不良文字图片识别技术研究

杜刚[1]，戴晶[2]，张晨[1]，杜雪涛[1]

(1 中国移动通信集团设计院有限公司，北京 100080；2 中国移动通信集团有限公司，北京 100053)

摘　要　5G消息服务的开展为不良图片信息的传播提供了便利条件。不良文字图片作为一种特殊的不良图片信息给通信网络的内容安全带来了挑战。为了能够对不良文字图片进行有效治理，运营商需要借助人工智能技术对图片中的文字信息进行识别和提取。本文详细介绍了不良文字图片治理整个技术过程需要引入的3个处理步骤，并对3个处理步骤涉及到的深度学习模型结构和原理进行了深入的研究。本文的研究内容对运营商进行不良文字图片治理具有很大的技术参考价值。

关键词　目标检测；文字区域检测；OCR

中图分类号　TN918　　**文献标识码**　A　　**文章编号**　1008−5599（2021）06−0032−06

DOI:10.13992/j.cnki.tetas.2021.06.007

不良图片信息通常是指那些散播谣言、扭曲事实和煽动情绪的图片信息。这类信息严重影响社会舆论，甚至引发线下群体活动，扰乱社会秩序。随着未来 5G 消息服务的上线，图片的传播分享将更加便利。不法分子也会借势而起，更加猖獗地传播不良图片。为了有效遏制不良图片传播势头，需要研究有效的不良图片识别技术。

1　不良文字图片识别技术

目前一种比较有效的方法是将图片中的文字信息识别并提取出来，再对提取出的文字做语义上的判断。可以借助人工智能领域中自然场景文字识别技术来达到上述目的。如图 1 所示，使用自然场景识别技术，可将不良文本图片识别分为 3 个步骤。

（1）文字区域检测。该技术的主要任务是在一幅图片中锁定出现文字的区域，此处限定文字区域为任意四边形。其主要原理是通过在图片中寻找文字视觉特征的稠密区域，并使用四边形对文字区域进行范围圈定。

图1　文本图片识别流程

（2）文字形态矫正。由于图片的拍摄角度问题，文字在图片中可能会有三维旋转效果。这种情况下定位的四边形文字区域很可能不是矩形（可能由于近大远小形成梯形或由于角度偏斜形成平行四边形等）。文字形态矫正将不规则四边形文字区域通过透视变换校正为矩形文字区域。其过程类似于将观察文字的角度调整到文字

收稿日期：2020−03−22

的正对面，从而大幅提高后续光学字符识别（OCR）的效率。

（3）图片文字识别。该技术主要利用 OCR 技术将图片中的文字还原为文本信息。由于许多不良文字信息都偏好使用繁体字，故需要同时考虑对简体汉字和繁体汉字的识别。同时，由于文字在旗帜和横幅等不平整物体上，可能会发生扭曲变形，甚至会被物体遮挡，故需要 OCR 模型具有较强的抗干扰能力。

下面将从如上 3 个阶段所涉及技术进行详细介绍。

2　文字区域检测技术

文字区域检测属于目标检测中的一个特例，可以使用目标检测算法进行文字区域检测。目标检测领域主要形成了两个研究分支，一个是以仅需一次视觉处理（YOLO）算法为代表的基于滑动窗口的目标检测方法；另一种是以卷积特征区域（RCNN）为代表的基于语义分割的目标检测方法。基于这两个分支分别衍生出了一系列目标检测算法，也衍生出一系列文字区域检测方法。其中基于滑动窗口分支的文字区域检测方法以高效高精度场景文字检测（EAST）为代表；基于语义分割分支的文字区域检测方法以遮罩文本检测器（Mask TextSpotter）为代表。

从文字区域的形状上来说，文字区域检测可分为矩形文字区域、任意四边形文字区域和不规则文字区域 3 种情况。若使用目标检测算法进行文字区域识别，其可识别的文字区域通常是矩形框。当文字由于拍摄角度发生扭曲或旋转时，无法精确地圈定文字区域。使用 EAST 算法可支持识别不规则四边形的文字区域，从而实现更加精确的定位。Mask TextSpotter 更进一步可以实现任意形状文字区域的检测，从而更好地圈定文字区域。YOLO 算法、EAST 算法、Mask TextSpotter 算法圈定同一文字区域的情况，如图 2 所示。由图可以看出，Mask TextSpotter 圈定的文字区域最佳。

