面向语义分析的文本识别研究与实现

答辩人：肖文韬 160800224
指导老师：万燕

东华大学, July 15, 2020

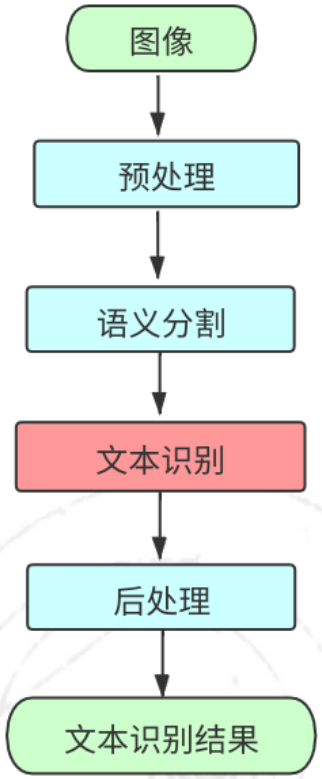
目 录

目 录

光学文本识别系统概览

光学文本识别（OCR）：

专业深度学习调试
caffee安装10元
CNN5元/层
RNN8元/层



常见的 OCR 识别错误：

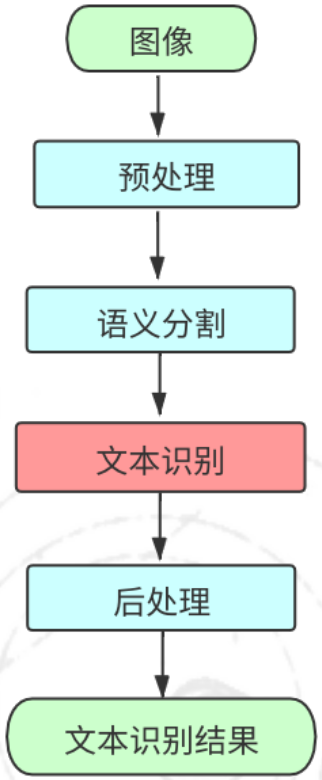
1. 替换错误：“通货膨服” ⇒ “通过膨胀”
2. 冗余错误：“休体育总局” ⇒ “体育总局”
3. 遗漏错误：“根据国” ⇒ “根据国际”

光学文本识别系统概览

发现：许多 OCR 任务的识别结果是一段自然语言，例如“专业深度学习调试”

常见的 OCR 识别错误：

1. 替换错误：“通货膨服”⇒“通过膨胀”
2. 冗余错误：“休体育总局”⇒“体育总局”
3. 遗漏错误：“根据国”⇒“根据国际”



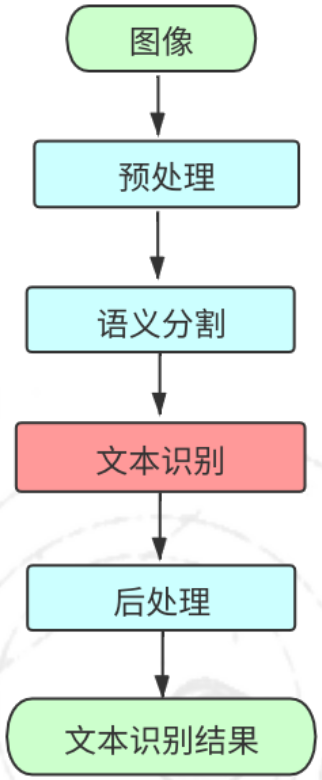
光学文本识别系统概览

发现：许多 OCR 任务的识别结果是一段自然语言，例如“专业深度学习调试”

思考：能否语义分析 OCR 识别结果并修正其识别错误?

