

第 3 章

AI 与智能物联网

物联网（Internet of Things, IoT）是能够让所有具有独立功能的实体实现互联互通的网络，它被视为继计算机、互联网之后的又一次信息化浪潮。在这个网络上，人和人之间（Human-to-human, H2H）、人和物之间（Human-to-machine, H2M）以及物和物之间（Machine-to-machine, M2M）都可以通过唯一的电子标识连接起来，进行信息交互和通信，实现智能化识别、定位、跟踪、监控和管理。近十年来，物联网行业快速发展。根据中国信息通讯研究院发布的《物联网白皮书（2018 年）》显示，全球物联网产业规模从 2008 年的 500 亿美元增长到 2018 年的 1510 亿美元。时至今日，物联网正在深刻影响着人们的生活，它已经渗透进了诸多领域，如智能家居、医疗、公共安全、交通、电力等。

本章在分析 5G 时代海量数据的实时处理需求的基础上，提出了基于 AI 能力的边缘计算和云边协同的技术必要性，并进一步结合不同的业务应用场景，详细阐述了 AI 技术如何在实际的业务场景中进行应用并解决实际的业务问题。

3.1 5G 时代 IoT 海量数据实时处理

在国际电信联盟（International Telecommunication Union, ITU）召开的 ITU-RWP5D 第 22 次会议上，确定了未来 5G 应具有的三大类使用场景：增强型移动宽带 eMBB、超高可靠与低时延的通信 uRLLC 和大规模（海量）机器类通信 mMTC，前者主要关注移动通信，后两者则侧重于物联网。如果说 5G 的三大应用场景中，eMBB 场景追求的是人与人之间极致的通信体验，强调的

是人与人的连接，那么 mMTC 和 uRLLC 则是物联网的应用场景，mMTC 主要是人与物之间的信息交互，uRLLC 主要体现物与物之间的通信需求。

5G 三大场景的技术参数如图 3-1 所示。

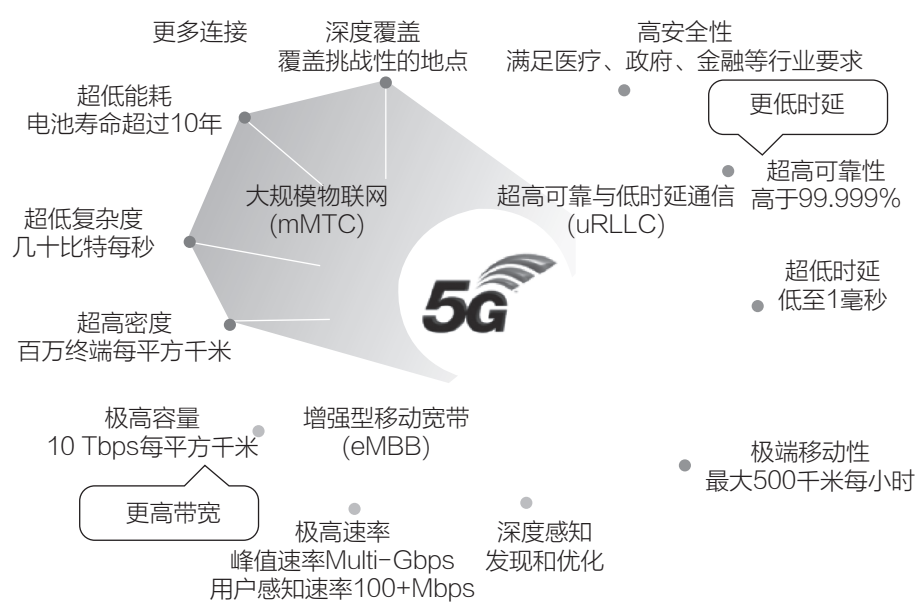


图 3-1 5G 三大场景的技术参数

5G 时代，一些典型场景如表 3-1 所示。

表 3-1 5G 业务典型应用场景

	业务场景	5G 技术需求
增强型移动宽带（eMBB）	超高清视频	8K（3D）视频传输需要传输速率超过 1 Gbps，4G 技术下的传输速率平均只有几百 Mbps
	高铁或飞机联网	解决高铁信号差、飞机上网的问题，需要提高网络覆盖，增强通信能力
大规模机器类通信（mMTC）	大型体育场馆	密集场所内，超过百万的连接数，且设备传输数据量较小
	环境监测	环境监测对传输时延不敏感，且传输速率较低，而传感器数量较多
	智能家居	智能家居类产品对时延要求不敏感，且数据传输量小
	智慧城市	被连接的“物体”多种多样，有超强的连接密度，且各物体的传输数据量较小



续表

	业务场景	5G 技术需求
超高可靠性与低时延通信（uRLLC）	AR/VR	20 ms 以下才能有效地缓解眩晕状态，而 4G 时延大约 100 ms
	无人驾驶	为了保证用户的安全，传输时延需低至 1 ms，且要有超强的可靠性
	远程医疗	需要在短时间内处理大量的数据，且传输网络要有高可靠性

相比 4G 网络，5G 网络有着更深、更广的覆盖范围，有着更高的连接密度，更低的能源消耗，真正使物联网的发展进入了一个全新时代。网络边缘连接的设备数和数据量也在急剧增加，网络规模不断扩大。据互联网数据中心（Internet Data Center，IDC）数据统计，预计到 2020 年，全球物联网设备和终端数量将超过 500 亿，这些设备和终端将会产生超过 80 ZB 的数据量。以云计算为代表的集中式处理模式虽然可以处理如此庞大的数据，但它在能耗、带宽、安全性、隐私性、实时性和智能性等方面存在不足。比如，远程医疗需要毫秒级的传输速度，从端到云距离太远，技术上难以实现。因此，云端处理能力将会下沉到更加贴近数据源头的地方，也就是网络的边缘。

3.2 边缘计算与云边协同

3.2.1 边缘计算

边缘计算（Edge Computing，EC）应运而生，它指的是在网络边缘执行数据存储、处理和程序运行等功能的一种分布式计算形式。据 IDC 预计，未来超过 50% 的数据需要在网络边缘侧进行分析处理和存储。相比于传统云计算，边缘计算需要更少的网络流量、更低的维护成本，能够对数据进行更快速的处理，满足低时延的需求。此外，由于大部分数据保存在网络边缘侧，降低了隐私泄露的风险。物联网、边缘计算和云计算的关系如图 3-2 所示。

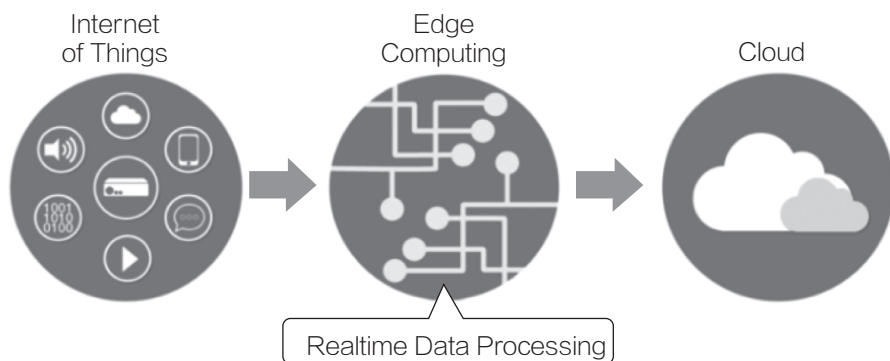
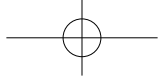
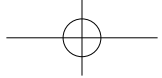


图 3-2 物联网、边缘计算和云计算的关系

边缘计算的应用场景触及社会生活的方方面面，包括能源、交通、农业、智慧社区、智能家居和工业制造等。根据不同行业的特点以及它们对时延、带宽和安全性的需求，边缘计算能够为各类垂直行业构建出针对性的解决方案。在城市交通管理中，用来监测交通路况、违章车辆等的高清摄像头遍布各个路口，如果在边缘服务器上运行智能交通控制系统对这些数据进行路况的实时分析和交通信息灯的控制，可以有效缓解车辆拥堵的状况。在智能家居领域，智能控制系统需要使用大量的物联网设备来监控整个家庭的内部状态（如安防系统、湿温传感器、照明系统、窗帘控制系统，信息家电等），完成对家居环境的调控和执行家庭人员的指令等。家庭数据往往涉及隐私，比如室内的相关视频，用户并不愿意将它们上传到云端处理。使用边缘计算可以防止数据泄露，提升系统的安全性。在游戏领域，尤其是在增强现实（Augmented Reality，AR）、虚拟现实（Virtual Reality，VR）场景中，边缘计算可以提供足够的存储资源，并在用户侧进行音频和视频的渲染，而不用将大量的数据上传到云端处理，降低了带宽需求，也减少了时延。

受边缘计算的巨大潜力驱动，各大型互联网、云计算、电信、芯片和系统集成等企业，纷纷推出了各自的边缘计算解决方案。它们中既有商业化的，也有开源的。前者包括 G 公司的 Google Cloud IOT、M 公司的 Azure IoT Edge、A 公司的 Lambda@Edge、GE 公司的 GE Predix、C 公司的 IOx、ARM 公司的 Mbed Edge 等；后者主要是 Linux 基金会的 EdgeX Foundry 和开放网络基金会（ONF）的 CORD 平台等。在国内，各大互联网公司、通信设备提供商、电信运营商也在布局边缘计算，H 公司和 Z 公司都有完整的移动边缘计算解决方



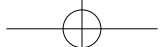
案，中国移动成立了边缘计算开放实验室，B公司也开源了自己的边缘计算框架 OpenEdge 等。从 Gartner 2019 年发布的技术成熟度曲线上看，目前边缘计算正处于技术研究的热点区，在未来具有不可估量的市场潜力。

3.2.2 云边协同

边缘计算的发展必须依赖云边协同。物联网的发展有两大支柱：一个是云计算；另一个是边缘计算。在互联网时代，计算都是在本地计算机上进行的。后来物联网兴起，计算逐渐迁移到云端。具体运作过程是：物联网设备产生和收集的数据发送到云端数据中心，云端运行计算模型并把结果返回终端。这种集中化的处理和管理模式受到市场的青睐，互联网和软件巨头纷纷推出了自己的云服务，比如 A 公司在 2006 年就推出了 AWS 云服务，之后，M 公司、G 公司和 A 公司等也分别推出了自建的云服务和产品等。如今，几乎所有规模稍大的企业都在使用云计算服务。根据面向的服务对象不同，云服务分为 3 种模式，即公有云、私有云和混合云。其中公有云是在开放的公有网络中提供服务，价格低廉，无须客户自己维护，但是安全性不高。私有云是为客户单独构建的，满足隐私和安全的要求，但是安装和维护成本较高。混合云是二者的结合，即部分使用公有云，部分使用私有云。

数据上传至云端必然会带来一定的延时，在对时效性要求不高的场合，云计算模式不会带来问题。但有些时候，比如自动驾驶中的路况分析、远程医疗和 AR/VR 游戏场景渲染等，稍微长一点的延时就会造成难以挽回的损失或者糟糕的体验。在数据量特别大（比如涉及音频、视频等）的时候，数据上传至云端需要足够并且可靠的带宽，这也限制了云计算的使用。另外，在云计算模式下，数据的存储和安全完全由云提供商负责。随着人们对数据隐私的日益关注，云端的安全问题也引起了人们的忧虑。在这样的背景下，让计算重回数据本地成为了新的趋势。所谓边缘计算，就是在靠近数据端的地方进行数据的处理、分析、存储和计算。因为不再需要把数据上传到云端处理，边缘计算释放了数据中心的计算压力，减少了数据传输带来的延时，可以提升用户体验。

物联网设备产生的数据通常是总量大但细小（即单个设备产生的数据量小），并且周期短，需要实时响应。如果数据都上传到云端处理，会对云端造成巨大压力。在未来，物联网产生的数据大约有一半的量会转移到边缘侧处理。



