**摘要**

CTC和注意机制是最近场景文本识别工作中使用的两种主要方法。 与基于注意力的方法相比，CTC解码器的推理时间要短得多，但准确性较低。 为了设计有效的模型，我们提出了GTC， 利用一个更强的注意力机制指导CTC模型能学习到更好的对齐方式和特征表示。 利用指导训练的优势，CTC模型可以对规则和不规则场景文本进行可靠而准确的预测，同时保持快速的推理速度。利用指导训练的优势，CTC模型能在保持快速的推理速度的同时，对规则和不规则场景文本做出更鲁棒而且准确的预测。此外，为了充分利用CTC解码器的潜力，提出了一种图卷积网络（GCN），以学习提取特征的局部相关性。在标准基准上进行的大量实验证明，端到端模型实现了针对规则和不规则场景文本识别的最好的结果，并且推理时间比基于注意力的方法短6倍。

**引言**

场景文本识别已经在学术界和工业界中进行了多年的研究，并且在各种现实生活中起着重要的作用，例如车牌识别，身份认证和文本分析。 近年来，提出了许多方法来识别自然场景文本。 但是，由于场景文本的大小，字体，颜色和字符位置不同，因此场景文本识别仍然是一项艰巨的任务。

**当前的识别框架**： 通常场景识别的模型框架是编码解码结构。最近的方法主要是用俩种技术来训练序列识别模型，即CTC和注意力机制。受到语音识别的启发，引入了CTC，将帧概率和标签对齐。在基于CTC的方法中，CNN用来提取特征序列，RNN用与对特征序列进行建模。他们经过CTC损失训练，使用并行解码实现快速预测。基于注意力的方法使用注意力机制来捕捉文本行中每个字符的依赖性，和基于CTC的方法相比，它能更好的学习对齐方式和深层网络特征。此外，一些校正方法被提出用于文本图像预处理。校正模块可根据Thin Plate Spline（TPS，薄板样条）转换来转换输入文本图像并校正字符对齐方式。 然后将校正后的图像传递到编码器-解码器结构以进行识别。 该模块可以添加到CTC或基于注意力的方法中，并且以端到端的方式进行培训以学习自适应转换。

**动机**：尽管先前的方法在常规或不规则场景文本识别方面都得到出了不错的结果，但是它们仍然在平衡识别精度和推理时间之间取得平衡方面存在局限性（见图1）。 由于基于注意力的方法根据先前时间步长中的特征进行预测，因此这种非并行解码方案将大大减慢推理过程。

尽管基于CTC的方法相对高效，但不如基于注意力的方法准确，因为基于CTC的方法会误导其特征对齐和特征表示的训练。 为了设计一种高效且高准确率的场景文本识别器，我们旨在优化CTC模型。 为了克服CTC的局限性，我们有两个动机：（1）通过更准确的指导中学习更好的特征表示，以及（2）建立局部特征之间的相关性。

