机制使DBN成为深度学习中有效且快速的网络之一。

9) 阶梯网络 (Ladder Network)



两层阶梯网络

阶梯网络在无监督和半监督学习任务中达到了先进的水平。阶梯网络由两个编码器和一个解码器组成。编码器作为网络的有监督部分，解码器进行无监督学习。训练目标是最小化有监督部分和无监督网络的损失和。

快速实时深度学习结构

使用深度学习模型对数据流进行快速实时的处理仍在起步阶段。早期工作【1】是对超限学习机（Extreme learning machine, ELM）的扩展——OS-ELM，将一个实时序列学习算法应用到单隐层的前馈神经网络。Ren等人【2】提出的Faster-RCNN在图片中的目标检测中达到了接近实时的速度。他们的目标检测框架的运行时间为5–17fps。然而对于图像处理任务，真正的实时效果需要系统的处理和分析时间达到30fps或更高。Redmon等人【3】提出了YOLO，将目标检测的速度提高到45fps，以及更小版本的YOLO，速度更是达到了155fps，已经适用于智能相机。

深度学习与其他方法结合

1) 深度增强学习 (Deep Reinforcement Learning)

深度增强学习是将增强学习和深度神经网络相结合的产物。其目标是创建能自主学习的个体（agent），通过建立成功的交互过程以获得长期的最大正反馈（reward）。当环境（environment）可由大量状态表示时，传统的增强学习方法稍显不足，而深度神经网络则弥补了这一点。在IoT领域，【4】使用深度增强学习实现了半监督条件下智能校园环境中的定位。

2) 迁移学习与深度模型 (Transfer Learning with Deep Models)

迁移学习主要应用在域适应和多任务学习的领域。迁移学习对于许多难以收集训练数据的IoT应用来说都是一个可用的解决方案。例如训练一个通过智能手机的低功耗蓝牙和Wifi fingerpringting的定位系统，同一时间，在同一地点的RSSI值（Received Signal Strength Indication接收的信号强度指示）对于不同的平台来说可能不同。如果我们对一个平台训练了一个模型，该模型可以迁移到其他平台，而不需要对新平台再收集

训练数据。

3) 深度学习与在线学习算法

由于IoT的应用产生的数据流会上传到云平台来分析，在线机器学习算法的角色变得越来越重要，因为训练模型需要随数据的增加而更新。

框架

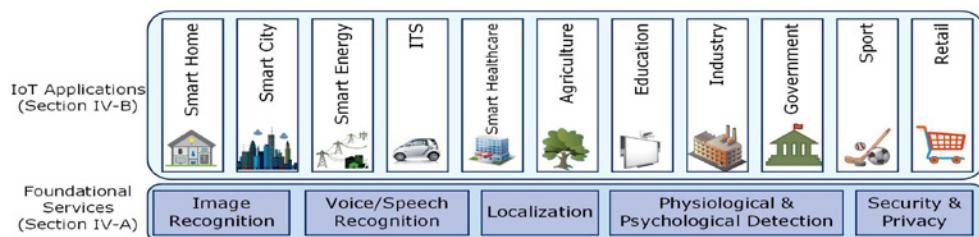
近几年，随着深度学习在各个领域的应用热潮，各种深度学习框架也应运而生。

- Tensorflow:** Tensorflow是机器学习系统的开源库，可以使用多种深度神经网络。Tensorflow使用图表示来建立神经网络模型。开发人员也在使用TensorBoard，能可视化神经网络模型，并且观测学习过程，包括参数更新。
- Torch:** Torch是一个机器学习开源框架，包含大量深度学习算法，可用于深度神经网络模型的简单开发。它基于Lua语言开发，是训练深度学习算法的轻量级快速框架。支持在CPU和GPU上开发机器学习模型，并且提供了训练深度神经网络的并行计算库。
- Caffe:** Caffe是一个深度学习算法和参考模型集的开源框架。基于C++，支持CUDA进行GPU运算，并且提供Python和Matlab接口。Caffe通过配置文件定义模型，而不需要在源代码中定义参数，将模型表示和实现分开。