边缘计算节点可以负责所属范围的数据存储和计算，分担云端压力。同时，对于非一次性数据，边缘处理后仍需汇集到云端做大数据挖掘和进行算法模型的训练升级，之后再推送到边缘侧，升级更新前端设备，形成自主学习闭环。同时，数据备份在云端也能防止边缘出现问题时造成数据丢失。也就是说，边缘计算和云计算之间并不是竞争和取代的关系，而是协作的关系。边缘计算可以提供无所不在且可靠的本地服务，这些服务可以以不同的方式集成到云环境中。物联网场景下的云边协同过程如图 3-3 所示。

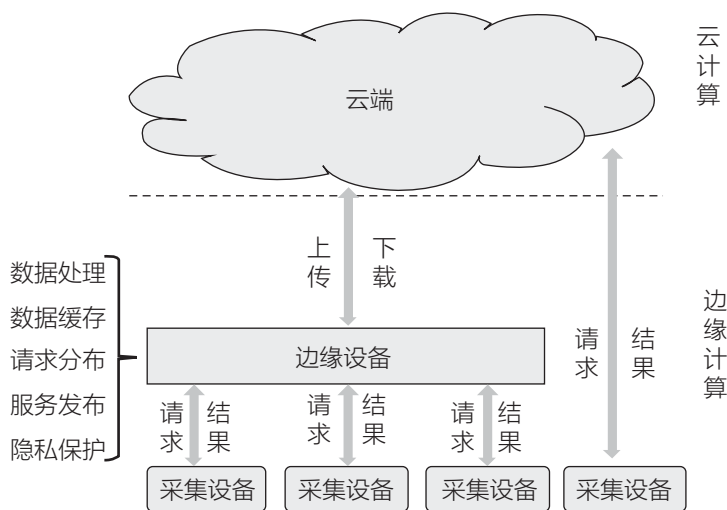


图 3-3 物联网场景下的云边协同过程

2018 年，H 公司在其全连接大会（HC2018）上发布了智能边缘平台 IEF（Intelligent Edge Fabric），明确提出了云边协同一体化服务的概念。同年的 S 公司提出了工业边缘概念，通过在云端部署工业边缘管理服务实现云边的协同。A 公司也明确提出打造了“云 + 边 + 端”三位一体战略。如图 3-4 所示，云边协同的内涵涉及 3 层次：云边 IaaS 协同、云边 PaaS 协同和云边 SaaS 协同。具体而言，IaaS 协同可实现网络、虚拟化资源和安全等方面的协同；PaaS 的协同可实现数据、智能（AI 模型推理）、应用管理和业务编排等的方面协同；SaaS 的协同可实现服务的协同。云边的全面协同可以提供连续的，无处不在的 ICT 资源服务，使计算和决策可以随时随地进行。

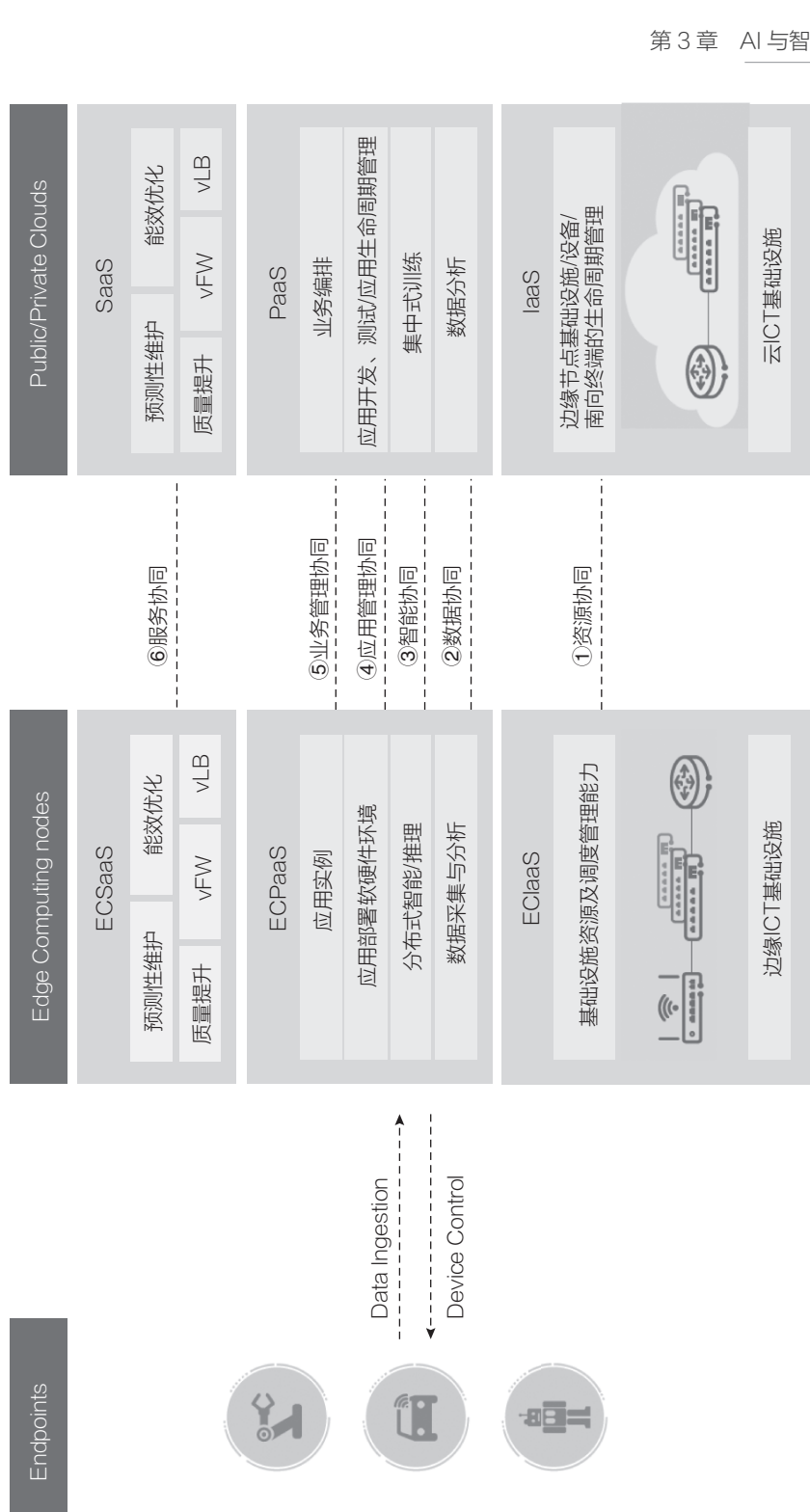
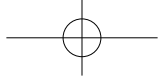


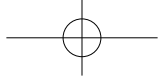
图 3-4 云边协同内涵



随着 5G 和物联网技术的普及，万物互联的时代已经来临。越来越多的应用如智能制造、智慧医疗和智慧城市等需要在边缘侧（数据采集端）部署计算设备和网关设备，这些设备配合分布式数据库构成了一个完整的边缘计算体系，再与云计算中心进行数据和应用的互动。这种方式既降低了数据传输的带宽成本，减少了延时，对边缘设备来说，也降低了能耗，增加了使用寿命。边缘计算本质上是一种去中心化的分布式计算模型，具有低功耗、低时延、保护隐私、支持本地交互、自主独立运行的优势，被 Gartner 列为 2019 年十大战略技术之一。但是，无论是边缘设备还是云计算中心，要处理海量的物联网数据都需要 AI 技术的加持。这里面既包括 AI 芯片，又包括 AI 算法。

AI 芯片是对人工智能算法，尤其是深度学习算法里大量的矩阵、卷积、积分等运算，做了特殊加速处理的芯片，主要用来处理人工智能应用中有大量计算任务的模块。AI 芯片可以分成两类：一类是同时面向模型训练（Training）和推断（Inference）的；另一类是只面向推断的加速芯片，只处理模型运行的问题。前者的代表是 G 公司的 TPU，后者的参与方比较多，比如 C 公司的 NPU、H 公司的 BPU、X 公司的 DPU 等。AI 芯片能够极大地提高人工智能算法的效率，降低能耗。在云端市场，多数厂家采用的是“GPU + CPU”“FPGA + CPU”方案，G 公司使用的是自研的“TPU+CPU”方案。云端市场上 N 公司一家独大，它主推的 GPU 单卡或多卡集成主机方案覆盖不同算力和数据结构的需求。此外，国内也有一些企业加入了云端训练芯片的竞争，比如 A 公司的含光、H 公司的昇腾系列、Y 公司的 QuestCore 等。近年来，AI 芯片的市场规模持续扩大，2017 年的全球 AI 芯片规模为 62.7 亿美元，到 2020 年，这个数字将达到 160 亿美元。

深度学习算法模型通常有几百上千万甚至上亿的参数量，模型的训练和推理需要消耗大量的 GPU 资源。在边缘侧，终端设备的能量、存储资源和算力都有限，难以运行深度学习算法模型。为了让 AI 模型能够部署在边缘侧，需要有专门的 GPU 和轻量级的算法模型。目前，A 公司的 Bionic 处理器、Q 公司的骁龙处理器都加入了神经网络引擎。在国内，H 公司的麒麟 970 也已经搭载了神经网络处理元件。针对移动端，主流的深度学习框架都推出了精简版本，比如 Tensorflow 的精简版本 Tensorflow Lite、Caffe2 和 Pytorch 1.3。算法层面，G 公司提出了深度可分离卷积（Depthwise Separable Convolution）用于移动端和嵌入式设备，并据此推出了 MoblieNet V1、MoblieNet V2、MoblieNet V3 模型。



之后，G公司又提出了EfficientNet，不仅模型的参量大幅减少，模型准确率也得到了提高。轻量化网络已经是当前深度学习的研究热点之一，研究者通过知识蒸馏、剪枝等技术来压缩模型，在保证准确率的同时减少所需要的算力和存储资源。

在万物互联时代，云边智能协同将为传统应用带来变革，并衍生出新的应用场景。比如身份认证，利用人脸识别技术和边缘设备的算力，可以为车站、机场和金融机构等提供高效的用戶身份认证手段，提升安全性。在智慧交通领域，越来越多的城市正在部署智能交通系统，图3-5展现了基于云边协同智能的智慧交通系统，利用部署在各交通要塞的边缘计算服务器和遍布路口的高清摄像头、传感器等，结合人工智能算法，可以对交通状况进行实时监控、预警和调度。在加拿大多伦多市，Alphabet旗下智慧城市子公司Sidewalk Labs推出了智慧社区项目。该项目计划在加拿大滨海地区打造成包含自动驾驶、办公、医疗、居住、零售、制造等一系列解决方案的高科技城市。在未来，云边智能协同将进一步改变人类的衣食住行。

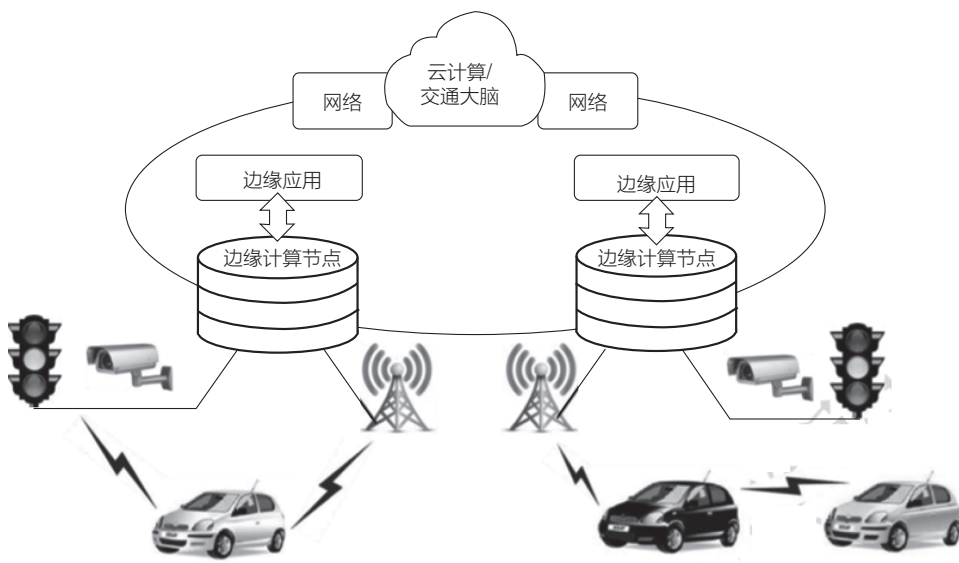
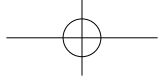


