WOTA

51CTO

World Of Tech 2017

全球架构与运维技术峰会

2017年4月14日-15日　北京富力万丽酒店

ARCHITECTURE

WOTA

51CTO

出品人及主持人：

王朝成

饿了么 首席移动架构师

移动端架构演进

WOTA

51CTO

On Device AI
架构及案例分析

WOTA

51CTO

梁宇凌

Google美国总部
高级Android架构师

分享主题：

On-Device AI架构及案例分析

议题

当今火热的AI技术，大多需要服务器端强大的运算能力才能被有效运行起来。然而今天，移动端都在收集大量的用户数据，如何有效地在计算能力薄弱的设备上让AI落地，是一个很有挑战的课题。这次演讲，我尝试结合具体案例做一些相关方面的介绍。

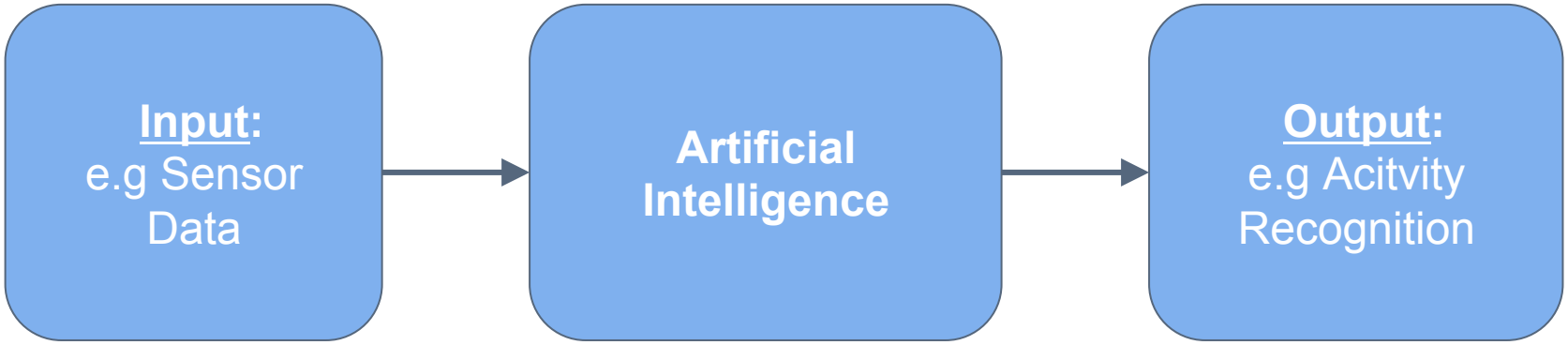
议题

架构　　应用案例　　经验心得

- 主线：
  - 如何合理地让训练，预测在移动上落地
  - 移动端上的AI，都在使用什么算法

什么是机器学习

机器学习是近20多年兴起的一门多领域交叉学科，主要设计和分析一些让计算机可以**无需经过定性**编程，能够**自动“学习”**的算法。



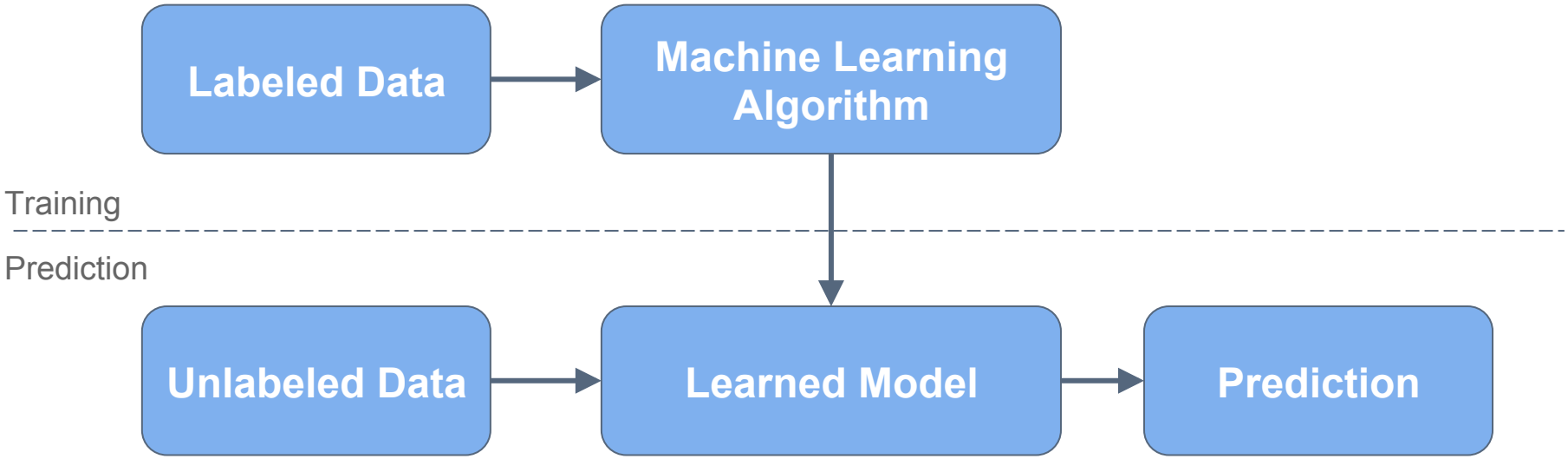
为什么要关注机器学习

- 能有效根据数据进行分类和预测。
  - 面对海量数据，无法一一通过人工规则定制业务逻辑
  - 人工规则能应付主要应用场景，还有大量的长尾场景无法手动满足
  - 数据在不断更新，手动更新人工规则满足新数据太昂贵