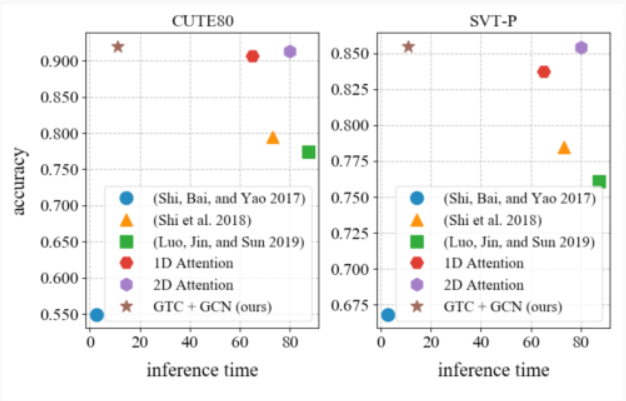
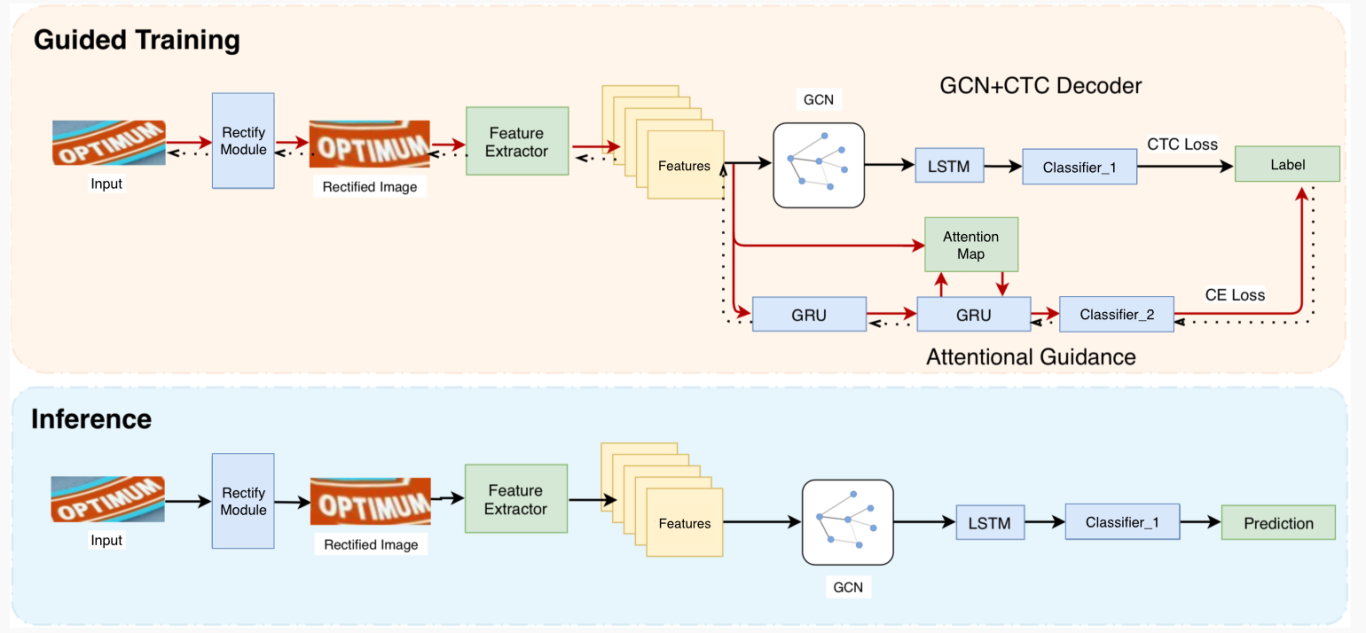


图1：不同方法的精度与推理时间（ms /图像）的权衡表。 第一个是CTC，第二个是ASTER，第三个是MORAN，1D注意和2D注意表示两种基于注意的方法。 GTC+GCN方法更快，更有效。



图二：本文提出的方法概述。 箭头的不同颜色表示不同的计算图。 注意，在反向传播过程中，从CTC损失计算出的梯度仅用于更新GCN + CTC解码器的权重。指导网络解码器和GCN + CTC解码器的真实标签相同。

**工作：**提出的GTC是一种CTC优化的新方法。 所提出的方法使用如图2所示的端到端可训练框架。校正模块是一个简单的转换网络，将校正应用于文本图像。CTC解码器用于训练和评估，让推理具有高效性。作为指导，它由有高准确率的注意解码器组成，该解码器将不会在推理中使用。一个强大的网络的编码器可以更好的学习对齐方式和特征表示，并且其特征图更易于解码。因此，CTC模型可以从指导中学习，并变得更加有效。根据交叉熵损失计算出的梯度将用于直接优化校正模块、ResNet CNN和注意力解码器，从而提供有力的指导。 指导的选择是灵活的，这使得此方法非常普遍。

由于CTC解码允许重复字符和空白标签，因此可以在多个时间步长中预测一个标签。 我们假设相邻的时间步具有补充特征，并且特征序列内部存在某些相关性。 例如，如果只考虑字母“ H”的一半，则字母“ H”可能会误分类为“I”。 为了合并特征序列的补充特征，我们进一步提出了一种新颖的GCN模块，以捕获每个序列切片的依赖性，并根据相关性合并属于同一标签的特征。在保持了快速的推理速度同时，进一步提高了CTC解码器的鲁棒性和准确性。