图 3-5 基于云边协同智能的智慧交通系统



3.3 应用于智能 IoT 中的 AI 技术

3.3.1 联邦迁移学习

随着计算机技术的快速发展,越来越多的可穿戴设备进入人们的日常生活,比如 A 公司的智能手表 iWatch、G 公司推出的智能眼镜 Google Glass、X 公司推出的智能手环、N 公司推出的智能运动鞋等。这些设备可以很轻易地获取人们的健康信息,比如活动、睡眠、运动等。利用可穿戴设备获取的数据,可以训练出检测人们健康状态的机器学习模型。但是这会面临两个挑战。首先,在现实生活里,可穿戴设备的数据经常以孤岛的形式存在。尽管不同公司、组织和机构拥有大量的数据,但受隐私和安全方面考量,这些数据不可能共享。目前,中国、美国、欧盟都颁布了严格的数据隐私保护条例,因此不可能获得大量的用户数据来构建模型。另一个是个性化的问题,不同的用户有不同的身体条件和活动规律,用统一的云端训练的模型无法满足用户的个性化需求。来自中科院计算所、M 公司和深圳鹏城实验室的团队提出了一种联邦迁移学习框架 FedHealth,它用联邦学习的方式来处理数据,再用迁移学习来构建个性化的模型。FedHealth 能在保证安全和隐私的前提下向用户提供准确、个性化的医疗保健服务。FedHealth 的整体架构如图 3-6 所示。

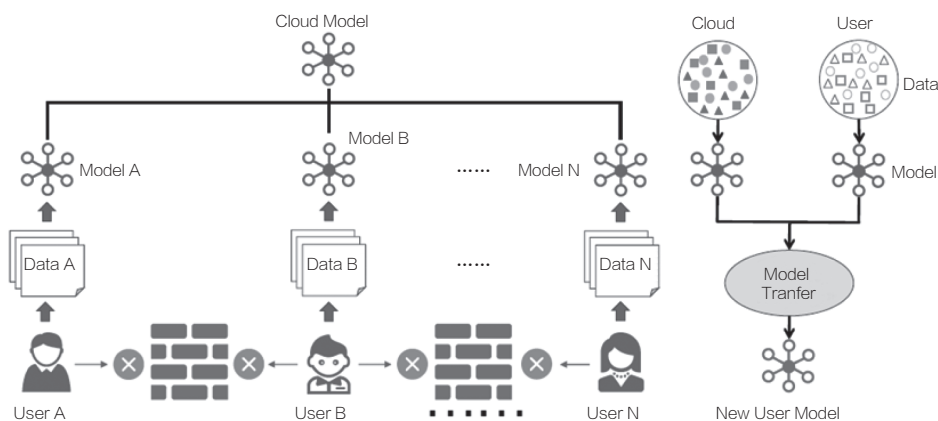
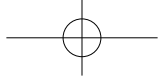


图 3-6 FedHealth 的整体架构



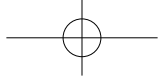
具体的思路是：首先，利用公共的数据训练出一个云端的模型，并用同态加密的方式分发给所有用户；然后用户使用自己的数据训练出一个模型，并用同态加密的方式上传至云端；接着云端通过对齐用户模型更新云端模型，并分发给所有用户；最后用户使用迁移学习得到个性化的模型。因为用户的数据不断产生，以上步骤也是持续进行的。在这里，云端模型保证了模型的通用性，而用户模型兼顾了个性化。如果直接使用云端模型，对于特定的用户，效果可能不太好，因为云端数据的分布和用户的数据分布是有差别的。联邦学习解决了数据孤岛的问题，迁移学习解决了个性化的问题，二者结合在准确性上也不弱于传统的使用公有模型的方法。

为了验证 FedHealth 方法的效果，研究团队使用了一个称为 UCI 智能手机的公开的人类活动识别数据集进行实验。这个数据集包含 30 个年龄在 19~48 岁的用户的 6 种类型的活动，分别是行走、向上行走、向下行走、坐、站立、躺下。总共有 10299 个样本，其中 70% 作为训练集，30% 作为验证集。为了适应 FedHealth 框架，使用 25 个用户的数据作为公有数据训练云端模型，剩下 5 个用户的数据独立构建个性化的模型。表 3-2 给出了测试集上的实验结果，P1 至 P5 表示 5 个独立的用户。作为对比，传统的 K-近邻算法 (KNN)、随机森林 (RF) 和支持向量机 (SVM) 的结果也列在了表中。模型的参数通过交叉验证的方法确定，为了消除随机性，所有的实验都独立进行 5 次取平均结果。

表 3-2 FedHealth 方法测试集上的分类准确率

Subject	KNN	SVM	RF	NoFed	FedHealth
P1	83.8	81.9	87.5	94.5	98.8
P2	86.5	96.9	93.3	94.5	98.8
P3	92.2	97.2	88.9	93.4	100.0
P4	83.1	95.9	91.0	95.5	99.4
P5	90.5	98.6	91.6	92.6	100.0
AVG	87.2	94.1	90.5	94.1	99.4

在这个实验中 FedHealth 使用模型是卷积神经网络，如图 3-7 所示。网络包含 2 个卷积层、2 个池化层和 3 个全连接层，在迁移学习过程中，卷积层和池化层冻结，全连接层对齐微调训练。表 3-2 中 NoFed 模型使用的也是这个网络，不过只用到了云端模型，没有用联邦学习和迁移学习。结果表明，FedHealth



的分类效果最好，平均准确率达到了 99.4%，相比于 NoFed 模型，FedHealth 提高了 5.3%。同时可以看出，对于人类活动的识别，深度学习方法（FedHealth 和 NoFed）要好于传统模型（KNN、SVM 和 RF），这得益于深度神经网络强大的表征能力，而传统方法依赖于手工构造的特征来学习。深度学习的另一个优点是可以在线增量更新，不必重新训练。而传统的方法需要进一步的增量算法。

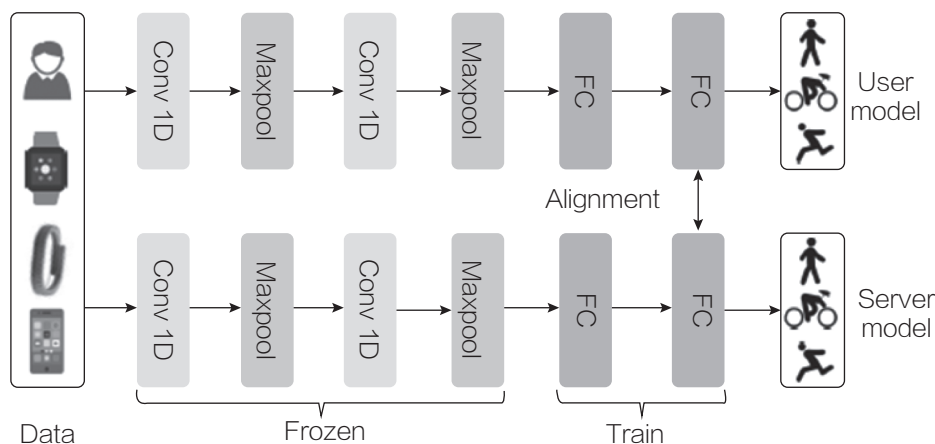
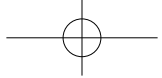


图 3-7 FedHealth 中的迁移学习过程

3.3.2 RPnet网络与车牌识别

车牌的识别和检测（License Plate Detection and Recognition, LPDR）是智能交通系统的关键功能，被广泛用于各种交通场景，比如交通路口监控、高速公路收费站和停车场的出入管理等。虽然车牌的检测识别算法层出不穷，越来越准确，然而对于复杂现实场景，挑战依然存在。比如雾雪天气、扭曲、不均匀光照、倾斜等，如图 3-8 所示，在这样的情况下，人眼都很难分辨出车牌号。车牌识别模型的实际效果不理想的另一个原因是数据集太小。目前公开的 LPDR 数据库如 AOLPE、Azam、SSIG、ReID、UFPR 等，受限于人工采集车牌图片的人力成本，既缺乏数量（少于 1 万张），又缺乏种类（都是在固定监控摄像头采集）。为了激励更优秀的 LPDR 算法出现，中国科技大学的团队公开了一个中国城市泊车的数据库（CCPD）。这些图片由手持设备拍摄，对



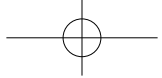
应的车牌号是收费员手动输入的。CCPD 每个图像都有详细的注解，包括车牌编号、车牌边界框（Bounding Box）、四个顶点位置、水平倾角和竖直倾角，以及其他相关信息，如车牌区域、光照程度、模糊程度等，具有很高的质量。



图 3-8 复杂条件下车牌

通常 LPDR 被拆分为检测、识别两个阶段或检测、分割与字符识别 3 个阶段，将车牌图像分步骤处理。然而，从识别中分离检测不利于整个识别过程的准确性和效率。一个低质量的预测包围框（Bounding Box）可能会导致车牌的部分缺失，进而造成后续识别的错误。中国科学技术大学的团队在车牌识别阶段利用车牌检测阶段中提取的卷积特征来识别车牌字符，并据此设计了一个命名为路边停车网络（RPnet）的全新架构，仅用一次前向通路就能完成车牌的检测与识别。RPnet 统一了车牌检测与识别，可以进行端到端的训练。因为检测和识别任务共享了特征图，联合优化了损失函数，所以 RPnet 可以更快、更准地检测和识别车牌。即使在较低端的 NVIDIA Quadro P4000 上，RPnet 的速度运行也可以超过 60 fps。

RPnet 的模型结构由两个模块组成，第一个模块命名为“检测模块”，是一个 10 层深的卷积神经网络，用来提取车牌图像中不同层级的特征图，然后将特征图输出到 3 个平行的全连接层中（称为“框预测器”，用于检测边框）。第二个模块为“识别模块”，利用感兴趣区域（ROI）汇聚层来提取感兴趣的特征图，然后通过若干分类器来预测区域内的车牌号码。在检测模块中，随着层数的增加，特征图逐渐变小，提取的特征层级也越高，对车牌的识别和边框的预测越有利。模型的损失函数分为定位损失和分类损失，联合优化这两部分损失可以提取更丰富的车牌字符信息，提高检测和识别性能。RPnet 的识别精度可以达到 95.5%。



3.3.3 对抗生成网络与移动目标检测

完整运动物体检测是许多计算机视觉任务的基础，比如人类活动分析，智慧城市的交通监控、安防等。然而，在存在伪装、噪声、照明条件变化、背景动态变化等情况时，运动物体检测依然具有挑战性。为了解决这些问题，基于鲁棒主成分分析（Robust Principal Components Analysis, RPCA）的算法被用来把数据矩阵分解成低秩的背景部分和稀疏的运动部分。这些算法在过去十几年里表现良好，许多基于 RPCA 的算法已经被采用并成功用于运动物体检测。然而，在稀疏部分，运动物体的每个元素都是单独处理的，这会导致运动物体所在的区域不完整。此外，在存在动态背景时，稀疏分量往往是分散的。另外，当运动物体被伪装时，基于 RPCA 的算法性能会下降。利用深度神经网络可以解决基于 RPCA 方法的局限性，因为鲁棒的深度自编码能够实现无遮挡的完整运动目标检测。

为了估计无遮挡的完整运动目标，需要一个具有有效编码特性的判别模型和一个生成模型来重建运动目标中遮挡的缺失信息。来自韩国庆北国立大学、巴基斯坦拉合尔信息技术大学和法国拉罗谢尔大学拉罗谢尔 MIA 实验室的研究人员提出了一个基于条件生成对抗网络（cGAN）的解决方案，该方案采用一种基于无遮挡运动目标检测场景训练的鲁棒深度自动编码器结构。他们在不同挑战环境下训练出了一个完全的 cGAN 模型，称为 CcGAN。CcGAN 的无遮挡训练是通过动态背景信息覆盖场景的所有语义信息，完成与 RPCA 机制有一定相似性的运动目标检测来学习跨空间的子空间。如图 3-9 所示是运动的车的检测场景，图中的树也是在动的。DCP 和 DECOLOR 都是基于 RPCA 的算法，从图中可知，基于 RPCA 的方法把动态的树也包含了进来，而车是不完整的。但 CcGAN 却能把运动的背景去掉了，而且重建出完整的运动的车。