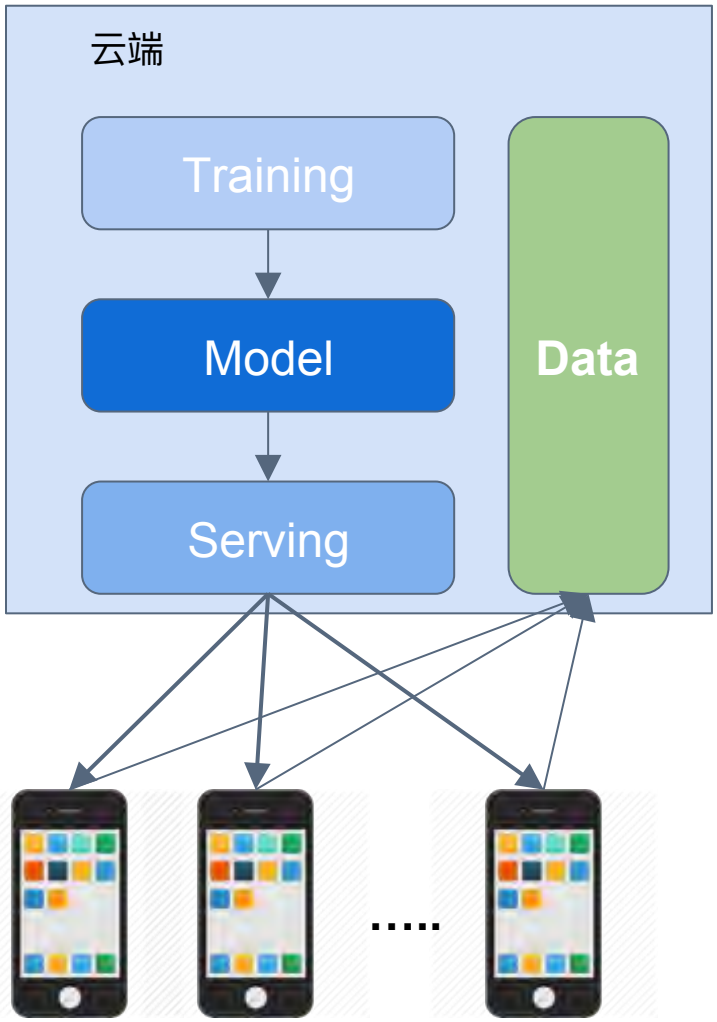
为什么要关注机器学习

- 能有效根据数据进行分类和预测。
  - 面对海量数据，无法一一通过人工规则定制业务逻辑
  - 人工规则能应付主要应用场景，还有大量的长尾场景无法手动满足
  - 数据在不断更新，手动更新人工规则满足新数据太昂贵
- 能让你的产品做到真正的个人定制。
  - 模型根据个人数据，产生真正属于用户本身的预测
  - On Device AI能提高反应速度和保护用户隐私

AI/ML 常用架构


Labeled Data
Machine Learning Algorithm
Training
Prediction
Unlabeled Data
Learned Model
Prediction

