AiCon
全球人工智能与机器学习技术大会

AI驱动下的移动输入革新之路

姚从磊
CTO, kika Tech

主办方 Geekbang. 极客邦科技 InfoQ

kika

TABLE OF CONTENTES


kika



Go
2018


2018





2018



3 Needs

Globalization

- 140+ Countries
- 173 Languages

Localization

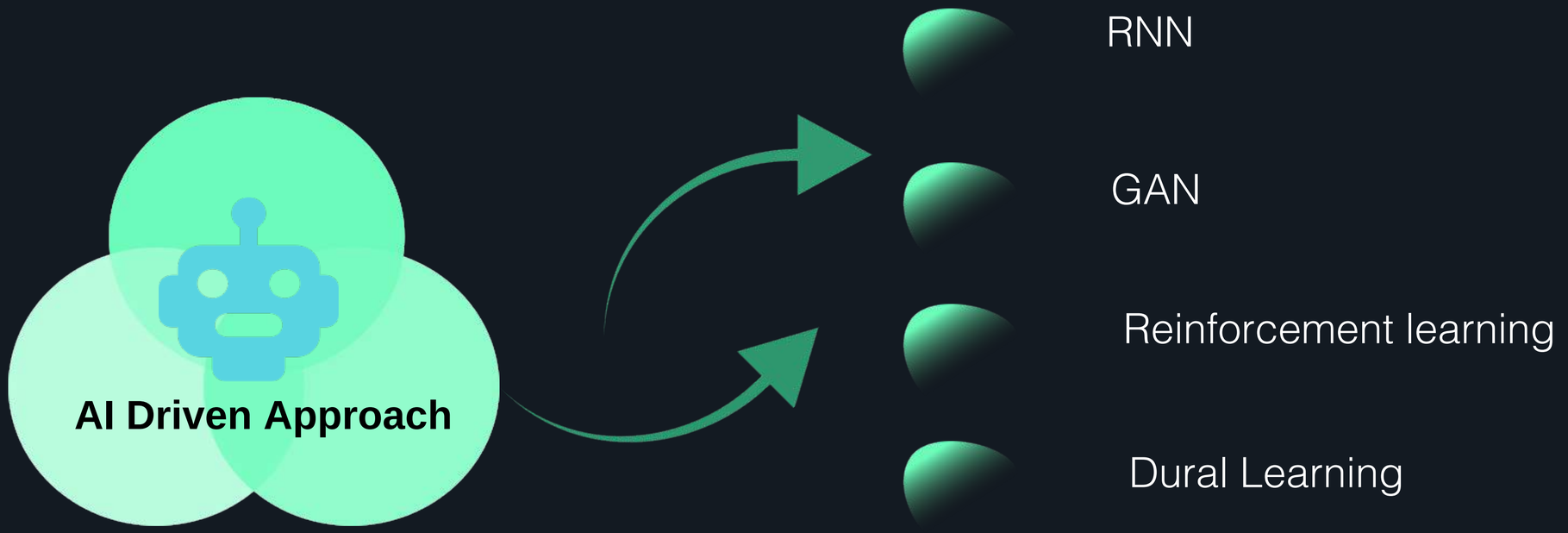
- Location Oriented
- Simultaneously Multilingual Typing

Multimedia

- Text/Emoji
- Voice /Gif/Stickers

kika

How to Fulfill the Needs ?