Frameworks	Core Language	Interface	Pros	Cons	Used in IoT Application
H2O	Java	R, Python, Scala, REST API	▪ Wide range of interfaces	▪ Limited number of models are supported ▪ Is not flexible	[54]
Tensorflow	C++	Python, Java, C, C++, Go	▪ Fast on LSTM training ▪ Support to visualize networks	▪ Slower training compared to other Python-based frameworks	[55]
Theano	Python	Python	▪ Supports various models ▪ Fast on LSTM training on GPU	▪ Many low level APIs	[56]
Torch	Lua	C, C++	▪ Supports various models ▪ Good documentation ▪ Helpful error debugging messages	▪ Learning a new language	[55] [57]
Caffe	C++	Python, Matlab	▪ Provides a collection of reference models ▪ Easy platform switching ▪ Very good at convolutional networks	▪ Not very good for recurrent networks	[58]–[60]
Neon	Python	Python	▪ Fast training time ▪ Easy platform switching ▪ Supports modern architectures like GAN	▪ Not supporting CPU multi-threading	[61]
Chainer [62]	Python	Python	▪ Supports modern architectures ▪ Easier to implement complex architectures ▪ Dynamic change of model	▪ Slower forward computation in some scenarios	[63]
Deeplearning4j	Java	Python, Scala, Clojure	▪ Distributed training ▪ Imports models from major frameworks (e.g., TensorFlow, Caffe, Torch, and Theano) ▪ Visualization tools	▪ Longer training time compared to other tools	[64], [65]

深度学习框架对比

深度学习在 IoT 领域的应用



IoT应用和基础服务

基础服务

1) 图像识别

IoT的一大应用场景中，输入深度学习的数据是图片或视频。每天，每个人都在用手机的高清摄像头拍摄者图片和视频，除此之外，家居、校园或工厂也在使用智能摄像头。所以，图像识别、分类、目标检测是这类设备的基础应用。

2) 语音识别

随着智能手机和可穿戴设备的普及，语音识别也成了人们和自己的设备互动的一种自然而方便的方式。Price等人【5】搭建了一个专用的低功耗深度学习芯片，用于自动语音识别。这种特制芯片的能量消耗要比目前手机上运行的语音识别工具的能量消耗低100倍。

3) 室内定位

室内定位在IoT领域有许多应用，例如智能家居、智能校园、或智能医院。例如DeepFi系统，在线下训练阶段，通过深度学习用之前储存的WiFi通道状态信息数据来训练网络权重，在线上定位阶段通过fingerprinting来测定用户位置。

4) 生理和心理状态检测

IoT与深度学习的结合也应用在了检测各种生理或心理状态中，例如

姿态、活动和情绪。许多IoT应用都在交付的服务中整合了人体姿态估计或活动识别模块，例如智能家居、智能汽车、XBox、健康、运动等等。

5) 安全和隐私

安全和隐私是所有IoT领域应用所关注的一个重要问题。事实上，系统功能的有效性取决于是否能保护机器学习工具和处理过程不受攻击。虚假数据注入（False Data Injection, FDI）是数据驱动系统的一种常见攻击类型。He等人【6】提出用条件DBN从历史数据中提取FDI特征，然后利用这些特征进行实时攻击检测。作为物联网数据和应用程序的一大贡献者，智能手机也面临着黑客攻击的威胁。Yuan等人【7】提出用深度学习框架来鉴别安卓应用中的恶意软件，准确率达到了96.5%。深度机器学习方法的安全性和隐私保护是能否在IoT领域应用的最重要因素。Shokri等人【8】提出了一种解决分布式学习的深度学习模型隐私保护问题的方法。

应用

1) 智能家居

智能家居的概念涉及广泛的基于IoT的应用，它有助于提高家庭的能源使用和效率，以及居住者的便利性、生产力和生活质量。如今，家电可以与互联网连接，提供智能服务。例如微软和 Liebherr的一个合作项目，对从冰箱内收集的信息应用了Cortana 深度学习。这些分析和预测可以帮助家庭更好地控制他们的家庭用品和开支，并结合其他外部数据，可用于监测和预测健康趋势。