AI 架构演化



训练和预测都在云端进行

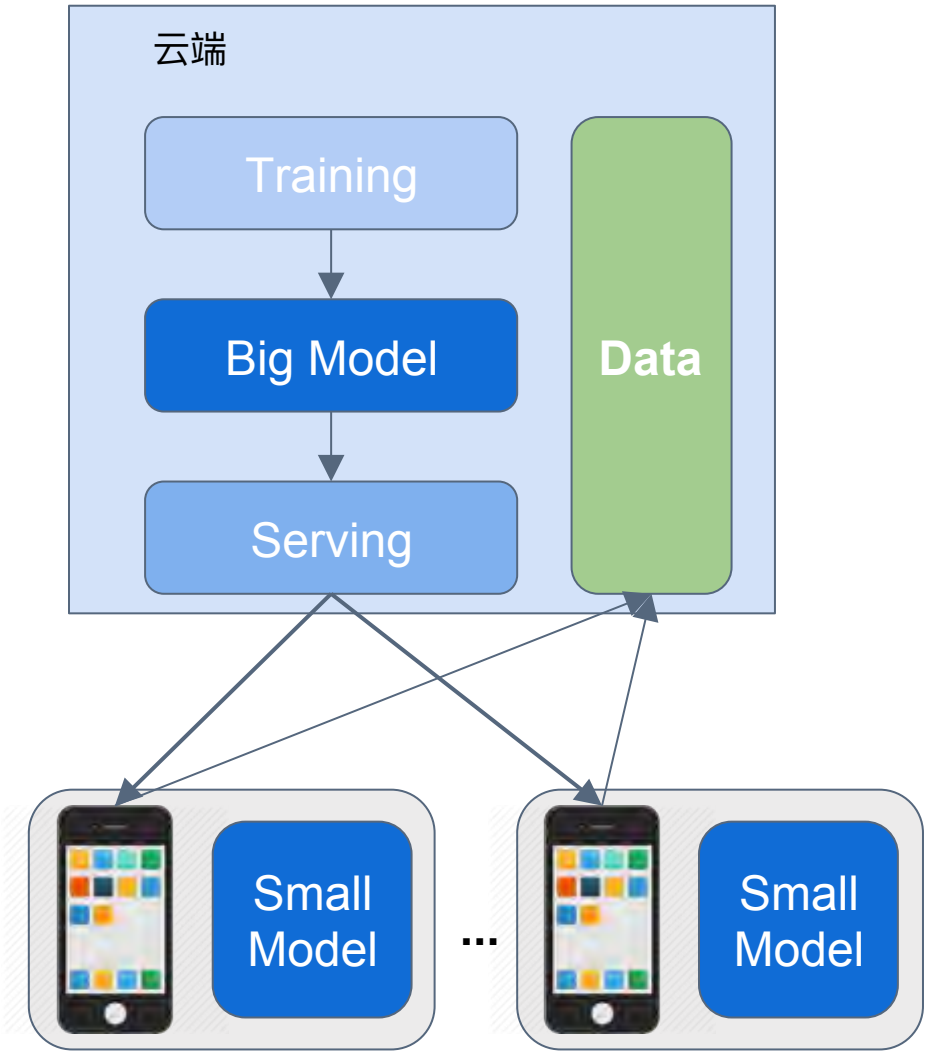
优点：

- 云端有海量存储量和计算量
- 模型迭代和发布低延迟
- 方便实现Experiment

缺点：

- 需要随时联网
- 预测速度响应慢
- 数据上传浪费带宽

AI 架构演化



训练和复杂预测在云端进行，简单预测在客户端进行

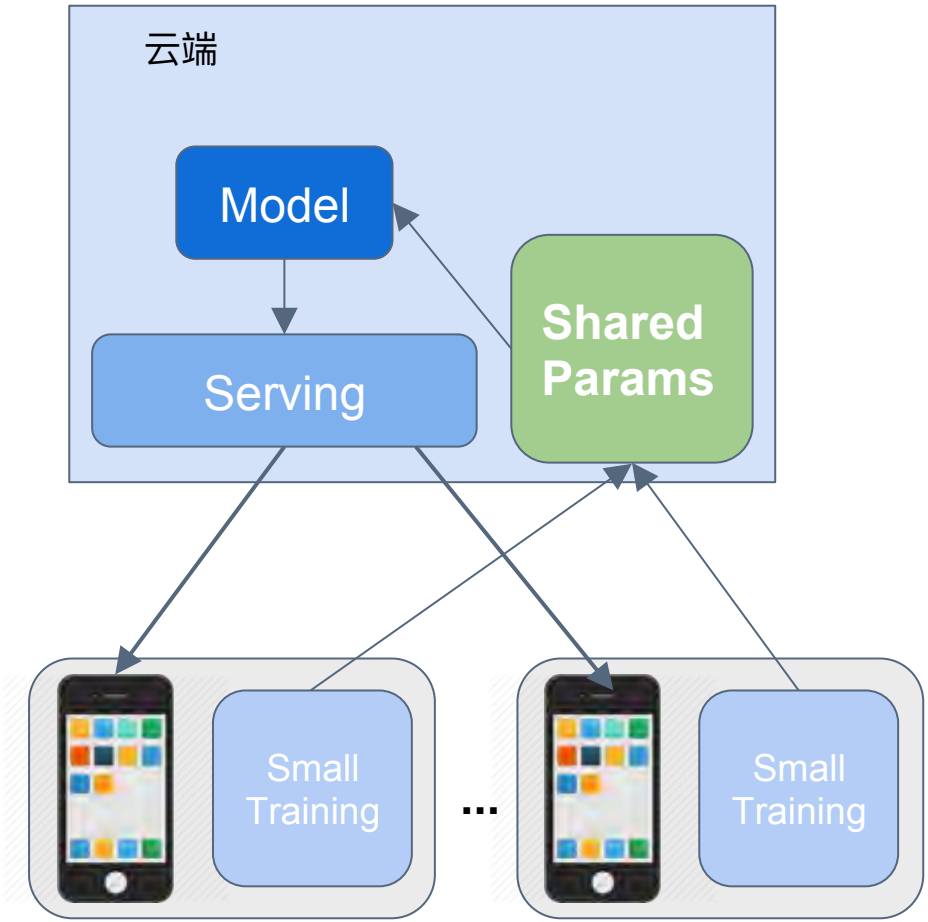
优点：

- 简单模型响应速度更快
- 对联网依赖性减弱

缺点：

- 两套模型，架构更复杂
- 客户端存储空间要求增大
- 客户端模型需要定期更新
- 数据上传依然浪费带宽

AI 架构演化



预测在云端或客户端进行，训练在客户端进行

优点：

- 用户隐私性保护最佳
- 数据上传量大幅减少，只传递参数修正
- 云端计算量大幅减少，只负责更新参数

缺点：

- 客户端数据质量不一，存在平衡性等各种问题
- 海量数据并行更新模型，架构更复杂
- 客户端计算量要求增大

AI 常用架构，Cloud vs On-Device

- 目前Machine Learning需要的计算量还非常巨大，因此大部分还停留在云端训练和预测。

AI 常用架构，Cloud vs On-Device

- 目前Machine Learning需要的计算量还非常巨大，因此大部分还停留在云端训练和预测。
- 移动端上进行ML，还处于初级阶段，但有云端不能比拟的优势：
  - 用户隐私得以保障（照片，短信，个人位置）
  - 实时反应，无需连接网络（外国旅游时的文字翻译）
  - 海量数据，在移动端进行精简处理后，能大幅减少服务器端存储和计算压力

AI 常用架构，Cloud vs On-Device

- 目前Machine Learning需要的计算量还非常巨大，因此大部分还停留在云端训练和预测。
- 移动端上进行ML，还处于初级阶段，但有云端不能比拟的优势：
  - 用户隐私得以保障（照片，短信，个人位置）
  - 实时反应，无需连接网络（外国旅游时的文字翻译）
  - 海量数据，在移动端进行精简处理后，能大幅减少服务器端存储和计算压力
- 然而，移动端AI面临着更明显的劣势：
  - 计算能力差
  - 容量限制（ImageNet的model如果不经过裁剪，要96MB）
  - 电池容量有限

