DataFunSummit

# 自然语言处理

## 在线峰会

## 机器翻译与同传 论坛


2021.07.10（周六） 09：00~17：30

2021

2021

有道 youdao | DataFunSummit

# AI技术在有道词典笔上的应用实践


程　桥 资深算法工程师
张广勇 高性能计算专家

# 目录

CONTENTS

# 01 题目

Subject

## 有道词典笔

# 有道词典笔


BUSINESS
By WANG ZHUOQIONG and LIU KUN in Wuhan
Yeast Co Ltd, one of the
's largest yeast makers and a
solution provider, is bank-
ring facilities in
further expand
overseas mar-
arch and develop-
el Yeast, said
in Egypt is
ational from
ly built factory
st yeast extract
o the European
costs. The factory
production capac

# 有道词典笔

- 跨行整句扫描
  - 扫描准确率98%
- 整句翻译
  - 离线也可以翻译
- 超快点查
- 互动点读


# 有道词典笔上的 AI 技术

- 扫描和点查技术
- 离线翻译
- 高性能端侧机器学习计算库（EMML）

02

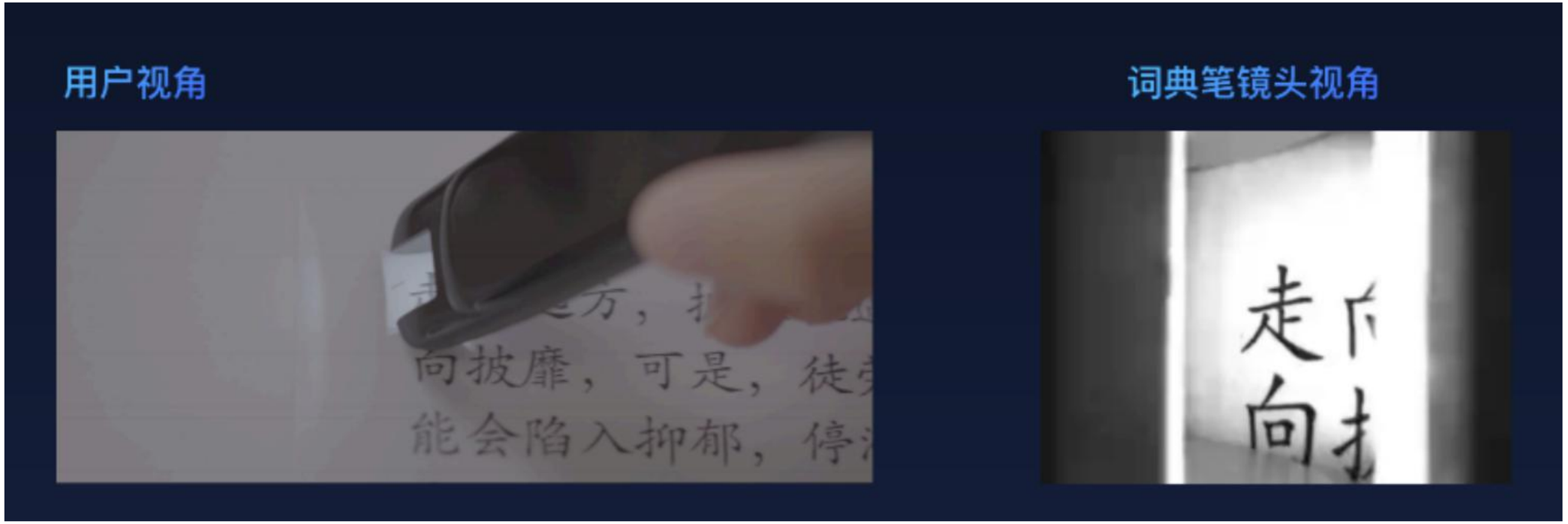
# 题目

Subject

## 扫描和点查

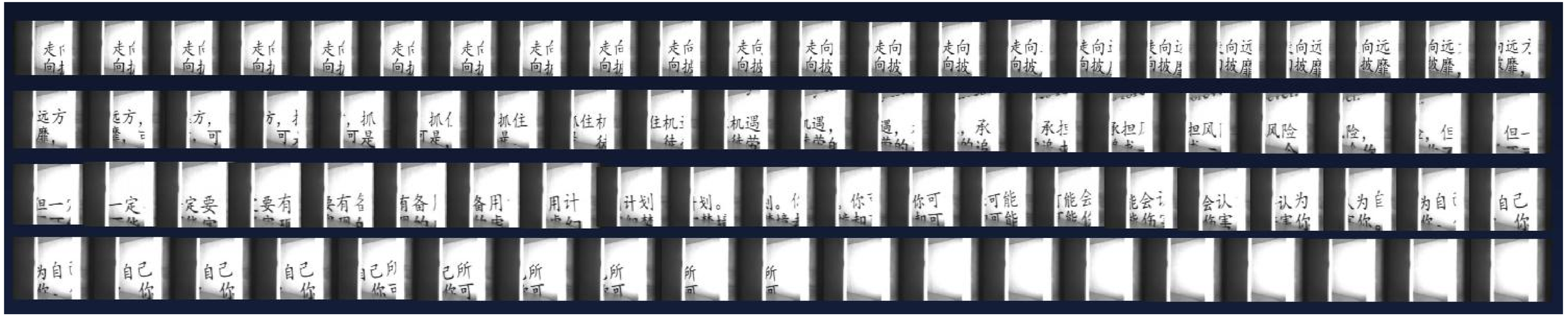
# 扫描识别

- 扫描识别和常见的字符识别场景不一样



# 扫描识别

- 一秒钟100张图像
  - 算法需要从快速从拍摄的图像中提取文字

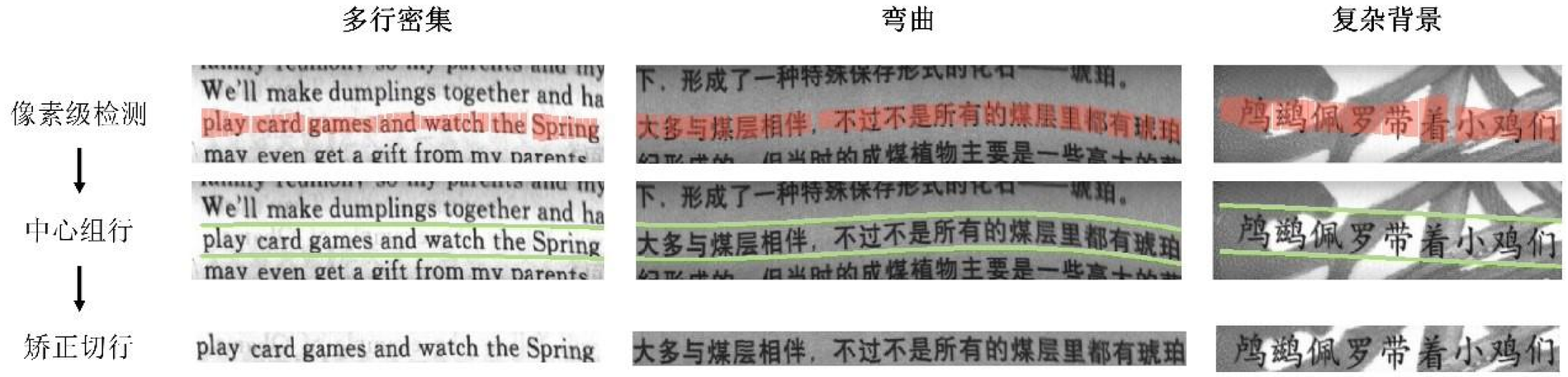


# 扫描识别

- 全景拼接
  - 拼接效果对识别影响很大

# 扫描识别

- 全景拼接
  - 像素级检测：对每个像素位置进行文字和背景分类
  - 中心组行：基于分类结果和位置信息，将扫描的中心文字连接并组合成行
  - 矫正切行：将文本行从复杂的背景中切分出来



