边缘人工智能

边缘人工智能

- 边缘场景中的人工智能
- 人工智能在边缘场景的应用
- 边缘网络的人工智能
- 移动端开源机器学习框架
- 边缘人工智能展望

边缘场景中的人工智能

- 人工智能简述
- 常用概念简介
- 边缘计算与人工智能结合的意义

人工智能简述

- 人工智能主要关注与构建能够处理通常需要人类智能才能处理的任务的智能机器。如今人工智能已经成为技术行业的重要组成部分，能够帮助解决计算机科学、软件工程以及运筹学中的众多难题。

常用概念简介

- 机器学习
- 深度学习
- 联邦学习

机器学习

- 机器学习被视为人工智能的子集，是对通过经验自动改进的计算机算法的研究。如今有关机器学习的技术与应用已经涉及到生活的方方面面，自动驾驶、指纹解锁等技术的兴起也为生活带来更多便利。

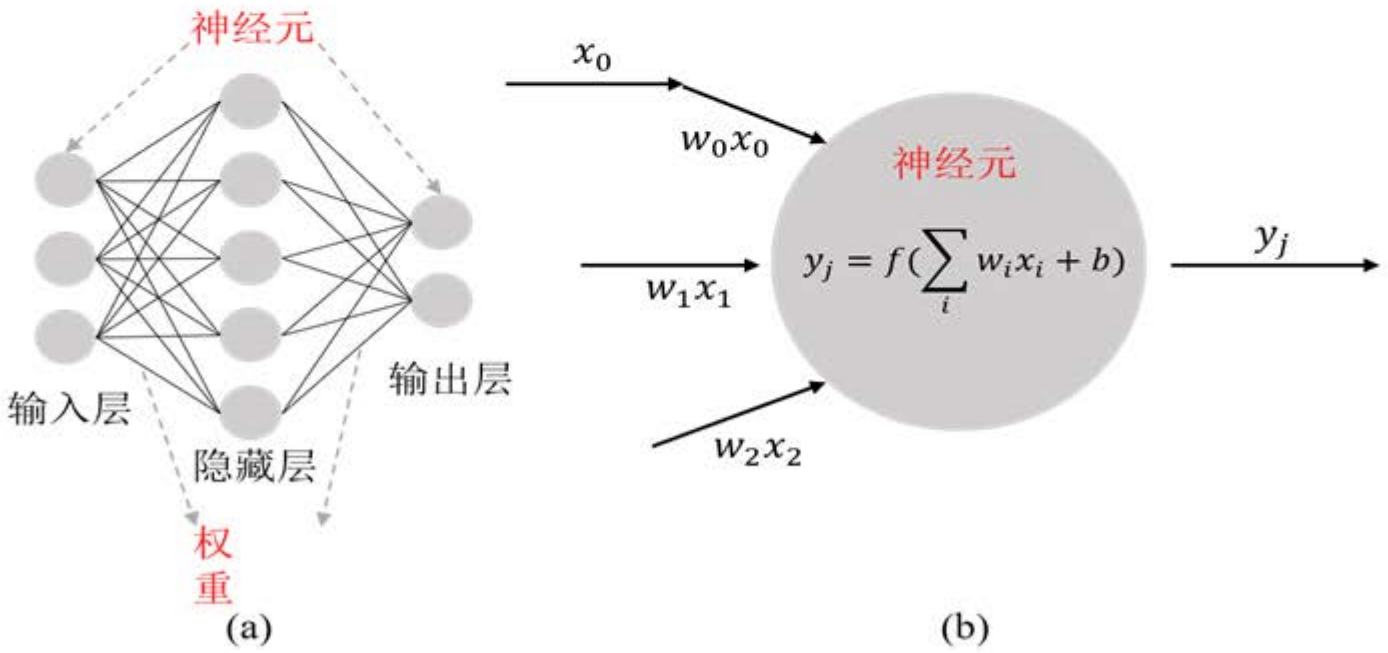
机器学习

机器学习方法分类
监督学习
无监督学习
强化学习

机器学习常见算法
决策树
支持向量机
朴素贝叶斯
K-近邻（KNN）
线性回归
人工神经网络（ANN）

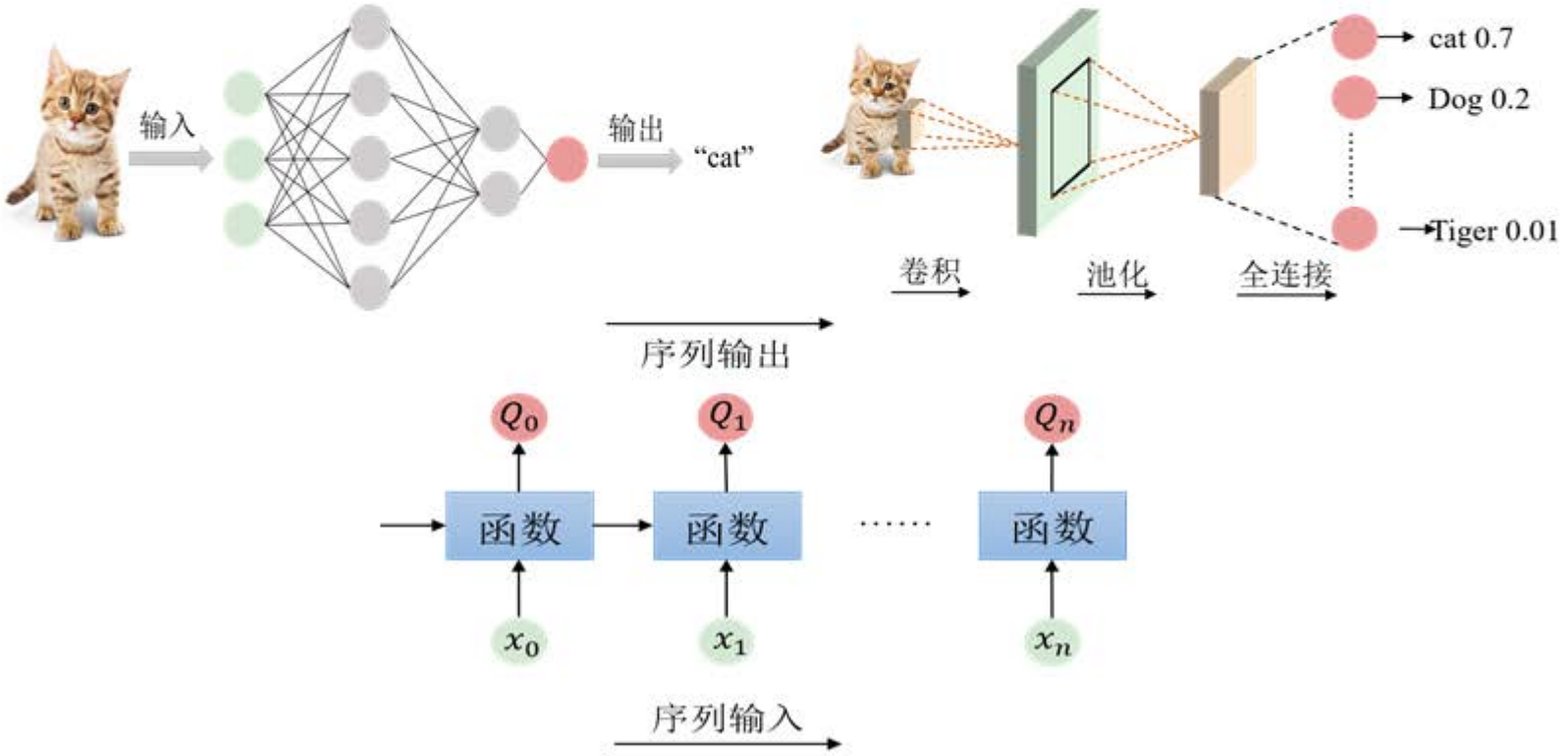
深度学习

- **深度学习通过利用人工神经网络（ANNs）来学习数据的深度表示，由于人工神经网络通常由一系列的层次组成，因此该模型被称为深度神经网络（DNN）。**