On Device AI案例分析

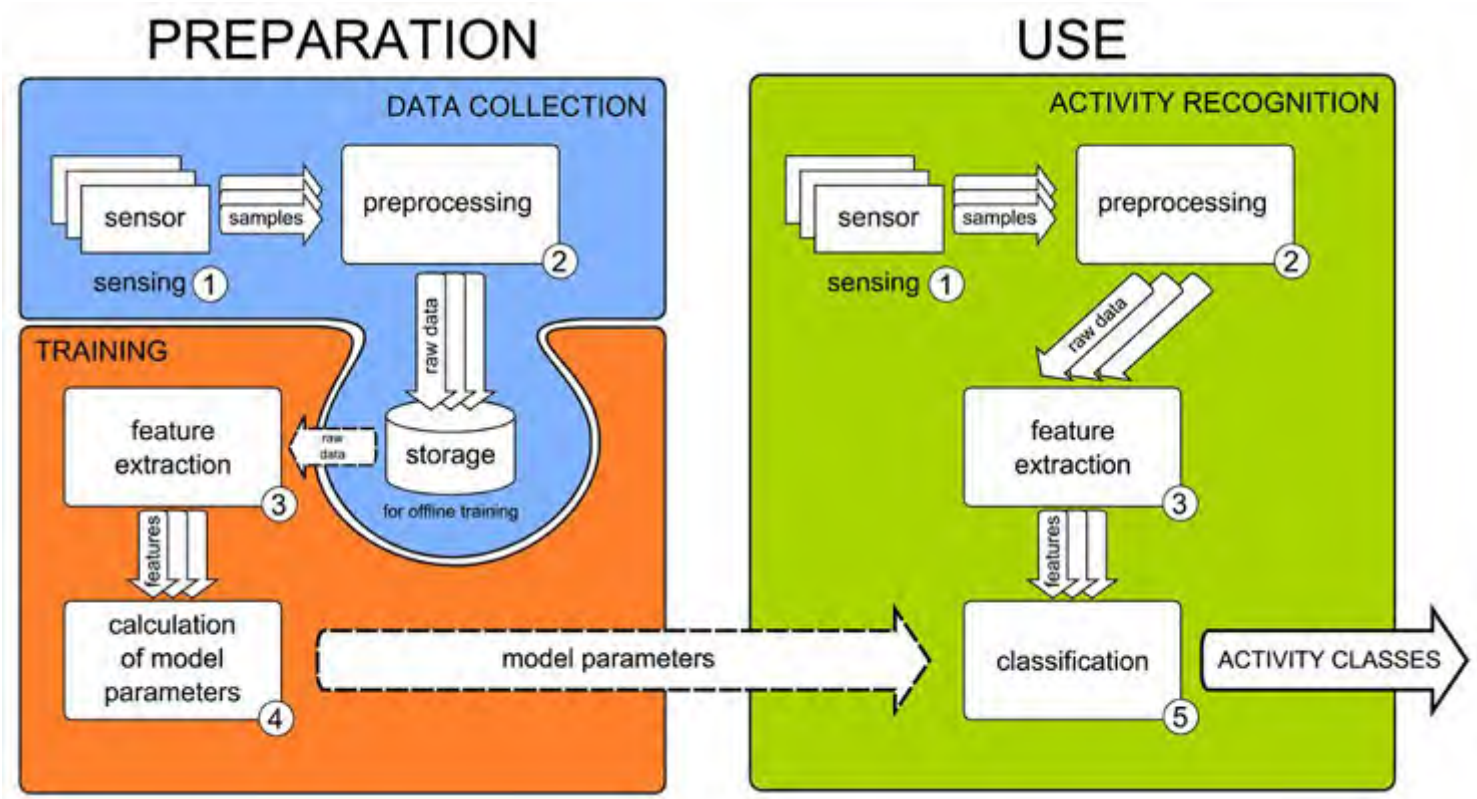
案例 1：用户行为检测

On Device AI案例分析

案例 1：用户行为检测

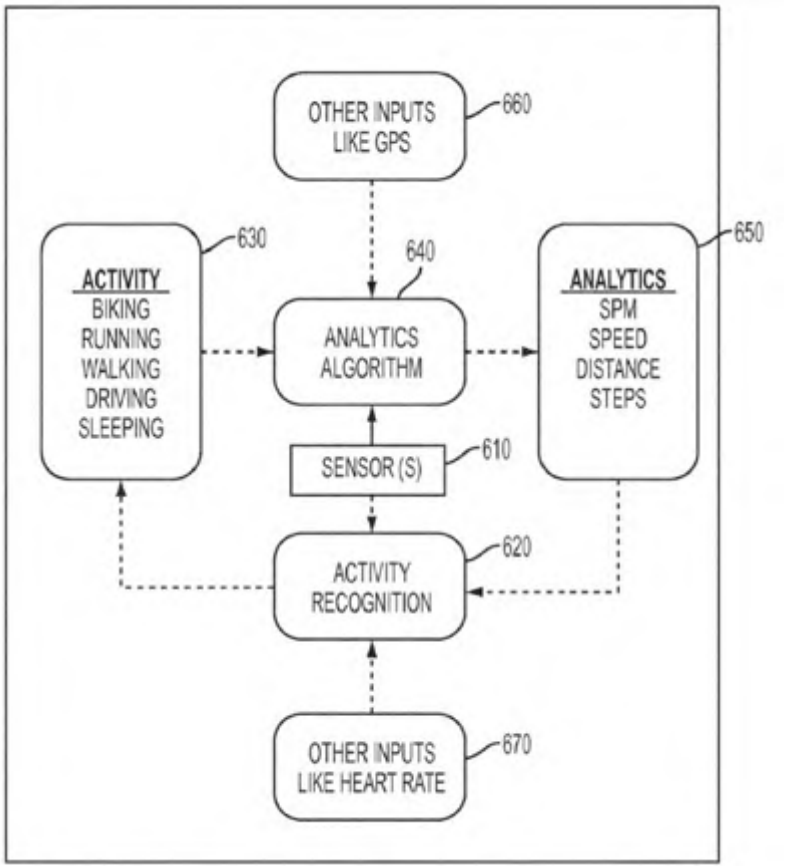
On Device AI案例分析

案例 1：用户行为检测



On Device AI案例分析

案例 1：用户行为检测

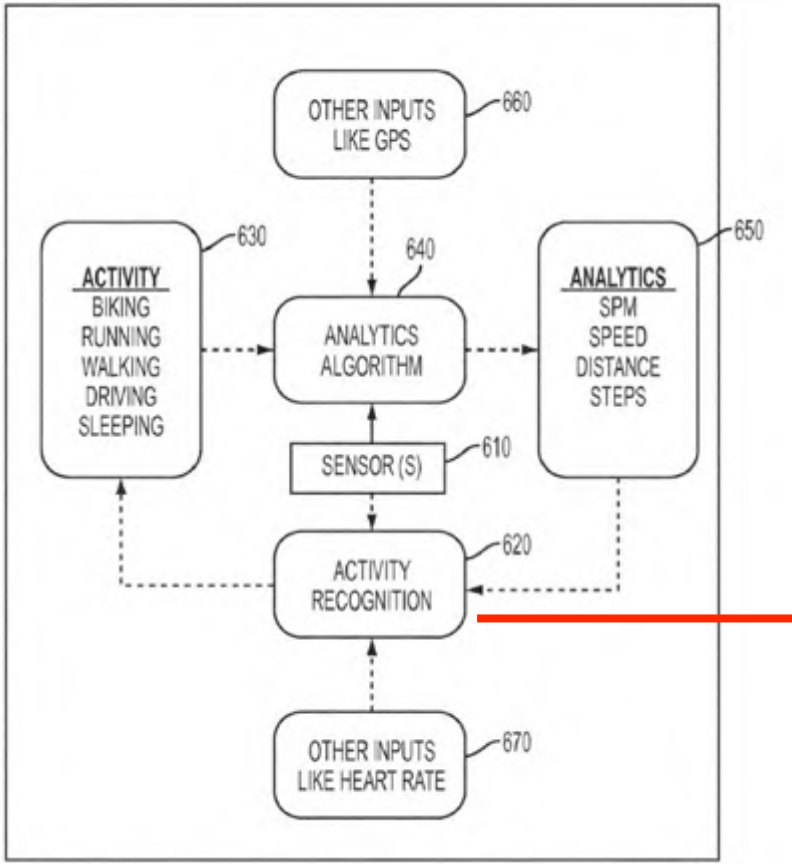


InvenSense 专利 EP2868273 A1

- 也许是最早被广泛应用的On-Device AI

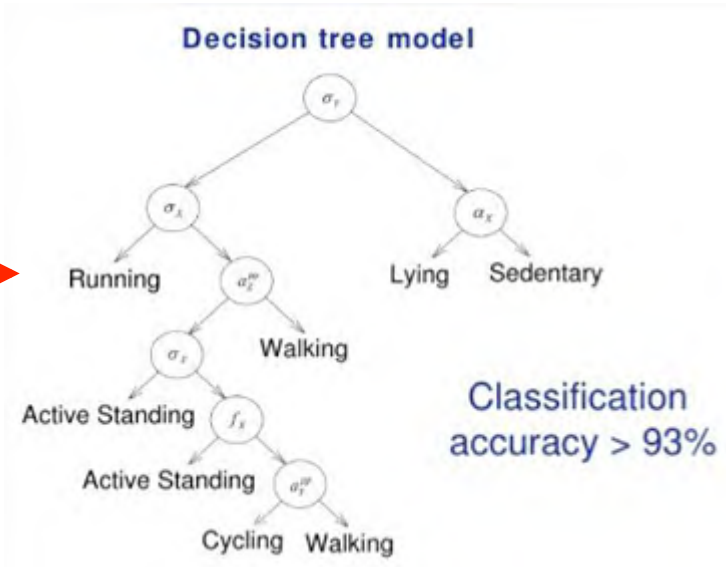
On Device AI案例分析

案例 1：用户行为检测



InvenSense 专利 EP2868273 A1

- 也许是最早被广泛应用的On-Device AI
- 主要利用传统机器学习算法，如决策树
- 算法最早被继承到硬件中



On Device AI案例分析

案例 2: Google翻译

On Device AI案例分析