2) 智慧城市

智慧城市服务跨越多个物联网领域，如交通、能源、农业等。智慧城市的一个重要问题是预测群体移动模式，Song等人【9】开发了基于深度神经网络模型的系统，在城市级别实现了这一目标。Liang等人【10】基于RNN模型搭建了实时群体密度预测系统，利用移动手机用户的通信数据

对交通站的群体密度进行预测。废物管理和垃圾分类也是智慧城市的一个相关任务，可以通过基于视觉分类任务的CNN模型来实现自动化。Amato等人【11】基于智能相机和深度CNN开发了检测停车场的使用中和空闲车位的系统。

3) 能源

消费者与智能电网之间的双向通信是IoT大数据的来源。能源供应商希望学习当地的能源消费模式、预测需求，并根据实时分析做出适当的决定。在智能电网方面，从太阳能、风能或其他类型的自然可持续能源中预测电力是一个活跃的研究领域，深度学习在这一领域的许多应用中越来越多地被使用。

4) 智能交通系统

来自智能交通系统（ITS）的数据是大数据的另一个数据源。Ma等人【12】采用RBM和RNN结构设计了一个交通网络分析系统，模型输入是参与该系统的出租车GPS数据。该系统通过一小时内的累积数据预测交通拥堵的准确率高达88%。ITS也带动了交通标志检测和识别的发展，这一技术在自动驾驶、辅助驾驶系统中都有很重要的应用。除此之外，许多初创公司应用深度学习来完善自动驾驶汽车系统的检测行人、交通标志、路障等任务。

5) 医疗和健康

IoT结合深度学习也在为个人和组织提供医疗和健康方案中得到应用。例如，开发基于移动应用程序的精确测量饮食摄入量的解决方案，可以帮助提升个人健康和幸福感。Liu等人【13】采用CNN开发了识别食物图片和相关信息的系统。用深度学习对医学图片进行分类和分析是医疗领域的研究热点。Pereira等人【14】通过CNN识别手写图片来鉴定早期帕金森症。除此之外，深度学习与IoT的结合在声音异常检测、乳腺血管疾病检测中也得到了应用。

6) 农业

生产健康作物和发展有效的种植方式是健康社会和可持续环境的要求。使用深度神经网络进行植物病害识别是一个可行的解决方案。深度学习也被用于遥感，进行土地和作物的检测与分类。研究显示，使用CNN进行作物识别准确率达到了85%，相比于MLP或随机森林有很大提高。自动耕作中的预测和检测任务也应用了深度学习。

7) 教育

IoT和深度学习的结合有助于提高教育系统的效率。移动设备可以收集学生的数据，深度分析方法可以用来预测和解释学生的进步和成就。增强现实技术结合可穿戴设备和移动设备也是深度学习在这一领域的潜在应用，激发学生的兴趣，让教育学习方法更有效。此外，深度学习可以用于个性化推荐模块，向教育者推荐更多相关内容。利用深度学习对大型开放式网络课程数据（MOOC）进行分析，可以帮助学生更好的学习。除此之外，利用CNN监测教室占用率是深度学习在教育方面的另一个应用。

8) 工业

对于工业部门来说，IoT和网络物理系统（CPS）是推动制造技术迈向智能制造（工业4.0）的核心要素。工业中的广泛应用均可以受益于深度学习模型的引入。通过将装配线中生产车辆的图像及其注释都输入深度学习系统，可以利用AlexNet、GoogLeNet等网络实现视觉检测。

9) 政府

许多涉及市政的各种任务需要精确的分析和预测。【15】利用美国地质调查局网络的历史数据训练LSTM网络，可进行地震预测。【16】利用极端气候的图片训练CNN，进行极端气候事件探测。此外，城市的基础设施，如道路、供水管道等的损害检测，是IoT和深度学习可以为政府提供便利的另一个领域。

10) 运动和娱乐

运动分析近年来发展迅速，为团队或运动员带来了竞争优势。【17】

提出了深度学习方法打造智能篮球场。【18】采用RNN识别NBA比赛中的球员违规。【19】结合了可穿戴设备传感数据和CNN进行排球运动员活动识别。【20】采用层级结构的LSTM模型研究排球队的整体活动。