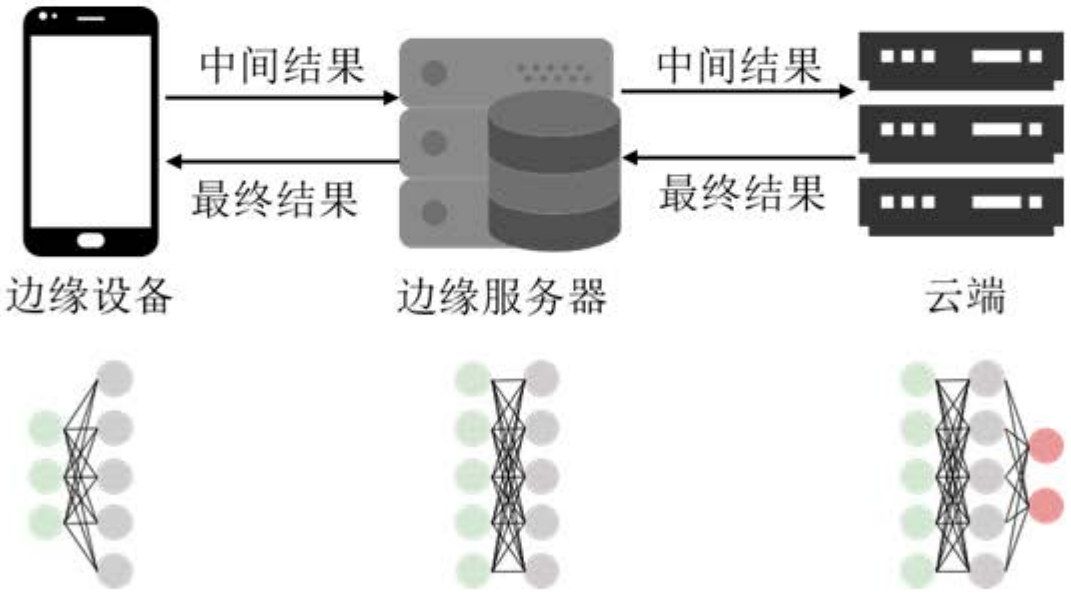
深度学习

- DNNs的三种常用结构：多层感知器（multilayer perceptrons，MLPs）、卷积神经网络（convolution neural networks，CNNs）和递归神经网络（recurrent neural networks，RNNs）。



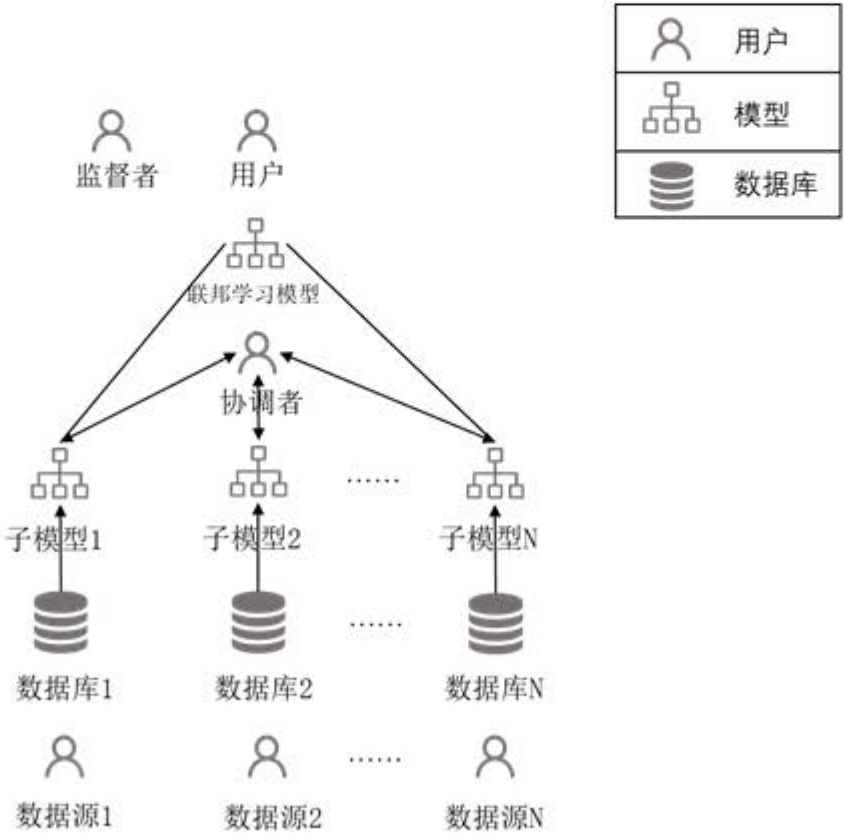
深度学习

- **在边缘网络中，除了能够将DNN部署在边缘服务器以外，还能利用DNN的结构进行模型分割。**



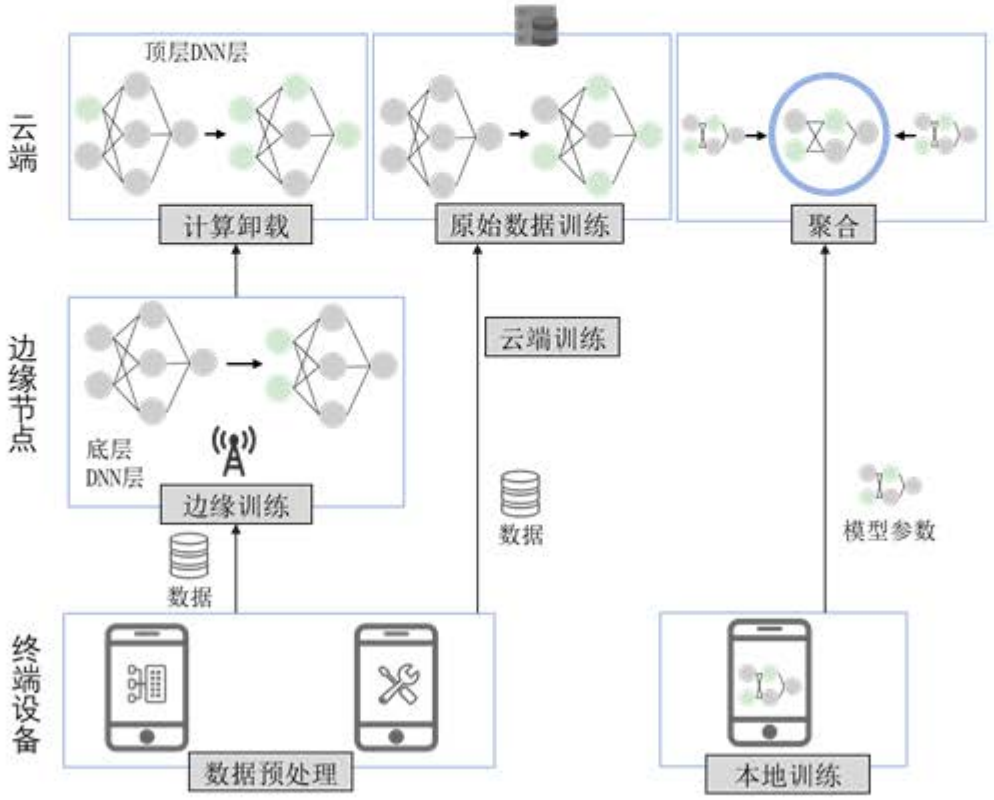
联邦学习

- **联邦学习(Federated learning）是谷歌在2016年提出的一种机器学习技术，有三大构成要素：数据源、联邦学习系统、用户。**



边缘计算与人工智能结合的意义

- **边缘计算能够为人工智能应用带来更低的延迟以及更低的带宽消耗。一种新的MEC中的模型训练协作模式如图。**



边缘计算与人工智能结合的意义

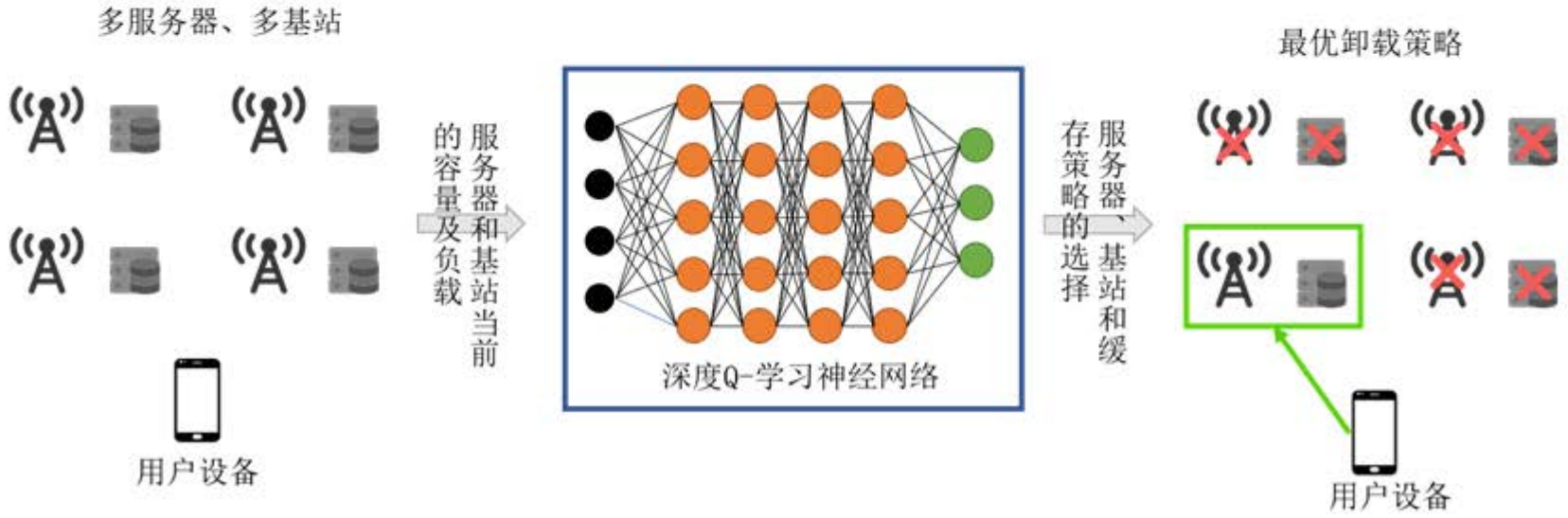
- 利用人工智能能够更有效的利用边缘结点获取到的各种数据，边缘计算可以通过人工智能应用来普及。
- 由于在边缘运行人工智能应用的优越性和必要性，因此边缘人工智能近年来已经受到了广泛的关注。

