像人类一样阅读:自主、双向和迭代的语言场景文本识别建模

Shancheng Fang Hongtao Xie* Yuxin Wang Zhendong Mao Yongdong Zhang
中国科学技术大学
{fangsc, htxie, zdmao, zhyd73}@ustc.edu.cn, wangyx58@mail.ustc.edu.cn

抽象的

语言知识对场景文本大有裨益
识别。然而,如何有效地模拟语言
端到端深度网络中规则的发现仍然是一个研究挑战。在本文中,
我们认为,容量有限的
语言模型的来源有:1)隐式语言
建模;2)单向特征表示;3)
具有噪声输入的语言模型。相应地,我们提出了一个自主、双向、
迭代的 ABINet,用于
场景文本识别。首先,自主建议
阻止视觉和语言模型之间的梯度流
强制执行明确的语言建模。其次,一种新颖的
双向完形填空网络(BCN)作为语言模型
是基于双向特征表示提出的。
第三,我们提出了一种语言模型迭代修正的执行方式,可以有效缓解
噪声输入的影响。此外,基于集成
迭代预测,我们提出了一种自我训练方法
可以有效地从未标记的图像中学习。广泛的
实验表明,ABINet 在低质量图像上表现出色,并在多个主流基准测试
中取得了最佳结果。此外,ABINet 训练的

集成自训练显示出良好的改善
实现人类水平的识别。代码可访问

1. 简介

拥有从场景图像中读取文本的能力对于人工智能来说是必不可少的[24,41]。
为此,早期的尝试认为字符毫无意义
符号并通过分类模型识别符号[42, 15]。然而,当面对具有挑战性的

遮挡、模糊、噪声等环境,它变得
由于视力丧失而昏厥。幸运的是,
文本承载着丰富的语言信息,字符可以
根据上下文进行推理。因此,一堆
方法[16, 14, 29]将注意力转向语言建模,并取得了毋庸置疑的进步。

然而,如何有效地模拟语言行为
人类阅读能力的提高仍然是一个悬而未决的问题。根据心理学的观察,我们可以对以下问题做出三个假设:
人类阅读认为语言建模是自主的、双向的和迭代的:1)作为聋哑人和盲人
可以分别拥有完全功能的视觉和语言,
我们用“自主”一词来解释
视觉和语言之间的学习。自主性也
意味着视觉和语言之间的良好互动
独立学习的语言知识可以有助于
视觉中字符的识别。2)
推理字符上下文的行为类似于完形填空任务,因为
难以辨认的字符可以被视为空白。因此,可以利用

左右两侧的不可辨认字符同时显示,相当于双向显示。3)

迭代描述了在具有挑战性的环境下,
人类采取渐进策略来提高预测能力
通过迭代修正识别结果来增强信心。

首先,将自主原则应用于场景
文本识别(STR)是指识别模型
应该解耦为视觉模型(VM)和语言
模型(LM),子模型可以作为独立的功能单元,单独进行学习。最近的
基于注意力机制的方法通常基于以下方面设计语言模型:
RNN 或 Transformer [39],其中的语言规则是
在耦合模型中隐式学习[19, 36, 33]
(图1a)。然而,LM 是否以及如何
了解人物关系是不可知的。此外,这
这类方法无法捕捉丰富的先验知识
通过直接从大规模未标记文本中预训练 LM。

其次,与单向语言模型[38] 相比,
具有双向原则的LM捕获量翻倍
信息。构建双向模型的一个直接方法是合并一个从左到右的模型和一个从右到左的模型[28, 5],无论是在概率层面[44, 36] ,还是在

特征级[49] (图1e)。然而,它们严格来说
因为它们的语言特性是单向的

*通讯作者

事实上,整体模型意味着
计算和参数都很昂贵。最近
NLP 领域的杰出成果是 BERT [5],它引入了一种深度
通过掩盖文本标记学习的双向表示。
直接将 BERT 应用于 STR 需要屏蔽所有
文本实例中的字符,而这非常
由于每次只能屏蔽一个字符,因此成本较高。

第三,使用迭代原理执行的语言模型可以根据视觉和语言线索重新细化预测,这
目前的方法还没有探索。
执行 LM 是自回归[44, 3, 45] (图1d),在
错误识别被积累为噪声并被视为
输入进行以下预测。为了调整 Transformer
架构， [25, 49]放弃自回归并采用
并行预测（图1e)来提高效率。然而，
并行预测中仍然存在噪声输入,其中错误
虚拟机输出的预测结果会直接损害语言模型的准确率。此外,SRN
[49]中的并行预测存在长度不对齐的问题,导致 SRN 难以推断出
正确的字符
如果 VM 错误地预测了文本长度。

考虑到当前方法的不足
内部交互、特征表示和
执行方式,我们提出了以自主、双向和迭代原则为指导的ABINet。
首先,我们
探索一种通过阻断梯度的解耦方法（图1b）
VM 和 LM 之间的 BGF,强制 LM
明确地学习语言规则。此外,VM 和 LM
是自主单元,可以通过图像进行预先训练
和文本。其次,我们设计了一个新颖的双向完形填空网络（BCN)作为语言模型,它消除了
结合两个单向模型的困境（图1c）。
BCN 以左右上下文为条件,通过指定注意掩码来控制访问

两侧字符。此外,跨步骤访问
以防止信息泄露。第三,我们建议
一种LM迭代修正的实现方式（图1b）。
通过将 ABINet 的输出反复输入到 LM,可以逐步完善预测,并且未对齐长度
问题可以得到一定程度的缓解。此外，
将迭代预测视为一个整体,探索一种基于自训练的半监督方法,该方法
探索一种实现人类水平识别的新解决方案。

本文的贡献主要包括:1)我们提出
自主、双向和迭代原则来指导
STR 中 LM 的设计。根据这些原则,LM
一个功能单元,用于提取双向
表示并迭代地纠正预测。2)一种新颖的
引入了 BCN,它使用双向表示来估计字符的概率分布,例如完形填空任务。3)提出的 ABINet 达到了最先进的