11) 零售

随着移动设备的普及，网上购物的人数大大增加了。最近出现了通过视觉搜索技术向产品图像检索的转变。CNN一直用于服装和时尚市场的视觉搜索，帮助你在网店中找到在电影中看到的或在街上看到的商品。IoT结合深度学习可以搭建视觉购物辅助系统，包括智能眼镜、手套和购物车，目的是帮助视障人士购物。此外，智能购物车的开发可以实现实时自结账的功能。

IoT 设备上的深度学习

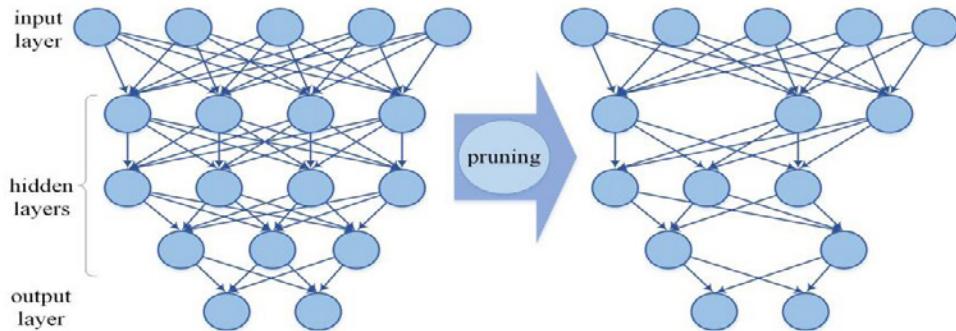
深度学习在语音和视频方面的成功为IoT的基础服务打下了良好的基础，如何将它们的模型和方法部署在资源受限的设备上成了IoT领域的一个重要研究方向。到目前为止，深度学习方法难以应用于IoT和资源受限设备，因为它们需要大量的资源来运行，如处理器、电池能量和存储器。幸运的是，近期研究显示，深度神经网络的许多参数是冗余的，有时也不需要大量的隐层。有效的去除这些参数或层可以减少网络的复杂度，同时对输出不会有太大的影响。

方法和技术

1) 网络压缩

在资源受限设备上应用深度神经网络的方法之一是网络压缩，将密集的网络转化为一个稀疏的网络。主要局限性在于，它不足以支持所有类型的网络。它只适用于具有这种稀疏性的特定网络模型。另外，修剪多余的和不重要的参数或神经元，是在资源受限的设备上运行深度神经网络的另一个重要途径。

2) 近似计算



深度神经网络剪枝整体概念图

近似计算是实现在IoT设备上部署机器学习工具的另一种方法，并有助于主机设备的节能。在许多IoT应用中，机器学习的输出不一定是精确的，而是在可接受的范围内提供所需的质量。实际上，将深度学习模型与近似计算相结合，可以为资源受限设备提供更有效的深度学习模型。

3) 加速器

设计特定的硬件和电路来优化IoT设备中深度学习模型的能量效率和内存占用是另一个活跃的研究方向。目前已有工作为DNN和CNN设计加速器，并且应用Post-CMOS技术进行电子自旋加速。

4) 微处理器

除了之前所提方法，开发具有强深度学习能力的小尺寸处理器也是研究热点。微处理器的设计尺寸在一立方毫米的范围内，可以用电池驱动，进行深度神经网络分析只消耗大约300毫瓦。通过这种技术，许多对时间要求较高的IoT应用程序可以在设备上执行决策，而不是将数据发送到高性能计算机，等待它们的响应。

IoT 的雾和云中心深度学习

最近，人们提出了雾计算，使计算和分析更接近终端用户和设备，而不是仅仅停留在云计算上。实验表明，通过对雾计算节点进行数据分析，可以避免向遥远的云节点传输大量原始数据，从而提高整体性能。还可以