人工智能在边缘场景的应用

- 计算卸载决策优化
- 服务器部署决策优化
- 资源分配决策优化
- 深度学习在边缘网络中的应用

计算卸载决策优化

- 在大多数边缘网络场景中，需要合理安排每个移动设备任务的卸载决策，使用基于学习的算法不仅能够做出最小化计算延迟的卸载决策，也能对边缘系统的可靠性、安全性带来提升。
- Carleton University的He等人在智能城市的环境中使用神经网络来选择用户将连接到哪个虚拟网络。

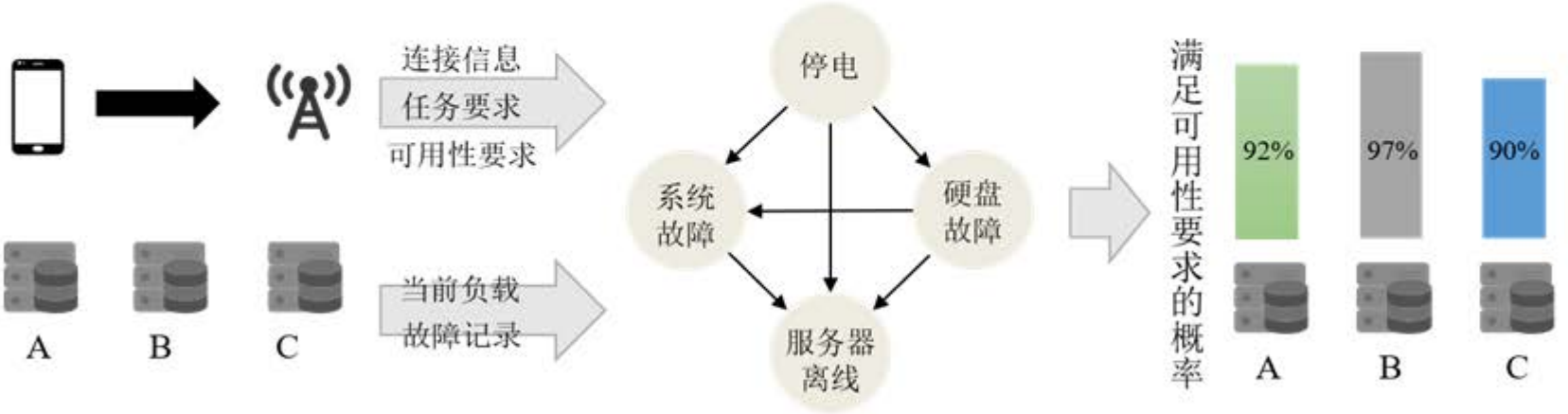


计算卸载决策优化

- 首尔科技大学的Rathore等人提出使用hesitant fuzzy set来根据资源的状态以及用户所需的安全级别等级来决定用户任务是否应该卸载到某一台边缘服务器。
- 美国明尼苏达大学数字技术中心的李等人的工作主要关注IoT设备任务卸载到边缘服务器场景中的安全方面问题。
- 清华大学电子工程系的孙等人考虑了一个高度动态化的移动边缘网络场景。
- 意大利电信网络和远程信息处理实验室的Carrega等人在文中提出中间件的构想。

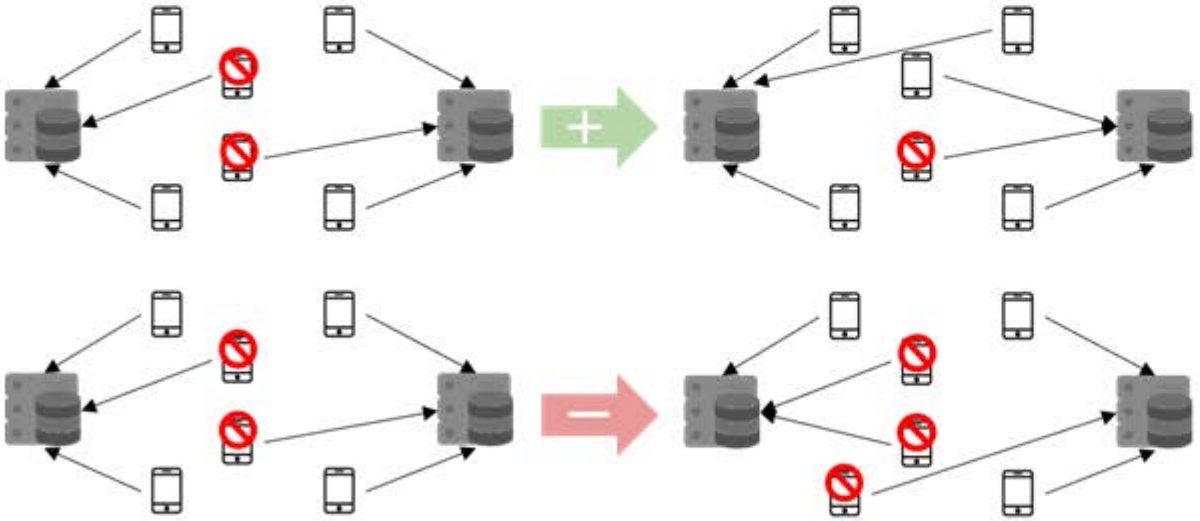
计算卸载决策优化

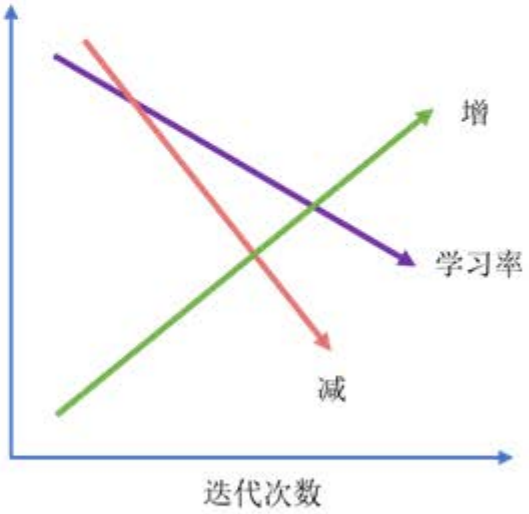
- **奥地利维也纳理工大学软件技术与交互系统研究所的Aral等人使用基于树的朴素贝叶斯来决策每个用户应该将任务卸载到哪台边缘服务器与虚拟机。**



计算卸载决策优化

- InterDigital Communications 的Kim等人提出了一种用于确定边缘网络场景中用户卸载目标的算法。



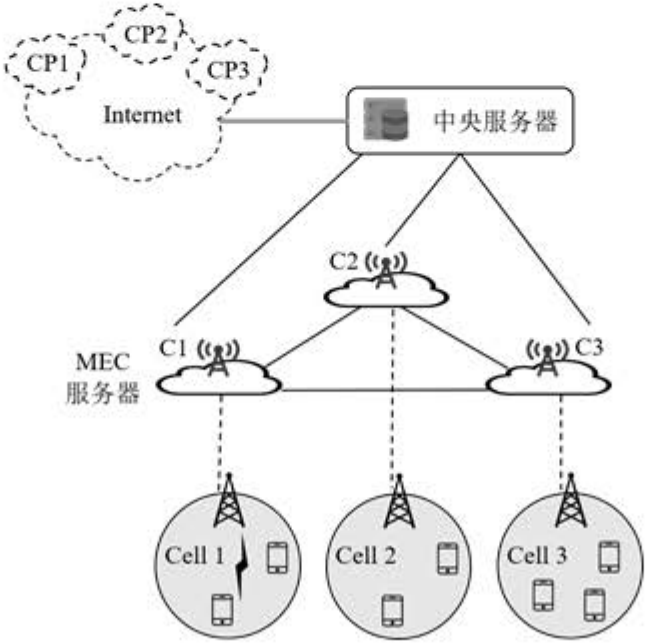


计算卸载决策优化

- 伊利诺伊大学香槟分校的Gu团队使用模糊控制模型来决定子任务应该在本地还是在边缘服务器上执行。
- 拉马尔大学的He等人通过受约束的马尔可夫决策过程对MEC中的安全需求问题进行建模。
- 新加坡技术与设计大学的Dinh等人研究了用户任务应该卸载到哪里的决策问题。
- 美国犹他州立大学的Tan和Hu等人考虑移动边缘场景中的移动设备对象是车辆。
- 东南大学国家移动通信研究实验室的Wu等人在一个边缘服务器为车辆中的用户提供服务的场景中使用了支持向量机 。