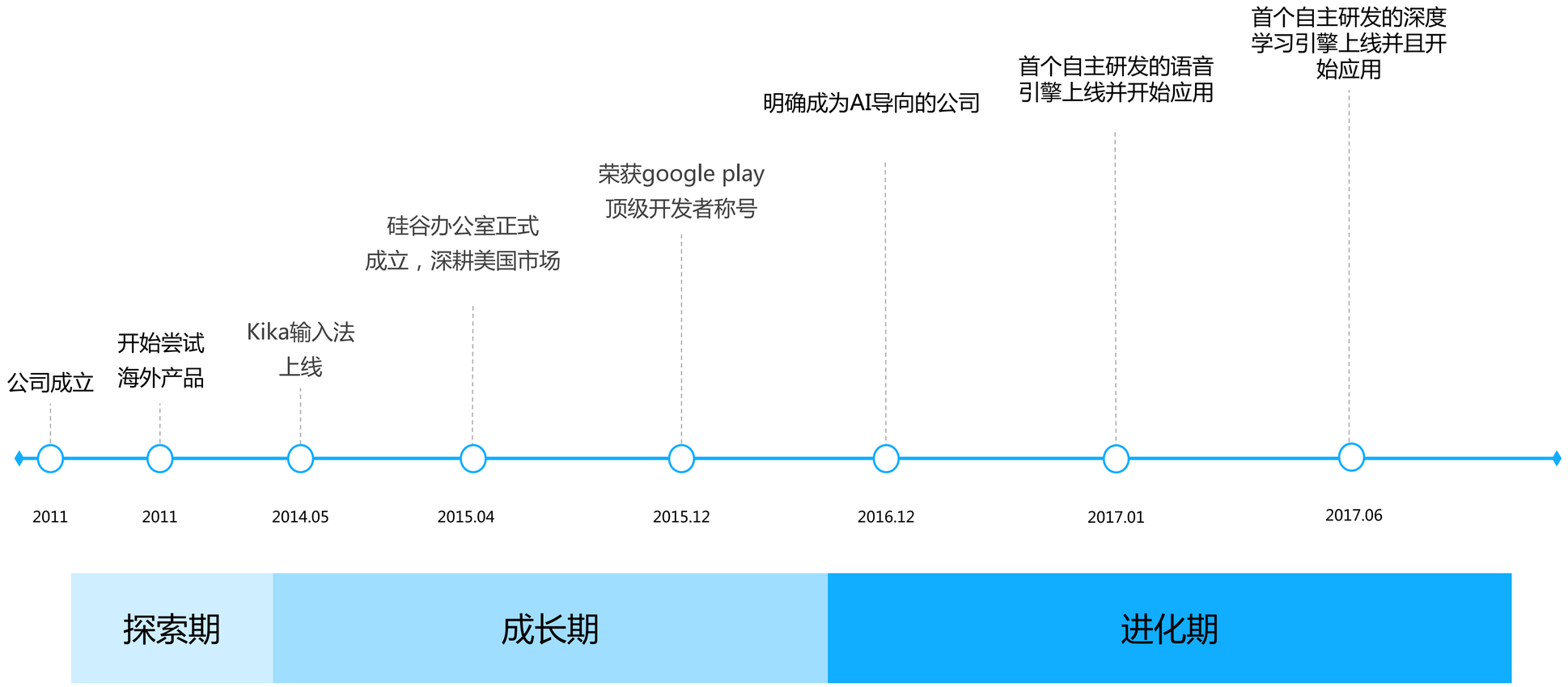


kika

TABLE OF CONTENTES

kika

Kika持续成长进化，不断积累AI势能



kika

全、准、快的沟通

全

- 文本
- 语音
- 图像

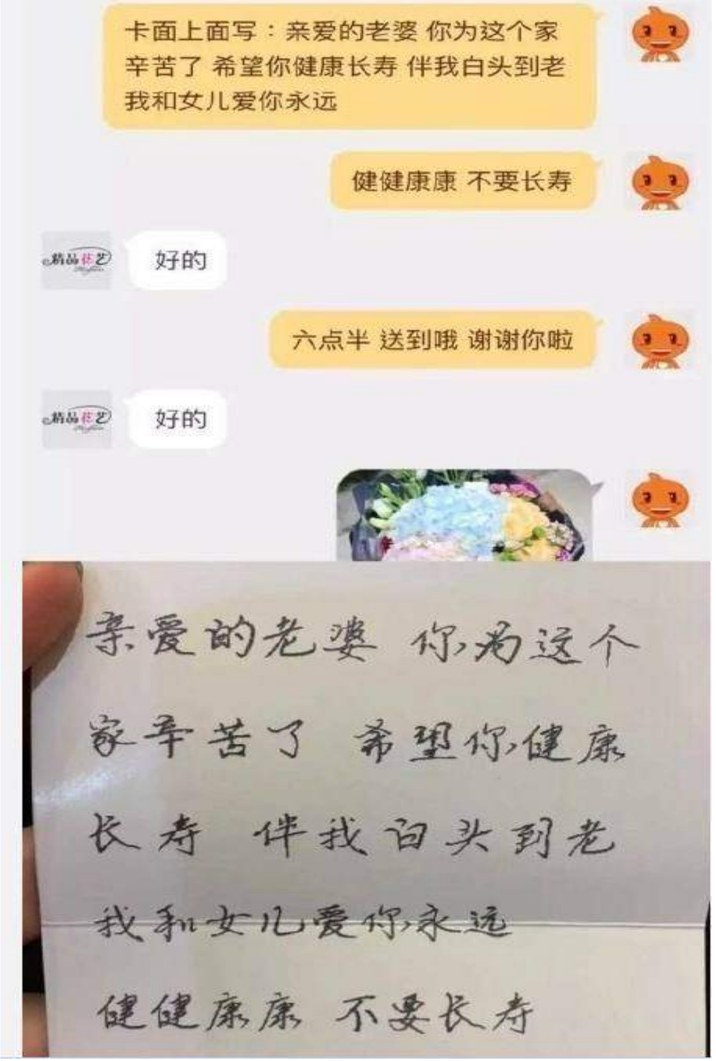
准

- 准确输入
- 准确输出
- 没有误解

快

- 输出快
- 输入快
- 效率为王

kika


卡面上面写：亲爱的老婆 你为这个家辛苦了 希望你健康长寿 伴我白头到老 我和女儿爱你永远
健健康康 不要长寿
好的
六点半 送到哦 谢谢你啦
好的
亲爱的老婆 你为这个家辛苦了 希望你健康长寿 伴我白头到老 我和女儿爱你永远 健健康康 不要长寿


买家留言：刻一个黑黑的风
订单编号：
更多
商家：
更多
一个黑黑的风

kika

kika

“3A” 引擎

- Engine Alps – Kika Keyboard Engine
- Engine Appalachian – Kika Voice Engine
- Engine Andes – Kika Content Recommendation Engine