在一定程度上进行实时分析，因为雾计算在本地，靠近数据源。

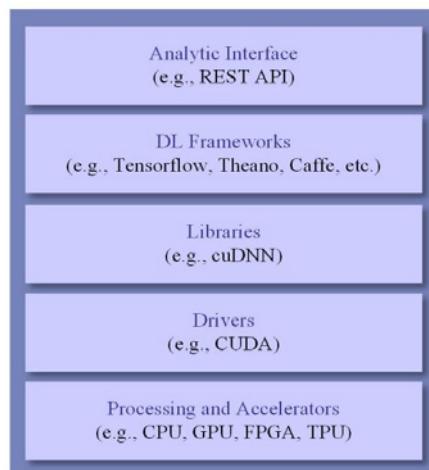
Product	Description	Application	Platform
Amazon Alexa	Intelligent personal assistant (IPA)	Smart home	Fog
Microsoft Cortana	IPA	Smart Car, XBox	Fog
Google Assistant	IPA	Smart Car, Smart home	Fog
IBM Watson	Cognitive framework	IoT domains	Cloud

一些用于在雾或云上使用深度学习和服务的IoT领域的产品

技术和平台

尽管在雾计算架构上引入了深度学习分析，云计算仍然是许多无法在雾计算中处理的IoT应用的唯一可行的解决方案。因此，设计的可扩展的和高性能的云中心的DNN模型和算法，对大量的IoT数据进行分析，仍然是一个重要的研究领域。

除了在云平台上托管可扩展的深度学习模型基础设施的进步，还需要研究使深度学习模型通过API访问的机制和方法，以便容易地集成到IoT应用程序中。



在云平台中作为服务的深度学习模型

挑战

在雾计算节点上进行深度学习分析时，也会面临一些挑战：

- 深度学习服务发现：设备需要通过深度学习分析的某种扩展服务发现协议，来识别适当的分析提供者的来源。
- 深度学习模型和任务分布：在雾节点之间划分深度学习模型和任务的执行，以及在可用节点之间优化数据流分配，对于时间敏感的应用程序是至关重要的。
- 设计因素：研究如何雾计算环境的设计因素，以及在这种环境中部署深度学习模型如何影响分析服务的质量是很有必要的。
- 移动端：在设计终端辅助的深度学习分析系统时，需要考虑移动端计算环境的动态性，因为移动设备可能会加入或离开系统。

深度学习带来的 IoT 挑战，以及未来的研究方向

挑战

1) 缺少大型 IoT 数据集

缺乏可用的实际IoT应用大数据集将深度学习模型引入IoT的一个主要障碍，因为深度学习需要更多的数据来实现更高的精度。此外，更多的数据也可以防止模型过度拟合。

2) 预处理

许多深度学习方法需要对数据进行预处理以产生更好的结果，对于IoT应用，预处理会更复杂，因为系统处理的是来自不同数据源的数据，可能有多种格式和分布，而且还可能有数据丢失。

3) 安全和隐私

确保数据安全和隐私是许多IoT应用的一个主要问题，因为IoT大数据将通过互联网进行分析，因此世界各地都有可能看得到。此外，深度学习训练模型也容易受到恶意攻击，如虚假数据注入或对抗性样本输入，其中

IoT系统的许多功能或非功能性要求可能无法得到保证。

4) IoT 大数据”6V“特性

Volume (数据量) 对于深度学习模型的时间消耗和结构复杂性提出了很大的挑战。并且数据量巨大也带来了包括噪声和未标注数据的挑战。

Variety (多样性) 带来了管理不同数据源之间冲突的挑战。在数据源没有冲突的情况下，深度学习能够有效处理异质数据。

Velocity (速率) 带来了高速处理和分析数据的要求，增强深度学习的在线学习和序列学习的技术仍需进一步研究。

Veracity (可信度)，当输入数据不是来自可信的数据源时，IoT的大数据分析则是无用的。

Variability (可变性)，IoT大数据的流速可变性对在线分析提出了挑战。

Value (价值)，企业经理采用大数据的一个主要挑战是，他们不清楚如何使用大数据分析来获得价值，并改善他们的业务。

5) IoT 设备上的深度学习

在IoT设备上开发深度学习是一个新的挑战，要考虑在资源受限的设备上处理深度神经网络的需求。

6) 深度学习局限

尽管深度学习模型在许多应用中显示出令人印象深刻的成果，它仍然有局限性。研究发现，深度网络会将无法识别的图片分类到熟悉的种类中。并且深度神经网络的回归能力有待增强。