（SOTA)在主流基准上的表现,以及
经过集成自训练的 ABINet 表现出良好的前景
实现人类水平的识别的提高。

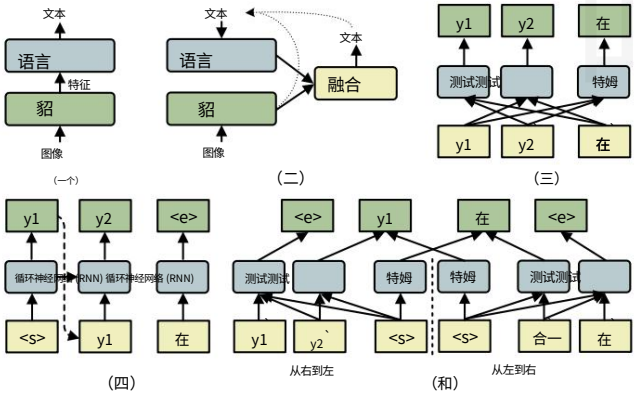

图 1. (a) 耦合语言模型。(b) 我们的自主语言模型（带有迭代校正）。(c) 我们的双向结构。(d) 自回归中的单向 RNN。(e) 集成

两个单向 Transformer 进行并行预测。

2.相关工作

2.1. 不依赖语言的方法

无语言方法通常利用视觉特征
不考虑字符之间的关系,例如基于 CTC 的[7]和基于分割的[21]

方法。基于 CTC 的方法采用 CNN 提取视觉特征,采用 RNN 建模特征序列。
然后使用 CTC 对 CNN 和 RNN 进行端到端训练
损失[34, 11, 37, 12]。基于分割的方法应用 FCN 在像素级上分割字符。廖等人。
通过将分割像素分组来识别字符
文本区域。Wan 等人[40]提出了一种额外的顺序
分割图,将字符转录为正确的
顺序。由于缺乏语言信息,非语言方法无法解决低质量

图像值得称赞。

2.2. 基于语言的方法

视觉与语言之间的内在互动。在
一些早期的作品中,文本字符串的 N-gram包由 CNN 预测,该 CNN 充当显式 LM [14, 16, 13]。
此后,基于注意力的方法开始流行，
它使用更强大的隐式模型语言
RNN [19, 36]或 Transformer [43, 33]。基于注意力机制的
方法遵循编码器-解码器架构,其中编码器处理图像,解码器生成字符

通过关注一维图像特征[19, 35, 36, 3, 4]或二维图像特征[48, 45, 23, 20]中的相关信息。
例如， R2AM [19]使用递归 CNN 作为特征提取器,使用 LSTM 作为学习的 LM,在字符级别隐式建模语言,从而避免使用

N-gram。此外,这类方法通常通过
积分整流模块[36, 51, 47]用于不规则
在将图像输入网络之前。不同的
从上述方法中,我们的方法致力于构建一个更
通过明确的语言建模来实现强大的语言模型。在尝试

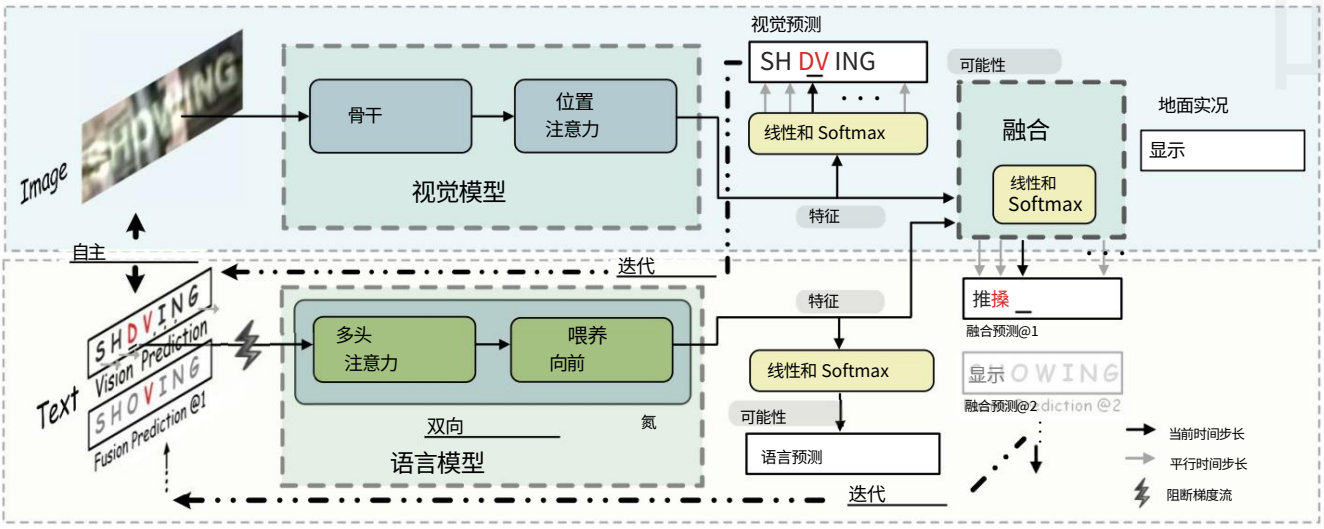

图 2.ABINet 的示意图。

为了提高语言表达能力,一些作品引入
多个损失,其中一个额外的损失来自语义[29, 25, 49, 6]。其中,SEED [29]
提出
使用预先训练的 FastText 模型来指导 RNN 的训练,
这带来了额外的语义信息。我们偏离了
因为我们的方法直接在未标记的文本中预训练 LM,
这在实践上是比较可行的。
语言特征的表示。基于注意力机制的方法中的字符序列通常被建模
为
从左到右的方式[19, 35, 3, 40]。例如,Textscanner [40]继承了基于注
意力机制的单向模型
方法。不同的是,他们采用了额外的位置
分支来增强位置信息,并减少无上下文场景中的错误识别。为了利用
双向
信息,像[8, 36, 44, 49]这样的方法使用集成
两个单向模型的模型。具体来说,为了捕捉
全局语义上下文,SRN [49]结合了
一个从左到右和一个从右到左的 Transformers,用于进一步
预测。我们强调集合双向
模型本质上是一种单向的特征表示。
语言模型的执行方式。目前,
语言模型的网络结构主要基于 RNN 和
Transformer [39]。基于 RNN 的 LM 通常在
自回归[44, 3, 45],它采用最后一个
字符作为输入。DAN [44]等典型工作得到
首先使用提出的
卷积对齐模块。之后,GRU 预测
通过对最后一个字符进行预测嵌入,对每个字符进行预测
时间步长和当前时间步长的特征
作为输入。基于 Transformer 的方法具有优势
在并行执行中,每个时间步的输入
要么是视觉特征[25] ,要么是来自
视觉特征的预测[49]。我们的方法属于
并行执行,但我们试图缓解噪音问题
输入存在于并行语言模型中。