常见的 OCR 识别错误：

1. 替换错误：“通货膨服” ⇒ “通过膨胀”
2. 冗余错误：“休体育总局” ⇒ “体育总局”
3. 遗漏错误：“根据国” ⇒ “根据国际”

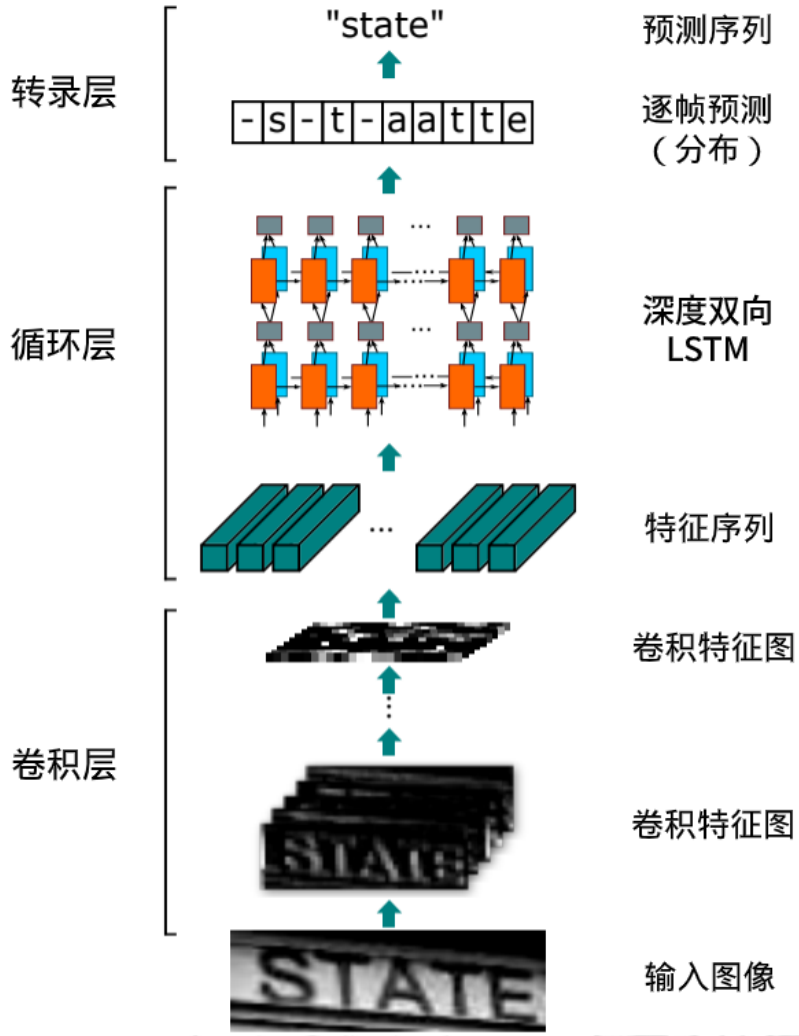


CRNN 文本识别模型

简要介绍

CRNN[1] 思路：

1. CNN[2]：提取图像特征序列
2. RNN：获取上下文信息
3. CTC：变长序列映射
   $\mathcal{B} : \mathbb{R}^{|L|} \rightarrow \mathbb{R}^{<|L|}$



[1] Shi B, *el al.* An EndtoEnd Trainable Neural Network for Image Based Sequence Recognition and Its Application to Scene Text Recognition, **TPAMI**.

[2] 本文使用的 CRNN 实现与原文略有不同：瓶颈层

基于语义分析的后处理

基于短语统计机器翻译（SMT）的方法[3]：

1. 中文句子首先被分词，识别错误会导致一串单字
2. n-gram 检测成串单字是否为识别错误
3. SMT 从候选结果中将错误翻译为正确形式

基于 n-gram 统计特征和迷惑集的方法[4]：

1. 使用迷惑集枚举所有候选句子
2. 使用动态规划找到 n-gram 分数最高的句子
3. 使用 Laplace 平滑解决 n-gram 稀疏问题

[3] Chiu H w, *el al.* Chinese Spelling Checker Based on Statistical Machine Translation. **ACL**.

[4] Huang Q, *et al.* Chinese spelling check system based on trigram mode. **SIGHAN**.