kika

TABLE OF CONTENTES

kika

The Best Keyboard Engine

反应快

- 响应时间＜60ms
- 键盘弹起＜60ms
- 内容推荐＜100ms

预测准

- 滑行输入
- 键盘输入
- 高度容错

个性化

- 用户习惯个性化
- 输入场景个性化
- 地理位置个性化

kika

Challenges

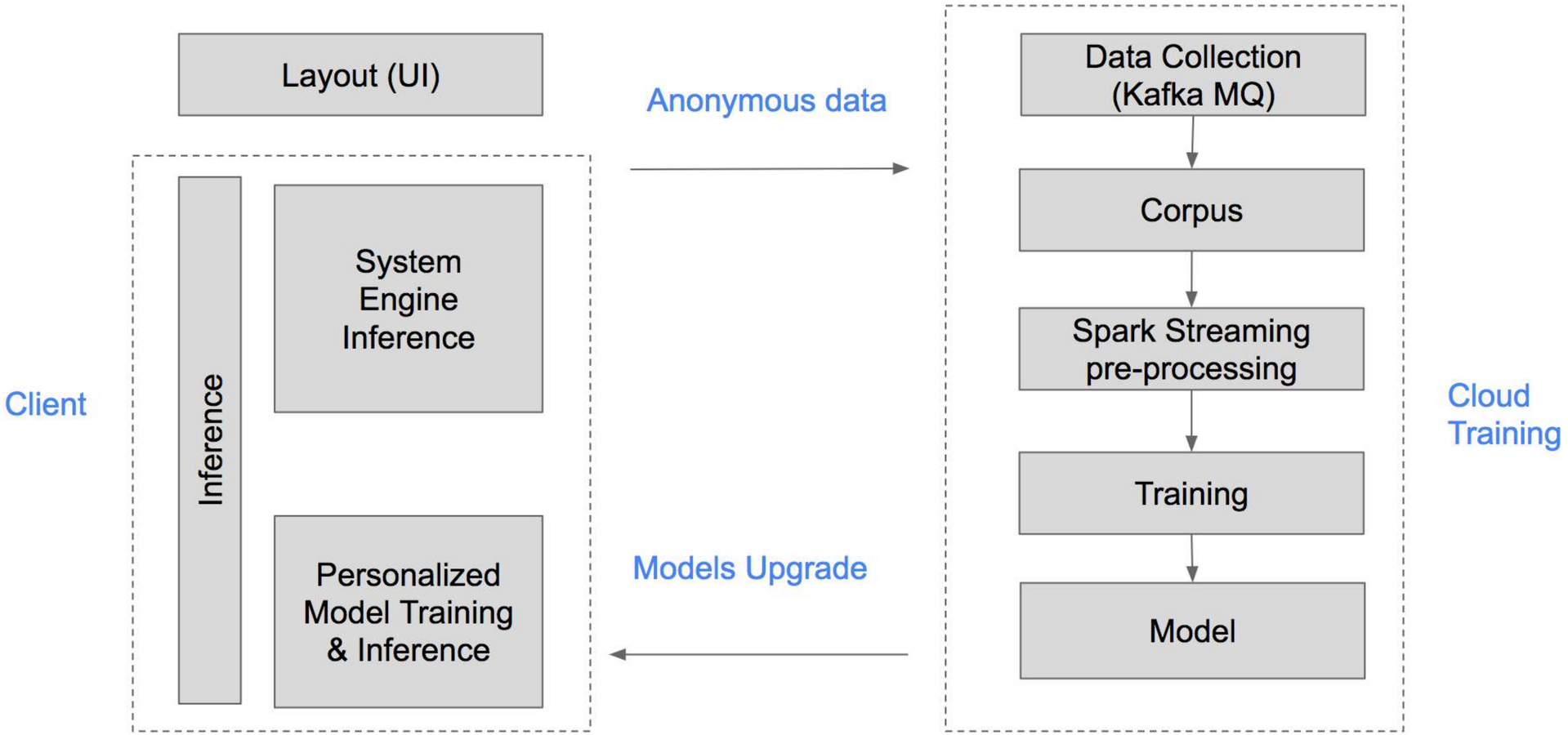
反应快

预测准

个性化

kika

Architecture



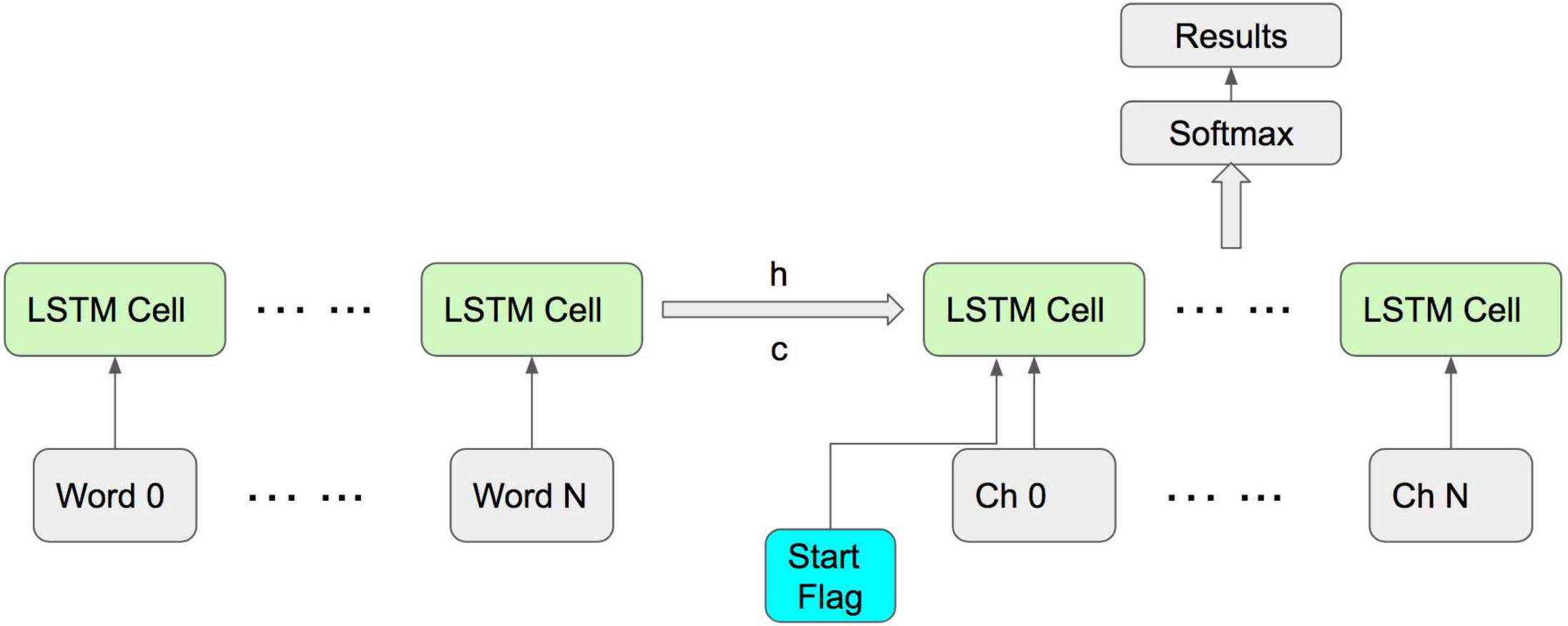
kika

Cloud Training

- 关键问题
  - 数据中的大量噪音
    - 语言检测和概率生成模型，过滤数据噪音
  - 模型部署的目标 – Android 平台
    - 平台自带轻量化需求：输入法，安装包体积受限，内存受限，Android Go
    - 解决方案：基于 TF Mobile & TF Lite 二次开发，并优化模型
  - 模型结构
    - 同时解决候选词(单词+词组+Emoji)推荐和按键纠错的问题

kika

Cloud Training – Model Structure



kika

Model – Compress & Quantize

- 稀疏表示与学习
  - 目的：压缩 word/ch embedding矩阵及输出端softmax向量矩阵
  - 原理：巨大的向量矩阵 => 少量的过完备基向量组合
  - Feature
    - 存储空间压缩
    - Inference计算量压缩
    - 过完备基向量可自动学习获得
    - 精度损失小