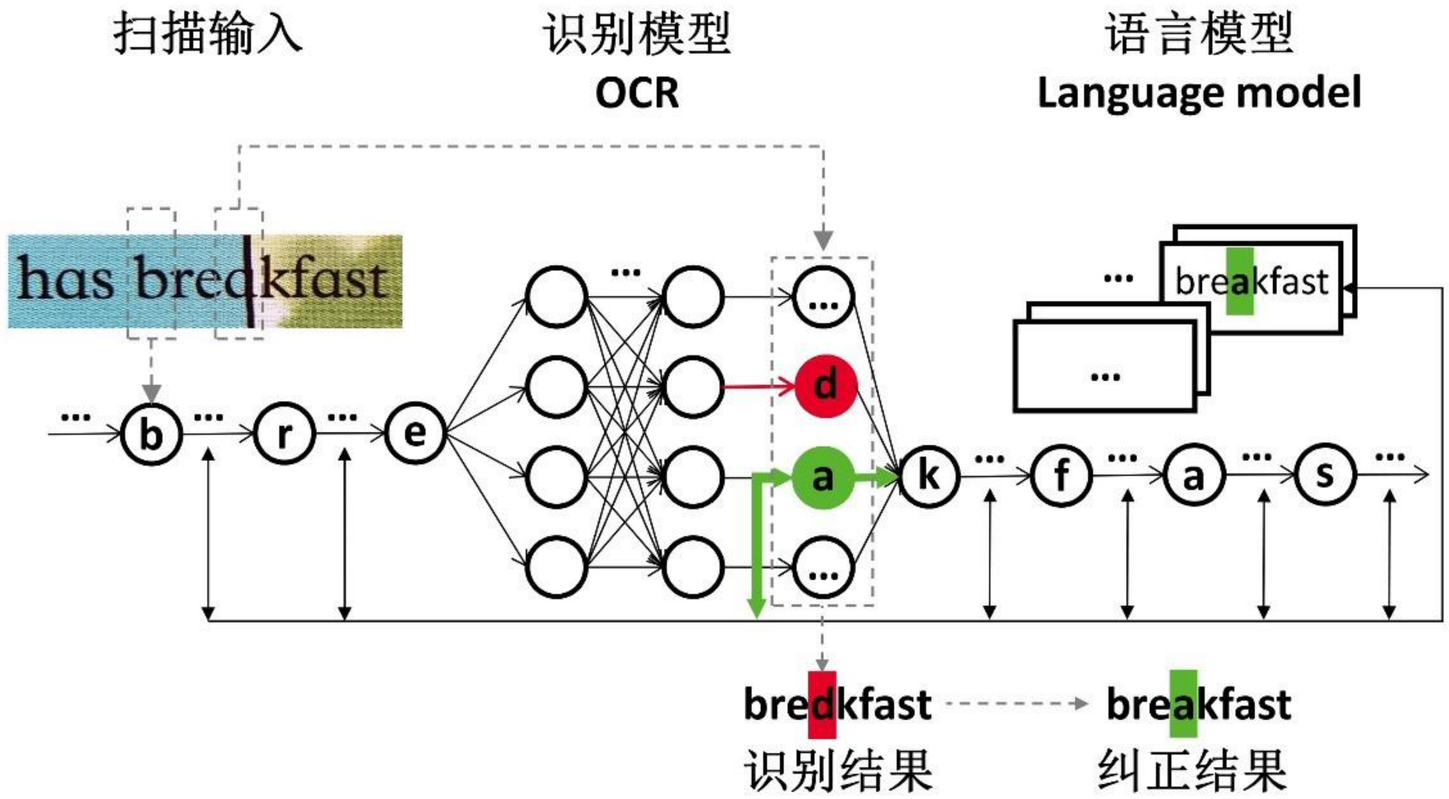
# 扫描识别

- 复杂的应用场景
  - 特殊字体，形近字，背景都会干扰识别


特殊字体、形近字、背景干扰

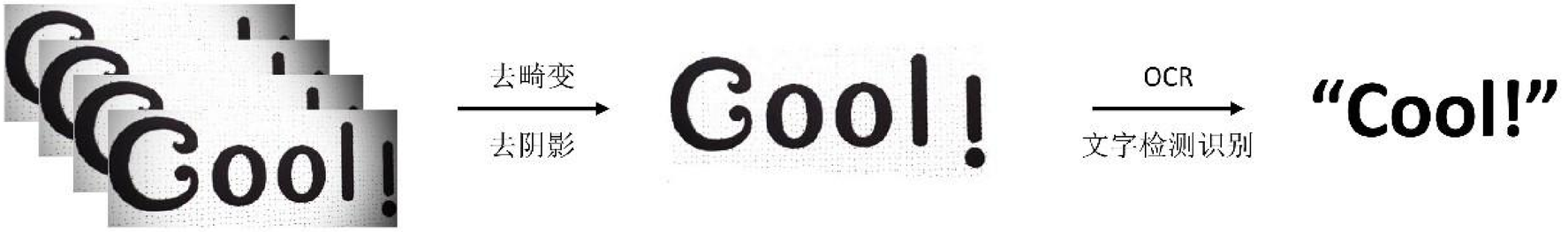
# 扫描识别

- 检测模块+识别模块+纠正模块



# 超快点查

- 问题
  - 超大广角点查导致广角畸变、光照不均
- 超快点查
  - 根据采集图像预设变换参数
  - 将采集图像逆变换得到无畸变图像
  - 对阴影进行补偿