图2　YOLO算法、EAST算法、Mask TextSpotter算法圈定文字区域示意图

文字区域定位的准确性直接影响到了进一步的 OCR 效果，需要根据文字信息分布的实际情况进行合理的选取。在不良文字图片中，文字的区域大部分情况下为矩形区域，一些情况下会出现不规则四边形区域，极少情况下出现更复杂的文字区域。故可以选择任意四边形文字区域作为主要的检测对象。

EAST 模型是目前最成熟的检测任意四边形文字区域的模型。其特点是利用卷积操作实现滑动窗口效果，使用多层卷积实现多种尺度的滑动窗口，从而实现对图片中各种大小的四边形文字区域进行快速识别。

如图 3 所示，EAST 模型可以分为编码器层、解码器层和输出层 3 个部分。

编码器层通过多层卷积和池化层的堆叠提取图片

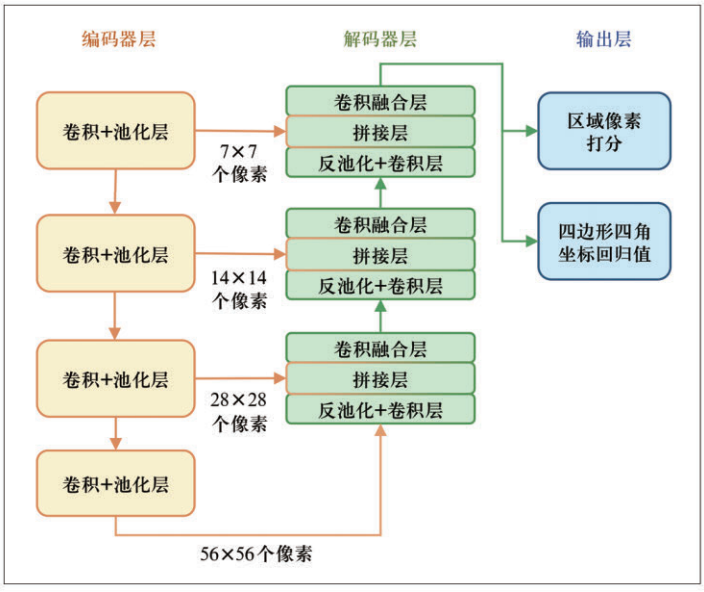

图3　文字区域锁定模型网络结构

中各种尺度滑动窗口下的视觉特征。其中每一次池化操作后，滑动窗口尺度就会放大一倍，相应能够识别的文字区域的字体大小也会随之增大一倍。窗口尺寸随着不断卷积和池化由 7×7 个像素逐步放大到 56×56 个像素。编码器通常使用 VGG16 和 ResNet 等模型。通过选取网络中不同卷积池化层输出，可以获得不同尺度滑动窗口大小下的视觉特征。每个选定的卷积层加池化层输出的视觉特征都会分阶段输入到解码器层进行特征融合。

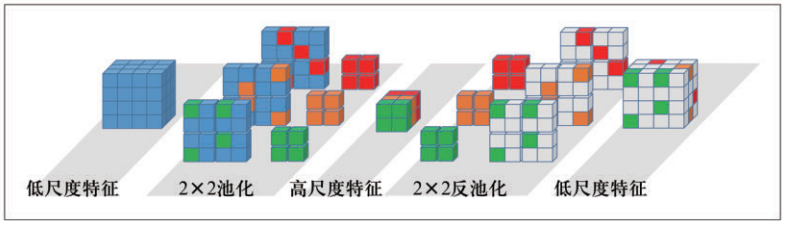

图4　池化与反池化示意图

解码器层是由若干特征融合单元构成，每个特征融合单元包含反池化层、拼接层和卷积融合层 3 个部分。其中反池化层负责将高尺度特征扩展到更低尺度，使其能够与低尺度特征进行拼接。