CcGAN 模型的结构如图 3-10 所示，它包含了一个生成器和一个判别器。判别器的输入是 256 像素 × 256 像素 × 3 像素的大小，包含 4 个下采样卷积层，第一层有 64 个通道，最后一层有 512 个通道，输出大小为 30 × 30。生成器是一个带有残差连接的 Unet 编码—解码网络，包含了 4 个下采样层，每个下采样层之前都有一个卷积层。解码器具有 4 个上采样层，然后进行反卷积操作，以重建与输入图像。生成器总共有 8 个卷积层，编码器的最后一层产生 512 个特征图，解码器的最后一层产生 64 个特征图。在训练阶段，生成器和判别器共同训练。在测试阶段，只需使用编码器来检测运动物体。

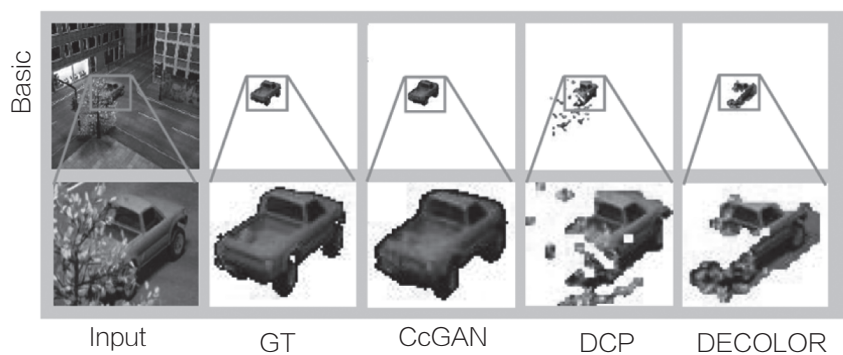


图 3-9 基于 SABS 基准数据集的 CcGAN 完全运动目标检测可视化演示

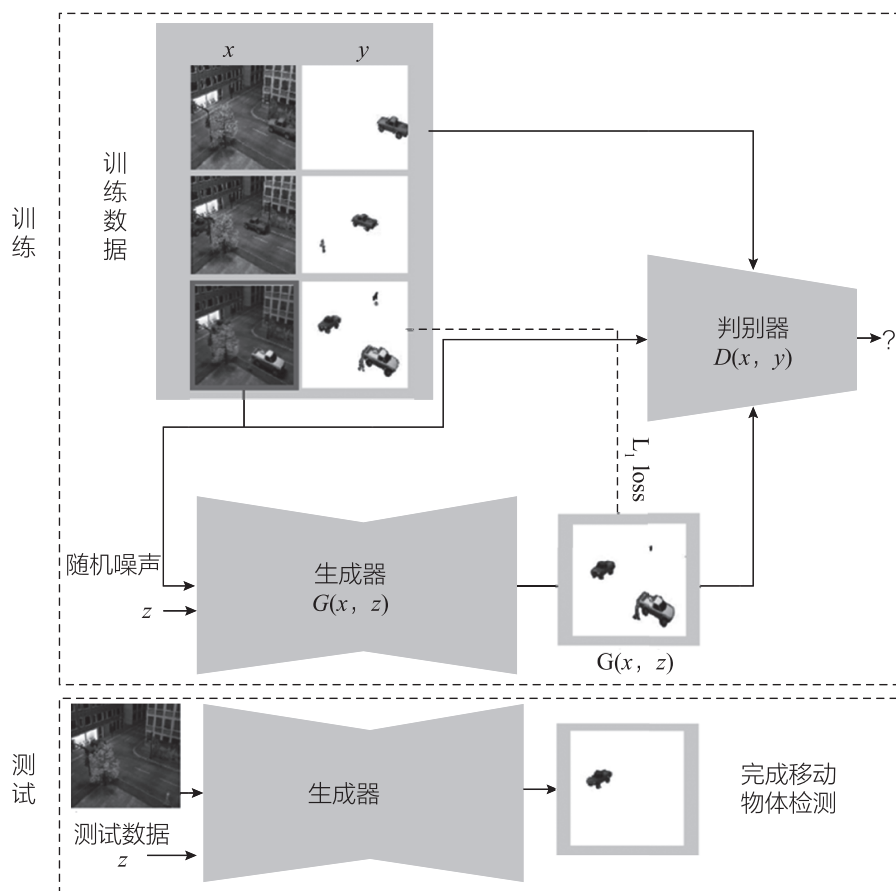
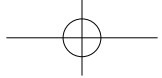


图 3-10 CcGAN 模型的训练和测试过程



3.3.4 Android手机去中心化的分布式机器学习

机器学习一般是在单机或者集群上进行数据处理和模型训练，G 公司已经建立了世界上安全性和鲁棒性都顶级的云基础平台设施来为机器学习提供云服务。为了在移动设备上基于用户交互来训练模型，G 公司在 2017 年推出了联邦学习。联邦学习使得在移动设备上协同训练一个共享模型成为可能，训练数据全部在移动端存储，使得机器学习的模型训练和云端存储能够解耦。解决了以往模型只能云端下发训练好的模型，而无法在本地训练的问题。联邦学习的工作原理如图 3-11 所示：用户先用手机下载云端最新的共享模型，然后用用户手机上的历史数据来微调这个模型，之后将用户个性化的模型抽取为一个小的更新文件，最后仅将模型的差异部分使用加密方式上传到云端，云端整合其他用户上传的模型差异做平均化的更新，以改善共享模型。这样做的好处是，所有的训练数据都在用户的手机上，并不发送隐私信息到云端，发送的仅仅是模型的变更部分。

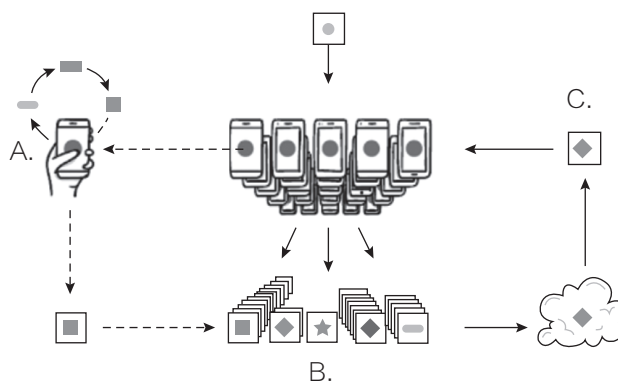
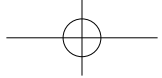


图 3-11 联邦学习示意图

联邦学习能够训练出更好的模型，而且延时低、功耗小，同时保护了用户的隐私。它除了能够不断优化云端共享模型外，还能够更新用户本地模型。因为本地模型是根据用户历史数据训练的，它具有个性化的特征。目前，谷歌已经在 Android 设备上的多个任务部署了联邦学习系统。比如 Google Pixel 上的搜索设置排序，可以免去不断服务器查询，并且搜索内容和用户选择等隐私信息都是保存在本地，安全性更高。另一个应用是对用户在移动设备上的键盘输入内容提供建议，根据用户的特征进行建议项目的排名，提升用户的体验。G



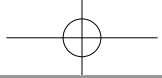
公司移动键盘团队还使用联邦学习平台训练递归神经网络（RNN）用于下一次预测。整个模型大约有 140 万个参数，使用了 150 万用户的 6 亿个句子迭代训练了 5 天，该方法把基线 n-gram 模型的召回率从 13.0% 提高到了 16.4%。

3.3.5 “AI + 移动警务”

近年来，伴随着我国城市经济的快速发展，城市人口和机动车辆基数增长加速，给执法环境带来重大挑战。而传统的移动警务无论是从产品性能还是专业应用上，都很难满足当前移动警务的执法需求。在警务信息化建设不断深入的新形势下，全国各级公安机关已经陆续开展“AI + 移动警务”新模式，以提高城市一线警务工作人员的实战水平和核心战斗力。

“AI + 移动警务”主要是由多元化的前端移动警务设备、系统软件和应用软件构成的。前端移动设备主要包括移动警务通、单兵执法记录仪、移动执法车载系统等。应用软件以客户端 APP 呈现，不同的警种可根据实际业务下载相应的 APP 应用，而系统软件则对接公安部人口、出入境、车辆、在逃、盗抢等大型资料库。移动警务终端系统可以覆盖到社区警务、治安视频监控、巡逻防控、车辆核查等领域。在“AI + 移动警务”模式中，人工智能的能力主要体现在对人、车、物进行快速识别和比对分析，直接在前端实现视频数据的迅速结构化处理，为警务人员的行动决策提供参考依据。在“AI + 移动警务”模式下，警务人员可以利用移动警务终端产品实现人车定位、分析研判、指挥调度等警务工作，并随时随地获得公安业务信息的支持并实现实时比对。也可以现场采集信息及时回传公安内部信息中心，有效解决通缉、协查、堵截、搜查、处罚等一线公安工作中对嫌疑人的照片识别问题。

H 公司在“AI + 移动警务”方面有积极的实践。具体来说，就是把 AI 能力前置到移动终端上，结合云端的数据分析处理能力和业务流程优化，为警务人员构建一套高效、实用的系统。比如 H 公司的“深眼”移动警务系统，实现了基于车载行车记录仪的前端车牌识别、黑名单比对以及警情接收、互动等功能，带来了缉查、布控、报警以及出警、拦截、勤务模式的变革，提升了报警、拦截效率，在速度上，要比传统移动警务快 5 ~ 10 倍。此外，对于难以达到城市级的全域覆盖的固定卡口，移动警务可充当移动的卡口，弥补固定点位的覆盖盲区。面对城市精细化治理的迫切需求，“AI + 移动警务”把固定设施与移动警务动静结合，能够有效解决警务工作难的问题。



第5章

AI 与客户体验管理

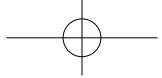
5G 时代，AI 技术不仅可以为网络智能切片、物联网和运营商 5G 网络的多量纲计费等业务场景进行注智赋能，同样，在客户服务、客户管理、业务流程管理、商业智能分析（BI）等业务领域也有着重要的应用前景。

随着越来越多的企业通过采用人工智能来提升客户体验，客户体验领导者和从业人员开始认识到，人工智能具备改变客户体验本质的潜力，覆盖了从客户体验战略到设计内容和设计方式的各个环节。企业管理者意识到 AI 技术在提升客户体验上可以使企业获得更高的客户评价和收益。人工智能系统能够尝试模仿人类思维的方式理解非结构化信息。除此以外，还能够以更快的速度利用大量数据，从交互中学习，累积学习的能力。由于人工智能系统已经具备看、说、听等能力，这将帮助现有的客户体验团队进入了一个全新的时代。

本章从客户网络感知与运营商实际网络 KPI 指标的差异角度切入，引入了客户体验管理（CEM）的理念，并引入了具体的业务应用案例，读者可通过本章内容阅读，详细阐述 AI 技术如何为客户的体验管理提供注智赋能。

5.1 客户感知网络质量与客观 KPI 指标差异

5G 网络承诺有极快的上网速率，更低的延迟以及可实现大量的机器对机器的通信交流，但这些是否真的就能够保障消费者可以愉快地使用 5G 服务呢？为了保障在各种各样新场景中的服务质量（Quality of Service, QoS），5G 服务使用了端到端的网络切片架构。这样人们就可以实现创建虚拟网络，



用来接收动态分配的端到端虚拟资源，从而满足特定服务的性能目标。这样一来，一些紧急的服务可以相对于其他没那么关键的服务在资源受限的情况具有更高的优先权来接收网络切片。理论上来说，这样应该可以保证更高的QoS。QoS是实现用户QoE的技术面因素之一。QoS针对一项具体的业务，由特定业务的功能、性能、可用性、可操作性、安全性及其他因素来描述。用于实现QoS保证的参数有优先级、延迟、延迟抖动、包丢失率、吞吐量、承载可靠性、帧误码率、业务响应时间、包重排序、会话阻塞率、会话可用性、会话持续性和会话接入时间等。如图5-1所示为网络性能KPI指标、QoS、QoE的关系。