采集图像　　增强图像　　识别结果

# 03

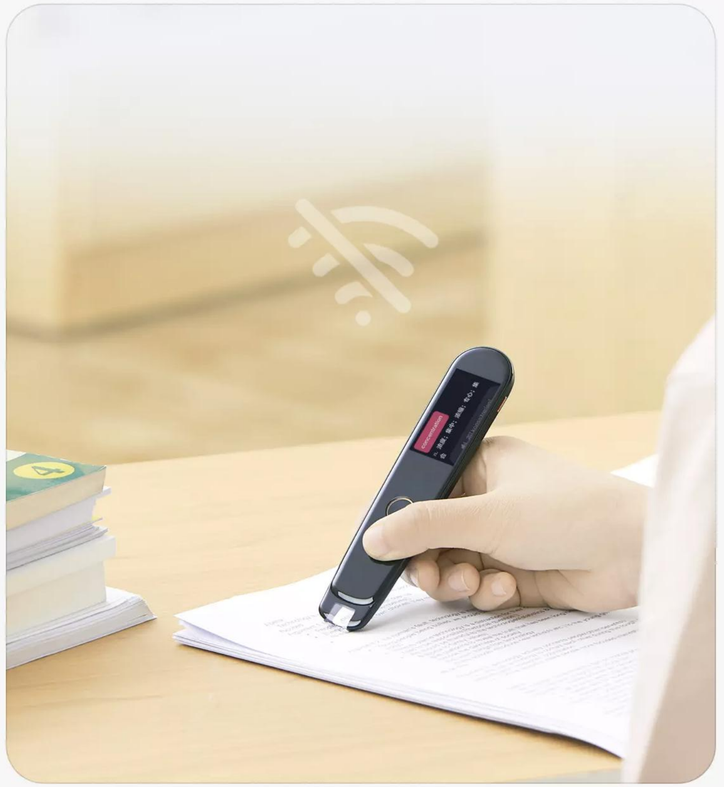
## 题目

Subject

# 离线翻译

# 离线翻译

- 离线翻译的需求
  - 无网络环境
  - 低时延
  - 节省带宽
  - 隐私

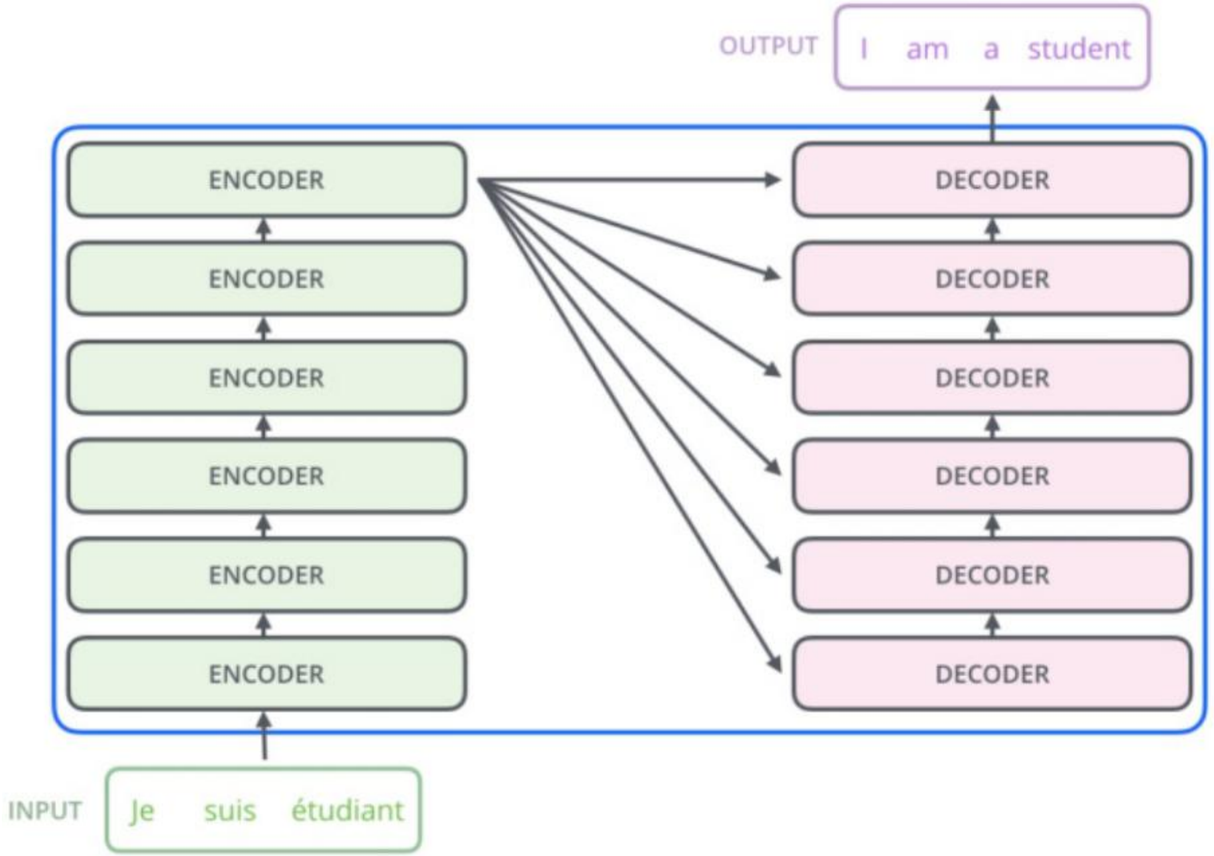


**不用网络也能翻译**
**让学习更专注**

无需介入WiFi，提笔即用
沉浸式学习，开学也能带去学校

# 在线翻译模型

- 编码器-解码器架构
- 多个编码器层和解码器层
- 很宽的维度
- 参数量达到上亿规模



# 模型压缩

- 神经网络模型存在一定冗余

| Dim | newstest2013 | Params |
|---|---|---|
| 128 | $21.50 \pm 0.16$ (21.66) | 36.13M |
| 256 | $21.73 \pm 0.09$ (21.85) | 46.20M |
| 512 | $21.78 \pm 0.05$ (21.83) | 66.32M |
| 1024 | $21.36 \pm 0.27$ (21.67) | 106.58M |
| 2048 | **21.86** $\pm 0.17$ (22.08) | 187.09M |

Table 1: BLEU scores on newstest2013, varying the embedding dimensionality.

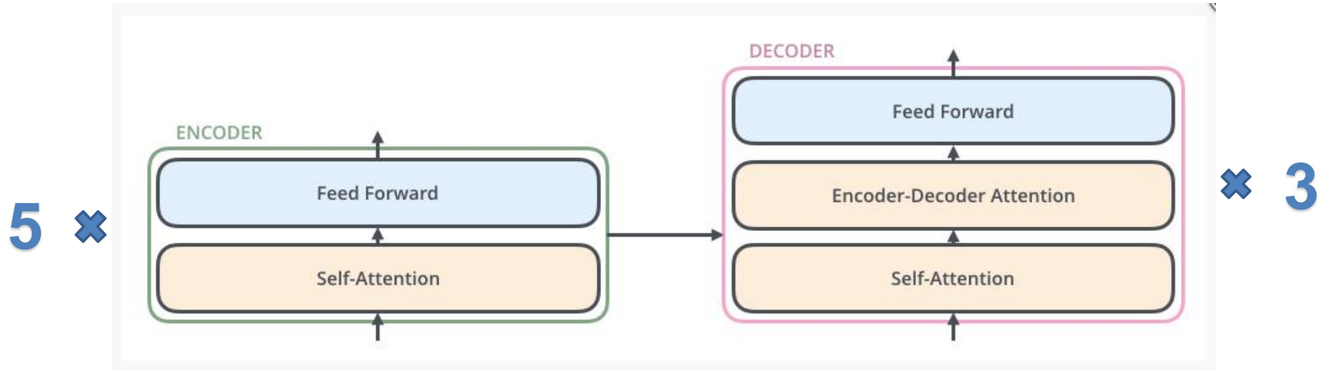
Britz D, Goldie A, Luong M T, et al. Massive exploration of neural machine translation architectures[J]. arXiv preprint arXiv:1703.03906, 2017.