具体方式如图 4 所示，左侧的低尺度特征图的维度为 4×4×3。2×2 池化操作首先将特征图的 3 个通道分开；其次将每个通道下每 2×2 个像素分成一组。针对每一组像素，取该组像素最大值作为该组的代表，每个通道选取 4 个代表（不同通道的代表用不同颜色标出）。各个通道选出的代表再次组合形成高尺度特征。经过 2×2 池化后，高尺度特征的一个像素点代表了低尺度特征 4 个像素点。注意，在各个通道选出代表时，模型记住了代表在原通道中的位置，这样便于在反池化时恢复代表原先的位置。

在反池化操作时，高尺度特征首先按通道拆分；其次构造一个宽高各扩大一倍的矩阵，将高尺度特征中的像素还原到其原先所在位置，其余元素（图 4 中灰色的元素）全部填写 0；最后，各个通道进行合并，形成低尺度特征。经过反池化操作，高尺度特征的尺度转化为低尺度特征的维度，从而可以实现低尺度特征之间的拼接。

卷积融合层将拼接后的特征进行降维和融合，保留突出的文字视觉特征。多个特征融合单元融合了各种尺度的视觉特征，为后续输出层输出各种尺度的文字区域打下基础。

输出层包含两个输出，第一个输出是对特征图中每个像素使用 sigmoid 函数进行打分，看其是否属于文字区域；第二个输出是对应文字区域的四边形四角坐标。结合两个输出，可以找到所有文字区域的四边形四角坐标。最后，通过特征图与原图的比例关系，可以推得文字区域在原图中的四角坐标。

3　文字形态矫正技术

如图 5 所示，图片中文字的呈现角度可能任意。一方面文字可能有轻微的旋转角度；另一方面，文字可能由于拍摄角度问题产生近大远小的效果，从而产生不规则的四边形文字区域。文字区域矫正是将图中不规则的四边形文字区域进行透视变换，从而使文字姿态更加标准，进而提高后续 OCR 的准确率。

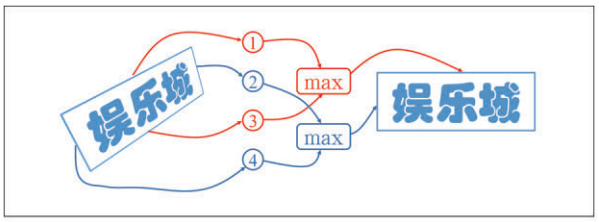

图5　文字形态矫正示意图

在进行透视变换前，需要先确定目标矩形区域的高和宽。目标矩形高度可以设定为四边形两条竖向边长度的最大值，如图 5 中蓝色线所示。目标矩形宽度可以设定为四边形两条横向边长度的最大值，如图 5 中红色线所示。在确定目标矩形的高和宽后，则可以确定目标矩形 4 个顶点坐标。通过使用 opencv 提供的空间透视转

换功能，可以根据原图中四边形 4 个顶点与目标图 4 个顶点坐标的映射关系计算透视变换矩阵，进一步根据透视变换矩阵得到矫正后的文字图像。

4　图片文字识别

图片文字识别指的是将图片化的文字转换为文字本身。这项任务从起初基于卷积神经网络逐步发展到基于结构学习的神经网络。卷积神经网络重点学习的是图片的纹理特征，而结构学习网络在学习图片的纹理特征基础上进一步学习纹理特征之间的关联性，从而达到对文字结构进行学习的目的。

目前基于结构学习神经网络的经典模型是 CRNN 模型。CRNN 模型是一个多层卷积网络与双向长短期记忆（LSTM）网络组合而成的模型。如图 6 所示，其基本思路是使用多层卷积网络获取图片中的视觉纹理特征，再通过双向 LSTM 网络分析纹理特征之间的横向关联性，对不同纹理特征进行恰当的组合从而识别出文字。基于该模型，衍生出了多种更加复杂的结构化模型。如将 CRNN 中的卷积网络增加 Inception 结构，将双向 LSTM 替换为 Attention 结构或者 Transformer 结构等。