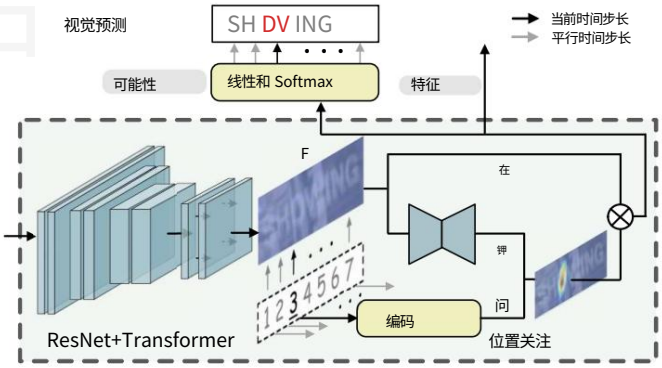

图3.视觉模型架构。

3.提出的方法

3.1. 视觉模型

视觉模型由骨干网络和
位置注意模块（图3）。继上文
方法,ResNet1 [36, 44]和Transformer单元[49, 25]
分别作为特征提取网络和序列建模网络。对于图像 x,我们有:

$$Fb = T(R(x)) \in R^{\frac{W}{4} \times \frac{H}{4} \times C}, \quad (1)$$

其中 H、W 是 x 的大小,C 是特征维度。
位置注意模块将视觉特征并行地转录为字符概率,其基于
关于查询范式[39]:

$$Fv = softmax(\frac{Q.K.T.}{\sqrt{C}})V. \quad (2)$$

具体而言， $Q \in R^{\text{温度} \times \text{碳}}$ 是位置编码[39]
字符顺序， T是字符序列的长度。

1总共有5个残差块,并进行了下采样
在第 1 和第 3 个区块之后。

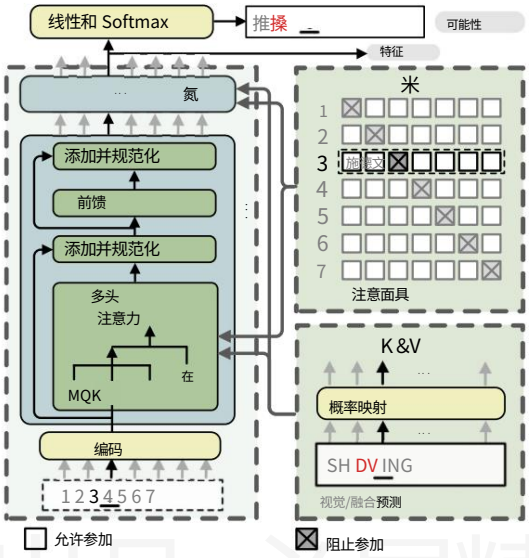

图 4.语言模型（BCN)的架构。

K=G(Fb) ∈R $^{\text{硬件}}_{16}$ ×C ，其中G(·)由一个迷你 U-Net2 [32]。V =H(Fb) ∈R $^{\text{硬件}}_{16}$ ×C ，其中H(·)是身份映射。

3.2. 语言模型

3.2.1 自主策略

如图2 所示,自主策略包括以下特点:1)LM 被视为一个独立的采用概率向量的拼写纠正模型字符作为输入并输出概率分布预期字符。2)训练梯度的流程是在输入向量处阻塞（BGF）。3)LM 可以进行训练与未标记的文本数据分开。

遵循自主策略,ABINet 可以被划分成可解释的单元。通过将概率作为输入,LM 可以被替换（即,被替换为更强大的模型)和灵活（例如执行在第 3.2.3节中迭代）。此外,还有一点很重要BGF 强制模型学习语言知识不可避免地,这与隐含的模型究竟学到了什么,我们无法得知。此外,自主策略允许我们直接分享NLP社区的先进进展。例如,对 LM 进行预训练可以有效地提升性能。

3.2.2 双向表示

给定一个文本字符串y = (y1, . . . , yn) ,其中文本长度n和类别数c,双向和单向模型的yi的条件概率为

P(yi |yn, . . . , yi+1, yi−1, . . . , y1) and P(yi |yi−1, . . . , y1),

分别。从信息论的角度来看,双向表示的可用熵

可以量化为Hy = (n − 1) log c。然而,对于单向表示,信息是

$\frac{1}{n}\sum_{i=1}^{n}(i-1)\log c = \frac{1}{2}$Hy.我们的见解是,先前方法通常使用两个单向模型的集成模型,本质上是单向表示。单向表示基本上捕获$\frac{1}{2}$Hy信息,导致功能能力受限与双向对应物相比,抽象。

受益于第 3.2.1 节中的自主设计，现成的具有拼写纠正能力的 NLP 模型可以迁移。一个可行的方法是利用BERT [5]中的掩码语言模型 (MLM)通过将yi带有 token [MASK]。然而,我们注意到这是不可接受的,因为 MLM 应该被分别调用n次每个文本实例,导致效率极低。而不是屏蔽输入字符,我们通过指定注意力面具。

总的来说,BCN 是L 层 Transformer的一个变体解码器。BCN 的每一层都是一系列多头注意力和前馈网络[39] ,随后是残差连接[10]和层归一化[1],如下图所示图4. 与 vanilla Transformer 不同,特征向量被输入到多头注意力模块,而不是网络的第一层。此外,多头注意力机制中的注意力掩码是为了防止“看到自己”而设计的。此外,BCN 中没有使用自注意力机制,以避免泄漏跨时间步长的信息。注意力操作多头块可以形式化为:

$$\text{我}=\begin{cases}0, i=j\\ -\infty, i=j\end{cases}, \tag{3}$$

$$Ki = Vi = P(yi)Wl, \tag{4}$$

$$Fmha = \text{softmax}\left(\frac{Q.K.T.}{\sqrt{C}} + M\right)V, \tag{5}$$

其中Q ∈ R $^{\text{温度×碳}}$ 是字符的位置编码第一层的订单,否则为最后一层的输出。K ,V ∈ R $^{\text{温度×碳}}$ 由特征概率c×C和Wl ∈ R得到性P(yi) ∈ R c ， 是线性映射矩阵。M∈R $^{T\times T}$ 注意力遮罩矩阵可以防止参与当前角色。堆叠 BCN 层后进入深层架构,双向表示Fl文本 y 已确定。

通过以完形填空的方式指定注意力掩码,BCN能够学习更强大的双向表示比单向表示的集合更优雅。此外,得益于类似 Transformer 的架构,BCN可以独立、并行地进行计算。此外,它比集成模型更有效,因为需要一半的计算和参数。

2A 网络具有 4 层编码器、64 通道,添加融合和插值上采样。

3.2.3 迭代校正

Transformer 的并行预测采用噪声输入,这些噪声输入通常是来自视觉预测 [49] 或视觉特征 [25] 的近似值。具体而言,如图2 所示,在双向表示下,P(“O”)的理想条件是“SH-WING”。然而,由于环境模糊和遮挡,从虚拟机获得的实际条件是“SH-VING”,其中“V”成为噪声并损害了预测的置信度。由于虚拟机中的错误预测增加,它对语言模型 (LM) 的影响往往更大。