本文贡献

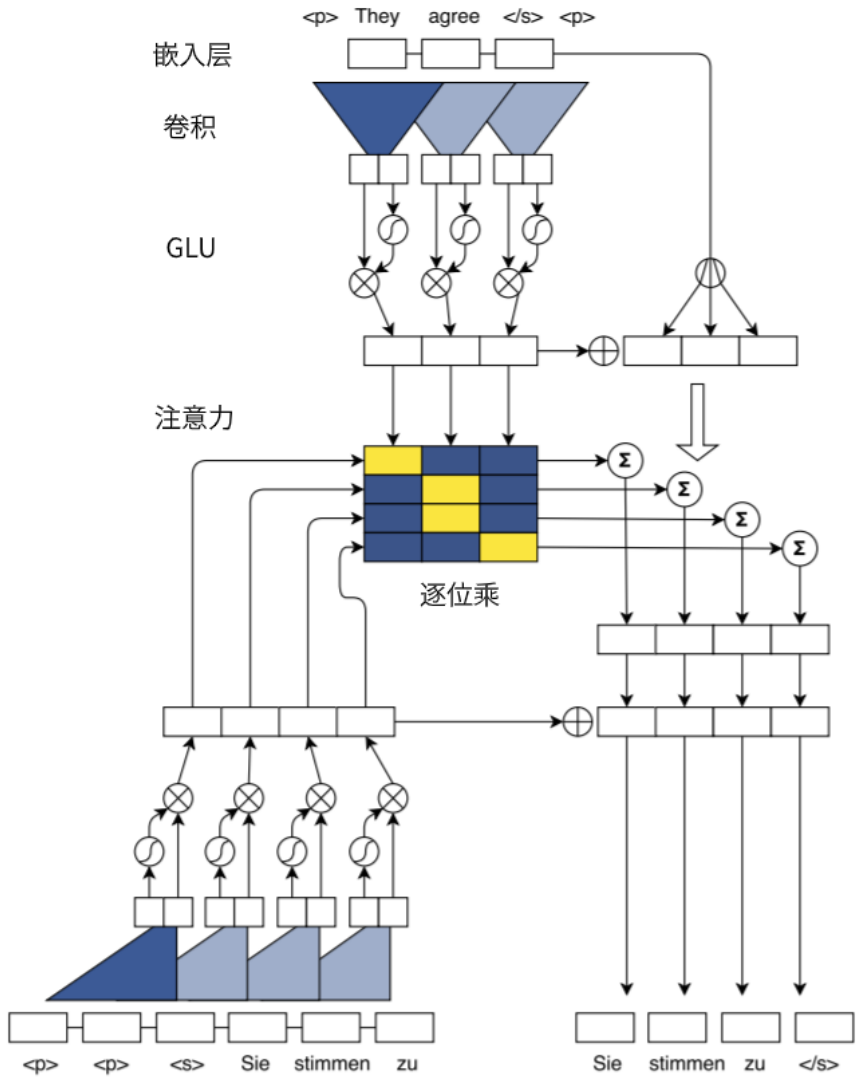
1. 构建了一个大规模合成数据集（160w）
2. 研究并实现了基于 ConvS2S 字卷积的 OCR 后处理模块
3. 研究并实现了基于 BERT 和 Transformer 的 OCR 后处理模块
4. 与其他已有研究成果（n-gram+ 迷惑集）进行对比实验

目 录

ConvS2S 架构

ConvS2S[5] 为机器翻译模型，后在 NLPCC2018[6] 中用作中文语法修正

1. 输入和上文引入绝对值位置嵌入
2. CNN 激活单元：
   $\mathrm{GLU}([A, B]) = A \otimes \sigma(B)$
3. 编码器解码器架构：
   $p(y_{n+1}|y_1, \ldots, y_n, \mathbf{e})$
4. 多步注意力：
   结合编码层和解码层的信息



[5] Gehring J, *et al.* Convolutional Sequence to Sequence Learning. **ICML**.

[6] Ren H, *el at.* A Sequence to Sequence Learning for Chinese Grammatical Error Correction. **NLPCC**.