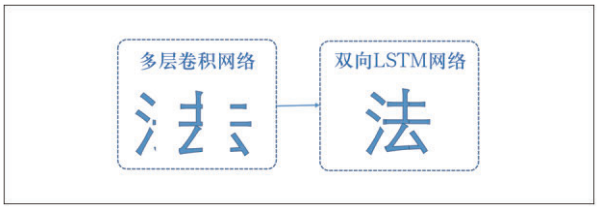

图6　CRNN文字识别过程示意图

4.1 CRNN 模型

CRNN 模型是最经典的结构学习模型，其基本框架一直沿用至今。一个标准的 CRNN 框架结构包含了特征提取层和特征关联层两部分，如图 7 所示。

首先，CRNN 模型对输入图片有如下假设。第一，图片中仅能包含一行文字，不能包含多行文字。第二，

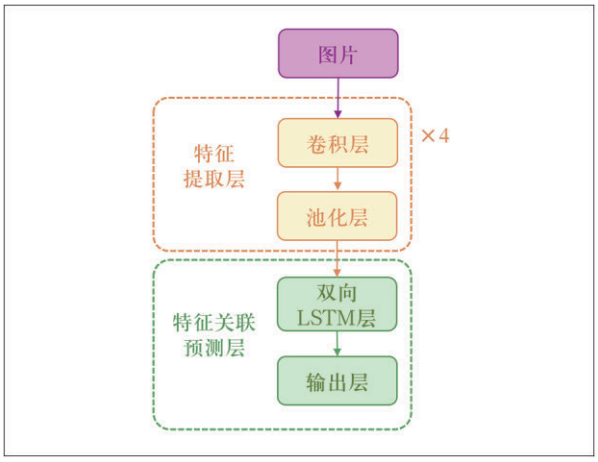

图7　CRNN模型结构

图片的高度必须为特定高度，如 16 个像素。若图片不满足此高度，可以将图片缩放至此高度，图片的宽度不受限制。

标准的 CRNN 模型中，特征提取层为一个标准的卷积模型，其由若干个卷积加池化层叠加而成。随着不断的卷积和池化操作，输入图片被概括归纳为特征图。特征图可以看作是广义上的图像，普通图像每个像素一般包含 3 个通道（红、黄、蓝）。特征图每个像素可以有任意多个通道，用来表达更加丰富的图像特征信息。如图 8 所示，输入图经过特征提取层后，图像的尺寸会成倍缩小（由 16×20 个像素变为 4×5 个像素），但图像像素的通道数量会加大（图中由 3 变成 8，现实中可达到 512）。特征图中的一列像素区域按比例对应输入图中的 4 列像素区域（图中红框标出）。

为了将特征提取层的特征图正确地输入到特征关联层中，需要将特征图转换为特征序列。特征图每个向量为 8 维，每一列包含 4 个向量。将每一列向量进行拼接形成 24 维向量作为特征序列中的一个元素，则特征序列中每一个元素对应特征图的一列像素区域，同样对应原输入图中 4 列像素区域。由于输入图片的高度固定，故该区域的大小固定。采用此方法可以方便后续特征关联预测层对图中的横向特征进行关联分析，从而将相关纵向特征合并成字。