为了应对噪声输入问题,我们提出了迭代语言模型(如图2 所示)。语言模型会重复执行M次,并对y 进行不同的赋值。对于第一次迭代,yi=1是虚拟机的概率预测。对于后续迭代, yi≥2是最后一次迭代中融合模型(见第3.3 节)的概率预测。通过这种方式,语言模型能够迭代地校正视觉预测。

另一个观察结果是,基于 Transformer 的方法普遍存在长度不对齐问题[49],这表明 Transformer 很难纠正视觉

如果字符数与基本事实不一致,则进行预测。

长度不对齐问题是由于不可避免的填充掩码(padding mask)造成的,该掩码用于过滤文本长度之外的上下文。我们的迭代语言模型可以缓解这个问题,因为它将视觉特征和语言特征多次融合,从而逐步优化预测的文本长度。

3.3. 融合

从概念上讲,在图像上训练的视觉模型和在文本上训练的语言模型来自不同的模态。

为了对齐视觉特征和语言特征,我们简单地使用门控机制[49, 50]进行最终决策:

$$G = \sigma([Fv, Fl]Wf), \tag{6}$$

$$Ff = G\,Fv + (1-G)\,Fl, \tag{7}$$

其中$Wf \in R^{2C\times C}$且$G \in R^{T\times C}$。

3.4. 监督训练

ABINet 使用以下多任务目标进行端到端训练:

$$L = \lambda_v Lv + \frac{\lambda_l}{M}\sum_{i=1}^{M} Ll + \frac{1}{M}\sum_{i=1}^{M} Lf \tag{8}$$

其中Lv、 Ll和Lf分别是Fv、 Fl和Ff的交叉熵损失。具体来说, L和Lf是第 i 次迭代时的损失。λv和λl是平衡因子。是

3.5. 半监督集成自训练

为了进一步探索我们的迭代模型的优越性,我们提出了一种基于

算法1集成自训练

要求:带有标签 Y 的标记图像 X 和未标记图像 U 1:使用公式 8 训练 ABINet 的参数θ0 (X, Y) 。
2:使用θ0为 U 生成软伪标签 V 3:通过使用 C < Q (公式9)过滤 (U, V) 来获取 (U , V) 4:对于i = 1, . . . , Nmax do 5:如果i == Nupl then 6:使用θi更新 V 7:通过使用 C < Q (公式9)过滤 (U, V) 来获取 (U , V) 8:结束 if
9:采样Bl = (Xb, Yb) (X , Y), Bu = (U 10:使用公式8将Bl、 Bu更新为θi。11 :结束

b, 在b) (在, 在)

通过迭代预测集成进行自我训练[46] 。

自训练的基本思想是先由模型自身生成伪标签,然后利用额外的伪标签重新训练模型。因此,关键问题在于构建高质量的伪标签。

为了过滤噪声伪标签,我们提出以下方法:1)选择文本实例中字符的最小置信度作为文本确定性。2)将每个字符的迭代预测视为一个整体,以平滑噪声标签的影响。因此,我们定义过滤函数如下:

$$\text{碳} \quad 1\leqslant t\leqslant T \quad E[\log P(yt)] = \min e$$

$$P(yt) = \text{最大值}_{1\leqslant m\leqslant M} \text{下午}(yt), \tag{9}$$

其中C是文本实例的最小确定性, Pm(yt)是第 m 次迭代时第 t 个字符的概率分布。

训练过程如算法1所示,其中Q为阈值。Bl 、 Bu分别为来自标记数据和未标记数据的训练批次。Nmax为最大训练步数, Nupl为更新伪标签的步数。

4.实验

4.1. 数据集和实现细节

为了公平比较,实验按照[49]的设置进行。具体而言,训练数据集是两个合成数据集 MJSynth (MJ) [13, 15]和 SynthText (ST) [9]。六个标准基准包括 ICDAR 2013 (IC13) [18]、 ICDAR 2015 (IC15) [17]、 IIIT 5K- Words (IIIT) [27]、街景文本 (SVT) [42]、街景文本视角 (SVTP) [30]和 CUTE80 (CUTE) [31] 作为测试数据集。这些数据集的详细信息可以在以前的工作[49] 中找到。此外,删除标签的Uber-Text [52]被用作未标记数据集来评估半监督方法。

模型维度C全程设置为512。BCN共有4层,每层8个注意力头。平衡因子λv和λl分别设置为1和1。图像直接缩放至32×128 ,并进行数据增强处理,例如几何变换(即旋转、仿射和透视)、图像质量下降和颜色抖动等。我们使用

表 1. VM 的消融研究。Attn 是注意力机制，
Trm Layer 是 Transformer 的层数。SV、 MV1、 MV2和
LV 是四个不同配置的虚拟机。

模型 姓名	收件人	Trm IC13 SVT IIIT 层 IC15 SVTP CUTE 94.2 89.6 80.6 82.3 93.6 89.3 80.8 83.1				平均值	参数时间3 （×106）（毫秒）	
SV （小的）	平行线	2	94.5 89.5		93.7 85.1	88.8	19.6	12.5
MV1 （中间）	位置	2			94.2 85.4	89.0	20.4	14.9
MV2 （中间）	平行线	3	81.1	83.7	94.3 86.8	89.4	22.8	14.8
LV （大号）	位置	3	94.9 90.4 81.7 84.2		94.6 86.5	89.8	23.5	16.7

表 2. 自主策略的消融研究。PVM 是在 MJ 和 ST 上以监督的方式预训练虚拟机。PLMin是在 MJ 和 ST 上以自监督的方式使用文本预训练语言模型 (LM)。
PLMout在 WikiText-103 [26]上对 LM 进行预训练，采用自监督
方式。AGF表示允许VM和LM之间的梯度流。

PVM PL最小PLM输出AGF				IC13 SVT IIIT IC15 SVTP 可爱			平均值
-	-	-	-	96.7 84.5	93.4 86.8	95.7 86.8	91.7
✓	-	-	-	97.0 85.0	93.0 88.5	96.3 89.2	92.3
-	✓	-	-	97.1 83.6	93.8 88.1	95.5 86.8	91.6
✓✓		-	-	97.2 84.9	93.5 89.0	96.3 88.5	92.3
✓	-	✓	-	97.0 85.3	93.7 88.5	96.5 89.6	92.5
✓	-	-	✓	96.7 83.3	92.6 86.5	95.7 88.5	91.4

4 个 NVIDIA 1080Ti GPU 用于训练我们的批量模型
384. 采用ADAM优化器进行初始学习
速率 1e $^{-4}$，衰减至 1e $^{-5}$ 经过 6 个时期。