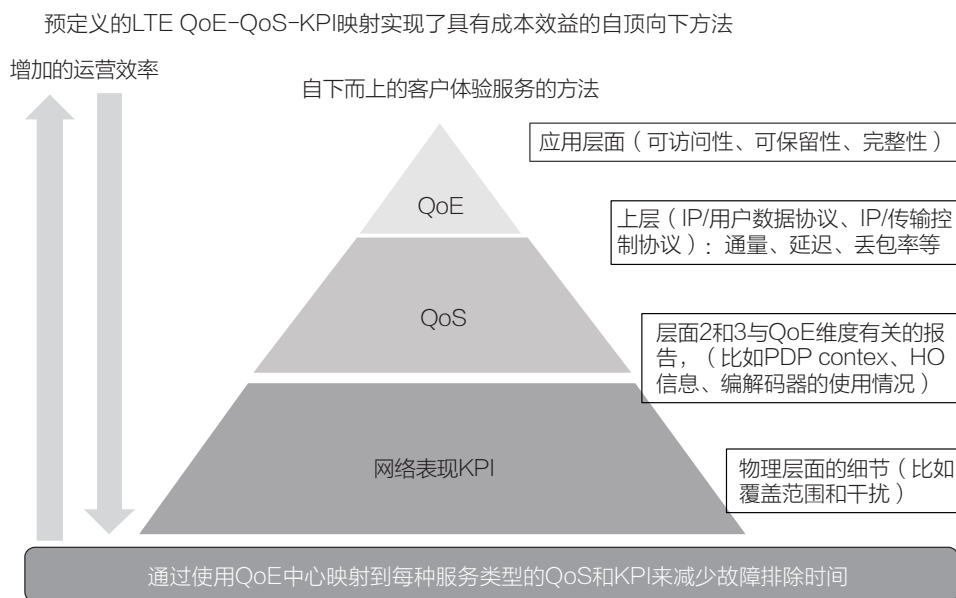
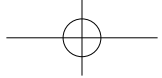


图5-1 网络性能KPI指标、QoS、QoE的关系

但是要注意，这里可能会存在一个问题，在网络评估中一般通过监测各种关键性能指标 (Key Performance Indicator, KPI) 来计算 QoS，而这些 KPI 指标反映了关于网络质量参数的监测。而这些 KPI 指标并不能保证用户端感知的服务质量 (Quality of Experience, QoE)，QoE 或称客户体验质量是定义在用户感受层面上的。好比你对家中的自来水管道做了各种检修，最后能够保证水管不漏水，但你却不能以此来保证得到管道中的水很好喝的结论。通信服务在提供商日常的网络运维过程中时常会发现，传统的网络 KPI



很难完全反映终端用户的感受。传统的网络规划方法通常是从测试到网管记录的网络性能指标（KPI）分析形成规划方案，最后进行资源投放；这样会造成 KPI 指标好，客户感知却不理想的情况，且无法获得解决，甚至无从知晓问题在哪里。客户感知的网络质量与客观上使用的网络 KPI 指标的不一致性，举例说明有：

- ✓ 流媒体电影可能满足某些KPI指标，但是客户感知的QoE可能还会依赖于他/她使用的是电视观看或者是小型的移动设备。
- ✓ 只专注于平均KPI指标的通信服务提供商可能错失对一些高级游戏玩家进行个性化服务微调的机会。
- ✓ 整体网络的KPI指标表现良好，但是单个用户的投诉却依然很多甚至有上升趋势。或者说，KPI指标提升了，但客户端的网络服务感知却没有明显的提升。

在很多情况下，资源总会不可避免地变得受限，在一个网络切片或者不同切片之间，一些服务之间会出现资源争夺的现象，网络切片技术将不足以解决所有出现的问题。网络 KPI 指标对于服务、应用、用户、设备以及地理位置这些概念不能实现更细粒度的感知，而这些信息对于更精确地计算和优化终端用户的 QoE 至关重要。

下面我们举一个具体的例子来说明这种 KPI 指标和用户感知的不一致性。对于一个专用网络来说，比如无线电接入网（RAN）、核心网（CN）或者承载网络，通常具有自己专用的元素管理系统（Element Management System, EMS）来确保网络的可用性以及 QoS。传统意义上的针对网络运营与维护不一定能意味着较好的用户感知网络质量。图 5-2 和图 5-3 分别展示了很高网络 KPI 指标的网络系统以及较差的 QoE。

图 5-2 中，CN、RAN 和 IP 承载网的 KPI 指标都非常高，这意味着网络的质量相当好；然而，呼叫质量却仍然很差。这可能是因为呼叫在异常的单元中会掉线。在这个单元中的用户可能会体验到非常差劲的服务，但是这种低品质单元可能被其他高 KPI 的单元所掩盖，如图 5-3 所示。因此这告诉我们网络的 KPI 指标并不一定反映出端对端的服务质量，并且不应该作为改进服务质量的唯一基准来使用。

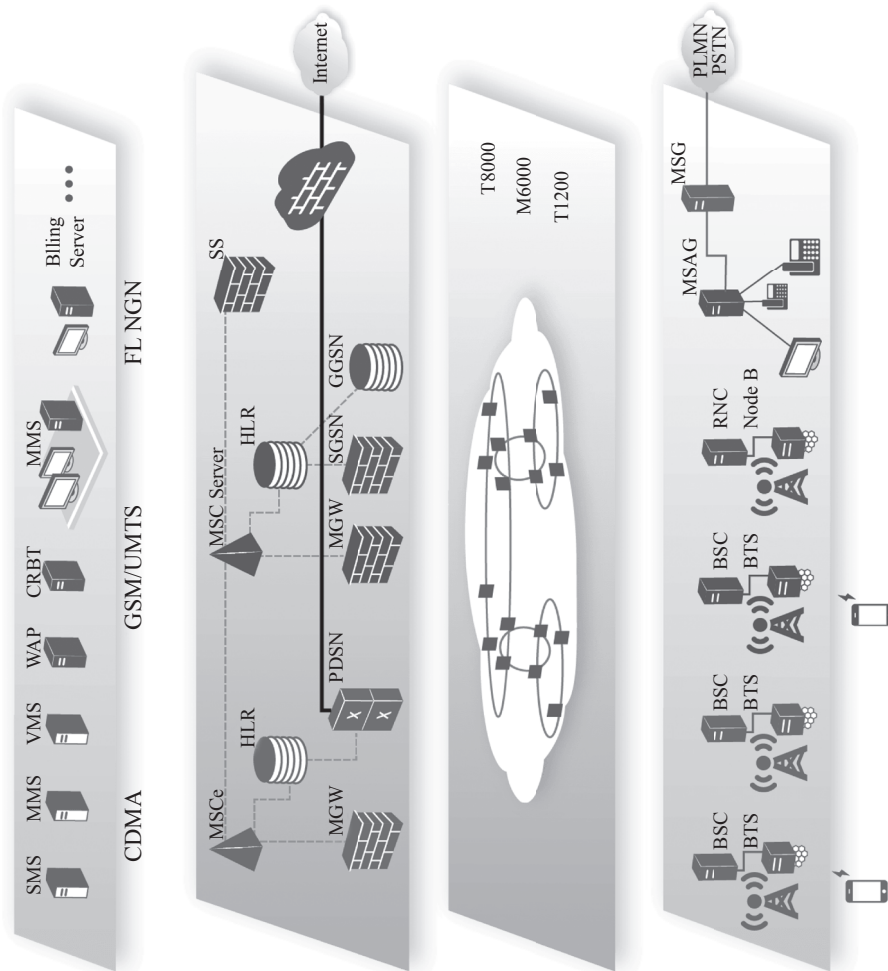


图 5-2 良好的网络 KPI 表现

CN KPI	
Through Connection Ratio	Over 98%
Successful Location Update Ratio	Over 95%
Successful Handover Ratio	Over 97%
Successful Voice Paging Ratio	Over 95%
Successful SMS MO/MT	Over 97%

IP Bearing KPI	
Packet Loss	<1%
Delay	<150ms
Jitter	<30ms

RAN KPI	
Call Setup Success Ratio	Over 98%
Handover Success Ratio	Over 97%
Call Drop Ratio	0.18%
TCH Availability	Over 97%
DL RO(93) Ratio	Over 96%

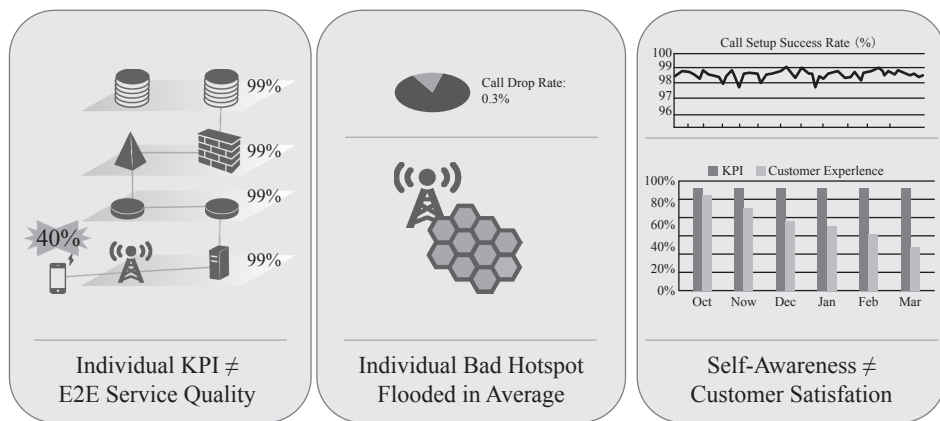
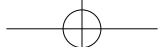


图 5-3 实际的用户端的感知质

5.2 CEM 概述

5.2.1 CEM基本概念

通常人们讨论客户体验（Customer Experience, CX）是指客户对自身与品牌之间关系的一种感受；它不仅具有直接意识上的感知，还包含潜意识中的感受，这些感受自然是客户在和品牌产品的互动接触中获得的。客户体验管理（Customer Experience Management, CEM）则是指在和客户交互的活动中设计并响应从而达到或者超出客户预期水平的一种实践，品牌商通过 CEM 实现较高的客户满意度、客户忠诚度以及客户拥护度。我们也可以认为 CEM 实际的目的在于消除公司期望实现的客户体验和实际发生的客户体验之间的差距，如图 5-4 所示。

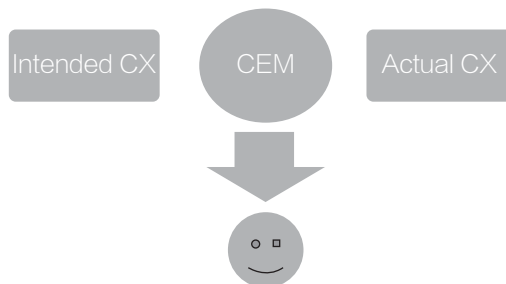
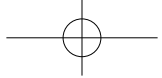


图 5-4 CEM 实现对用户真实体验管理



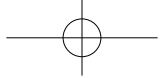
CEM 的概念听起来可能过于理想化或者难以捉摸，但可以肯定的是任何商业公司忽视了 CEM 都将会感受到明显的挫败感。实际上，CEM 在今日高度竞争化、广连接的全球市场环境中已经成为一个关键性的区分指标。麦肯锡分析指出聚焦于 CEM 的企业将会在两到三年内得到 5% ~ 10% 的收入增加以及 15% ~ 25% 的成本减少。但是在当下的社会环境中，实现超高的客户体验比以往任何时候都要困难。社交媒体平台的爆炸性发展以及巨大的成本压力迫使服务提供商们不得不寻求一些新的方式控制成本，并在可接受范围的情况下尽可能提升客户的体验感受。设计交付无缝对接的敏捷性系统，建立客户忠诚并减少客户流失，在所有的数字通道端利用数据来重新定义电信服务提供商和客户之间的关系。电信服务提供商必须面对着激增的用户数量以及所提出高质量的需求和更好的客户体验的挑战。增强用户服务体验，提升运营效率，控制交易成本。