ConvS2S 架构

多步注意力

带位置信息的嵌入表示：

- 输入 $E = (w_1 + p_1, \cdots, w_m + p_m) \in \mathbb{R}^{m \times f}$
- 上文内容 $G \in \mathbb{R}^{n \times f}$

对解码层第 l 层，多步注意力结合卷积层输出 $\hat{H}_D^l$ 和编码层最终输出 H_E^L：

$$Z_E^L = \text{affine}_{h \to f}(H_E^L) \tag{1}$$

$$C^l = \text{Attention}(\text{affine}_{h \to f}(\hat{H}_D^l), Z_E^L, Z_E^L + E) \in \mathbb{R}^{n \times f} \tag{2}$$

$$H_D^l = \hat{H}_D^L + C^l W_c \tag{3}$$

缩放点乘注意力：

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^\top}{\sqrt{d_k}})V \tag{4}$$

$$\text{softmax}(X)_{ij} = \frac{\exp(X_{ij})}{\sum_j \exp(X_{ij})} \tag{5}$$

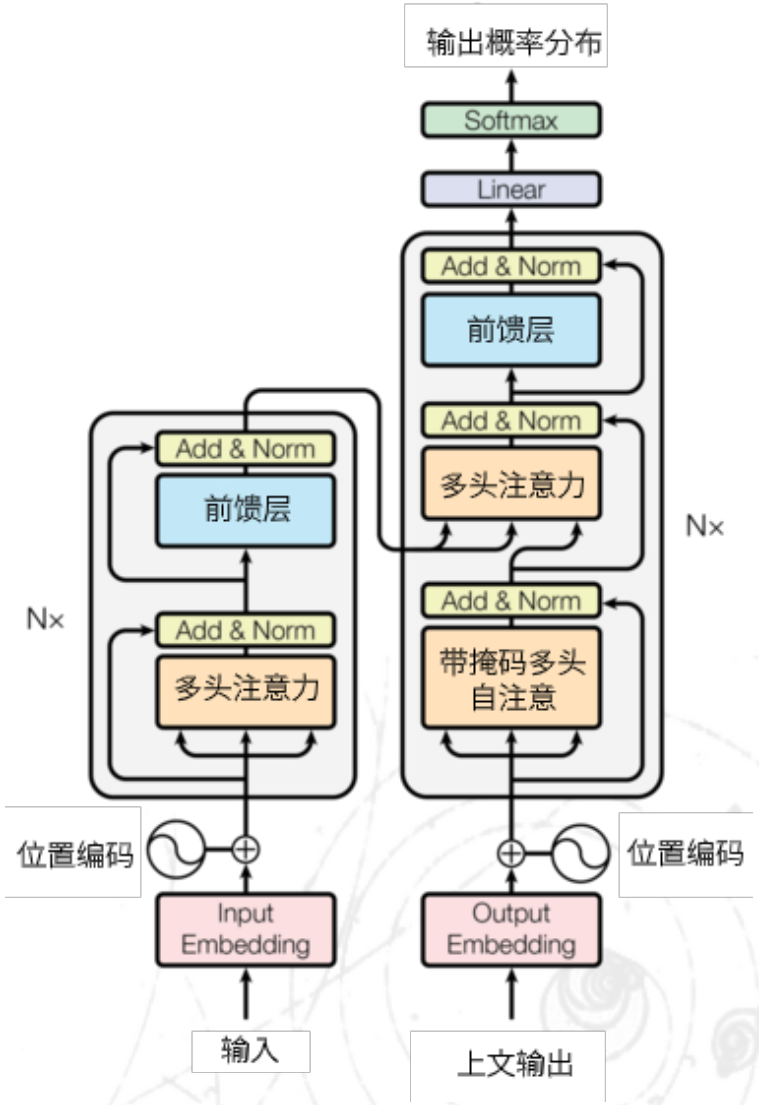
目 录

Transformer 机器翻译

简要介绍

Transformer[7] 特点：

1. 编码器解码器架构
2. 只使用自注意力
   $Q = K = V$
3. 注意力中引入掩码
4. 使用多头注意力学习更多特征
5. 三角函数族位置编码



[7] Vaswani A, *et al.* Attention is All you Need. **NIPS**.