未来研究方向

1) IoT 移动数据

IoT数据的一大部分来自移动设备。研究利用移动大数据与深度学习方法相结合的有效方式，可以为IoT提供更好的服务，特别是在智慧城市场景中。

2) 结合环境信息

单靠IoT的传感数据不能理解环境的情况。因此，IoT数据需要与其他数据源融合，即环境信息，以补充对环境的理解。

3) IoT 分析的在线资源供应

基于雾和云计算的深度学习快速数据分析部署需要在线配置雾或云资源来承载数据流。由于IoT数据的流特性，无法提前知道数据序列的容量。因此，我们需要一种新的基于当前数据流的算法，并且不依赖于数据流的先验知识。

4) 半监督分析框架

为半监督学习而设计的先进的机器学习算法非常适合于智慧城市系统，可以使用少量的训练数据集训练模型，然后使用大量未标记数据来提高模型的准确性。

5) 可靠的 IoT 分析

深度学习方法可以通过分析大量的信息物理系统（CPS）和IoT系统的日志，以识别和预测可能受到攻击的系统的薄弱点。这将有助于系统防止或从故障中恢复，从而提高CPS和IoT系统的可靠性水平。

6) 自组织通信网络

由于IoT设备的数量庞大，配置和维护他们的基本物理M2M通信和网络变得越来越难。虽然大量的网络节点及其相互关系对传统的机器学习方法是一个挑战，但它为深度学习体系结构提供了一个机会，通过提供自配置、自优化、自修复和自负载平衡等一系列的自我服务足以证明它们在这一领域的能力。

7) 新兴 IoT 应用

无人机：无人机被用于许多实时图像分析任务，如监视、搜索和救援行动，以及基础设施检查。这些设备的采用面临包括路由、节约能源、避免私人区域和避障等挑战。深度学习对于该领域的预测和决策任务有很大

的影响，可以推动无人机达到最佳性能。

虚拟/增强现实：虚拟/增强现实是受益于IoT和深度学习的另一个应用领域。增强现实可以用于提供诸如目标跟踪、行为识别、图像分类和对象识别这样的服务。增强现实可以极大地影响如教育，博物馆，智能车等几大领域。

结论

深度学习和IoT近年来受到研究人员和商业领域的广泛关注，这两项技术对我们的生活、城市和世界都产生了积极的影响。IoT和深度学习构成了一个数据生产者-消费者链，其中IoT生成由深度学习模型分析的原始数据，深度学习模型产生高层次的分析，反馈给IoT系统，以微调和改进服务。

参考文献

- 【1】N.-Y. Liang, G.-B. Huang, P. Saratchandran, and N. Sundararajan, “A fast and accurate online sequential learning algorithm for feedforward networks,” *IEEE Transactions on neural networks*, vol. 17, no. 6, pp. 1411–1423, 2006.
- 【2】S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards realtime object detection with region proposal networks,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.
- 【3】J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.
- 【4】M. Mohammadi, A. Al-Fuqaha, M. Guizani, and J.-S. Oh, “Semisupervised deep reinforcement learning in support of IoT and smart city services,” *IEEE Internet of Things Journal*, vol. PP, no. 99, pp. 1–12, 2017.
- 【5】M. Price, J. Glass, and A. Chandrakasan, “A scalable speech recognizer with deep-neural-network acoustic models and voice-activated power gating,” in *Proceedings of the IEEE ISSCC2017*, 2017.
- 【6】Y. He, G. J. Mendis, and J. Wei, “Real-time detection of false data injection attacks in smart grid: A deep learning-based intelligent mechanism,”

IEEE Transactions on Smart Grid, 2017.

【7】 Z. Yuan, Y. Lu, Z. Wang, and Y. Xue, “Droid-sec: deep learning in android malware detection,” in ACM SIGCOMM Computer Communication Review, vol. 44, no. 4. ACM, 2014, pp. 371-372.

【8】 R. Shokri and V. Shmatikov, “Privacy-preserving deep learning,” in Proceedings of the 22nd ACM SIGSAC conference on computer and communications security. ACM, 2015, pp. 1310-1321.

【9】 X. Song, H. Kanasugi, and R. Shibasaki, “Deeptransport: Prediction and simulation of human mobility and transportation mode at a citywide level.” IJCAI, 2016.