案例 2: Google翻译

- 传统CV和深度学习结合，用CV找出文字和覆盖，深度学习进行文字识别

- 通过控制训练数据样本，调节模型，找到质量和大小的最佳结合点

- 算方法加速性能

https://research.googleblog.com/2015/07/how-google-translate-squeezes-deep.html

On Device AI案例分析

案例 3：Android Wear手表系统上的智能回复(Smart Reply)

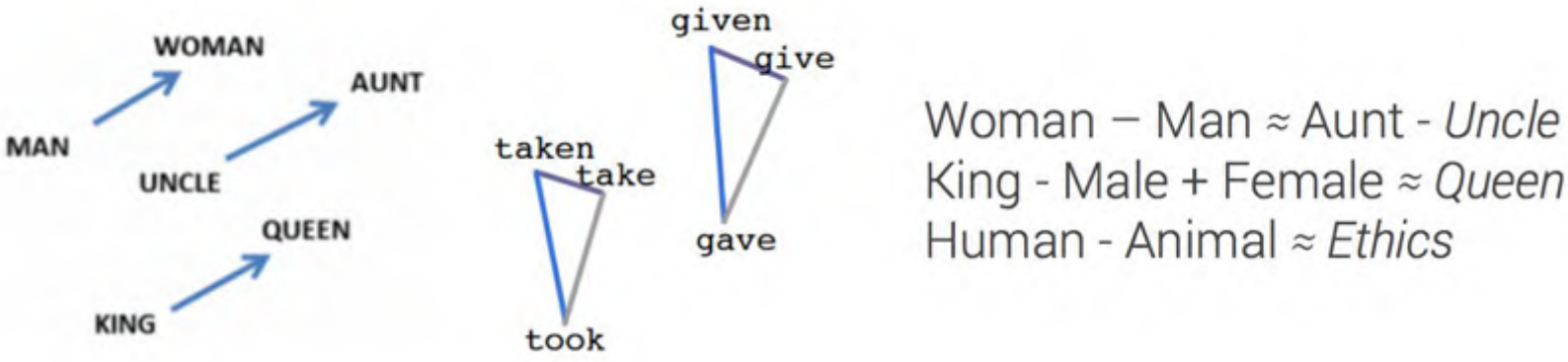
https://research.googleblog.com/2017/02/on-device-machine-intelligence.html

On Device AI案例分析

案例 3：Android Wear手表系统上的智能回复(Smart Reply)

流行的（非移动端）Retrival Based的做法：

- 训练时，利用深度学习训练出词和句子的向量 (word & sentence embeddings).
- 预测时，利用深度学习模型从回复句子集合中找到分数最高的若干回复