BERT 语言模型

简要介绍

BERT[8] 特点：

1. 类似 Transformer 编码层
2. 两个无监督任务，适合大规模预训练：
   - 2.1 带掩码的语言模型 (Masked LM)：预测被随机替换成 `[MASK]` 或随机词的内容
   - 2.2 下一句预测 (NSP)
3. 预训练后轻松微调到各种下游任务

ALBERT[9]：

1. 所有层全部共享相同的参数，极大减小模型大小
2. 使用句子顺序任务 (SOP) 替换掉 NSP 任务

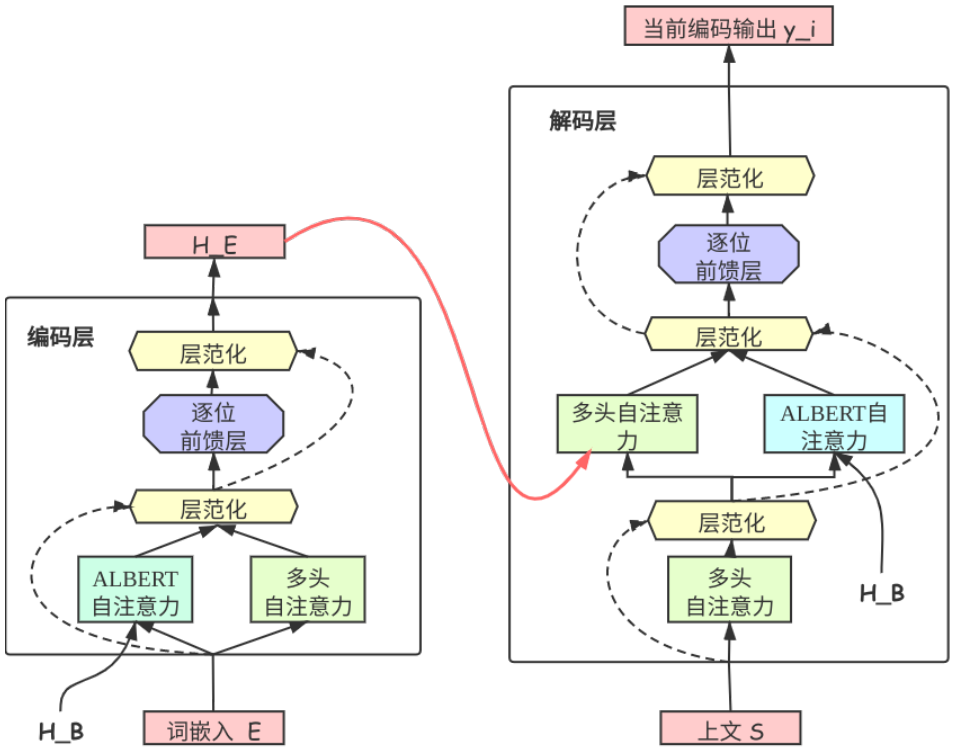
[8] Devlin J, *et al.* Bert: Pretraining of deep bidirectional transformers for language understanding.

[9] Lan Z, *et al.* ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. **ICLR**.

基于 ALBERT&Transformer 的方法

能否将预训练于海量数据的语言模型与机器翻译模型结合在一起?