客户的个人体验虽然因人而异，但必然存在很多共通的规律，正如心理学总可以将人的性格分类并指出描述不同群体的共通特性。电信服务提供商在与客户的交互活动中收集用户评价、对各种操作响应或者是问题解决方案中的互动，可以逐步获得客户对产品服务的感知数据，对客户进行分类、归因，从数据中感知客户需求，逐步实现客户保留的目的。

客户的满意度意味着必须能够迅速响应用户所提出的问题给出正确的信息反馈。在电信行业中，维持客户的满意度对于保留客户和减少用户流失是至关重要的。有调查表明，大约有 51% 的美国用户曾因为较差的用户体验转变了自己的服务提供商。只有 50% ~ 60% 的用户能够对自己的服务提供商感觉到满意。客户对服务体验的要求通常有：

- ✓ 快速的响应：约有 82% 的用户认为超快的问题解决速度是高级客户体验的首要原则。
- ✓ 56% 的电信用户使用自我服务选项来确定自己最合适的套餐计划；77% 使用自助服务来结账埋单或者是为账户充值。
- ✓ 个人化的体验：电信公司必须能提供吸引人的、个人化的体验，从而能够维持客户的忠诚度并得到他们的进一步支持。

在万物互联的时代，用户对于通信网络的诉求从以前的“处理更多的数据”已经逐步演化为“满足更多需求”。通信运营商之间的竞争固然可以通过降低资费来获得优势，但是真正意义上的网络服务质量提升才是获取客户并保持客



户拥护度的根本所在。因此关注用户的网络质量感知，可以帮助运营商降低投诉率，从客户终端的角度来提升体验满意度。

5.2.2 客户网络体验感知量化

我们通常可以认为用户对网络体验的感知，也称为 QoE，主要的决定性因素分为两个方面：一是技术面因素，端到端网络质量、网络 / 业务覆盖率和终端功能；二是非技术面因素，业务使用简易、业务内容、资费和客户支持。QoE 包含了所有参与网络通信过程中的各个方面，包括用户、运营商、内容提供商，或者应用提供商、设备制造商，或者系统集成商、终端设备和应用软件。

对用户感知的定量化，通常可以从以下 4 个方面来考量：

- ✓ 服务接入性质量，比如用户希望使用某应用服务，运营商应能尽快地保证服务接入，常见的如呼入成功率、建立呼叫时间。
- ✓ 服务完整性质量，即用户在使用应用服务过程的质量，比如语音、视频通话的质量。
- ✓ 服务保持性质量，它是用来描述一项应用服务的终止情况，比如掉线率。
- ✓ 网络覆盖性能，即最基础的用户能接受服务的覆盖性能。

从最基础的网络软硬件层面的基础性能出发，运营商可以抽取出关键性能指标 KPI，对数据进行分段特征性的量化提炼出业务相关的关键质量指标（Key Quality Indicator, KQI），最后通过所建立的模型（比如利用业务相关性和多样性决定的权重进行加权映射）计算得到 QoE 的一些核心指标，从而可以构建出一个全面、完整的客户感知的网络质量评估体系。

由网络端的 KPI 指标，可以进一步形成运营商业务的 KPI，再经由 KPI 映射到关键质量指标 KQI，KQI 是针对不同业务提出的贴近用户感受的业务质量参数。在非技术因素方面，可以将影响用户体验质量的非技术因素通过用户投诉来呈现，在进行聚合分类之后，根据各类投诉的内容及特点映射到技术因素方面相应的 KQI 上。由于 KQI 指标及用户投诉体现的只是针对某种业务的独立感知，与用户的真正体验仍存在一定差距，因此需要将各种 KQI 指标及用户投诉内容综合映射到用户体验质量 QoE。

如图 5-5 所示为一种 KQI 和 KPI 指标之间的关联示意图。在通常情况下，运营商可以从端到端业务实现的具体过程开始，根据经验提炼出影响关键服务质量指标 KQI 的网络影响因素，确定网络 KPI。

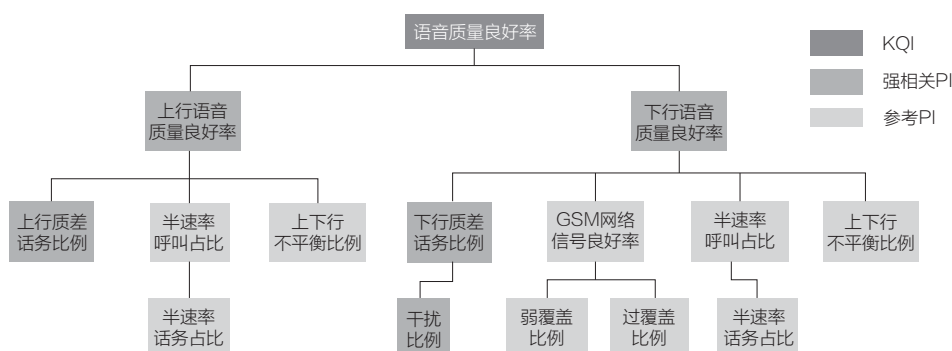


图 5-5 KQI 与 KPI 关联示意

图 5-6 是在通信领域中常用的指标分类方式，底层是原始数据参数，传输的是一些基本信息，比如带宽、抖动、网络等。第二层为 KPI，这些是在监控的时候抓取的指标，比如 bug 指标、延迟指标。再上层为 KQI，基于 KPI 生成的关键质量指标，还有 KPI 未涉及的音频、视频、用户交互等部分。最终这些综合起来就是 QoE。

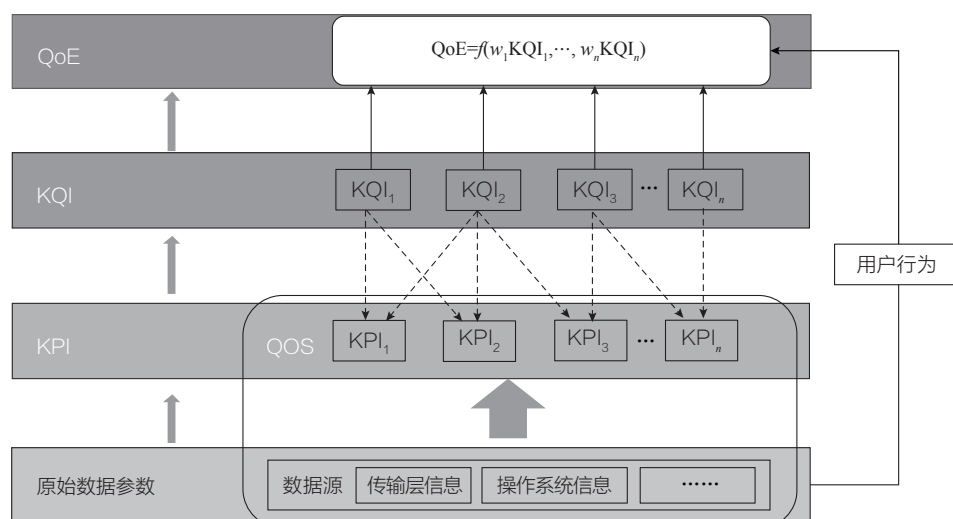
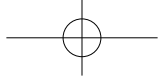


图 5-6 KPI、KQI、QoE 关系的示意图



如图 5-6 所示，服务提供商可以从 QoE 出发从上向下反推底层与之相关的技术因素 KPI，或设计权重和映射关系，也可以从原始的是网络性能参数关系出发设计 KPI，向上最相关的 KQI 指标，再寻找和 QoE 的关联关系。

5.2.3 CEMC与端到端客户服务体验改善

端到端流程是从客户需求端出发，到满足客户需求端去，提供端到端服务。通常通信服务提供商会设立端对端客户服务体验关系中心（Customer Experience Management Center, CEMC）来帮助运营商消除网络、服务和客户之间的距离，提供一套完整的端对端客户服务体验解决方案。

网络运营维护通常监测和管理 QoE 和 QoS 来完成优化改善。CEMC 一般包括服务质量管理（SQM）和 CEM。SQM 属于 QoS 管理的范畴，用于管理服务质量并提供可用于监测不同服务端对端质量的指标。CEM 属于 QoE 管理的范畴并用于从客户的视野角度来测量 QoE。通过分析客户所关心的问题，可以建立一个尽可能接近客户实际体验的用户感知 QoE 指标系统。CEM 通常用于管理用户服务体验感知。

端对端客户服务体验解决方案由关于客户、服务和网络的指标体系组成，这些指标用于测量和评估客户的感知，如图 5-7 所示。最顶层是客户体验指标（CEI），它提供了对客户体验的客观度量。CEI 用于描述客户对服务的体验。第二层是关键质量指标（KQI），它用于指示产品和服务的性能。KQI 可以通过不同的 KPI 指标来计算得到，通常被分为产品 KQI 和服务 KQI。第三层是基于网络的 KPI 指标，代表了某部分端到端的服务数据。KPI 对于评估网络的运营维护是很重要的，它还是 CEI 和 KQI 的数据基础。KPI 指标基于系统预警、性能表现以及网络配置的数据；或者是来自主动或被动监测和抓包获取的分析数据以及账单数据等。

网络运营与维护涉及发现、定位和解决问题。这三方面紧密连接形成了一个完整的网络运营维护的系统，如图 5-8 所示。发现问题意味着找到一个合适传感系统来确定客户的实时感知。定位问题意味着当检测出异常时及时确认问题出现的可能原因。解决问题意味着纠正出现的状况。

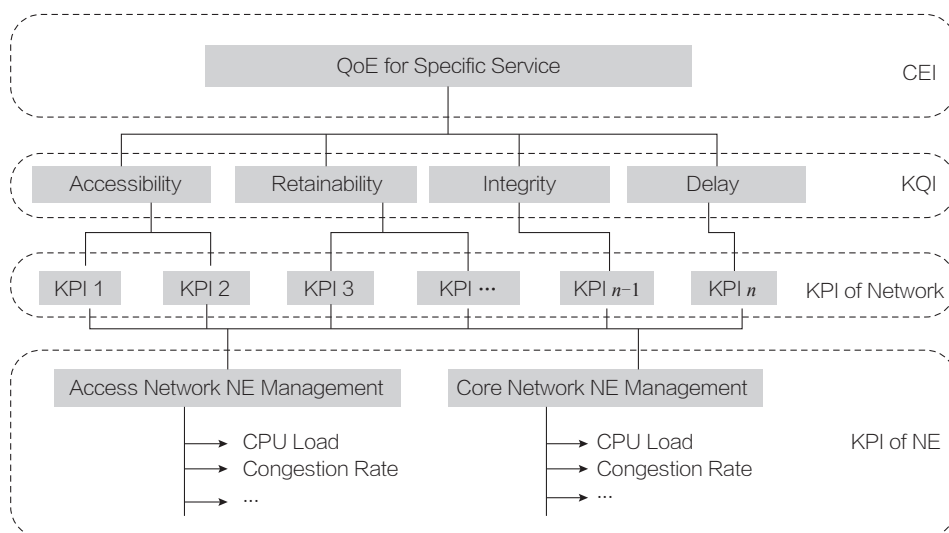


图 5-7 端对端客户服务体验方案

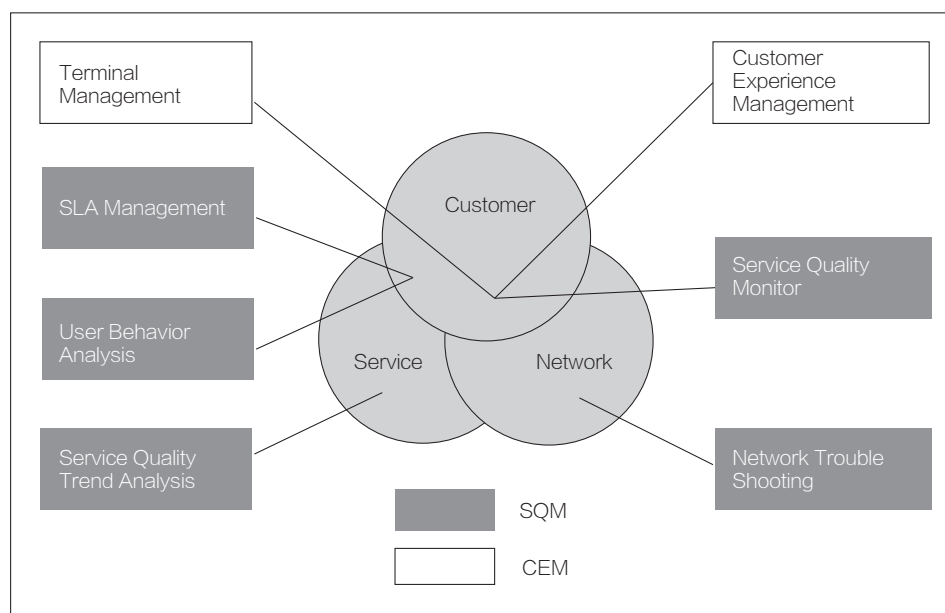
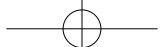