4.2. 消融研究

4.2.1 视觉模型

首先我们从两个方面来讨论VM的性能：
特征提取和序列建模。实验结果记录在表1中。并行注意力机制是一种

大众注意力方法[25,49],以及提出的立场
注意力机制对键/值有更强大的表示
向量。从统计数据我们可以得出结论：1)简单地升级虚拟机将大大提高准确率，但
代价是参数和速度。2)升级虚拟机，
我们可以在特征提取中使用位置注意力，并且
序列建模中更深层次的转换器。

4.2.2 语言模型

自主策略为了分析自主模型，我们分别采用 LV 和 BCN 作为 VM 和 LM。

从表2的结果中我们可以观察到：1)预训练
VM 很有用，可以提高准确率约0.6%-0.7%
平均而言；2)预训练 LM 对训练的益处

[3]推理时间是使用 NVIDIA Tesla V100 平均 3 来估算的
不同的试验。

表 3.双向表示的消融研究。

视觉语言		IC13 SVT IIIT IC15 SVTP 可爱		平均值	参数时间 （×106）（毫秒）	
SV	西德	96.0 90.3 81.9 96.3 86.0	94.9 85.4	90.2	32.8	19.1
	德国	90.9 82.6 86.4 96.7 91.7 83.1	95.0 87.5	90.6	45.4	24.2
	巴塞罗那	86.2	95.3 88.9	91.0	32.8	19.5
LV	西德	96.0 91.2 84.0 86.8 96.8 92.3	96.2 87.8	91.5	36.7	22.1
	德国	84.2 87.9	96.3 88.2	91.9	49.3	26.9
	巴塞罗那	97.0 93.0 85.0 88.5	96.3 89.2	92.3	36.7	22

表 4. 语言模型在文本拼写纠正中的 Top-5 准确率。

语言模型 字符准确率 词汇准确率		
德国	78.3	27.6
巴塞罗那	82.8	41.9

数据集（即 MJ 和 ST)可以忽略不计;3)在预训练
即使基础模型准确率很高，来自额外未标记数据集（例如
WikiText-103)的 LM也很有帮助。
上述观察结果表明，STR 有助于
对虚拟机和语言模型进行预训练。在额外的
未标记的数据集比训练数据集更有效
由于文本多样性有限且数据分布有偏差
无法促进表现良好的 LM 的学习。
此外，在未标记数据集上进行 LM 预训练的成本很低，因为
可以轻松获取附加数据。

此外，通过允许 VM 之间的梯度流（AGF）
和 LM 相比，性能平均下降了0.9% （表2）。
我们还注意到 AGF 的训练损失急剧减少
到较低的值。这表明过度拟合发生在
LM 作为 VM 有助于在训练中作弊，这也可能
发生在隐式语言建模中。因此
通过 BGF 强制 LM 独立学习至关重要。我们
注意，SRN [49]在 VM 之后使用了 argmax 操作，这
本质上是 BGF 的一个特例，因为 argmax 是不可微的。另一个优点是，自主
策略使模型具有更好的可解释性，因为我们
可以深入了解 LM 的性能（例如，
表4） ，这在隐式语言建模中是不可行的。

双向表示。由于 BCN 是
Transformer，我们将 BCN 与其对应物 SRN 进行比较。
基于 Transformer 的 SRN [49]作为单向表示的集合，表现出了优异的性能。
为了公平比较，实验采用
除网络外，其他条件相同。我们使用 SV 和 LV 作为
VM 验证不同准确率下的有效性
级别。如表3 所示，尽管 BCN 具有与单向版本的

SRN（SRN-U） ，在准确性方面取得竞争优势
在不同的VM下。此外，与集成中的双向SRN相比，BCN表现出更好的性能

尤其是在 IC15 和 CUTE 等具有挑战性的数据集上。
此外，配备 BCN 的 ABINet速度大约快20%-25%

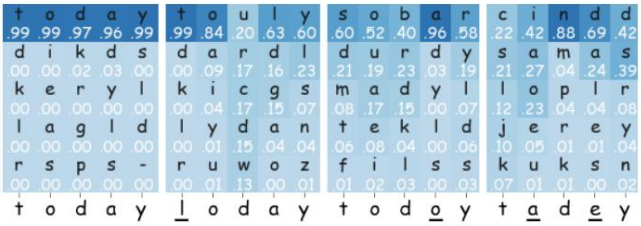

图 5.BCN 中 top-5 概率的可视化。

表5.迭代校正的消融研究。

模型	迭代 IC13 SVT IIIT 编号 IC15 SVTP CUTE				平均值	参数时间 （×106）（毫秒）	
SV	1	96.7 91.7 83.1 86.2 97.2 91.8		95.3 88.9	91.0	32.8	19.5
+	2	83.3 86.4 97.1 93.0 83.4 86.7		95.4 89.2	91.2	32.8	24.5
巴塞罗那	3			95.4 89.6	91.4	32.8	31.6
LV	1	97.0 93.0 85.0 88.5 97.1 93.4		96.3 89.2	92.3	36.7	22
+	2	85.2 88.7 97.3 94.0 85.5 89.1		96.3 89.6	92.4	36.7	27.3
巴塞罗那	3			96.4 89.2	92.6	36.7	33.9

比SRN更适合大规模任务。

第3.2.1节指出,LM 可以被视为
独立单位估计概率分布
拼写纠正,因此我们进行实验
这个视图。训练集是来自 MJ 和 ST 的文本。为了
模拟拼写错误,测试集为 20000 个项目,其中
是随机选择的,我们添加或删除一个字符
对于20% 的文本,替换一个字符,对于60% 的文本,保留
其余文本保持不变。从表4 的结果来看,我们
可以看到 BCN 的字符准确率比 SRN 高出4.5%
词汇准确率为14.3 % ,这表明 BCN 具有
字符级语言建模能力更强大。

为了更好地理解 BCN 在 ABINet 内部的工作原理,我们
图5 中可视化 top-5 概率,其中取“今天”
举个例子。一方面,因为“今天”是一个字符串
结合语义信息,将“-oday”和“tod-y”作为
输入,BCN 可以高置信度地预测“t”和“a”,并且
有助于最终的融合预测。另一方面,
错误字符“l”和“o”是其余预测的噪声,
BCN 信心下降,对最终结果影响不大
预测。此外,如果有多个错误字符,
由于缺乏
足够的背景。

迭代校正。我们再次应用 SV 和 LV,
BCN 展示迭代校正的性能
从不同层次。实验结果如表5所示。
其中迭代次数设置为 1、2 和 3
训练和测试。从结果可以看出,迭代
BCN 3 次可以分别提高准确率
0.4%,平均0.3% 。具体来说,
IIIT 是一个相对简单且特征清晰的数据集
外观。然而,当涉及到其他更难的数据集时