# 模型压缩

- 裁剪模型
- 共享参数
- 量化
- 知识蒸馏
- Lite Transformer

# 模型压缩

- 裁剪模型
  - 编码器相对更重要
  - 更多压缩解码器
  - 减少深度的同时减少宽度



# 模型压缩

- 共享参数
  - 词向量的共享

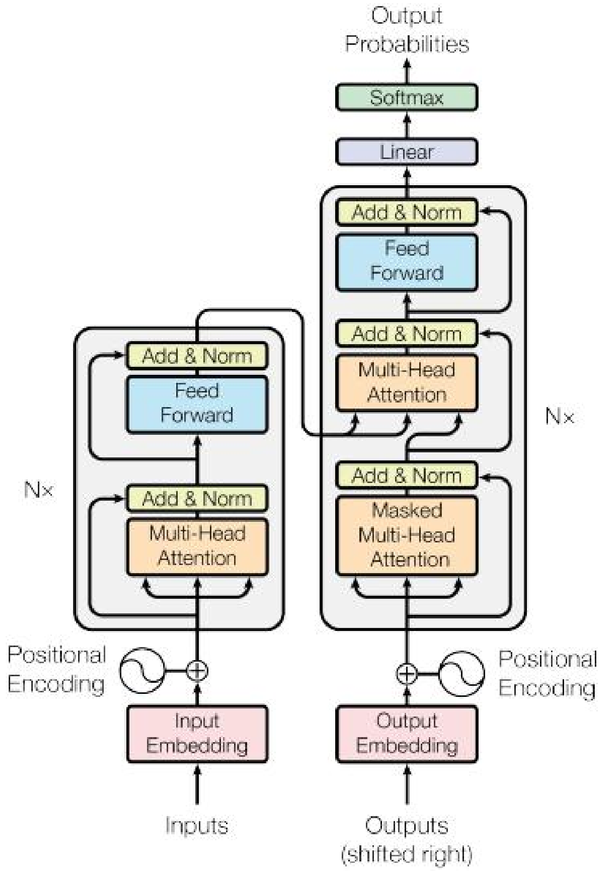
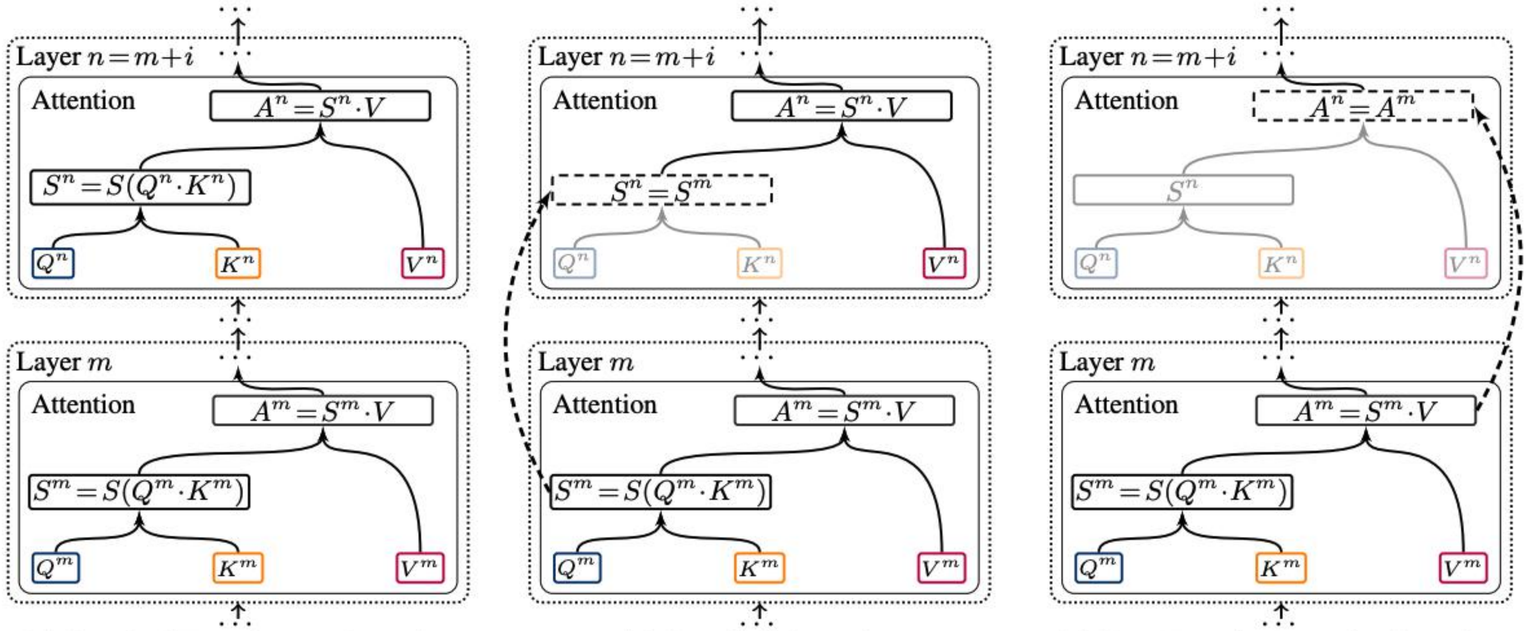



Figure 1: The Transformer - model architecture.

Press O, Wolf L. Using the output embedding to improve language models[J]. arXiv preprint arXiv:1608.05859, 2016.

# 模型压缩

- 共享参数
  - 词向量的共享
  - 不同层之间的共享




Xiao T, Li Y, Zhu J, et al. Sharing attention weights for fast transformer[J]. arXiv preprint arXiv:1906.11024, 2019.

# 模型压缩

- 量化
  - 高精度的浮点类型转化为低精度的整型计算

| Quantized | Float |
|---|---|
| 0 | -10.0 |
| 255 | 30.0 |
| 128 | 10.0 |

  - 浮点数运算使用量化运算

Jacob B, Kligys S, Chen B, et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 2704-2713.