图 5-8 CEMC 的核心功能

CEMC 提供完整的运维功能，从而可以帮助运营商提升 QoE，建立客户关爱机制，提供具有保障的服务水平协议（Service Level Agreement, SLA），并



同时能够吸引高价值企业客户。此外，端对端的客户服务体验解决方案还应当具有足够的可扩展性，在能适用当前服务运营商的用户感知管理的前提下还能够为未来可能需要的架构做足够的扩展性考虑，如图 5-9 所示为 Z 公司提供的一种 CEMC 方案的可扩展平台的示意图，横向的灰色模块表示功能性平台，垂直的白色模块表示服务支持包，这种组合性的架构具有很好的可扩展性能。

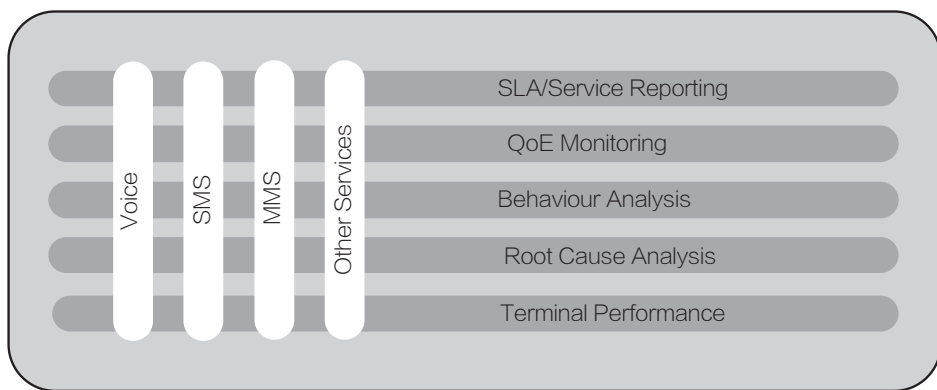
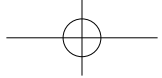


图 5-9 一种功能性平台和服务支持包的组合性架构

一套完整的客户服务体验解决方案要求网络的运营与维护必须是以客户为中心的，并且能够在提供服务中，前端的销售和客户部分能够通过后端网络的规划和优化完成有效的交互活动。这些解决方案的实现使得客户感知变得透明化、可控化，并且可溯源，同时能够帮助运营商在商业竞争中获取优势地位。

5.3 应用于 CEM 中的 AI 技术

AI 技术已经在各种工业场景中得到了广泛的应用，在电信行业的商业活动中 AI 也有自己的应用。具有创新思维的电信运营商已经在使用 AI 或机器学习技术来提高网络的可靠性，提升客户满意度和保留率，优化得到更高回报的商业过程。



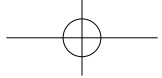
AI 技术在电信行业中最直接的应用场景就是用来增强客户服务体验。A 公司、C 公司和 V 公司等美国领先的电信公司在各种流程中都利用了人工智能技术，其中业务清单里包括了自动聊天机器人、个性化报价和有效简化的客户服务流程等。除了直接提高客户服务的交流体验之外，AI 技术还可以在电信网络分析和预测维护场景中得到应用，通过对网络状况的优化、预测和故障归因分析，提高网络服务质量，从而提高客户感知的服务体验。网络维护通常被认为是 AI 驱动的第二代的解决方案，其重点在于以软件为中心的方法来实现自修复、自优化以及自学习的网络。

几年前，网络提供商曾经派遣现场工作人员到站点来定期检查硬件，导致频繁的延误和错误，对客户的体验产生负面影响。这种方法在今天仍然很重要并且已被广泛使用，然而使用 AI 技术，网络提供商可以避免许多紧急和计划外的检查。如今，算法可以监视网络中数百万个信号和数据点，以实时检测即将发生的问题。根据这些数据，公司可以通过负载平衡做出反应，重新启动所涉及的软件，或派遣人员来解决该问题，从而避免在客户注意到之前就造成许多停机问题。A 公司实验室高级技术副总裁 Mazin Gilbert 预测，预测性网络维护将在未来几年继续推动有利的费用趋势。

5.3.1 ADS算法与用户网络感知原因定位

异常检测（Anomaly/Outlier/Novelty/Peak Detection）通常并没有普遍可接受的关于异常值定义，很多定义都依赖于所处理问题的具体情景。通常一个异常值是指和数据集中其他数据样本看起来相当不同的数据样本。在通信网络产业中，异常值通常和商业上的考虑紧密相关。一个异常的数据样本指向一个事件或者指向导致功能失调的组件（软件或硬件），通常直接或间接地造成经济损失。基于 AI 技术的异常检测系统（Anomaly Detection System, ADS）相较于传统的 ADS 具有更高的可推广性，可以集成于不同的环境中去。

通常 ADS 收集大量的数据并基于先验知识（或者是专家系统）来确认异常的数据点，或者是基于从数据中学习到的知识来确认异常的数据点。根据具体的应用领域和 ADS 的可靠性，可以将检测到的异常报告给专家以进行分析，也可以将其反馈给另一个自动运行缓解措施的系统。



根据具体的分析环境，ADS 可以具有不同的数据输入类型。数据输入的类型通常限制了异常值检测使用的技术或算法。

- ✓ ADS 的输入数据可以是由网络上不同设备运行过程中生成的系统日志文件，也可以是呼叫数据记录（Call Data Record，CDR）的通信日志。这些日志数据通常是一个高维的数据帧。
- ✓ ADS 也可以是带有计数标记的时间序列数据，比如某项服务的订阅使用数随着时间的变化函数。
- ✓ ADS 还可以以网络图作为输入，这些图可以是静态的或者动态的，比如可以考虑蜂窝网中的单元拓扑来检测干扰问题。空间位置的数据（比如用户坐标）也能揭示异常的存在，比如用户在某个接入点或者天线周围的异常分布可能是由于接入点内的问题引起的。

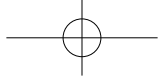
依赖于不同的输入数据类型，输出的异常类型也可以是多种多样的。比如，若输入数据为网络图，异常输出可以是边或者节点。如果输入是某种序列，比如网络日志等，异常可能就是一些事件序列。异常值可以分为异常点、集体异常和情景异常（Contextual Anomaly）。

- ✓ 异常点就是单个样本点具有异常行为。比如对一个服务器来说，短时间内的过载可能会对应一个反常的较长响应时间。
- ✓ 集体异常是指一堆数据样本点整体都表现异常。这种情况可能发生在序列数据中，比如时间序列数据中，比如一次错误的测试配置带来持续性的流量损失或丢包行为。
- ✓ 情景异常是指某个样本数据在特定的情景下是异常的，而在其他情景下可能是正常的。比如，在计划的维护操作期间，单元业务量的下降是正常的，而在其他时间段则是异常的。

ADS 的输出可以是分数或者是标签，灵活性的 ADS 通常返回一个异常的分数值。使用者可以有两种选择：① 选择具有最高分数的前 n 个样本；② 设置一个截断阈值，然后只考虑在阈值分数之上的样本为异常。ADS 当然也可以返回一个二元的标签（异常或正常）或者一个多分类的标签（正常、A-类异常、B-类异常等）。

检测异常值通常会在很多层面上面对多种挑战。

- ✓ 正常数据的定义是特定于具体领域的，没有统一普适的程序来对正常



的数据建模，定义为正常甚至是主观或难以校验的；并且数据的类别（正常或异常）的样本数一般都是不均衡的，很多统计的方法对小样本数据不适用，因此对异常数据建模也不是易事。

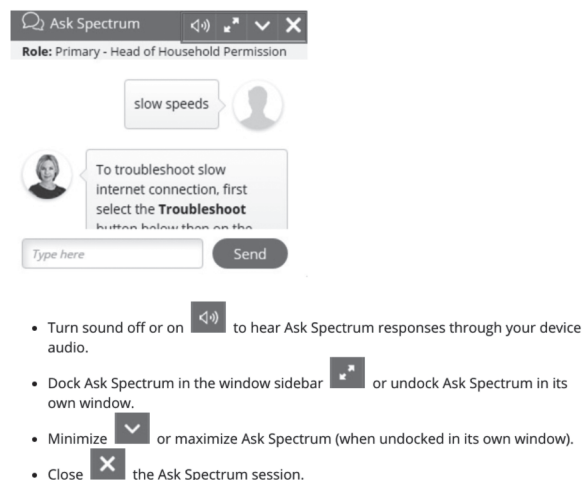
- ✓ 关于正常数据和异常数据的边界实际上是含混不清的，此外数据中可能存在噪声使得主观上假定的异常和正常的分别变得更加困难。
- ✓ 数据是否正常的判断可能是随时间改变的，并且模型中未定义过的新异常也可能会出现。
- ✓ 在很多领域中，拥有带标签的数据集来训练模型是很不常见或者不现实的。

用于异常值检测的算法可以是基于专家系统知识的方法，也可以是机器学习类统计方法，比如回归、分类、聚类、序列分析、规则推导等。现有的研究中有大量研究关于分类和预测，包括用于自修复的蜂窝网络。尽管异常值检测处理的是带有标签的数据，但是最终目的却是像监督学习那样做标签预测。异常值检测通常希望能够对网络异常值出现进行溯因，找到那些对用户网络使用具有负面影响的原因或者说是影响网络质量的问题进行定位。为了能够建立端对端的诊断系统来判断不同类型的问题，并同时考虑不同特征之间的依赖关系，通常需要使用一些规则归纳的算法来从数据中推断出一组指向无效性根源的规则集。在这种情况下，决策树的方法实际上并不是一个很好的选择。除了因为决策树需要编码好的分类数据（在数据量很大的情况下不现实）外，它需要一个已经处理好分类的数据库，这对成本来说是一个很大的负担，并且很难保持更新。

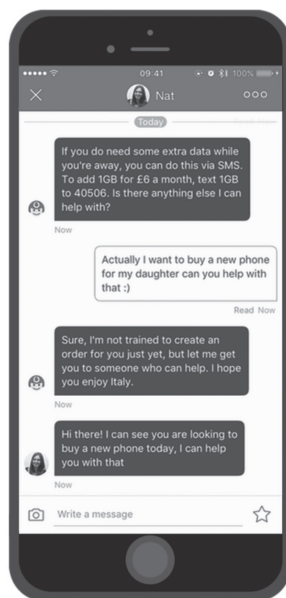
5.3.2 Chatbot技术与客服体验优化

AI赋能的客户服务交流解决方案通常以聊天机器人(Chatbot)的界面呈现，Chatbot 是一种计算机程序，能够通过音频或文本方法进行对话。即聊天机器人系统用于模拟人类对话，并且通常被公司用作与客户互动的方式，如图 5-10 (a) 和图 5-10 (b) 所示。

Options when using the Ask Spectrum Virtual Assistant:



(a)

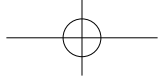


(b)

图 5-10 基于网页或移动端的聊天机器人客服

聊天机器人作为虚拟的聊天助手可帮助客户进行故障排除、账户信息或有关网络服务的一般问题。聊天助手管理的客户查询范围从确定服务中断到订购付费内容服务。聊天助手还可以为用户提供有用的提示和帮助中心的链接，或者在请求更复杂的情况下，将其推荐给实时聊天的人工客服。这样一来，一些工作任务就从人工客服的团队工作中分离出来，把人工客服用于处理更苛刻的案件。尤其当客户需要解决的问题数量非常大，以至于人类客服难以应付的时候，聊天机器人客服的优势将得到明显展现。服务提供商还可以通过多种方式使用聊天机器人，比如简单形式的站点指南或 FAQ 指南，或者以更专门的角色使用虚拟支持代理或虚拟销售代理。可以帮助客户填写表格或进行调查，甚至可以充当聊天室主持人的角色。尽管我们通常会在网站上遇到聊天机器人，但在其他地方也可以使用它们：