基于 ALBERT&Transformer 的方法



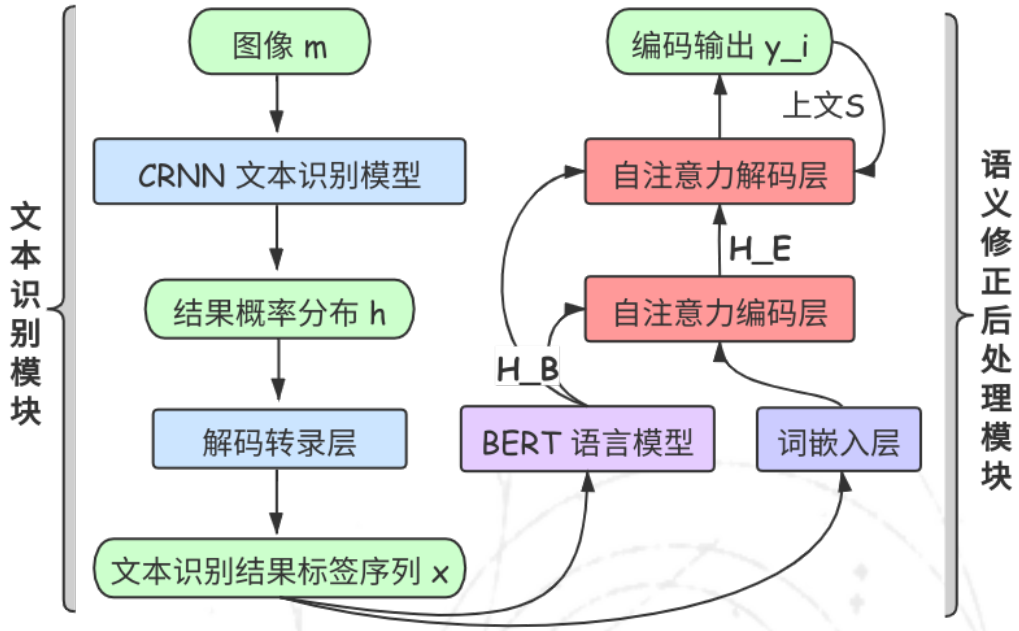
BERT-NMT[1] 机器翻译：

1. 在编码器解码器中引入 BERT 语言模型
2. 使用 drop-net 正则化 BERT 带来的过拟合

基于 ALBERT&Transformer 的方法

本文提出的改进：

1. 将 BERT-NMT[10] 迁移为 OCR 识别语义修正后处理
2. 将 BERT 替换为 ALBERT
3. 取消分词
4. 对输入和输出的嵌入层共享同一个 Lookup table



[10]Zhu J, *et al.* Incorporating BERT into Neural Machine Translation. **ICLR**.

目 录

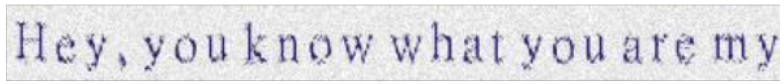
数据集和实验设计

没有合适的开源数据集，所以自己合成 OCR 识别数据集：

1. 文本来源：THUCNews[11] 新闻数据集
2. 均匀选择 1,637,012 个文本行，长度固定 $18 \sim 20$
3. 字典大小 6425
4. 多种数据增强：70 种不同字体，高斯噪声背景，随机文本畸变，色相抖动...

构造语义修正后处理数据集：

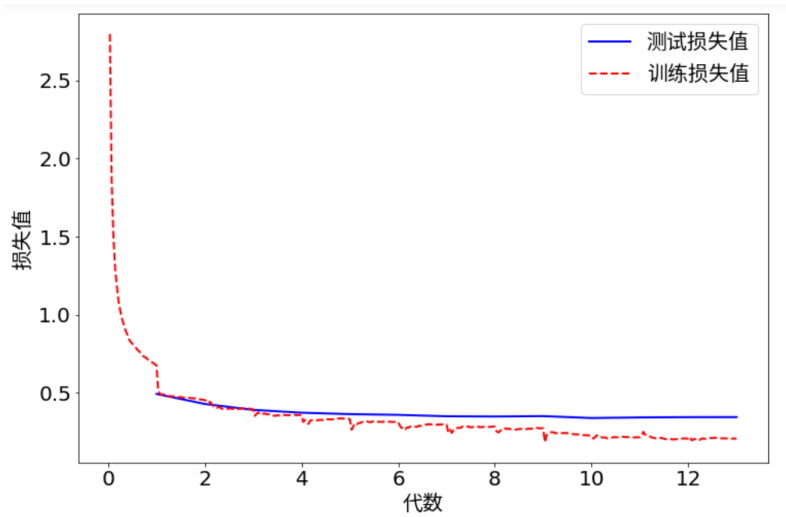
1. 按 $8:2$ 划分训练集和测试集 (327,403)
2. 使用 OCR 数据集训练 CRNN
3. 使用训练得到的 CRNN 识别整个数据集构造识别结果与正确结果数据对
4. 下采样正确识别的结果