服务器部署决策优化

- 使用基于机器学习的算法来完成服务器部署决策能够有效减小任务计算延迟、降低能耗以及提高边缘服务器的资源利用率。
- 电子科技大学通信科学技术国家重点实验室的Hou等人提出一种边缘网络场景下的MEC服务器缓存策略。



服务器部署决策优化

- 东南密苏里州立大学的Crutcher等人使用k-NN（K近邻，k-nearest neighbors）算法来最大程度地减少服务延迟和能耗。
- 云南大学信息科学与工程学院的Li等人使用了k-means聚类算法决定MEC服务器应该部署在何处。
- 上海电力大学计算机科学与技术学院的Du等人提出了一种基于CFSFDP进行聚类的方法。

服务器部署决策优化

- 北京邮电大学网络体系结构与融合重点实验室的Liu等人同样也使用了k-means聚类方法。
- 北京邮电大学网络与交换技术国家重点实验室的Li解决服务器部署问题。
- 墨西拿大学工程系的Vita等人虑一个具有多个边缘服务器和远端云服务器并且用户存在移动性的边缘网络场景。
- 西安工程大学计算机系的Gao等人解决了类似的问题。

资源分配决策优化

- **合理分配边缘服务器为用户提供的计算资源、通信资源，以及边缘服务器与云服务器之间的资源对于任务计算延迟、能量消耗等方面有决定性的作用。**

资源	影响
传输功率	用户与边缘服务器以及边缘服务器之间的传输功率，会直接影响传输能耗与传输速率
服务器计算资源	当有多个用户将任务卸载到服务器上，服务器需要为用户分配计算资源，会对用户任务计算速度、计算能耗以及服务提供商利润造成影响
服务器内存缓存	内存缓存越大，越多用户能够借助边缘服务器处理而无需将任务上传到云端，从而减小执行延迟
信道带宽	直接影响数据上传与下载速度
编码块	编码块长度太小会导致过高解码错误率，而太大会影响传输效率

资源分配决策优化

- 日本东北大学信息科学研究生院的Rodrigues等人使用粒子群优化算法。
- Rodrigues考虑一个用户具有移动性的边缘网络场景。
- 东南大学国家移动通信研究实验室的Zhang等人提出动态演化博弈。
- 浙江大学信息科学与电子工程学院的Wang结合遗传算法与模拟退火算法。
- 华南理工大学软件工程学院的Wen等人提出了一个用于向运行数据流应用程序的用户分配资源的方法。

资源分配决策优化

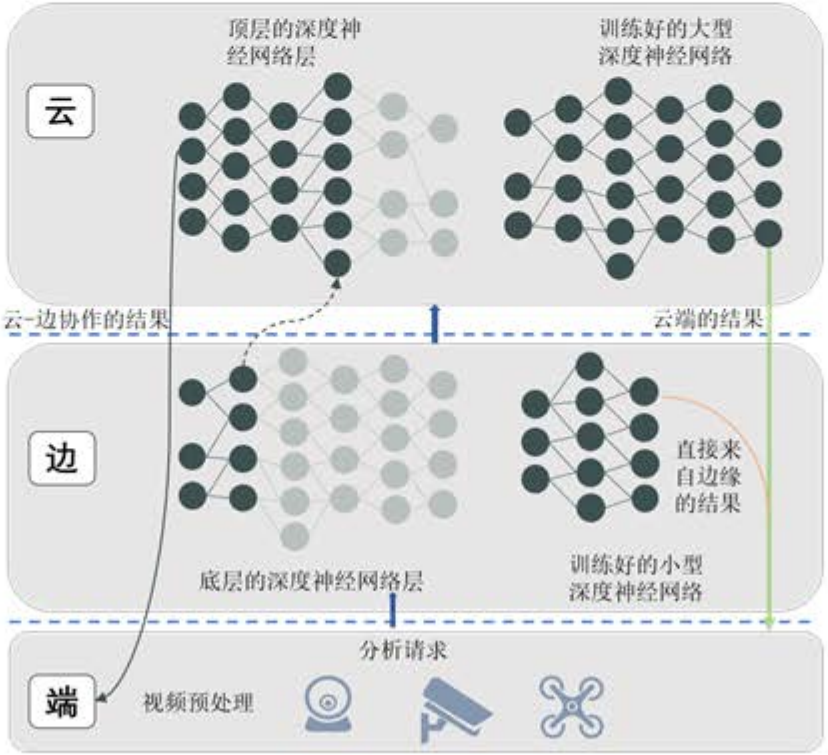
- 迈阿密大学电气与计算机工程系的Xu等人解决在非城市中心的区域部署边缘服务器的困难。
- 亚琛工业大学的Yang等人解决了将通信资源和计算资源从单个MEC服务器分配给多个用户的问题。

深度学习在边缘网络中的应用

- 实时视频分析
- 自主车联网
- 智能制造
- 智慧家庭与智慧城市

实时视频分析

- **实时视频分析在自动驾驶、虚拟现实（VR）、增强现实（AR）以及智能监控等领域都非常重要。 下图分别在终端、边缘和云端协作通过深度学习运行实时视频分析。**

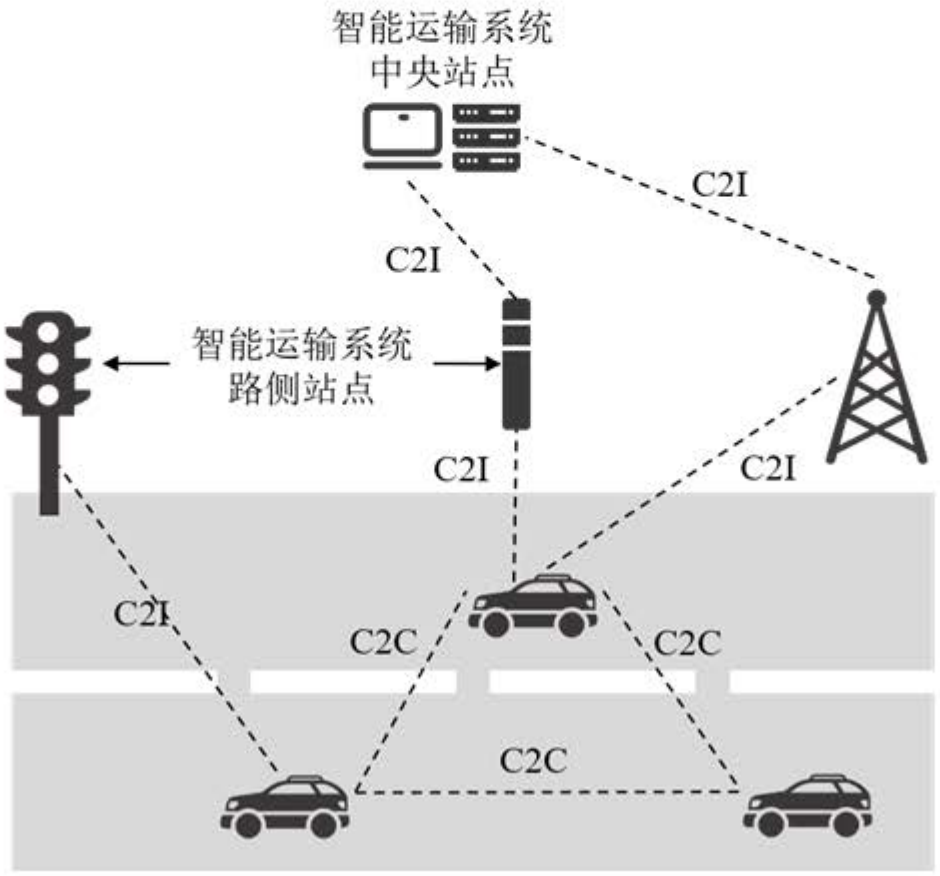


实时视频分析

- **终端**
  - 终端层在实时视频分析中主要用于视频获取、媒体数据压缩、图像预处理以及图像分割。
- **边缘端**
  - 在边缘层次，可以通过众多分布式边缘结点协作的方式提供更好的服务。
- **云端**
  - 在云的层次上，云服务器主要负责边缘层次内部的深度学习模型集成，并更新边缘结点上分布式深度学习模型的参数。

自主车联网

- **将车辆互相连接起来，能够提高车辆安全性、提高效率并且减少交通拥堵的发生。**



智能制造

- 智能制造时代的两个最重要的准则是自动化和数据分析，前者是主要目标，后者是最强大的工具之一。为了遵循这两个原则，智能制造首要任务便是解决响应延迟、风险控制和隐私保护方面相关问题，因此深度学习和边缘计算都是必不可少的。