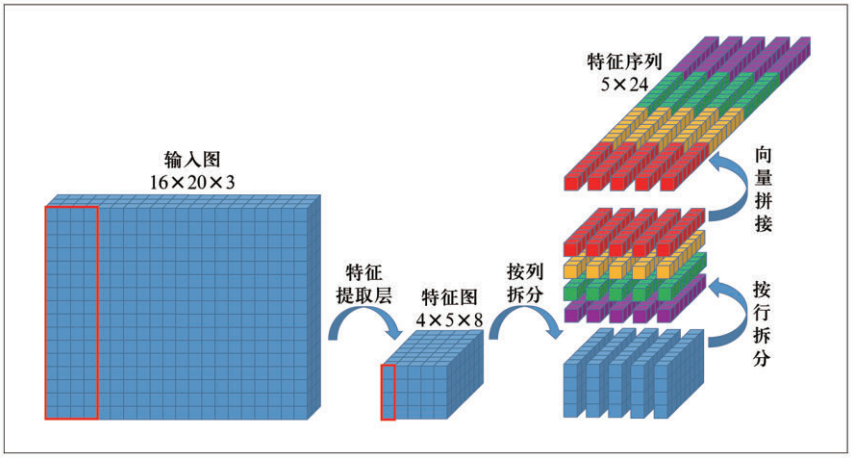

图8　特征图转换为特征序列示意图

特征关联预测层分析特征序列横向的关联关系，从而预测特征序列中包含的文字。其包含一个双向 LSTM 层和一个输出层。双向 LSTM 层可以从左到右、从右向左两个方向观察特征序列推断文字内容。双向观察的好处是能够大幅提高文字识别的鲁棒性。鲁棒性主要体现为两个方面：第一，当图片中一个文字发生小部分扭曲或遮挡，则无论遮挡位置在文字左或右，模型都可以给出更准确的预测，如“法”字 3 点水被遮挡了一个点，模型一样可以识别为“法”；第二，当图片中某个字大部分遮挡，当被遮挡字经常会与其周围字共同出现时，可以对被遮挡字进行更准确的预测，如“巧克力”中“克”字的“十”被遮挡，可通过“巧”和“力”以及未被遮挡的“兄”推断出“克”字。

4.2 CRNN 变体

为了进一步增强 CRNN 文字的识别能力，一些基于 CRNN 的变体应运而生。这些变体的改进思路主要有两方面。一方面是修改特征提取层的特征提取机制，从而提取出更多有用的文字视觉特征；另一方面是修改特征关联预测层的分析机制，加强特征横向关联分析能力和分析速度。

基于 CRNN+Inception 的变体识别模型，如图 9 所示。其基本思路是在特征提取层增加 Inception 机制从而更好的提取出各种尺度的文字视觉特征。当输入图片中的文字大小各不相同时，识别效果要优于标准的 CRNN。

基于 CRNN+Transformer 的变体识别模型，如图 10 所示。其思路将特征关联预测层的双向 LSTM 模型替换为 Transformer 模型。一方面 Transformer 模型的长距离关联分析能力要优于 LSTM；另一方面 LSTM 由于其网络结构特点，无法实现并行化运算，Transformer 模型在训练过程中可以被并行化运算，从而最大程度上利用 gpu 资源。