综上所述，本文的贡献包括三个方面：

1）设计了一种新颖的场景文本识别GTC方法，更灵活，有效。 该方法可以被各种基于CTC的方法采用，并且可以通过强大的指导下学习而使CTC模型更准确，这可能成为提高基于CTC的方法准确性的通用方法。我们还使用不同的指导说明GTC是CTC模型优化的通用方法。

2）这是将图形应用于场景文本识别并通过使用GCN建立序列相关性的首次尝试，这提高了CTC解码器的准确性和鲁棒性。 通过广泛的实验，我们证明了GCN模块是高准确率的。

3）我们提出的方法的推理时间比基于注意力的方法短6倍，并且在大多数规则和不规则场景文本数据集上都达到了最新水平，与其他的方法相比，该方法更加有效。

**相关工作**

**常规场景文本识别**：早期方法，Bissacco将场景文本识别视为字符分割和识别问题。 但是，复杂的背景和场景文本的不同对齐方式使字符分割成为要训练的最具挑战性的部分。 此外， Jaderberg还使用了多分类方法来直接预测每个单词，但是该方法在很大程度上受字典大小的限制。 Lee和Osindero提出了一种带有注意力解码器的递归模型R2AM，用于常规场景文本识别。受语音识别中序列到序列对齐的启发，Shi，Bai和Yao将CTC解码器引入了带有BiLSTM的场景文字识别中，用于对特征序列进行建模，称为CRNN。Wang and Hu提出了一个门控循环卷积神经网络（GRCNN），该训练通过CTC损失训练，可以进行常规文本识别。 另外，Liu提出了一种经过交叉熵损失训练的二进制卷积编码器/解码器网络（B-CEDNet），并实现了识别部分的快速推理。 但是，它是为常规场景文本识别而设计的，并且需要像素级的标签进行训练。 受CTC的启发，Bai提出了“编辑概率”来优化训练过程，因为缺少或过多的角色可能会误导CTC训练。 Zhang还引入了一种域自适应方法来变长文本识别。 最近的常规文本识别的主要方法仍然是基于CTC的，该方法能自动完成特征序列与标签之间的对齐。 但是，CTC无法预测不规则的场景文本，因为曲率会误导对齐。

**不规则场景文本识别：**近年来，识别不规则场景文本已引起越来越多的关注，因为这是一个更具挑战性的问题。 由于不规则文本的扭曲和曲率，Shi et al.2018和Shi et al.2016首先基于空间变换网络（STN）纠正了不规则文本，从而使文本图像更规则，更易于识别。Luo，Jin，Sun2019和Zhan and Lu2019专注于改进校正管道以获得更好的转换结果。这些方法都使用了基于注意力的解码器。Liu et al.2016使用了基于CTC的解码器，其中STN校正了输入图像。有些方法没有校正图像，而是直接识别不规则的文本。 由于特征图和注意力区域之间的位置不一致，Cheng 2017使用了聚焦网络来调整“注意力漂移”。 Liu，Chen，and Wong 2018还提出了字符感知神经网络（Char-Net）来纠正单个字符。 但是，这两种方法都需要字符级注释。 Cheng et al.2018在注意力网络之前使用多方向方法对特征进行编码。 除此，还有其他方法可将注意力机制扩展到2D特征图中，因为2D空间有更多的空间依赖性。Yang 2017和Liao 2019最近在他们的注意力网络中使用了2D局部特征，但是这两部作品也需要字符级注释。 Li 2019提出了一种2D注意解码器，显着提高了不规则文本识别的性能，尽管推理时间相对较长。Wang 2019提出了一种基于变换的解码器，该解码器也连接到2D特征图。 该方法实现了并行训练，但解码使用的非并行方式。 在提到的方法中，基于注意力的（Attention或STN + Attention）方法通常对不规则场景文本数据集具有更高的准确性，但是注意力解码器会减慢推理速度。 尽管STN + CTC保持更快的推理时间，但CTC解码器的容量限制了其性能。