# 模型压缩

- 量化
  - 计算量减少，对NPU，DSP芯片友好
  - 存储规模减少
  - 使用训练感知量化对质量影响也较小

# 模型压缩

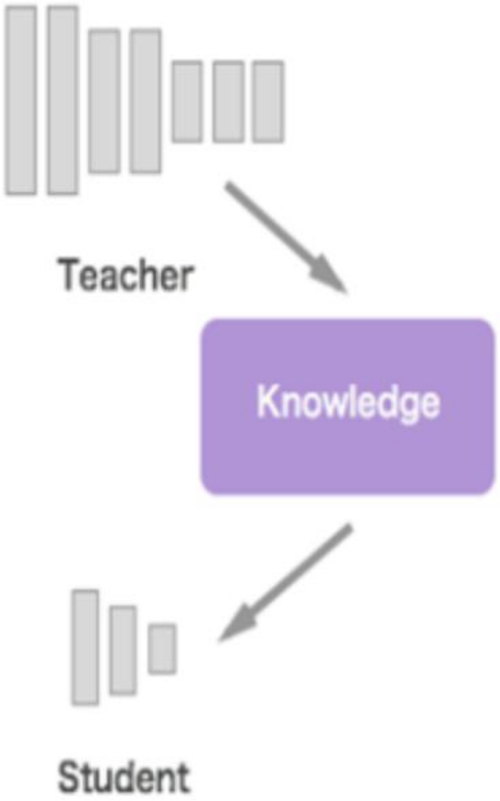
- 知识蒸馏
  - 模型压缩导致质量下降

| Model | $\text{BLEU}_{K=1}$ | $\Delta_{K=1}$ | $\text{BLEU}_{K=5}$ | $\Delta_{K=5}$ | PPL | $p(\mathbf{t} = \hat{\mathbf{y}})$ |
|---|---|---|---|---|---|---|
| *English → German WMT 2014* | | | | | | |
| Teacher Baseline $4 \times 1000$ (Params: 221m) | 17.7 | – | 19.5 | – | 6.7 | 1.3% |
| Baseline + Seq-Inter | 19.6 | +1.9 | 19.8 | +0.3 | 10.4 | 8.2% |
| Student Baseline $2 \times 500$ (Params: 84m) | 14.7 | – | 17.6 | – | 8.2 | 0.9% |

Kim Y, Rush A M. Sequence-level knowledge distillation[J]. arXiv preprint arXiv:1606.07947, 2016.

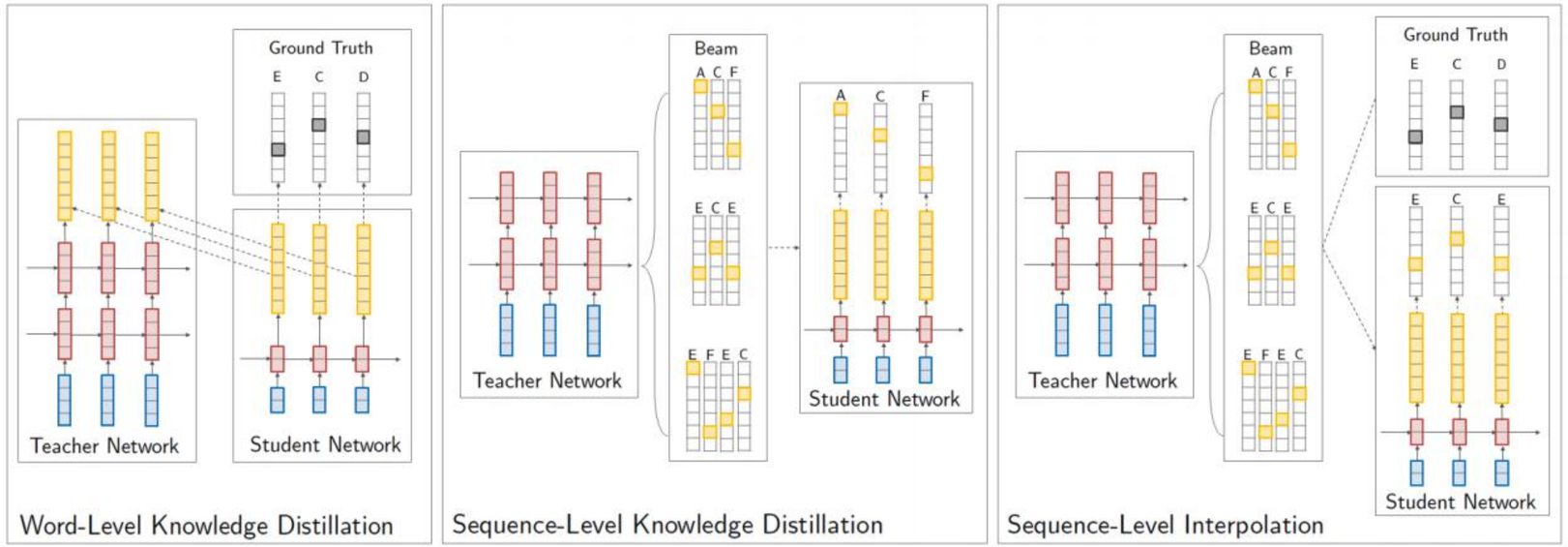
# 模型压缩

- 知识蒸馏
  - 利用教师模型提升学生模型性能
  - 教师模型：大而慢
  - 学生模型：小而快



# 模型压缩

- 知识蒸馏
  - 蒸馏的方法
    - Word-level KD
    - Sentence-level KD




Kim Y, Rush A M. Sequence-level knowledge distillation[J]. arXiv preprint arXiv:1606.07947, 2016.