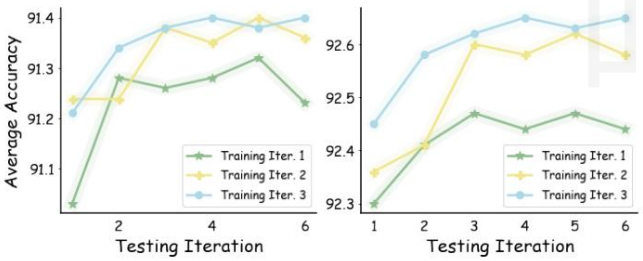

图 6. 训练和测试中迭代 BCN 的准确率。

图 7. 使用迭代修正的成功示例。文本
字符串是基本事实,视觉预测,融合预测,无需
迭代校正和迭代校正分别来自
从左到右,从上到下。

如 IC15、SVT 和 SVTP,迭代校正稳步
提高准确度,分别达到1.3%和1.0%
分别针对 SV 和 LV 对 SVT 进行了改进。
还注意到推理时间随着
迭代次数。

我们进一步探讨了
训练和测试。平均准确率的波动
图6表明:1)直接应用迭代校正
在测试中也表现良好;2)在训练中迭代
有益的,因为它提供了额外的训练样本
LM;3)迭代时准确率达到饱和状态
模型迭代超过 3 次,因此需要进行大规模迭代
号码是不必要的。

为了全面了解迭代校正,我们在图7 中将中间预测可视化。

通常情况下,视力预测可以在
在某些情况下,错误仍然存在。之后
经过多次迭代,最终可以修正预测。
此外,我们还观察到迭代校正能够
缓解长度不对齐问题,如最后所示
图7 中的列。

从消融研究中我们可以得出结论:1)双向BCN 是一种强大
的 LM,可以有效地提高
准确率和速度上的表现。2)通过进一步
为 BCN 配备迭代校正,噪声输入
问题可以得到缓解,建议处理
具有挑战性的例子,例如低质量图像
增量计算的费用。

4.3 与现有技术的比较

一般来说,公平地比较并不是一件容易的事
其他方法直接使用报告的统计数据[2],
主干（即 CNN 结构和

表6.与其他方法的准确度比较。

	方法	带标签 未带标签 常规文本						不规则文本		
		数据集 数据集 IC13 SVT IIIT IC15 SVTP CUTE								
效DTA	2019 吕等人。[25] （并行）	MJ+ST	-	92.7 90.1 94.0 76.3 95.3 90.6 93.9					82.3	86.8
	2019 Liao等人[22] （SAM）	MJ+ST	-	77.3					82.2	87.8
	2020 乔等人。[29] （SE-ASTER）	MJ+ST	-	92.8 89.6 93.8 80.0 81.4 92.9 90.1 93.9 79.4						83.6
	2020 Wan等人[40] （Textscanner）	MJ+ST	-	84.3						83.3
	2020 Wang等人[44] （和）	MJ+ST	-	93.9 89.2 94.3 74.5 94.8 88.1 95.3					80.0	84.4
	2020 Yue 等人[50] （RobustScanner)MJ+ST 2020 Yu 等人[49]		-	77.1					79.5	90.3
	（SRN）	MJ+ST	-	95.5 91.5 94.8 82.7 85.1 96.3 90.9 95.0 82.6						87.8
自	SRN-SV（复制品）	MJ+ST	-	86.4 96.8 93.2 95.4 84.0 87.0 96.8 92.3 96.3						87.5
	ABINet-SV	MJ+ST	-	84.2 87.9 97.4 93.5 96.2 86.0 89.3						88.9
	SRN-LV（复制品）	MJ+ST	-							88.2
	ABINet-LV	MJ+ST	-							89.2
	ABINet-LVst	MJ+ST Uber-Text 97.3 94.9 96.8 87.4 90.1 MJ+ST Uber-Text 97.7 95.5 97.2								93.4
	ABINet-LVest	86.9 89.9								94.1

图 8.ABINet-LVest 成功识别的困难示例。

参数）、数据处理（即图像校正和
数据增强)和训练技巧等。为了严格执行公平比较,我们复现了 SOTA 算法
SRN 与
ABINet,如表 6 所示。两个重新实现的
SRN-SV 和 SRN-LV 与报告的模型略有不同,它们替换了虚拟机,消除了副作用
多尺度训练,应用衰减学习率等
请注意,SRN-SV 的表现略优于 SRN,因为
上述技巧。从比较中可以看出,我们的
ABINet-SV 的表现比 SRN-SV 好0.5%、 2.3%、 0.4%，IC13、SVT、IIIT、IC15、SVTP 分别为1.4%、 0.6%、 1.4%和
CUTE 数据集。此外,ABINet-LV 具有
更强大的VM实现了0.6%、 1.2%的提升，IC13、SVT、IC15、SVTP 和 CUTE 分别为1.8%、 1.4%、 1.0%
基准高于其对应标准。

与最近在 MJ 上训练的 SOTA 作品相比
和 ST 一样,ABINet 也表现出色（表6）。
尤其ABINet在SVT、SVTP方面有显著优势
和 IC15,因为这些数据集包含大量低质量图像,例如噪声和模糊图像，
VM 无法自信地识别。此外,我们还
发现带有不寻常字体和不规则文本的图像可以
成功被认定为语言信息法案
作为视觉特征的重要补充。因此
ABINet 甚至可以在 CUTE 上取得第二好的结果
图像校正。

4.4. 半监督训练

为了进一步突破准确阅读的界限,我们
探索一种利用 MJ 和
ST 为标记数据集,Uber-Text 为未标记数据集
数据集。第 3.5节中的阈值Q设置为 0.9,并且
Bl和Bu的批量大小分别为 256 和 128。表6中的实验结果表明,所提出的自训练
方法 ABINet-LVst可以轻松胜过 ABINet-LV
在所有基准数据集上。此外,集成自训练 ABINet-LVest表现出更稳定的性能
通过提高数据利用效率。观察
在提升的结果中,我们发现稀缺的困难例子
字体和模糊的外观也可以经常被识别（图8），这表明探索半
/无监督学习方法是一个有前途的方向
场景文本识别。

5. 结论

在本文中,我们提出了 ABINet,探索在场景中利用语言知识的有效方法
文本识别。ABINet 1)自主,通过强化学习来提高语言模型的能力

明确地;2)学习文本表示的双向
通过共同调节双方的性格背景；
3)迭代逐步修正预测
减轻噪声输入的影响。基于ABINet
我们进一步提出了一种集成自训练方法
半监督学习。标准实验结果
基准测试证明了 ABINet 的优势,尤其是在低质量图像上。此外,我们还声称
利用未标记数据是可能的,并且很有前景
达到人类水平的识别。