kika

Model – Compress & Quantize

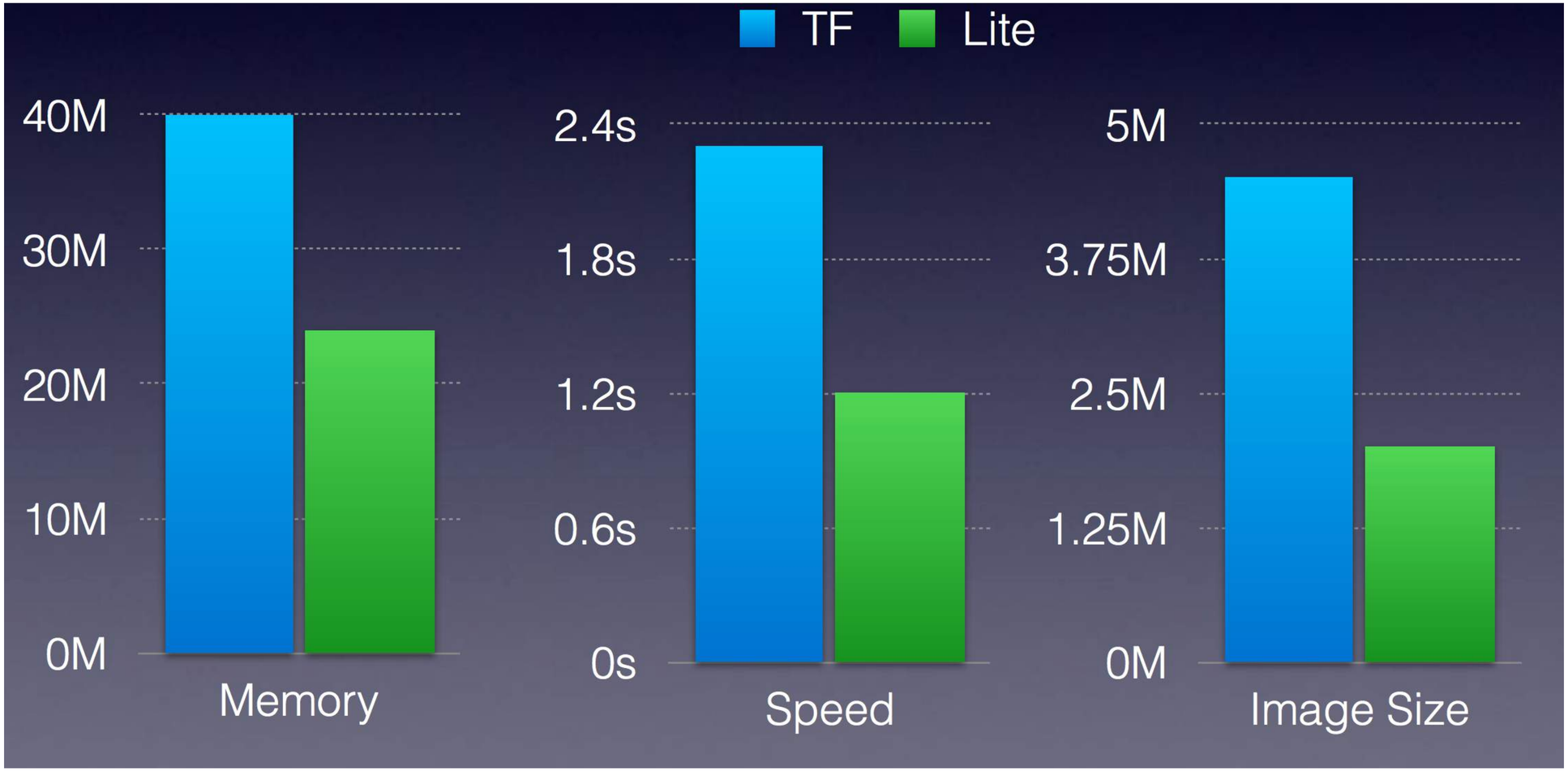
- 自适应学习量化
  - 基于 kmeans 进行自适应量化学习，与 tf 官方量化方法比，精度上略微占优
  - 去除 tf 官网的 Quantize/Dequantize Ops 进一步压缩客户端 so 的体积
  - 支持 dynamic rnn
  - 支持 node name 的混淆，保护知识产权
  - 支持 tf lite 的 flatbuffer 格式的 graph

kika

Runtime Inference Optimization

- 计算量控制
  - 如何在保证效果的前提下优化模型结构
- TF大量的优秀的底层库：
  - AOT (ahead of time): XLA(Accelerated Linear Algebra)，Eigen (tf的底层矩阵运算库)
  - Flatbuffer (降低内存占用) – 推荐基于 TF lite 做 inference 开发
  - 缺点：上层不够友好，需要大量的二次开发与深度优化以控制性能、内存以及 so 体积

kika


TF
Lite
40M
30M
20M
10M
0M
Memory
2.4s
1.8s
1.2s
0.6s
0s
Speed
5M
3.75M
2.5M
1.25M
0M
Image Size