# 04
# 题目

Subject

高性能端侧机器学习计算库

EMLL(Edge ML Library)

# 端侧AI面临的挑战

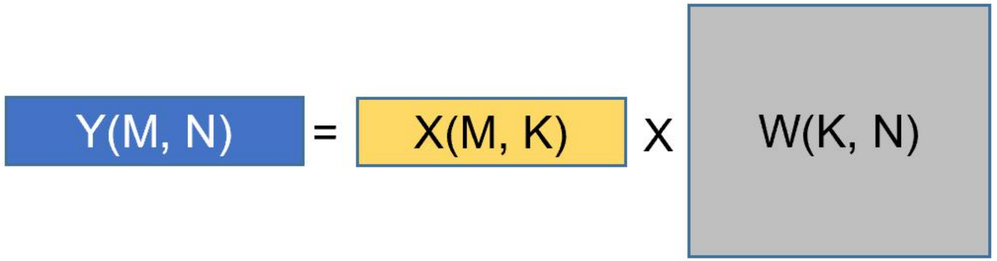
- 算力、内存有限
- 功耗限制
- 算法更新
- 多应用部署

# 端侧AI芯片

- 端侧AI芯片
  - ARM CPU
    - 当前端侧AI落地主流平台
  - NPU、DSP、GPU
    - 受生态环境影响，当前可落地的AI应用较少
    - 未来发展趋势

# 端侧AI核心计算

- 端侧AI底层主要耗时计算
  - gemm（全连接层、卷积层）
    - 扁平矩阵乘




  - 第三方blas库gemm针对端侧AI场景下计算性能较差

| 端侧AI中部分矩阵乘法 | Eigen | OpenBLAS | ARM Compute Library |
|---|---|---|---|
| **M = 128, N = 16000, K = 128** | 25% | 36% | 35% |
| **M = 7, N = 2048, K = 192** | 5% | 6% | 10% |
| **M = 23, N = 1536, K = 320** | 12% | 10% | 25% |

C(M, N) = A(M, K) * B(K, N)
ARM cortex-A53第三方库gemm计算效率

# EMLL

- EMLL(Edge ML Library)——高性能端侧机器学习计算库
  - 为加速端侧AI推理而设计
  - 为端侧AI常见的扁平矩阵的计算做了专门的优化
  - 支持fp32、fp16、int8等数据类型
  - 针对ARM cortex-A7/A35/A53/A55/A76等处理器进行汇编优化
  - 支持端侧运行OS：Linux和Android
- 已开源

**https://github.com/netease-youdao/EMLL**

# EMLL优化方法

- 访存
  - 展开外层循环 – 计算/访存比
  - 重排元素 – 顺序访存
  - 多级分块 – 利用缓存
  - 针对扁平矩阵的优化
- 计算
  - SIMD 指令
  - 指令顺序
  - 指令并发
  - 多线程(动态负载)