参考

[1]吉米·雷·巴、杰米·瑞安·基罗斯、杰弗里·E·欣顿。层规范化。arXiv 预印本 arXiv:1607.06450,2016 年。4

[2] Jeonghun Baek、Geewook Kim、Junyeop Lee、Sungrae Park、韩东润、尹相斗、吴成俊、和淑李。场景文本识别模型比较出了什么问题?数据集与模型分析。在

IEEE 国际计算机视觉会议,页数
4715–4723,2019年7月

[3] Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang
Pu 和 Shuigeng Zhou. 聚焦注意力:迈向精准
自然图像中的文本识别。在
IEEE 国际计算机视觉会议,页数
5076–5084,2017年。2,3

[4] Zhanzhan Cheng, Yangliu Xu, Fan Bai, Yi Niu, Shiliang
Pu 和周水庚。Aon:走向任意导向
文本识别。在 IEEE 会议论文集上
计算机视觉和模式识别,第 5571-5579 页,
2018. 2

[5] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina
Toutanova。Bert:用于语言理解的深度双向Transformer的预训练。载于NAACL-
HLT论文集,第4171–4186页,2019年。1 , 2, 4

[6] Shancheng Fang, Hongtao Xie, Zheng-Jun Zha, Nannan Sun,
Jianlong Tan, and Yongdong Zhang. Attention and language
基于卷积序列建模的场景文本识别集成。刊于第26届ACM多媒体国际会议论文
集,第248-256页,2018年。
3

[7] Alex Graves, Santiago Fernandez, Faustino Gomez, 和Jurgen
Schmidhuber. 联结主义时间分类:利用循环神经网络标记未分段序列数据
网络。第 23 届国际会议论文集
关于机器学习,第 369-376 页,2006 年。2

[8] Alex Graves, Marcus Liwicki, Santiago Fernandez, Roman Bertolami,
Horst Bunke, 以及 Jurgen Schmidhuber. 一种用于无约束手写识别的新型
联结系统. IEEE 模式分析与机器翻译学报, 2017, 19(1): 17-20.
情报,31(5):855–868, 2008.3

[9] Ankush Gupta, Andrea Vedaldi 和 Andrew Zisserman.自然图像中文本定位的合
成数据。IEEE计算机视觉与模式识别会议论文集
认可,第 2315-2324 页,2016 年。5

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun.
深度残差学习在图像识别中的应用。IEEE计算机视觉与模式识别会议论文集
认可,第 770-778 页,2016 年。4

[11] Pan He, Weilin Huang, Yu Qiao, Chen Change Loy, and
汤晓鸥,深度卷积序列中的场景文本读取,第三十届AAAI人工智能会议， 2016,2

[12] Wenyang Hu, Xiaocong Cai, Jun Hou, Shuai Yi, and Zhiping
Lin. Gtc:CTC引导训练,高效精准
场景文本识别。载于AAAI,第11005-11012页,2020年。2

[13] Max Jaderberg、Karen Simonyan、Andrea Vedaldi 和 An- drew Zisserman。
合成数据和人工神经网络
用于自然场景文本识别。在 NIPS 深度学习
研讨会， 2014.2,5

[14] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, 和 Andrew Zisserman.
深度结构化输出学习在无约束文本识别中的应用. 国际会议, 2017, 19(3):
19-23.
学习表征（ICLR）,2015. 1, 2

[15] Max Jaderberg、Karen Simonyan、Andrea Vedaldi 和 Andrew
Zisserman。利用卷积神经网络在野外阅读文本
神经网络。《国际计算机视觉杂志》，
116(1):1–20, 2016.1 , 5

[16]马克斯·贾德伯格、安德里亚·维达尔迪、安德鲁·齐瑟曼。
文本识别的深度特征。在欧洲会议上
计算机视觉,第 512-528 页。Springer,2014 年。1, 2

[17] Dimosthenis Karatzas、Lluis Gomez-Bigorda、Anguelos Nicolaou、
Suman Ghosh、Andrew Bagdanov、Masakazu Iwamura 、Jiri Matas、
Lukas Neumann、Vijay Ramaseshan Chandrasekhar 、Shijian Lu
等。 ICDARE 2015 强劲竞赛
阅读。第13届国际文档分析与识别会议,第1156-1160页。IEEE,2015年。5

[18] Dimosthenis Karatzas、Faisal Shafait、Seiichi Uchida，
岩村正和、路易斯·戈麦斯·比戈尔达、塞尔吉·罗伯斯
梅斯特、琼·马斯、大卫·费尔南德斯·莫塔、乔恩·阿尔马赞·阿尔马赞和
路易斯·佩雷·德拉斯赫拉斯。 Iddar 2013 稳健
阅读比赛。2013年第十二届国际会议
关于文档分析和识别,第 1484-1493 页。
IEEE， 2013.5

[19] Chen-Yu Lee 和 Simon Osindero。递归循环网络
并利用注意力模型实现 OCR 在野外的应用。论文集
IEEE 计算机视觉与模式识别会议
Recognition,第 2231–2239 页， 2016年。1、2、3

[20] Hui Li, Peng Wang, Chunhua Shen, and Guyu Zhang. Show,
参加并阅读:不规则的简单而强大的基线
文本识别。在 AAAI 会议论文集上
人工智能,第 33 卷,第 8610-8617 页,2019 年。2

[21] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei.
完全卷积实例感知语义分割。
在 IEEE 计算机视觉会议论文集上
和模式识别,第 2359-2367 页,2017 年。2

[22] Minghui Liao, Pengyuan Lyu, Minghang He, Cong Yao,
吴文浩和白翔。Mask textspotter:一种端到端可训练神经网络,用于识别任意
形状。IEEE 模式分析与机器翻译
情报， 2019.8

[23] Minghui Liao, Jian Zhang, Zhaoyi Wan, Fengming Xie, Jia-jun Liang, Pengyuan
Lyu, Cong Yao, and Xiang Bai. Scene
二维视角下的文本识别。在AAAI人工智能会议论文集上，
第 33 卷,第 8714-8721 页,2019 年。2

[24] Shangbang Long, Xin He, and Cong Yao. Scene text detection
与认知:深度学习时代。《国际期刊》
计算机视觉,第 1-24 页,2020 年。1

[25] Pengyuan Lyu, Zhicheng Yang, Xinhang Leng, Xiaojun Wu,
Ruiyu Li, and Xiaoyong Shen. 2d attentional irregular scene
文本识别器。arXiv 预印本 arXiv:1906.05708,2019 年。2、3 、
5、6、8