与上述方法不同，本文使用指导训练来优化具有更好图像表示的CTC模型。 我们还通过使用设计的GCN层，在序列之间建立空间和上下文相关性。据我们所知，这可能是第一次在场景文本识别中使用了指导的训练方法和GCN层。

尽管Cheng 2017使用大量实验表明，CTC和注意力网络的直接结合在场景文本识别中效果不佳，但他们没有给出解释。 其背后的原因是CTC降低了特征表示的学习。 我们使用可靠的实验表明，通过从有效的指导中学习，CTC模型可以实现更好的效果。与Kim，Hori和Watanabe 2017不同，他们通过CTC-attention共享编码器，并使用语音识别中的注意解码器进行评估，我们网络中的编码器和校正模型仅根据指导中计算出的梯度进行优化。在测试时，仅使用CTC解码器进行预测，因此推理时间比注意方法要短得多。该指导仅在训练阶段使用，该阶段监督CTC模型以学习更好的上下文对齐和特征表示。因此，GTC使CTC模型更加有效。

**方法**

**总览**

指导训练的提出是用于克服CTC本身的局限性。 在CTC模型的推理中，仅选择每个时间步长的最大概率作为最终预测。 但是，在训练中，单个时间步长中的不同概率会导致丢失不同的CTC路径。CTC训练将使特征表示能够忍受某些预测误差。 如果我们有一个真实标签“AB”，则对于3个时间序列输出，CTC路径（伪标签）可以是“A-B”或“-AB”或“AB-”或“AAB”或“ABB”。由于CTC损失计算的标签含糊不清，因此在每个时间步骤中学习特征表示都很混乱。缺少或多余的字符可能会降低其特征对齐和特征的学习。尽管我们对注意力模型和CTC模型使用了相同的编码器，但是在实验中我们发现CTC模型的编码器具有较差的特征表示。我们断言，CTC编码器的性能实际上受CTC损失本身的限制。

因此，指导训练可以为CTC模型提供更好的特征。

本节介绍了提出的GTC，其中一个常规的注意解码器用作指导，而GCN + CTC解码器用于训练和推理。 如图2所示，我们的网络包括四个部分。 第一部分是STN，与Shi et al.2016中的相同。它将输入图像转换为归一化图像。 第二部分是用于特征提取的ResNet主干网络，已广泛用于场景文本识别中。 第三部分是注意指导，它使用注意机制输出文本序列。 第四部分是由GCN支持的CTC解码器，它增强了特征序列的相关性。STN，ResNet-CNN和注意力指导网络仅通过交叉熵损失进行训练，而GCN + CTC解码器则通过CTC损失进行训练。

**STN**

由于自然场景中的许多文本图像都带有弯曲的文本和不同的视角，因此采用了转换模块以实现可靠而准确的识别，该模块将空间转换应用于文本图像并规范化了字符区域。 它是一个可微的模块，由本地化网络和网格生成器组成。本地化网络将预测转换参数并使用它们来创建网格。网格和输入图像将由生成器采样以生成转换后的输出。

**特征提取器**

我们选择ResNet50（He，2016年）作为网络的主干，如表1所示。为了提取更多的精确特征，我们将原始残差步长从2更改为1。我们还添加了两个最大池化层，用于对特征图进行下采样。提取的特征序列h1:T具有固定的高度和变化的长度，将用于解码。

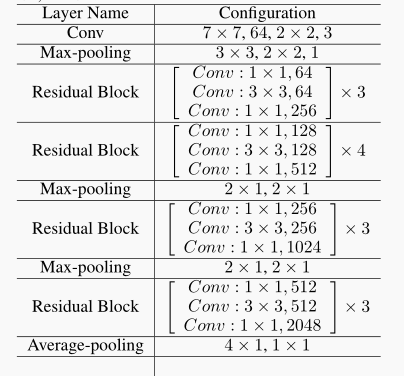


表1：ResNet50特张提取器的网络配置。“卷积”是指卷积层，包括其内核大小，输出通道，步长和填充。所有“Resudual Block”的步幅均设置为1。“最大池化”和“平均池化”的配置表示其内核大小，步幅和填充。 总体下采样率是W：1/4，H：1/16。