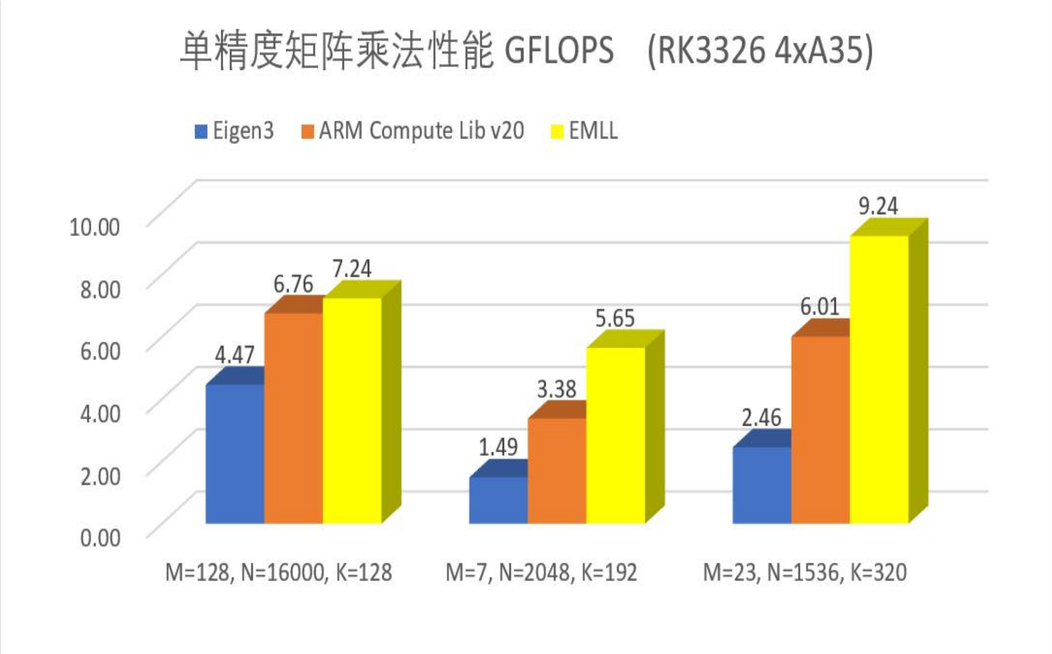
# EMLL功能

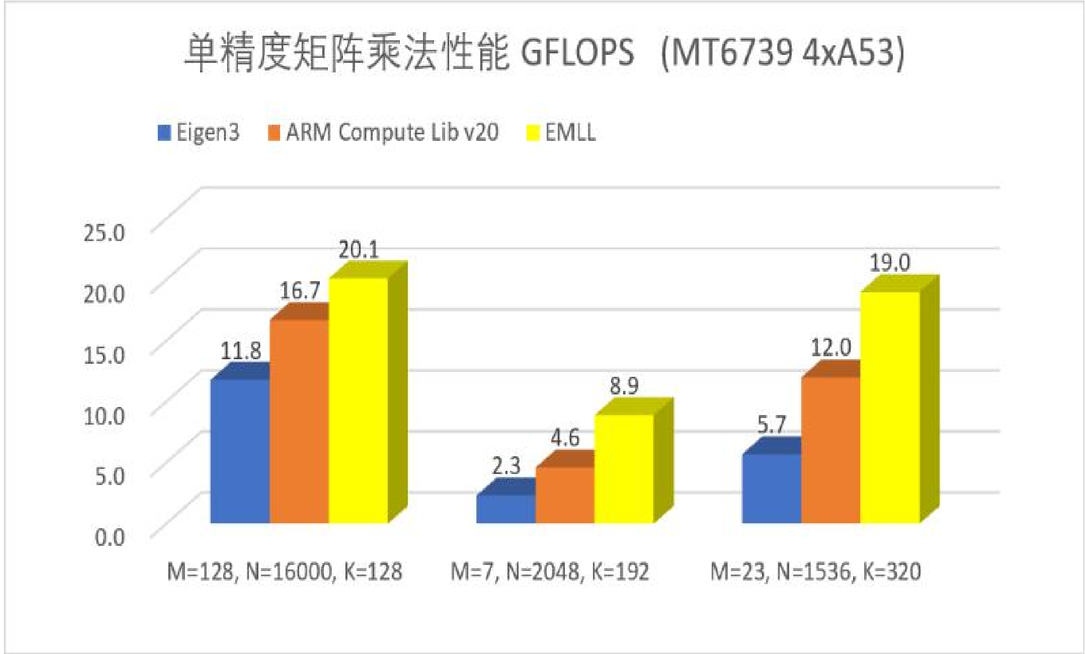
- 支持的计算函数

| 计算函数 | 支持的数据类型 |
| --- | --- |
| 矩阵乘法 | float32、float16、int8 |
| 全连接fc | float32 |
| 偏置 | float32、int32 |
| 量化 | float32 -> int8/int16 |
| 反量化 | Int8/int16/int32 -> float32 |
| 重量化 | int32 -> int16/int8，int16 -> int8 |

- 支持的架构
  - ARMv7a
  - ARMv8a
- 支持的端侧OS
  - Linux
  - Android

# EMLL GEMM 性能

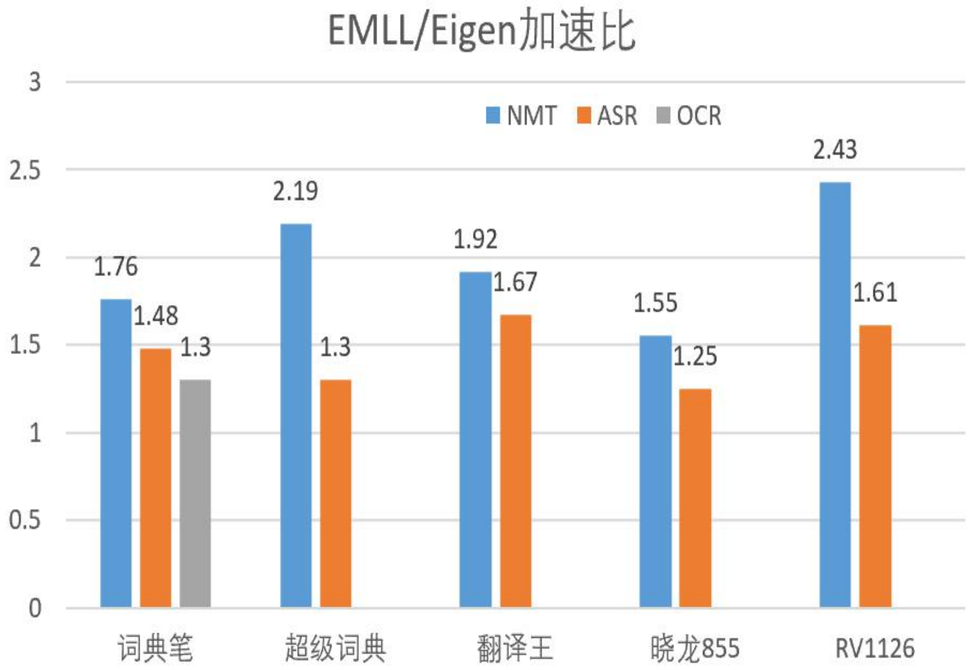



$$C(M, N) = A(M, K) * B(K, N)$$

# EMLL 在有道智能硬件中的性能

| 平台 | CPU型号 | 主频(GHz) |
| --- | --- | --- |
| 有道词典笔 | A35 | 1.2 |
| 有道超级词典 | A53 | 1.5 |
| 有道翻译王 | A53 | 2.0 |
| 某手机(晓龙855) | A76 | 2.8 |
| RV1126 | A7 | 1.5 |




NMT、ASR、OCR使用EMLL和Eigen端到端性能加速比

# 离线NMT量化效果

| | BLEU | 速度 | 内存 |
|---|---|---|---|
| 同模型int8 VS fp32 | 降低0.1以内 | 提升45~67% | 减少50~60% |
| 大模型int8 VS 小模型fp32 | 提升0.1 | 提升10% | 降低32% |

# 网易有道AI团队招聘

- NLP算法工程师
- 语音合成高级算法专家
- 图像算法工程师
- 算法研究员(自适应学习方向)
- 高性能计算研发工程师
- 数据工程师
- AI产品经理

简历发送：zhanggy@youdao.com

有道 youdao | DataFunSummit

# THANKS!

## 今天的分享就到这里...


Ending