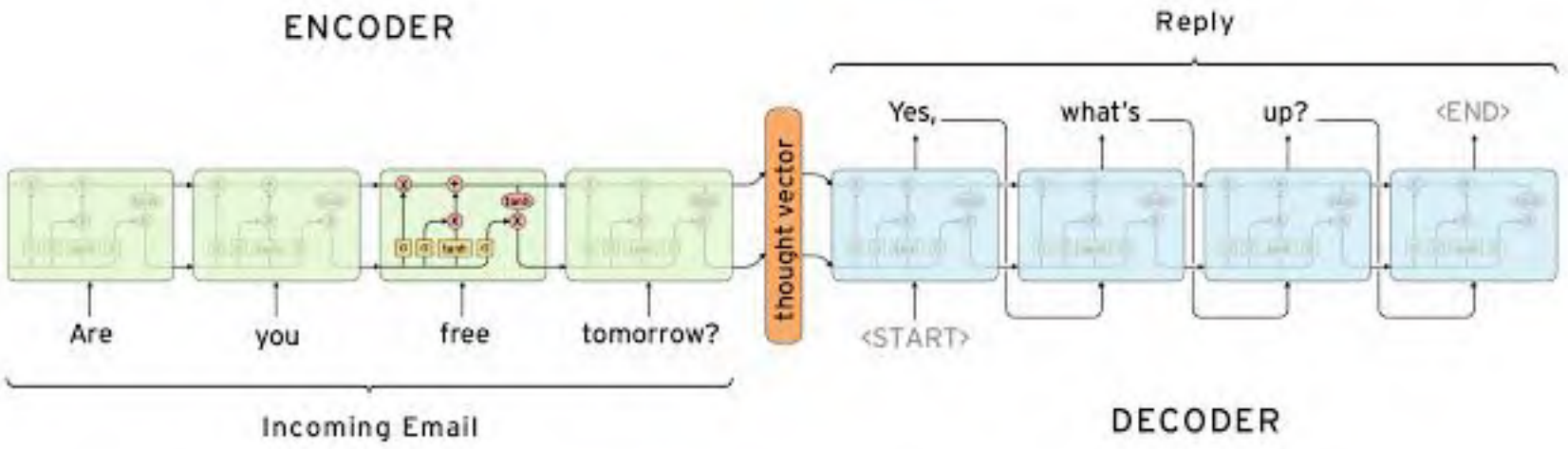


On Device AI案例分析

案例 3：Android Wear手表系统上的智能回复(Smart Reply)

流行的（非移动端）生成型做法：

- 利用深度学习里面encoder-decoder架构，自动生成回复。

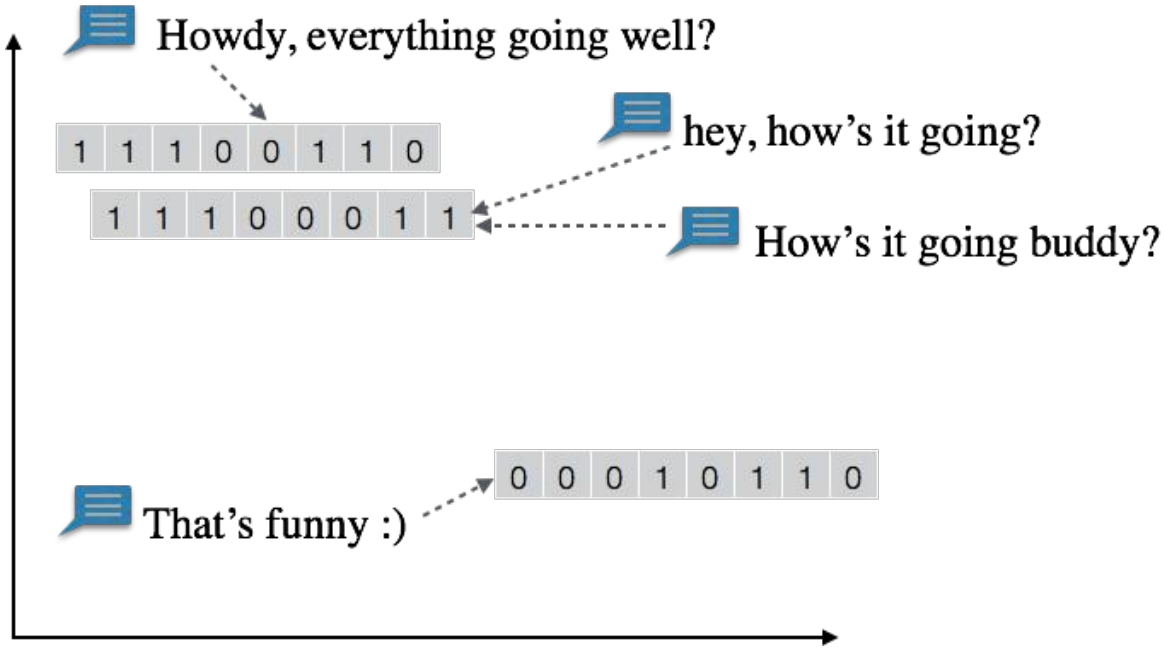


On Device AI案例分析

案例 3：Android Wear手表系统上的智能回复(Smart Reply)