[26]斯蒂芬·梅里蒂、熊才明、詹姆斯·布拉德伯里
Richard Socher。指针哨兵混合模型。arXiv
预印本 arXiv:1609.07843, 2016. 6

[27] Anand Mishra、Karteek Alahari 和 CV Jawahar。场景文字
使用高阶语言先验进行识别。在英国
机器视觉会议（BMVC）， 2012.5

[28] Matthew E Peters、Mark Neumann、Mohit Iyyer、Matt Gardner 、
Christopher Clark、Kenton Lee 和 Luke Zettlemoyer。
深度语境化的词语表征。在
NAACL-HLT,第 2227-2237 页,2018 年。1

[29] Zhi Qiao, Yu Zhou, Dongbao Yang, Yucan Zhou, and Weip-ing Wang.
Seed: Semantics enhanced encoder-decoder
场景文本识别框架。在
IEEE/CVF计算机视觉与模式会议
Recognition，第 13528–13537 页， 2020年。1、3、8

[30] Trung Quy Phan、Palaiahnakote Shivakumara、尚轩
Tian 和 Chew Lim Tan。通过透视识别文本
自然场景中的失真。在 IEEE
国际计算机视觉会议，第 569-69 页
576，2013.5

[31]安哈尔·里斯努马万 (Anhar Risnumawan)、帕莱亚汉科特·希瓦库马拉 (Palaiahankote Shivakumara)、Chee Seng
Chan 和 Chew Lim Tan。一种用于自然场景图像的鲁棒任意文本检测系统。专家系统
应用，41(18):8027–8048, 2014.5

[32]奥拉夫·罗纳伯格、菲利普·费舍尔和托马斯·布洛克斯。 U网：
用于生物医学图像分割的卷积网络。
在国际医学图像计算和
计算机辅助干预，第 234-241 页，2015 年。4

[33]盛芬芬，陈志能，徐波。Nrtr：一种用于场景文本识别的非循环序列到序列模型。
2019 年国际文档分析与识别会议 (ICDAR)，第 781-786 页。IEEE，2019 年。
1、2

[34] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end
用于基于图像的序列识别的可训练神经网络及其在场景文本识别中的应用。IEEE
模式分析和机器智能方面的交易，
39(11):2298–2304, 2016.2

[35] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao,
和 Xiang Bai。具有自动
整改。在 IEEE计算机视觉和模式识别会议论文集，第 4168-4176 页，2016 年。
2、3

[36] Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan
Lyu、Cong Yao 和 Xiang Bai。《Aster：引人注目的场景》
具有灵活校正功能的文本识别器。IEEE 交易
模式分析与机器智能，41（9）：2035-
2048，2018 . 1，2，3

[37]苏博兰、陆诗建. 汉语词汇准确识别
使用循环神经网络进行字符分割的场景
网络。模式识别， 63 :397–405，2017。2.

[38] Martin Sundermeyer、Ralf Schluter 和 Hermann Ney。用于语言建模的
LSTM 神经网络。在第十三届国际语音通信年会上
协会， 2012.1

[39] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit 、
Llion Jones、Aidan N Gomez、 ukasz Kaiser 和 Illia
波洛苏欣。你只需要注意力。在神经科学进展中
信息处理系统，第 5998-6008 页，2017 年。1，
3、4

[40] Zhaoyi Wan, Mingling He, Haoran Chen, Xiang Bai, and
丛瑶。Textscanner：按顺序读取字符
鲁棒场景文本识别。2020. 2, 3, 8

[41] Zhaoyi Wan, Jielei Zhang, Liang Zhang, Jiebo Luo, and Cong
姚。场景文本识别中的词汇依赖性。IEEE /CVF 计算机视觉会议论文集
和模式识别，第 11425-11434 页，2020 年。1

[42] Kai Wang、Boris Babenko 和 Serge Belongie。端到端
场景文本识别。2011 年国际会议上
计算机视觉，第 1457-1464 页。IEEE，2011 年。1, 5

[43] Peng Wang, Lu Yang, Hui Li, Yuyan Deng, Chunhua Shen,
以及张艳宁。一个简单而强大的卷积注意力网络，用于不规则文本识别。arXiv 预印本
arXiv:1904.01375, 6, 2019. 2

[44] Tianwei Wang, Yuanzhi Zhu, Lianwen Jin, Canjie Luo, Xi-aoxue Chen,
Yaqiang Wu, Qianying Wang, and Mingxiang
蔡。用于文本识别的解耦注意力网络。在
AAAI，第12216–12224 页， 2020年。1,2,3,8

[45] Zbigniew Wojna、Alexander N Gorban、Dar-Shyang Lee，
Kevin Murphy, Qian Yu, Yeqing Li, and Julian Ibarz.
基于注意力机制的结构化信息提取
街景图像。在2017年第14届IAPR国际文档分析与识别会议（ICDAR)上，
第 1 卷，第 844-850 页。IEEE，2017 年。2 , 3

[46] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le.
利用嘈杂学生模型进行自训练，提升 ImageNet 分类效果。载于《IEEE/CVF 计算机视觉与模式识别会议论文集》，第 10687–10698 页。
2020年5月

[47] Mingkun Yang, Yushuo Guan, Minghui Liao, Xin He, Kaigui
边松白、姚聪、白翔。用于场景文本识别的对称约束校正网络。
在 IEEE 国际计算机视觉会议论文集，第 9147-9156 页，2019 年。2

[48] Xiao Yang, Dafang He, Zihan Zhou, Daniel Kifer, and C Lee
Giles. 学习利用注意力机制阅读不规则文本。IJCAI，第1卷，第3页，2017年。2

[49] Deli Yu, Xuan Li, Chengquan Zhang, Tao Liu, Junyu Han,
刘静拓、丁尔瑞. 迈向精准场景文本
语义推理网络识别。在 IEEE/CVF 计算机视觉与
模式识别，第 12113-12122 页， 2020年。1、2、3、5、6 、8

[50] Xiaoyu Yue, Zhanghui Kuang, Chenhao Lin, Hongbin Sun,
以及 Wayne Zhang。Robustscanner：动态增强
位置线索在文本识别中的应用。
2020 年欧洲计算机视觉会议（ECCV）。
5、8

[51] Fangneng Zhan and Shijian Lu. Esir: End-to-end scene text
通过迭代图像校正进行识别。在会议论文集
IEEE 计算机视觉与模式识别会议
《认可》，第 2059-2068 页，2019 年。2

[52]张瑛，Lionel Gueguen，Ilya Zharkov，Peter 张，
Keith Seifert 和 Ben Kadlec。Uber-text：大规模
用于街道级图像光学字符识别的数据集。在 SUNw：场景理解研讨会 - CVPR
2017，2017.5