**注意力指导**

受机器翻译的启发，seq2seq模型用于将特征序列转换为字符序列，从而对齐输出和标签。这种模型中的注意力机制具有捕获输出依存关系并在每个时间序列上关注字符区域的能力。为了进行公平的比较，我们选择了一种常用注意解码器（Shi et al.2016; Cheng et al.2017; Zhan and Lu 2019）以证明GTC的有效性。我们在ResNet主干网的顶部采用注意的序列到序列解码器。它基于RNN生成长度为T的目标序列，表示为（y1，y2 ...，yT）。

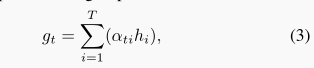
注意解码器可以预测字符或序列结束标记“ EOS”。当它预测为“ EOS”时，它将停止预测。 GRU用来学习注意力依赖性。 在时间序列t，输出yt为：yt = Softmax（WTst），（1）

其中，st是GRU单元的隐藏状态，W是可训练参数。

隐藏状态st通过GRU的循环过程更新：

st = GRU（yprev，gt，st-1），（2）

其中yprev先前输出yt-1的嵌入向量。 在训练过程中，yt-1被真实序列代替。gt为瞥见向量，计算如下：



hi是时间序列i时的h1:T特征序列向量。αt时注意力权重：

αt= Attention(st−1,hi),(4)

如Luong, Pham, and Manning 2015中介绍。

注意指力导与交叉熵损失一起训练，并且预测结果将不会用于评估中。

**GCN+CTC 解码**

给定长度为T的概率分布y1：T的序列，它寻找产生相同标签序列l的多个路径M-1(l)，从而允许重复连续字符或空白标签。yt表示时间序列t在分类标签L上的概率分布，包括了空白标签。M定义为将所有可能的路径π映射到目标标签的操作。例如，它将路径“-hh-e-lll-oo-”映射为“hello”。CTC 通过优化所有路径上的概率总和，达到训练网络的目的。

在CRNN中，BiLSTM用于通过从两个方向读取文本行来提取序列特征。 但是，由于字符出现在不同的位置，因此它缺乏集中于局部区域的能力。图卷积网络（GCN）（Kipf and Welling 2016）是图数据上CNN的有效变体，其中图的边缘表示数据中的隐式连接。给定一个由图定义的关系，图卷积将先别说从节点传递到其相邻接点。我们在BiLSTM之前使用了一个特殊的GCN层，其中相似邻接矩阵和距离矩阵相结合来描述空间上下文相关性。

给定ResNet CNN的特征图h1：T，邻接矩阵可以通过计算每两个序列切片之间的相似性来学习：

AS(i,j) = f(ci,cj), (8)

相似度投影函数定义为：

C:\Users\wlq\AppData\Local\Temp\1581536706(1).png

ci是hi的线性变换结果。 该公式计算成对的余弦相似度。 除了使用相似关系来关注相似特征外，距离矩阵还用于约束相似性以关注相邻特征。距离矩阵定义为：

C:\Users\wlq\AppData\Local\Temp\1581536839(1).png

其中di，j = | i − j |，β是比例因子。 因此，GCN + CTC的最终输出计算如下：

X = (AS∗ AD)HWg, (11)

Wc是可选的权重矩阵，X随后传给BiLSTM用于序列建模。

C:\Users\wlq\AppData\Local\Temp\1581537083(1).png

其中Wc是用于分类的权重矩阵，而Seq是BiLSTM，隐藏层大小为512。logits和标签l最终用于计算CTC损失，以训练GCN + CTC解码器。

总结，GTC使用功能更强大的模型来指导CTC解码器，在该模型中，根据CTC损失计算出的梯度将不会用于更新校正模块，ResNet CNN，特征图或注意力指导。CTC解码器会通过注意指导的训练过程进行自我更新，在该过程中，CTC解码器将学习根据更好的特征表示和更好的对齐方式进行预测。 GCN建立了特征之间的关联，并进一步提高了性能。

**实验结果**

