算式识别开题报告

项目背景

算式识别任务是一个计算机视觉任务，属于文字 OCR 识别，长度不固定，属于序列识别。

下表是 18 年以来主流会议中 OCR 识别取得的进展，IC 是 ICDAR（国际文档分析和识别大会）的缩写[1]。主流算法在自然场景中的识别率已突破 90%，OCR 识别技术已经发展的十分成熟，我们的项目相对而言场景比较简单，所以应该是可以解决的。

Conf.	Date	Title	IC03	IC13
'18-AAAI	18/01/04	Char-Net: A Character-Aware Neural Network for Distorted Scene Text Recognition	0.915	0.908
'18-AAAI	18/01/04	SqueezedText: A Real-time Scene Text Recognition by Binary Convolutional Encoder-decoder Network	0.931	0.929
'18-CVPR	18/05/09	Edit Probability for Scene Text Recognition	0.946	0.944
'18-TPAMI	18/06/25	ASTER: An Attentional Scene Text Recognizer with Flexible Rectification	0.945	0.918
'18-ECCV	18/09/08	Synthetically Supervised Feature Learning for Scene Text Recognition	0.947	0.94
'19-AAAI	18/09/18	Scene Text Recognition from Two-Dimensional Perspective		0.914
'19-CVPR	18/12/14	ESIR: End-to-end Scene Text Recognition via Iterative Image Rectification		0.913
'19-PR	19/01/10	MORAN: A Multi-Object Rectified Attention Network for Scene Text Recognition	0.950	0.924

个人对计算机视觉有兴趣，并且身边工作中也有这样的需求，因此想借此机会进一步了解 OCR 的相关技术。

问题描述

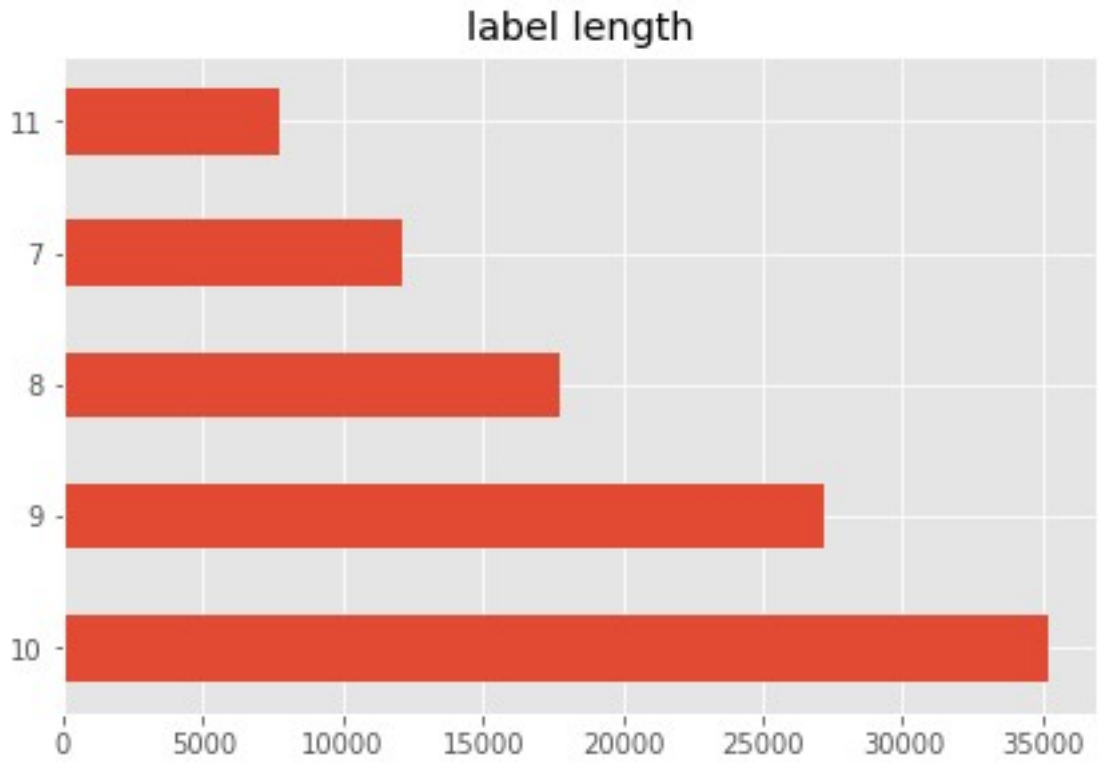
任务目标是：输入一张彩色图片，识别出图片中的算式文本。
这是一种经典的验证码识别场景，对于人来说很简单，程序而言需要准确顺序识别每个字符才算识别成功。