智慧家庭与智慧城市

- 为了保证家庭敏感隐私数据，智慧家庭的数据处理必须依赖边缘计算。智慧家庭能够扩展到社区甚至城市，则公共安全、健康数据、公用设施以及交通和其他领域都能够受益。

边缘网络的人工智能

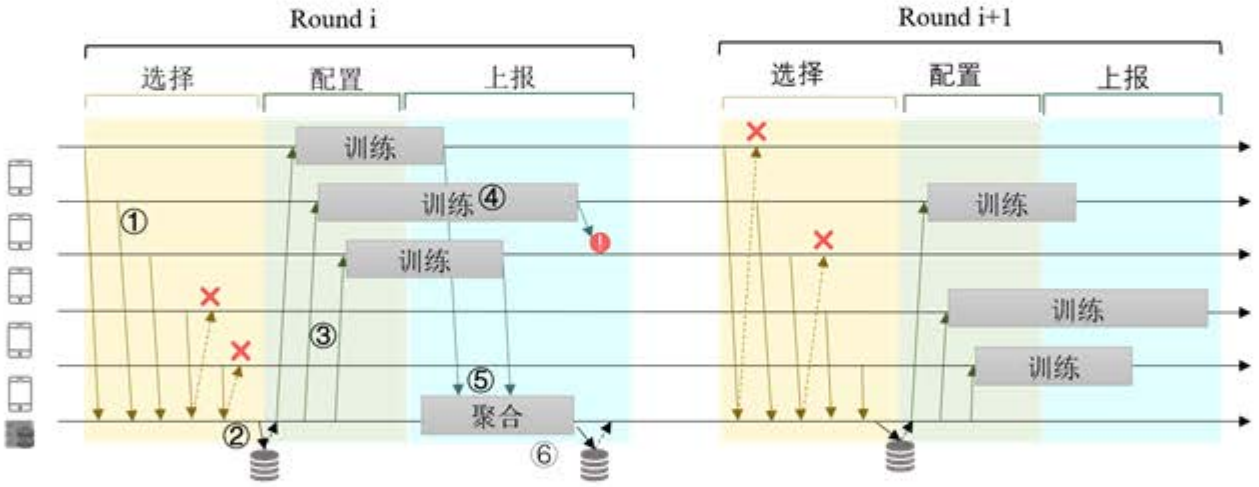
- 联邦学习与边缘网络
- TinyML
- Fregata
- AIoT系统

联邦学习与边缘网络

- 降低通信开销
  - 边缘和本地计算
- 模型压缩
- 基于重要性的更新

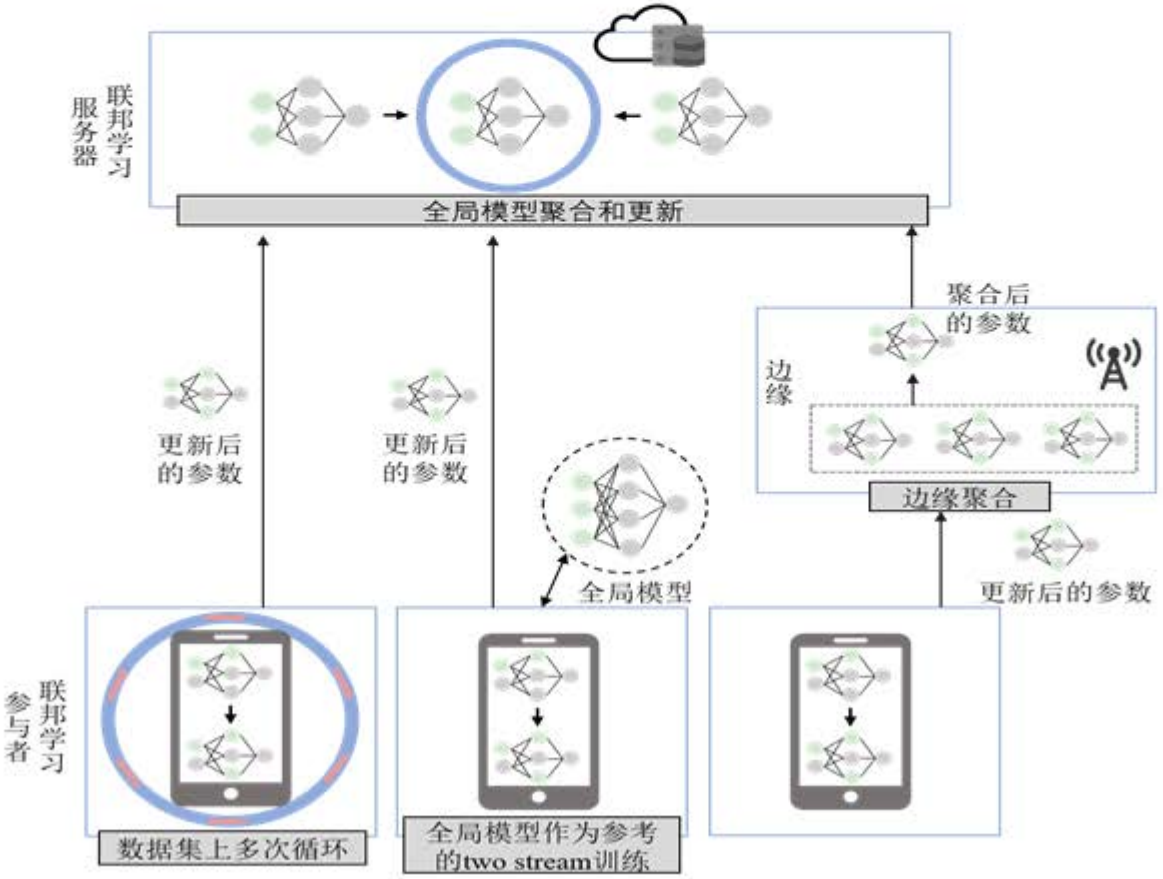
降低通信开销

- **边缘场景联邦学习中可能需要参与者和联邦学习服务器之间多轮通信来实现目标精度，提高通信效率是很有必要的。**



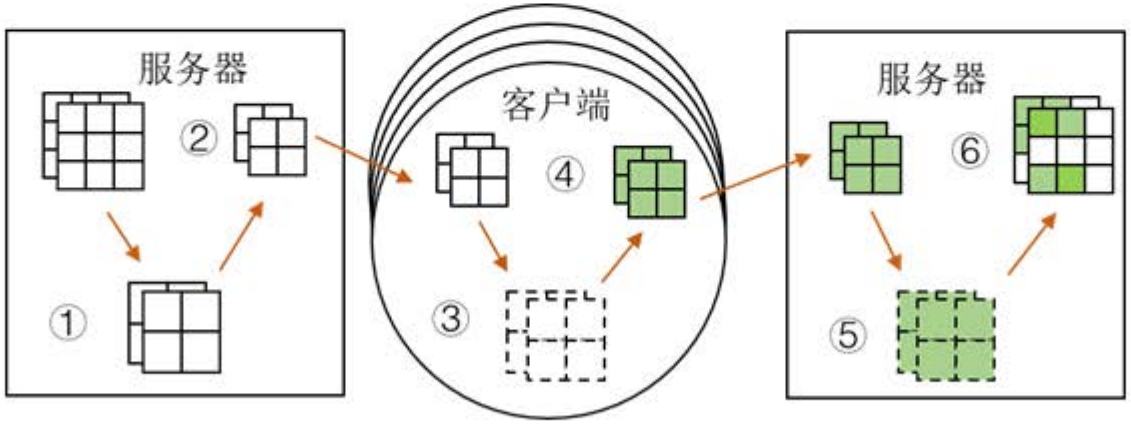
边缘和本地计算

- **边缘计算场景下的联邦学习中，通信开销带来的影响远大于计算开销，可在每次全局聚合前，在边缘节点或终端设备上进行更多的计算。**



模型压缩

- 模型压缩以及梯度压缩能够通过稀疏化、量化或二次采样将一次通信的更新变得更加紧凑。然而，由于压缩会引入噪声，因此此时的目标是保持训练模型质量的同时减小每轮更新的大小。