Android Wear智能回复的实际做法

- 利用LSH哈希算法，对类似句子进行映射。
- 映射结果中，相似的句子的映射结果距离相近。

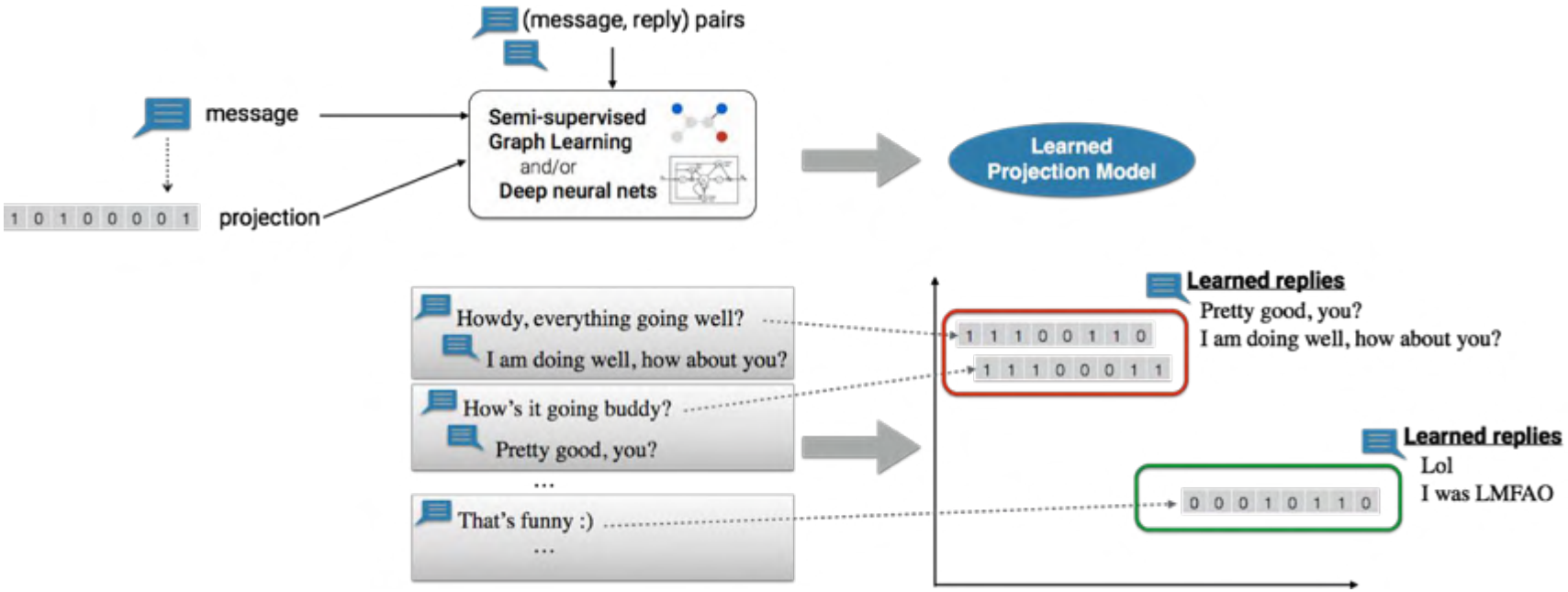


On Device AI案例分析

案例 2：Android Wear手表系统上的智能回复(Smart Reply)

Android Wear智能回复的实际做法

- 使用graph-based的半监督学习算法预测回复短语



On Device AI案例分析

案例 4：移动版Tensorflow实现Android端计算机视觉应用

图片识别

美术风格转移

物体检测

On Device AI案例分析

案例 4：移动版Tensorflow实现Android端计算机视觉应用

- 原生Inception V3要93MB
- 两种主要优化：
  - 经过计算图优化（把无用节点剪除）
  - quantization，利用8位整数代替模型中的32位的浮点数
- 优化后的Inception V3模型只需要24MB，V1只需要7MB，TF库只增加2MB的binary容量

心得

- 挑合适的算法，而不是最炫的算法
  - 没有Silver Bullet，必须因地制宜，随机应变
  - 决策树在mobile上就比深度学习要快很多，要小很多。

心得

- 挑合适的算法，而不是最炫的算法
  - 没有Silver Bullet，必须因地制宜，随机应变
  - 决策树在mobile上就比深度学习要快很多，要小很多。

- 算法很廉价，数据是王道。
  - 算法主要由学术界在推动，大公司在打包和优化。一般公司应该专注在结合具体应用上。
  - 一个问题是否能用AI解决，首先是有没有数据，够不够数据，然后是数据在预测阶段的完备性，最后是数据接下来在能不能形成闭环
  - 吴恩达曾经说过，百度有时候为了收集数据，会专门去发布一些小的产品或功能）

心得

- 挑合适的算法，而不是最炫的算法
  - 没有Silver Bullet，必须因地制宜，随机应变
  - 决策树在mobile上就比深度学习要快很多，要小很多。
- 算法很廉价，数据是王道。
  - 算法主要由学术界在推动，大公司在打包和优化。一般公司应该专注在结合具体应用上。
  - 一个问题是否能用AI解决，首先是有没有数据，够不够数据，然后是数据在预测阶段的完备性，最后是数据接下来在能不能形成闭环
  - 吴恩达曾经说过，百度有时候为了收集数据，会专门去发布一些小的产品或功能）
- 灵活运用移动端的硬件加速
  - 高通，英伟达，英特尔都在推移动AI芯片
  - 灵活运用，例如无人车的芯片就能用在安防摄像头上

总结

架构　　应用案例　　经验心得

- 主线：
  - 如何合理地让训练，预测在移动上落地
  - 移动端上的AI，都在使用什么算法


WOTA
51CTO
Thank you !