【10】 V. C. Liang, R. T. Ma, W. S. Ng, L. Wang, M. Winslett, H. Wu, S. Ying, and Z. Zhang, “Mercury: Metro density prediction with recurrent neural network on streaming cdr data,” in Data Engineering (ICDE), 2016 IEEE 32nd International Conference on. IEEE, 2016, pp. 1374-1377.

【11】 G. Amato, F. Carrara, F. Falchi, C. Gennaro, C. Meghini, and C. Vairo, “Deep learning for decentralized parking lot occupancy detection,” Expert Systems with Applications, 2017.

【12】 X. Ma, H. Yu, Y. Wang, and Y. Wang, “Large-scale transportation network congestion evolution prediction using deep learning theory,” PloS one, vol. 10, no. 3, p. e0119044, 2015.

【13】 C. Liu, Y. Cao, Y. Luo, G. Chen, V. Vokkarane, and Y. Ma, “Deepfood: Deep learning-based food image recognition for computer-aided dietary assessment,” in International Conference on Smart Homes and Health Telematics. Springer, 2016, pp. 37-48.

【14】 C. R. Pereira, D. R. Pereira, J. P. Papa, G. H. Rosa, and X.-S. Yang, “Convolutional neural networks applied for parkinson’s disease identification,” in Machine Learning for Health Informatics. Springer, 2016, pp. 377-390.

【15】 Q. Wang, Y. Guo, L. Yu, and P. Li, “Earthquake prediction based on spatio-temporal data mining: An lstm network approach,” IEEE Transactions on Emerging Topics in Computing, 2017.

【16】 Y. Liu, E. Racah, J. Correa, A. Khosrowshahi, D. Lavers, K. Kunkel, M. Wehner, and W. Collins, “Application of deep convolutional neural networks for detecting extreme weather in climate datasets,” Int’l Conf. on Advances in Big Data Analytics, 2016.

- 【17】W. Liu, J. Liu, X. Gu, K. Liu, X. Dai, and H. Ma, “Deep learning based intelligent basketball arena with energy image,” in International Conference on Multimedia Modeling. Springer, 2017, pp. 601-613.
- 【18】K.-C. Wang and R. Zemel, “classifying nba offensive plays using neural networks,” in Proc. MIT SLOAN Sports Analytics Conf, 2016.
- 【19】T. Kautz, B. H. Groh, J. Hannink, U. Jensen, H. Strubberg, and B. M. Eskofier, “Activity recognition in beach volleyball using a deep convolutional neural network,” Data Mining and Knowledge Discovery, pp. 1-28, 2017.
- 【20】M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, “A hierarchical deep temporal model for group activity recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1971-1980.



在微信上关注我们



InfoQ

国内最好的原创技术社区，一线互联网公司核心技术人员提供优质内容。订阅 InfoQ，看全球互联网技术最佳实践。做技术的不会没听过 QCon，不会不知道 InfoQ 吧？——冯大辉
从事技术工作，或有兴趣了解 IT 技术行业的朋友，都值得订阅。——曹政



关注「InfoQ」回复“二叉树”，看十位大牛的技术初心，不同圈子程序员的众生相。



聊聊架构

以架构之“道”为基础，呈现更多的务实落地的架构内容。

关注「聊聊架构」
和百位架构师共聊架构



细说云计算

探讨云计算的一切，关注云平台架构、网络、存储与分发。这里有干货，也有闲聊。

关注「细说云计算」
回复“群分享”，
看云计算实践干货分享文章



AI前线

提供最新最全AI领域技术资讯、一线业界实践案例、业界技术分享干货、最新AI论文解读。

关注「AI前线」
回复“AI”，下载《AI前线》
系列迷你书



前端之巅

紧跟前端发展，共享一线技术，不断学习进步，攀登前端之巅。

关注「前端之巅」
回复“京东”，看京东
如何做网站前端监控



移动开发前线

关注移动开发领域最前沿和第一线开发技术，
打造技术分享型社群。

关注「移动开发前线」
回复“群分享”，看移动
开发实践干货文章



高效开发运维

常规运维、亦或是崛起的DevOps，探讨如何
IT交付实现价值。

关注「高效开发运维」
回复“DevOps”，四篇精品
文章领悟DevOps