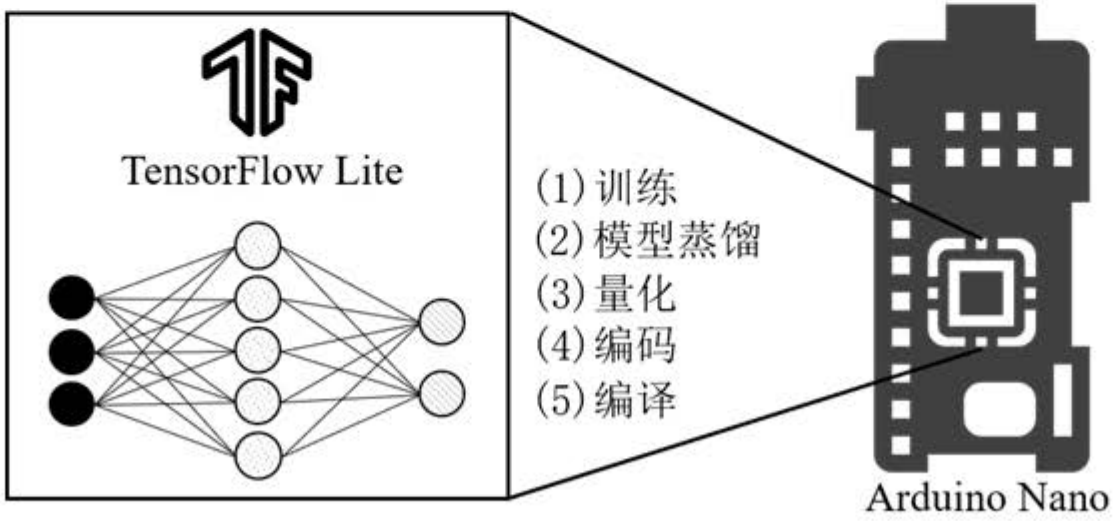


基于重要性的更新

- 这种类型策略包括选择性通信，即在每一轮中只传输重要或者相关的更新。实际上忽略参与者的部分更新除了能减小通信开销外，甚至能够提高全局模型性能。

TinyML

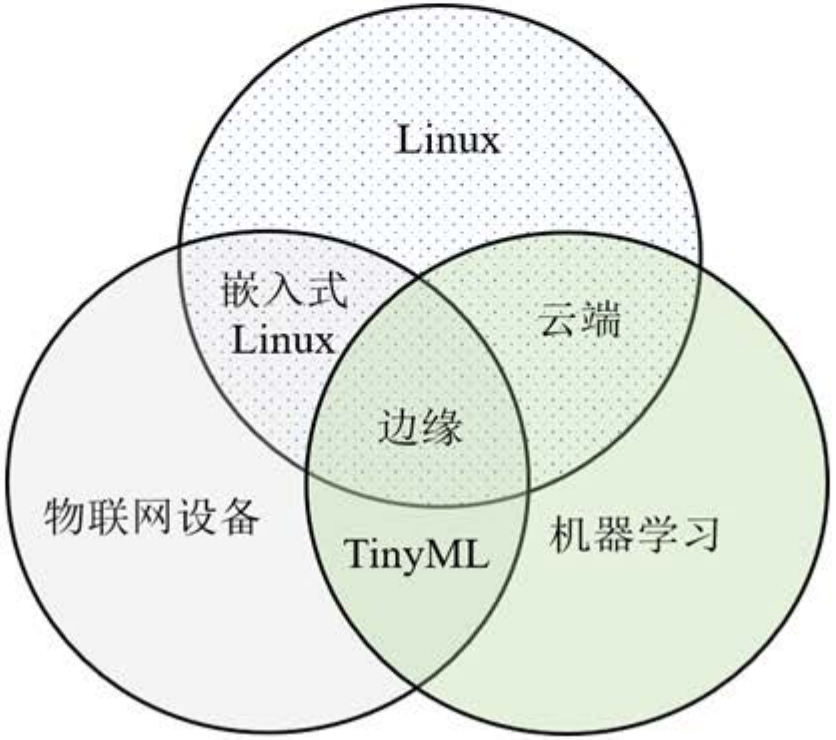
- 随着机器学习在工业中应用越来越多，物联网也迅速发展，如何在受制于低算力与低能耗的物联网设备以及嵌入式设备等终端硬件上长时间低功耗的运行AI应用亟待解决。
- TinyML,即微型机器学习，是指在终端、边缘端的微处理器上运行机器学习。



TinyML

TinyML

- TinyML相关工作涉及到多个领域的结合，包括机器学习领域相关技术的部署和使用、基于边缘计算的思路将计算和存储下沉到靠近设备一侧以及物联网设备的使用。



Fregata

- Fregata是TalkingData针对大规模机器学习中计算资源消耗大、训练时间长以及调参效率低下的问题提出的基于Apache Spark的轻量级、开源、超高速大规模机器学习算法库。

Fregata主要特点
精确
高速
无需调参或者调参较简单
轻量

AIOT系统

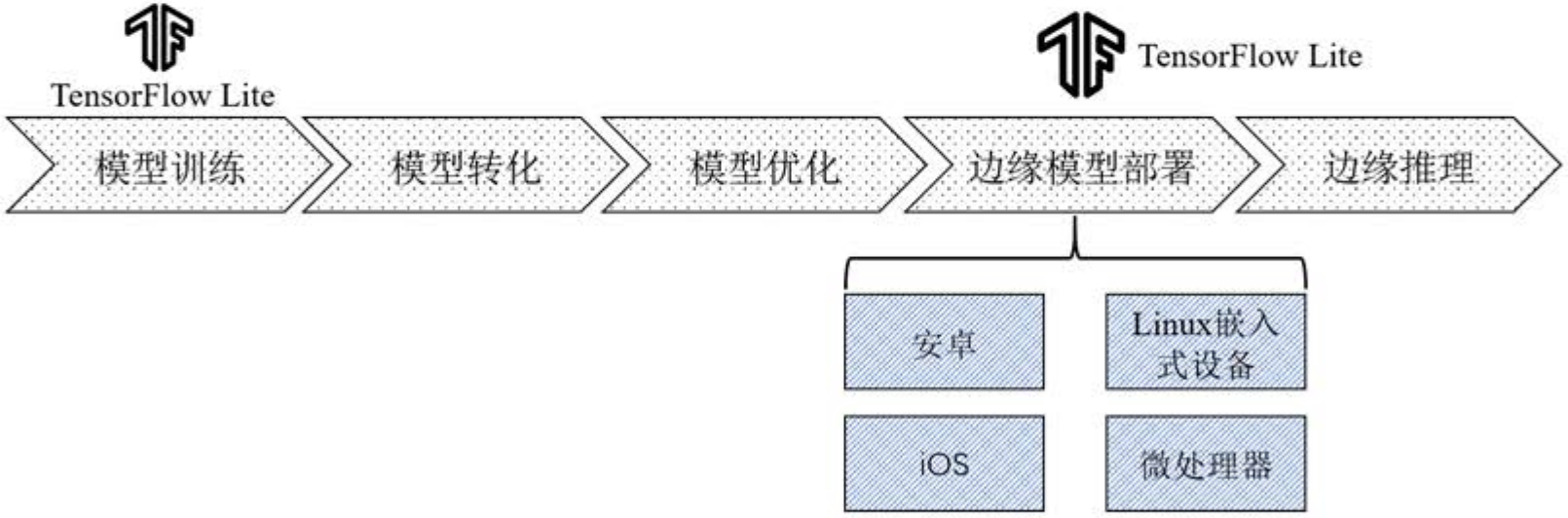
- AIoT，即人工智能物联网（Artificial Intelligence & Internet of Things），旨在通过物联网产生与收集海量数据并存储于边缘端与云端，然后通过人工智能与大数据分析，从而实现万能数据化与智联化。