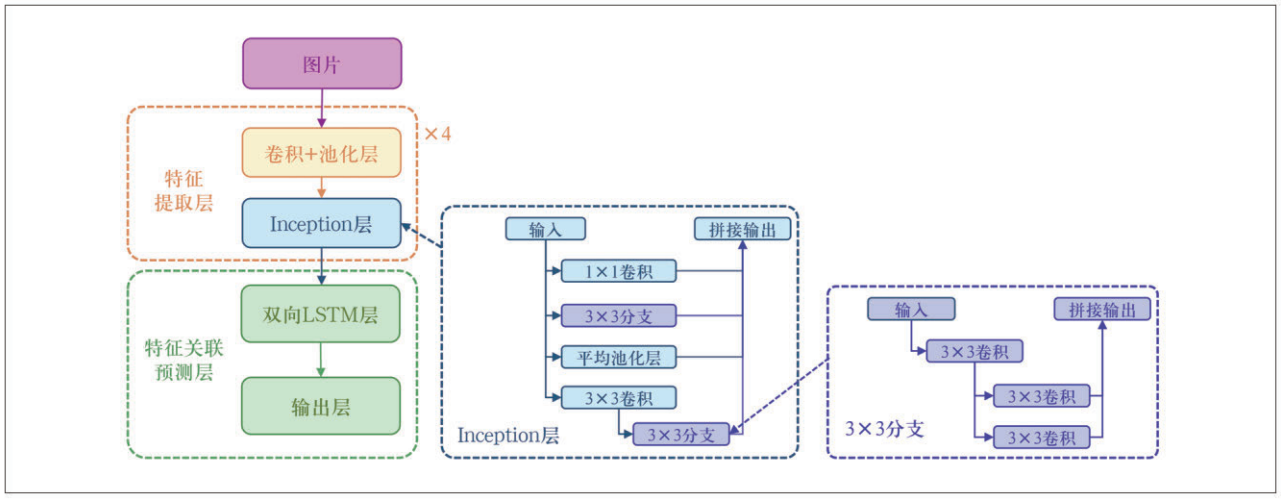

图9　CRNN+Inception网络结构

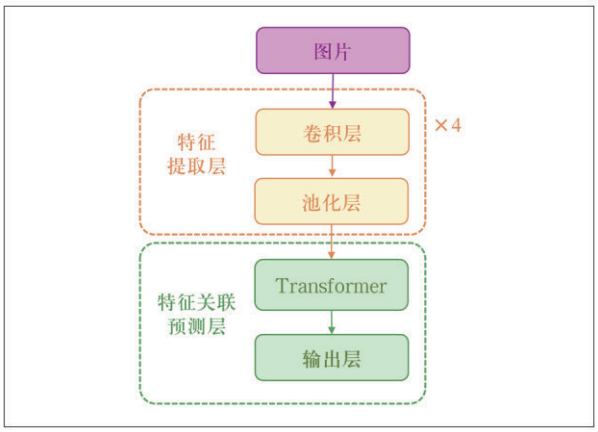

图10 CRNN+Transformer网络结构

4.3 数据增强及竖版文字识别

为了进一步提高 CRNN 对各种姿态文本的识别精度，可以通过程序生成各种姿态的文字图片进行训练。文字的变化因素包括但不限于字体、大小、旋转角度、颜色、背景和噪点等。

CRNN 模型只能对横向的文字图片进行有效识别。在对竖版文字进行训练时，需要将图片旋转为横向图片后再进行训练。竖版文字和横版文字模型最好单独训练，在识别过程中，可以先通过恰当手段判定文字方向后再通过调用横版或竖版文字识别的 CRNN 模型进行文字识别。

5 结束语

不法分子采用将不良文字信息图片化的方法逃避监管。这些图片需要通过提取其中的文字信息来实现自动化分析识别。运营商开发不良文字图片识别可以借助深度学习中的场景文字识别能力。具体地，可以从文字区域检测、文字形态矫正和文字 OCR 3 个方面入手。首先使用 EAST 等技术对不良文字图片中的文字区域进行识别，再通过文字形态矫正方法对文字区域进行透视变换，最后使用 CRNN 等模型对矫正后的文字图片进行识别。本文对 3 个部分所用到的深度神经网络模型的结构、原理和作用进行了深入的分析，对运营商进行不良文字图片治理具有很大的技术参考价值。

参考文献

[1] 蒋冲宇，鲁统伟，闵峰，等．基于神经网络的发票文字检测与识别方法[J]．武汉工程大学学报，2019(6)．
[2] 杨宏志，庞宇，王慧倩．基于改进Faster R-CNN的自然场景文字检测算法[J]．重庆邮电大学学报(自然科学版)，2019(6)．
[3] 姜维，张重生，殷绪成．基于深度学习的场景文字检测综述[J]．电子学报，2019(5)．
[4] 杨飞．自然场景图像中的文字检测综述[J]．电子设计工程，2016(12)．
[5] 金连文，钟卓耀，杨钊，等．深度学习在手写汉字识别中的应用综述[J]．自动化学报，2016(8)．

Research on bad text image recognition technology

DU Gang[1], DAI Jing[2], ZHANG Chen[1], DU Xue-tao[1]

(1 China Mobile Group Design Institute Co., Ltd., Beijing 100080, China; 2 China Mobile Group Co., Ltd., Beijing 100032, China)

Abstract The launch of the 5G messaging service provides convenient conditions for the dissemination of bad picture information. Bad text pictures, as a kind of special bad picture information, have brought challenges to the content security of communication networks. In order to effectively deal with bad text pictures, operators need to use artificial intelligence technology to identify and extract text information in pictures. This article describes in detail the three processing stages that need to be introduced in the entire technical process of bad text picture management, and conducts in-depth research on the structure and principle of deep learning models involved in the three processing stages. The research content of this article is of great reference value for operators to carry out bad text picture management.

Keywords object detection; text detection; OCR