

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, June 2, 2020



目 录

1. 背景介绍
2. 基于字卷积的方法
3. 基于语言模型和 Transformer 的方法
4. 实验验证
5. 总结与展望

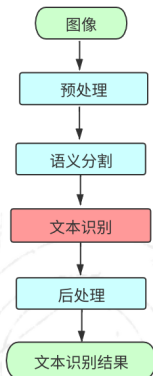
目 录

1. 背景介绍
2. 基于字卷积的方法
3. 基于语言模型和 Transformer 的方法
4. 实验验证
5. 总结与展望

光学文本识别 (OCR):



专业深度学习调试
caffe安装10元
CNN5元/层
RNN8元/层



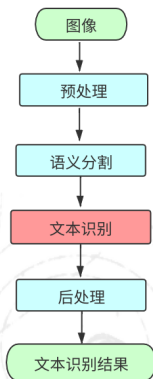
常见的 OCR 识别错误:

1. 替换错误: “通货膨**服**” \Rightarrow “通货膨**胀**”
2. 冗余错误: “**休**体育总局” \Rightarrow “体育总局”
3. 遗漏错误: “根据国” \Rightarrow “根据国**际**”

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**” \Rightarrow “通货膨**胀**”
2. 冗余错误：“**休**体育总局” \Rightarrow “体育总局”
3. 遗漏错误：“根据国” \Rightarrow “根据国际”

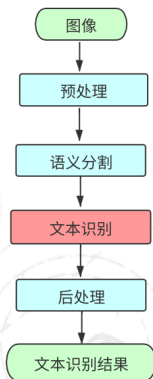


发现：许多 OCR 任务的识别结果是一段**自然语言**，例如“**专业深度学习调**
试”

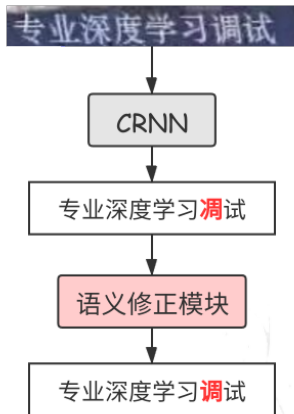
思考：能否语义分析 OCR 识别结果并修正其识别错误？

常见的 OCR 识别错误：

1. 替换错误：“通货膨**服**” ⇒ “通货膨**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**”



面向语义分析的文本识别研究与实现！



CRNN 文本识别模型

简要介绍

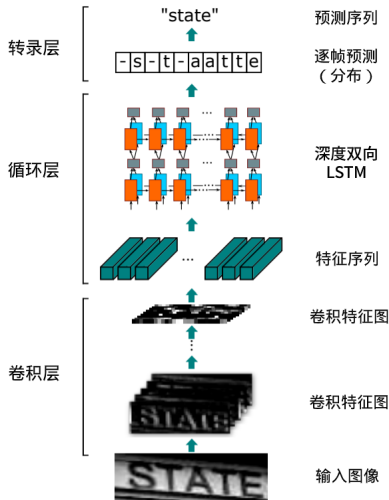


東華大學
DONGHUA UNIVERSITY

CRNN¹ 思路:

1. CNN²: 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

$$\mathcal{B} : \mathbb{R}^{|L|} \rightarrow \mathbb{R}^{<|L|}$$



¹Shi B, *et al.* An End-to-End 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

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

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



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 ALBERT 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

1. 背景介绍
2. 基于字卷积的方法
3. 基于语言模型和 Transformer 的方法
4. 实验验证
5. 总结与展望

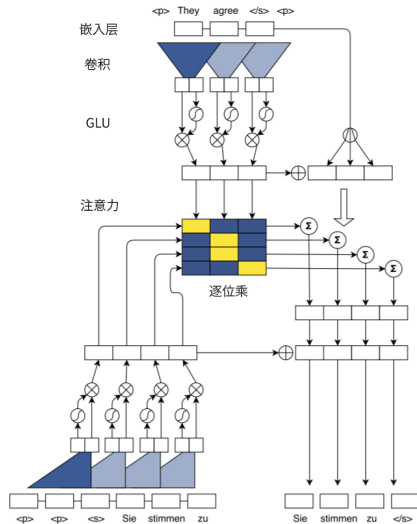
ConvS2S 架构



東華大學
DONGHUA UNIVERSITY

ConvS2S⁵ 为机器翻译模型，后
在 NLPCC2018⁶ 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：
$$\text{GLU}([A, B]) = A \otimes \sigma(B)$$
3. 编码器解码器架构：
$$p(y_{n+1} | y_1, \dots, y_n, \mathbf{e})$$
4. 多步注意力：
结合编码层和解码层的信息



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

⁶Ren H, *et al.* A Sequence to Sequence Learning for Chinese Grammatical Error Correction. **NLPCC**.

带位置信息的嵌入表示:

- 输入 $E = (w_1 + p_1, \dots, w_m + p_m) \in \mathbb{R}^{m \times f}$
- 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 :

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

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

$$H_D^l = \hat{H}_D^l + C^l W_c \quad (3)$$

缩放点乘注意力:

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

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

目 录

1. 背景介绍
2. 基于字卷积的方法
3. 基于语言模型和 Transformer 的方法
4. 实验验证
5. 总结与展望

Transformer 机器翻译

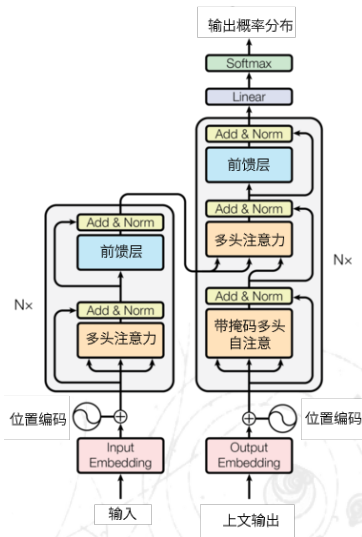
简要介绍



東華大學
DONGHUA UNIVERSITY

Transformer⁷ 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



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



BERT⁸ 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

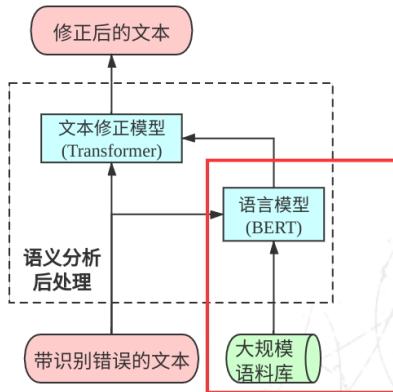
ALBERT⁹:

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.

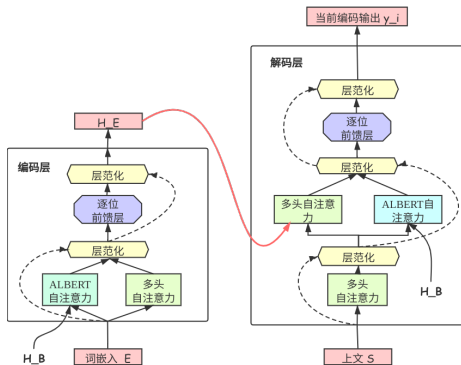
机器翻译模型 + 预训练语言模型？



基于 ALBERT&Transformer 的方法



東華大學
DONGHUA UNIVERSITY



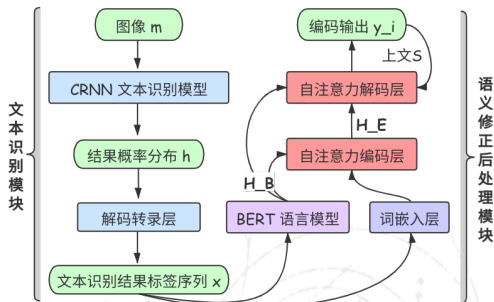
BERT-NMT¹⁰ 机器翻译:

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

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

本文提出的改进:

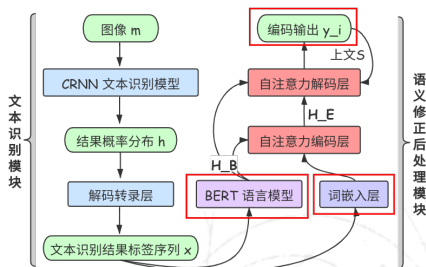
1. 将 BERT-NMT¹⁰ 迁移为 OCR 语义修正后处理
2. 将 BERT 替换为 ALBERT
3. 取消分词
4. 对输入和输出的嵌入层共享同一个词嵌入表



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

本文提出的改进:

1. 将 BERT-NMT¹⁰ 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



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

目 录

1. 背景介绍
2. 基于字卷积的方法
3. 基于语言模型和 Transformer 的方法
4. 实验验证
5. 总结与展望

数据集和实验设计



東華大學
DONGHUA UNIVERSITY

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：THUCNews¹¹ 新闻数据集
2. 均匀选择 1,637,012 个文本行，长度固定 18 ~ 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

作词：姚若龙 王力宏

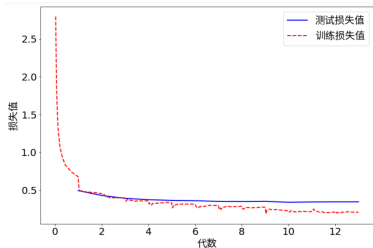
再也不要飘在人海里

Hey, you know what you are my

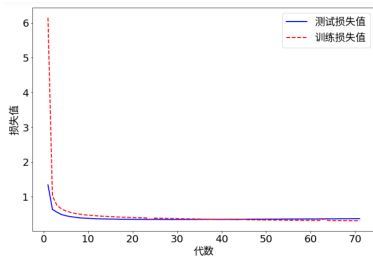
每个明天都会更感人

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

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胀目模为 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体制定和宵施。
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，零傳价约为每斤 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

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	中国外 论 交 马 中.心公布, 人民币 见 美元 9 日
识别错误类型	删除 + 替换
语义修正结果	中国外 汇 交 易 中心公布, 人民币 兑 美元 9 日中
OCR 识别结果	达 一点当 燃 丹益于饰 梅 者金范的端 止 外形,
识别错误类型	替换
语义修正结果	这 一点当 然 得 益 于饰 掩 者金范的端 正 外形, 不

基于语义分析的后处理修正模块失败案例

OCR 识别结果	更不知道到底欠了多少线, 不 遇退 他知道一 黏 ,
语义修正结果 (ALBERT-Transformer)	更不知道到底欠了多少线, 不过 过 他知道一点
语义修正结果 (ConvS2S)	更不知道到底欠了多少线, 不过他知道一点
实际正确结果	更不知道到底欠了多少线, 不过他知道一点
OCR 识别结果	王尽士念丽一闹门匿锯很风等
语义修正结果 (ALBERT-Transformer)	王女士表示, 一家人 数据等风险 。 他认为只要
语义修正结果 (ConvS2S)	王女士表示, 在 一家人 透露, 据很风险, 只要
实际正确结果	王女士表示, 这一家人的遭遇很让人感触,

目 录

1. 背景介绍
2. 基于字卷积的方法
3. 基于语言模型和 Transformer 的方法
4. 实验验证
5. 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. 近距离局部信息可能更加重要 (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！