移动端开源机器学习框架

- TensorFlow Lite
- Core ML
- NCNN
- Paddle Lite
- MNN
- MACE
- SNPE

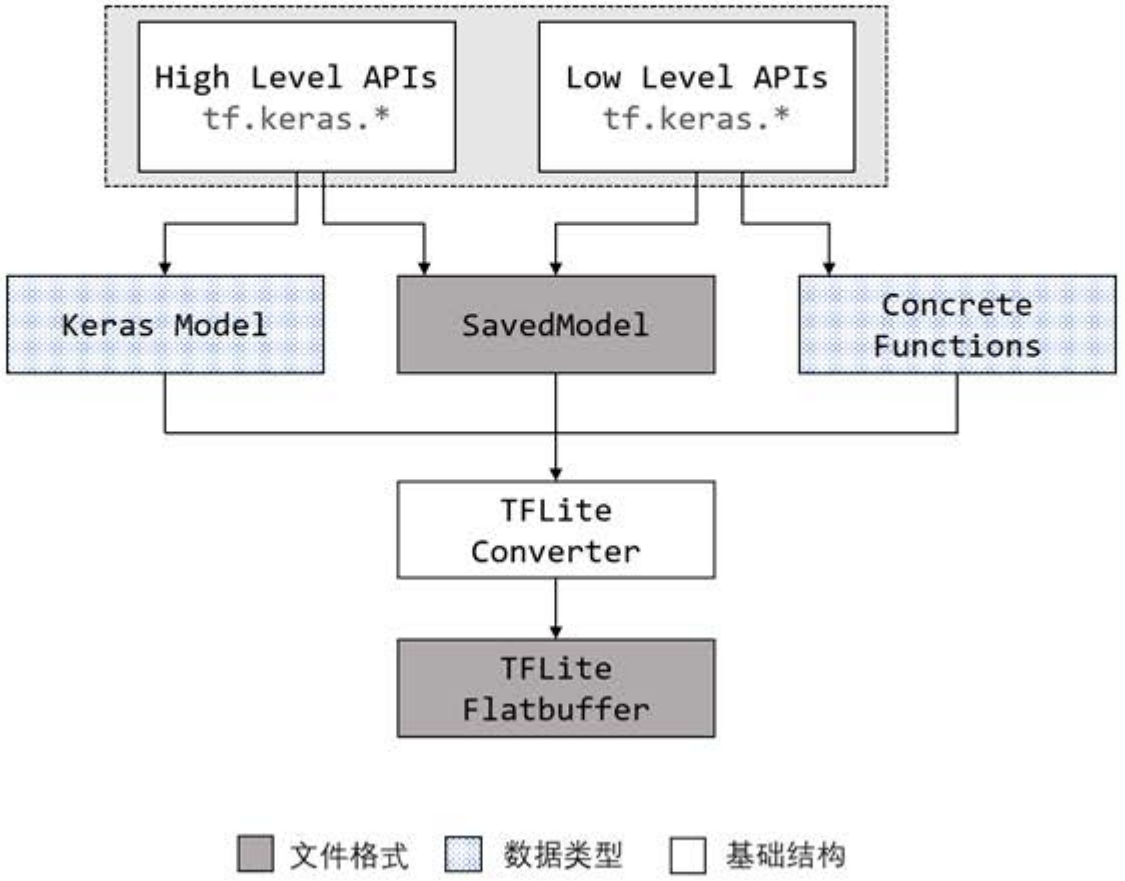
TensorFlow Lite

- TensorFlow Lite是一组能够帮助开发者在移动设备、嵌入式设备和物联网设备上运行TensorFlow模型的工具。使用TensorFlow Lite能够让开发者在网络边缘的设备上执行机器学习，而无需在设备与服务器之间来回发送数据。主要步骤如图。



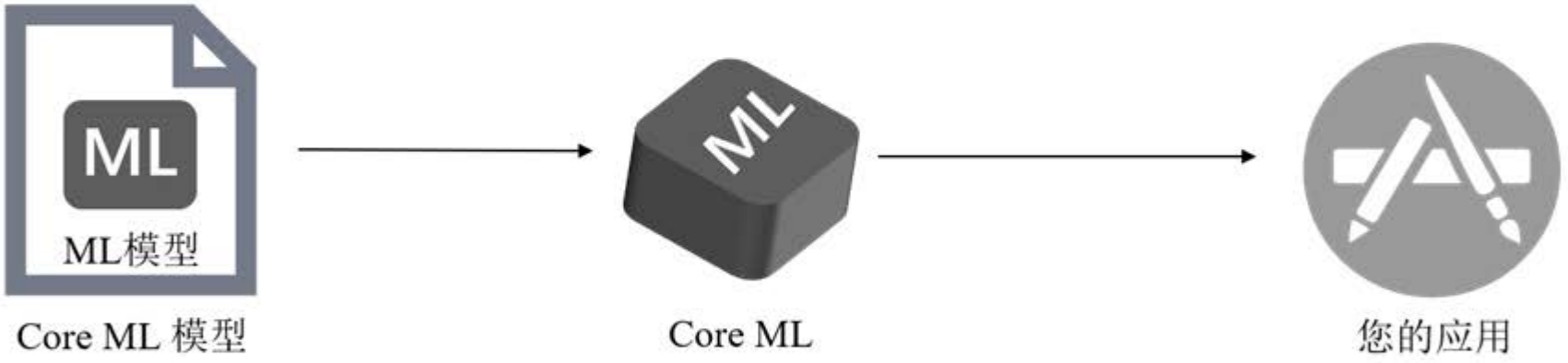
TensorFlow Lite

- TensorFlow Lite使用转换器进行模型转换时的更新过程如图。



Core ML

- Core ML是苹果在MLWWDC 2017开发者大会上推出的一款用于将机器学习集成到APP中的机器学习框架。



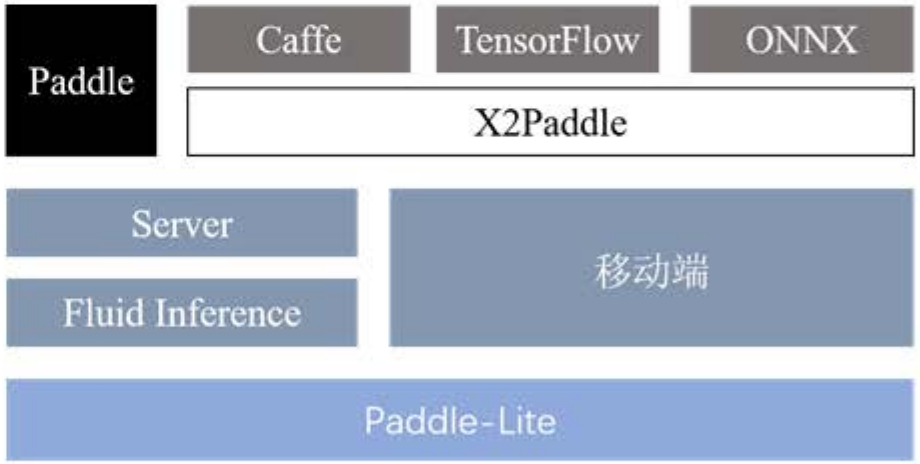
NCNN

- NCNN是为移动平台优化的高性能神经网络推理计算框架，在设计之初便是针对移动平台优化的高性能神经网络推理计算来开发。

Paddle Lite

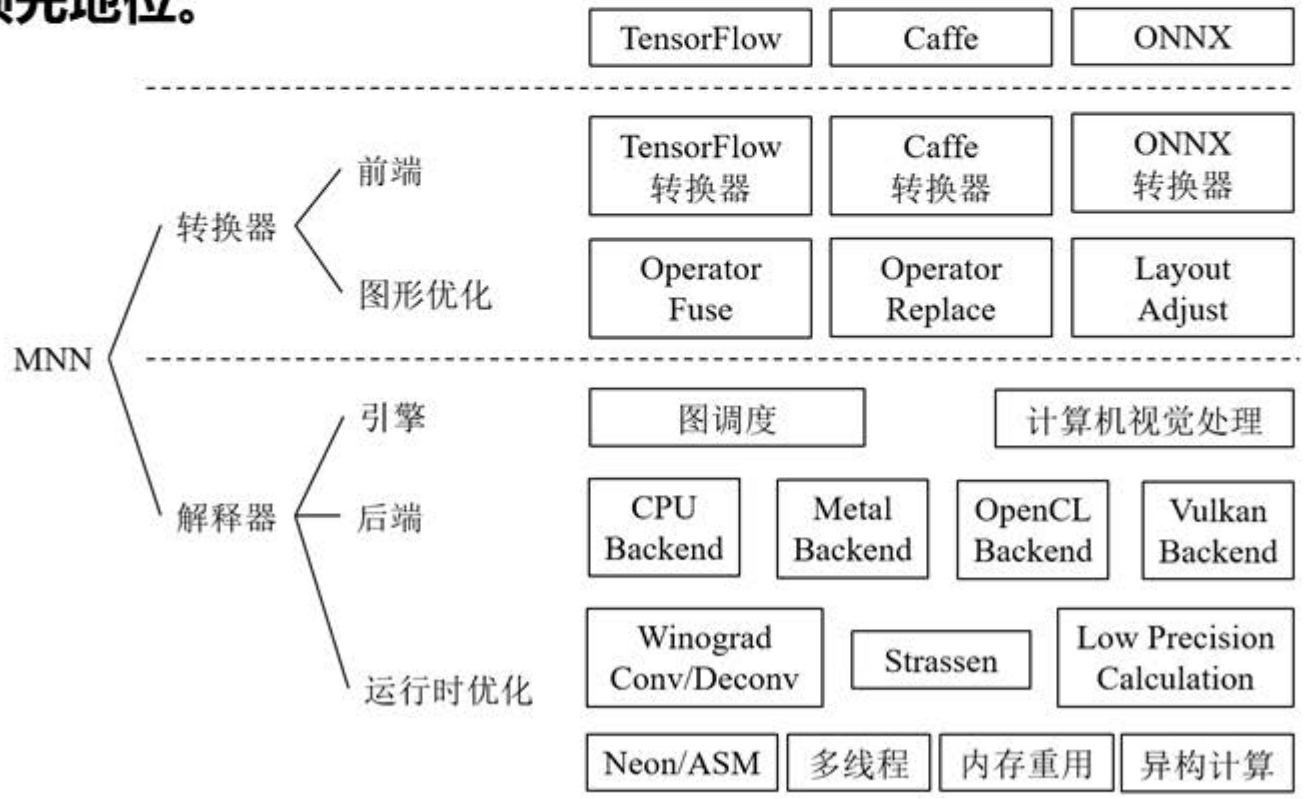
- **Paddle Lite是一个由百度飞浆推出的高性能、轻量级、灵活性强且易于扩展的深度学习推理框架，定位支持包括移动端、嵌入式以及服务器端在内的多硬件平台。**