- ✓ 在即时消息传递平台上：聊天机器人可以24×7全天候在线回答用户查询甚至进行讨论。
- ✓ 在SMS上：客户可以使用聊天机器人从手机中获得查询的答案。
- ✓ 在推特上：聊天机器人不仅可以回答查询，还可以代表公司发推文。



- ✓ 聊天机器人还可以部署在横幅广告、计算机亭或台式机中。

有时候这种服务也可能以后端的形式出现，来帮助客户服务部门的运行更有效。比如说通过分析广泛的后台数据帮助人工客服迅速确认客户的问题所在并能及时地找到对应的解决方案。以下是几个电信运营商使用 AI 算法在客户服务体验改善的例子：

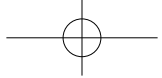
- ✓ 作为客户请求和客户帮助中心之间的网关。
- ✓ 将客户、客户请求路由到适当的代理商，并将具有购买意向的潜在客户直接路由到销售部门。
- ✓ 分析客户请求以及网络数据，更有效地找到解决客户问题的解决方案。
- ✓ 让客户通过语音而非远程控制来浏览或购买媒体内容。

研究表明，超过 50% 的客户相信无须客户服务代理商的帮助即可自行解决问题。聊天机器人将是此处的理想解决方案，因为它们可以引导客户完成整个过程，同时让客户带头。分析客户行为不仅对业务有益，而且对客户也有利。聊天机器人可以跟踪客户的使用方式并推荐合适的产品。通过提供有关客户感兴趣的产品的更多信息，聊天机器人可以帮助客户选择最适合其需求的产品。在每次交互过程中了解客户，可帮助聊天机器人有效地响应客户查询。

如今，客户可以使用其他渠道与服务提供商建立联系。在使用聊天机器人与世界进行无障碍通信之前，我们可能还有很长的路要走，但是该技术确实有可能彻底改变电信行业的客户体验管理。

5.3.3 基于KDtree、LSTM以及多算法融合的网络容量预测

通常运营商可以使用自建或者第三方提供的 AI 引擎平台来搭建自用的模型，有的平台还可以根据单元异常检测和容量预测方案帮助定义数据收集标准和元数据规范，并将多个数据源、多个周期和难以理解的低质量数据转换为统一的数据源以便轻松理解高质量数据。基于专家经验的预设异常阈值可检测异常数据并自动批量标记异常。对 5000 多个性能指标执行数据聚类，以识别 7 种类型的数据特征，比如数据周期性、趋势和突发等。这为高质量、高精度模型训练提供了保证。



短期容量预测：KD-tree 是一种分割 k 维数据空间的数据结构。主要应用于多维空间关键数据的搜索（如范围搜索和最近邻搜索）。使用 KD-tree 来构建网络单元在空间上的关系，有助于快速找到相邻网络单元。然后可以使用 LSTM Seq2seq 算法以统一的方式对网络单元进行建模，从而找出和用户行为密切相关的流量规律。此外，该算法还考虑了重要事件和天气因素，以有效预测意外事件。对于 7 天的容量预测和分析，该项目的预测有效性达到 97.21%，如图 5-11 所示。

长期容量预测没有很强的规律性，并且受网络更改和费率调整等大量外部因素的影响。容量预测的准确性始终是行业中的挑战。通过消除非人为因素引起的异常流量波动，基于小区特征聚类组对项目进行建模，并采用多算法融合收敛等关键方法提高基于数据特征聚类的容量预测精度。在分析了 20 000 个基站 6 个月的历史数据并预测未来 3 个月的容量之后，与传统的 Holt-winters 算法 39% 的有效性相比，融合收敛算法的预测有效性为 81%，如图 5-12 所示。

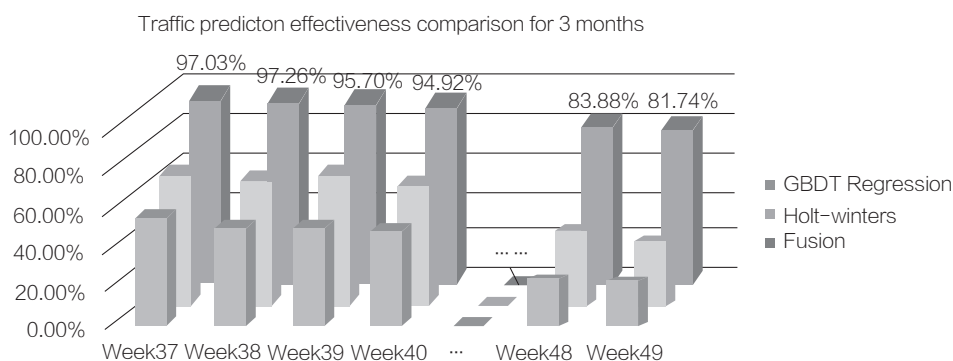


图 5-12 融合收敛算法效果

5.3.4 NPS度量与用户业务感知提升

2015 年世界移动通信大会（MWC）期间，E 公司发布了大数据分析套件 Expert Analytics 15.0。这套方案可以帮运营商预测 NPS，并且提出改进方案。NPS 是目前最流行的客户忠诚度分析指标，用以计量客户向其他人推荐企业业务的可能性。在同一个用户调查样本中，业务推荐者的比例减去业务贬损者的比例，即为 NPS。

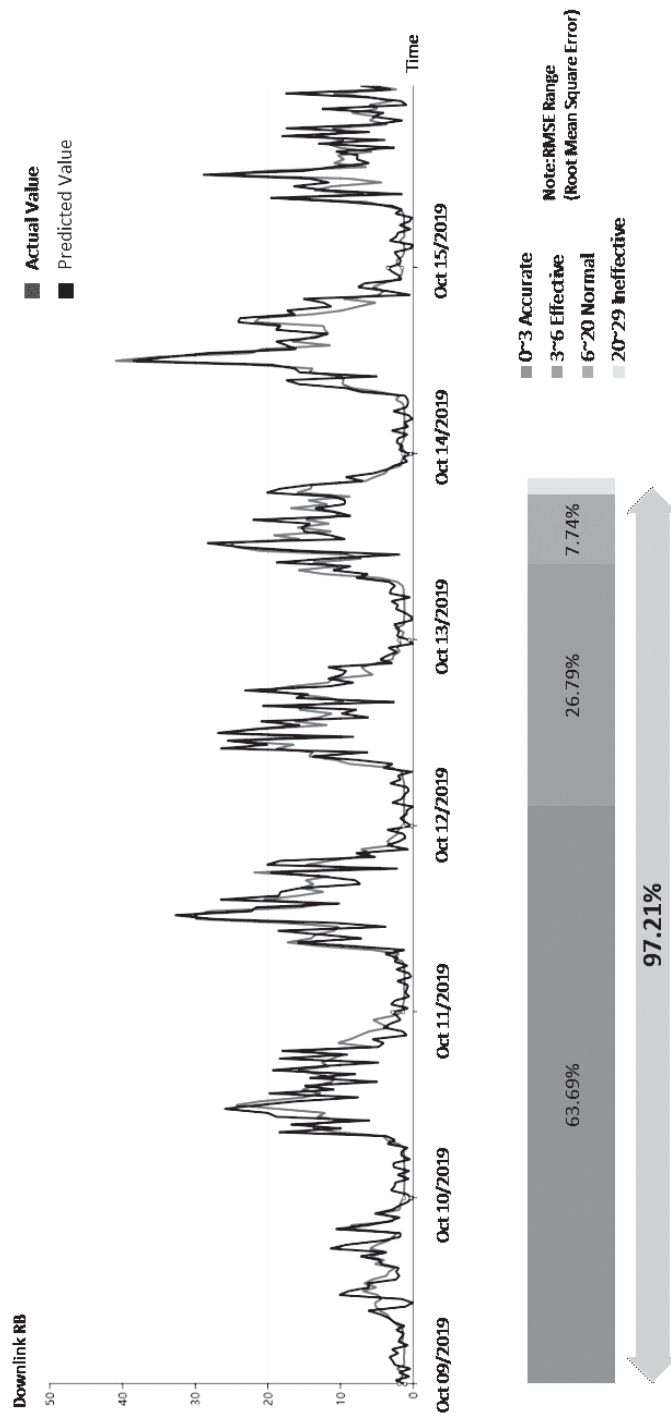
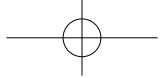


图 5-11 算法对网络单元容量的预测结果和真实结果对比



A 公司和 P 公司等很早就启动了 NPS 考核，并将其视为“未来利润”。对当前的运营商而言，更高的 NPS 不仅意味着领先于其他运营商，同样也可以在面对 OTT 冲击时降低用户流失率。当时 E 公司副总裁兼 OSS&CEM 产品管理主管 Shamir Shoham 在接受记者采访时表示，目前绝大部分电信运营商的 NPS 都低于 5%，甚至很多为负值。而相比之下，A 公司、G 公司等互联网公司的 NPS 超过 70%。运营商的用户忠诚度远不及互联网企业，Shamir Shoham 对记者表示，5% 是很危险的，运营商至少要提升到 30% 以上。

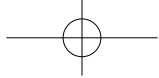
E 公司针对这一诉求推出了 Expert Analytics 15.0 分析套件，其可以在线、随时为运营商生成用户服务报告，并且根据 E 公司服务水平指数（SLI）预测 NPS，明确告诉运营商，是哪些因素影响了用户满意度。

以欧洲某家运营商为案例，该运营商在某小区的用户中有 25% 的高价值用户，E 公司收集并分析这些用户的行为，以及对应的网络要素。比如，在线收集用户观看视频的时间，此时的缓冲、下载速率以及时延指标。对应网页浏览，则采集网页打开时长；而对于用户通话行为，则采集通话时长、质量、计费等。整个过程不影响用户感知，根据 E 公司的分析模型，该运营商的 SLI 为 6.09，通过 SLI 测算出的 NPS 也非常低。

运营商可以针对 SLI 提出的问题加以改进，而前文提及的欧洲运营商，通过这套系统使得其高价值用户忠诚度明显提升，NPS 超过了 30%，整个过程，运营商都不需要进行问卷调查。对于高忠诚度用户，运营商可以进一步开展位置信息、定向广告等业务。

V 公司则发布了新的人工智能产品——数字化客户体验服务，供企业整合现有的客户支持渠道，并根据以往互动的支持体验，通过社交媒体、聊天服务、电子邮件、短信或电话，全天候提供虚拟援助。

V 公司的新平台由 4 个组件组成。此组件可以将用户升级为人工智能服务，以便在人工无法提供帮助的情况下使用。Live Agent 组件是一个桥接器，可通过文本、语音、视频方式以虚拟联络中心（Virtual Contact Center）等行业标准进行通信。此代理不仅将客户与人工智能代理连接起来，也为这些代理提供了有用的背景信息，包括历史记录、会话的详细信息和浏览状态等。Knowledge Assist 组件将创作工具与机器学习相结合，为人工智能代理提供相关的答案和指导。它不仅可以扫描内部和外部数据源，还可以使用客户历史数据来提供背景信息。最后，Social Engagement 组件是用于社交媒体推广的工具包，它不仅



可以帮助企业在社交媒体上查找关于其品牌的热门话题，还可用于举办社交媒体广告活动。

V 公司的新产品比通用虚拟助理技术（如 AWS 中的 Alexa Skills Kit 或 Amazon Lex 服务）更适合支持客户体验，使用起来更加方便。相反，虽然 Lex 具有灵活性，但是它需要更多的编程工作来将 Lex 和相关的 AWS 服务整合到所需要的解决方案中，这就在很大程度上增加了难度。

在不久的将来，人工智能在客户服务中的使用率可能会增加，根据 IDC 的报告显示：2020 年，40% 的数字化转型计划将使用人工智能服务，到 2021 年，75% 的企业应用程序也将实现数字化转型。