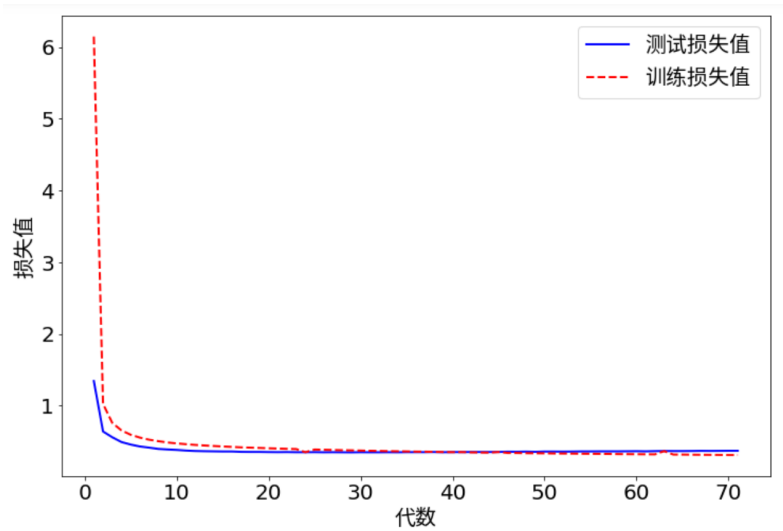
[11] Sum M, *et al.* THUCTC: an efficient Chinese text classifier.

模型训练

1. ConvS2S 训练收敛速度更快
2. ConvS2S 过拟合了
3. Transformer 训练更加平缓，没有过拟合



ConvS2S 的训练曲线



Transformer 的训练曲线

实验结果

评价指标：

1. 完整匹配 (EM) 精确度
2. Levenshtein 编辑距离归一化分数

结论：

1. 基于 n-gram 和迷惑集的传统方法效果不好
2. ConvS2S 修正效果明显，略好于 Transformer
3. ALBERT 对 Transformer 性能有帮助
4. drop-net 率推荐使用 0.4

drop-net 率参数调整实验结果

drop-net 率	EM 精确度	Levenshtein 分数
0.0	0.8320	94.3315
0.1	0.8347	94.2958
0.4	0.8483	94.3760
1.0	0.8453	94.3633

对比实验结果

方法名称	EM 精确度	Levenshtein 分数
CRNN	0.6420	94.0595
CRNN + n-gram	0.6427	94.0927
CRNN + ConvS2S	0.8461	94.3130
CRNN + Transformer	0.8421	94.2802
CRNN + ALBERT-Transformer	0.8483	94.3760

实验结果分析

基于语义分析的后处理修正模块成功案例

OCR 识别结果	英格兰银行的通货膨服目模为 2%。根据国
识别错误类型	替换 + 插入
语义修正结果	英格兰银行的通货膨胀目标为 2%。根据国际

OCR 识别结果	仍由民政部与罔家休体育鹅局具休制定和宵施。
识别错误类型	删除 + 替换
语义修正结果	仍由民政部与国家体育总局具体制定和实施

基于语义分析的后处理修正模块失败案例

OCR 识别结果	活家禽批发商会预计，零傅价约为每斤 2G6 元
语义修正结果 (ALBERT-Transformer)	活家禽批发商会预计，零售价约为每斤 66 元
语义修正结果 (ConvS2S)	活家禽批发商会预计，零售约为每斤 2G6 元
实际正确结果	活家禽批发商会预计，零售价约为每斤 26 元

OCR 识别结果	“我冈图密室部分，区(指因佩慈)因因不可
语义修正结果 (ALBERT-Transformer)	“我们紧密室部分，谢 (指导佩慈) 发现不可
语义修正结果 (ConvS2S)	“我们亲密室部分，谢 (指纹佩慈) 因为不可
实际正确结果	“我负责密室部分，她 (指吴佩慈) 负责不可

目 录

总结与展望

总结：

1. OCR 语义修正可以视为一种特殊的机器翻译
2. 基于海量文本的预训练语言模型可以提供有效的语义信息
3. 近距离局部信息可以更加重要（ConvS2S）

展望：

1. 将文本识别模型的中间特征加入语义修正模型
2. 更好的融合语言模型的特征
3. 优化基于 ALBERT 和 Transformer 的方法的训练速度

欢迎各位老师点评指导

谢谢观看！