Paddle Lite主要特点
多硬件支持
轻量级部署
高性能
支持量化计算



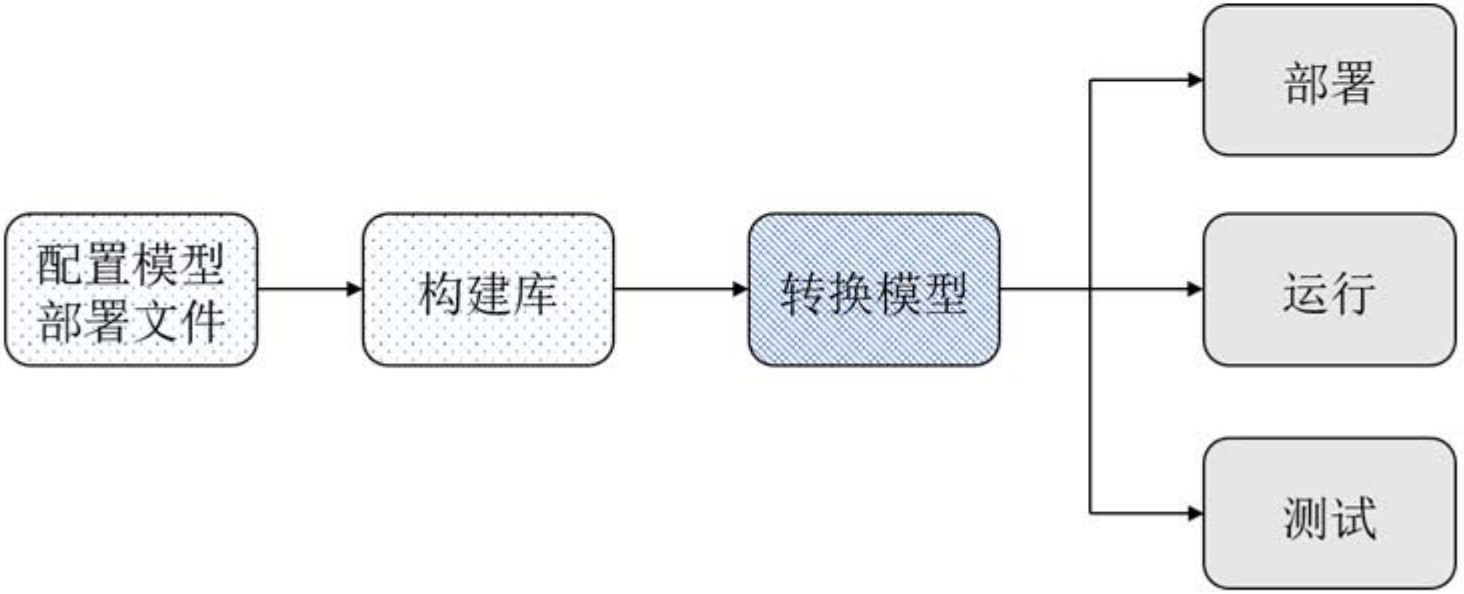
MNN

- **MNN是由阿里巴巴推出的一个轻量级的高效深度学习框架，支持深度学习模型的推理与训练，并在端侧推理与训练性能方面在业界处于领先地位。**



MACE

- MACE（Mobile AI Compute Engine）是小米推出的针对Android、iOS、Linux和Windows设备上针对移动异构计算优化的深度学习推理框架。 MACE工作流程如图。



SNPE

- SNPE（Snapdragon神经处理引擎）是由高通推出的用于深度神经网络执行的高通骁龙（Qualcomm Snapdragon）软件加速运行。SNPE完成了在Snapdragon移动平台上运行神经网络所需的许多繁重工作，可以帮助开发人员提供更多时间和资源来专注于构建新的创新用户体验。

移动端开源机器学习框架

框架	发布机构	发布年份	当前版本	集成模型	部署平台	特性
TensorFlow Lite	Google	2017	2.1.0	Mobilenet_V1_1.0_224_quant、COCO SSD MobileNet v1、Posenet、Deeplab v3、Style prediction model、Text Classification、Mobile Bert、recommendation等	Android、iOS、嵌入式Linux 设备、微控制器	多平台多语言支持、高性能、高效的模型格式、提供模型优化工具和预训练模型
Core ML	Apple	2017		FCRN-DepthPrediction、MNIST、UpdatableDrawingClassifier、MobileNetV2、Resnet50、SqueezeNet、DeeplabV3、YOLOv3、YOLOv3-Tiny、PoseNet、BERT-SQuAD	iOS、MacOS	针对设备性能进行了优化、最大限度上减少内存占用和功耗、确保用户数据的隐私、确保应用在网络连接不可用时保持功能和响应

移动端开源机器学习框架

框架	发布机构	发布年份	当前版本	集成模型	部署平台	特性
NCNN	腾讯	2017	ncnn-20200916		iOS、Android、Windows、Linux	无第三方依赖、跨平台、精细的内存管理和数据结构设计、支持多核并行计算加速、支持 8bit 量化和半精度浮点存储
Paddle Lite	百度	2017	2.7	ERNIE、ernie_tiny、lac、senta_bilstm、emotion_detection_textcnn等	ARM CPU、ARM GPU、Huawei NPU、Intel X86 CPU、NV GPU	多硬件支持、轻量级部署、高性能、支持量化计算

移动端开源机器学习框架

MNN	阿里巴巴	2019	1.1.0	DeepLab、DenseNet、Inception、LaneNet、LFFD、MnasNet、MobileNet、MobileNet SSD、Modified MobileNet SSD、MTCNN、Multi Person MobileNet、SqueezeNet、YOLO(s)	iOS 、Android	轻量级、通用性、高性能、易用性
MACE	小米	2018	1.0.0	convolutional-pose-machines、deeplab-v3-plus、fast-style-transfer、inception-v3、kaldi-models、micro-models/har-cnn、mobilenet-v1/v2、onnx-models、realtime-style-transfer、resnet-v2-50、shufflenet-v2、squeezenet、ssd-mobilenet-v1、vgg16、yolo-v3	Android、iOS、Linux和Windows设备	使用NEON、OpenCL和Hexagon对运行时进行了优化、引入Winograd算法加快卷积操作、支持图级别的内存分配优化和缓冲区重用

移动端开源机器学习框架

SNPE	高通	2017	1.43.0		Snapdragon TM CPU、AdrenoTM GPU、HexagonTM DSP	能够通过C++或者Java将深度神经网络集成到应用程序或者其他代码中、提供用于调试和分析深度神经网络性能的工具

边缘人工智能展望

- 资源友好型边缘AI模型设计
- 计算感知网络技术
- 任务卸载到IoT设备
- 动态预测
- ML集成
- DNN性能指标权衡
- 新型AI模型与技术探索

资源友好型边缘AI模型设计

- 大多数基于深度学习的AI模型都是高度资源密集型的，这意味着需要丰富的硬件资源（例如GPU、 FPGA、TPU）支持的强大计算能力来提高此类AI模型的性能。有一种方式是促进资源感知的边缘AI模型设计。

计算感知网络技术

- 在边缘AI场景中，基于AI的计算密集型应用程序通常运行在分布式边缘计算环境中。因此，需要使用计算感知的高级网络解决方案，以便计算结果和数据可以在不同的边缘节点之间有效地共享。

任务卸载到IoT设备

- 资源受限的移动设备上能够运行的任务种类越来越多，然而，将物联网设备作为潜在的服务器使得在选择前要考虑的因素会更多。为了选择请求卸载到何处，深度Q-learning神经网络，支持向量机以及贝叶斯网络都可能会运用到场景中。

动态预测

- 在动态性非常强的边缘计算场景下，用户随时都可能发生位置改变，这会导致服务器质量的波动。好的解决方案是预测动态变化并事先进行适当的修改。

ML集成

- 通常在机器学习的研究中需要算法充分了解全局系统，但在边缘计算场景中，数据可能产生于许多不同的位置，因此在执行算法之前要一个系统来集中数据，需要在全局信息集中而带来的额外开销和使用部分数据而带来的性能下降之间进行权衡。

DNN性能指标权衡

- 对于一个具有特定任务的边缘AI应用程序，通常有一系列候选DNN模型能够完成任务。然而，由于top-k精度或平均精度等标准性能指标无法反映边缘设备DNN模型推理的运行性能，我们需要研究这些指标之间的权衡，并确定各自的影响因素。

新型AI模型与技术探索

- 随着各种计算密集型的新型应用程序出现，硬件的算力的提升很难跟上满足对计算速度需求的增长，尤其是对于资源受限的移动设备与物联网设备。为了更好的在边缘侧部署深度学习应用，需要更加轻量的网络模型与算法来支持。

本章小测

- 何谓AIoT架构？与传统IoT架构相比有何异同？
- 人工智能能够在哪些问题场景为边缘计算提供优化
- 分布式机器学习中，常用的两种划分学习任务的方式是什么
- 列举四种深度学习在边缘网络的常用应用场景
- 介绍TensorFolw Lite的主要特点
- 在移动边缘网络中使用联邦学习时，需要额外关注哪个指标？谈谈如何优化这个指标