比如下图，考虑上下文后应识别为(0+0)+9=9。

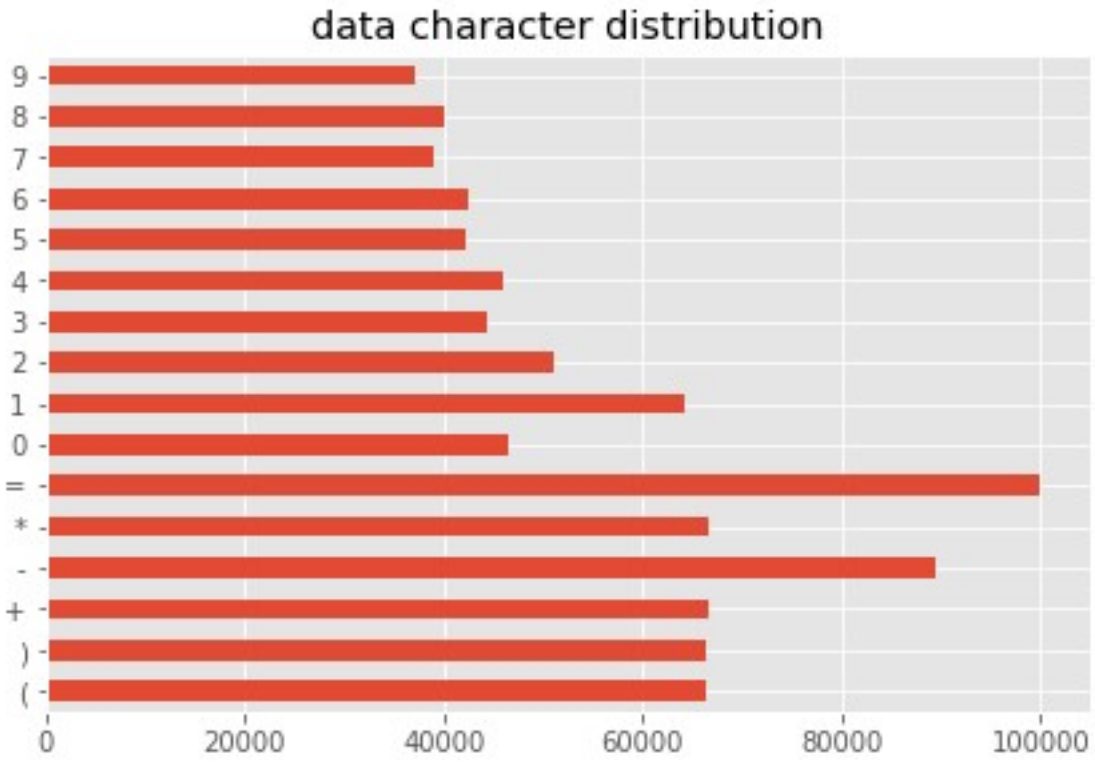
数据或输入

任务提供的数据集共 100,000 张图片，每张图片分辨率均为 300*64。算式由 3 个数字、2 个运算符、等号及运算结果组成，运算结果位数补丁，运算中还有括号，因此文本的长度不是固定的。需要注意的是，乘号只会以“*”的形式出现，所以十字符号只能是“+”。这个数据集将是本次任务中主要使用的数据集，首先会预留一部分用以作为测试集，查看程序的准确率；剩下的数据作为训练数据，并拟以 k 折的方式使用到训练中。

下面两图是对数据集的可视化统计。
Label length 直方图说明算式长度不定，分布从 7 到 11 都有，以 10 长度的数量最多，对模型的变长序列识别能力有较高要求。



data character distribution 直方图体现了整个数据集的字符分布情况，共 16 中字符，基于字典的算法是可以考虑的。算式中结果中“十几”运算结果会多一些，因此“1”的出现次数会多一些，到 60.000 多次，其他数字的分布差不多；运算符号中由于负数的存在，所以“-”的分布多一些，其他符号基本出现次数相同；“=”共 100,000，和数据集总数量一致，这是符合算式的特点的；“(”和“)”数量相等，这也是符合算式的特点的。从分布来看数据比较干净，呈均匀分布，但是也不排除其他的异常情况，这需要在训练调试过程中识别了。



对具体的图片进行观察后，发现图片有三个特点很可能影响模型的识别：1.字符不规律的倾斜，大小不一致；2.有噪点；3.由于文本长度不固定，右侧有不同程度的留白。在模型选择及预处理的时候应重点考虑这些特点。

评估标准

这里采用准确率来度量模型的好坏程度。
在算式识别这个任务中，我们关注的就是模型正确识别图片中文本的能力，并且数据集的观察结果没有明显的不均匀现象，因此采用准确率是可以有效度量模型的好坏的。

基准模型

本身模型会要求 99%的准确率。
另外会对比 tesseract 在相同测试集上的识别准确率。Tesseract 是一个 OCR 库,目前由 Google 赞助(Google 也是一家以 OCR 和机器学习技术闻名于世的公司)。Tesseract 是目前

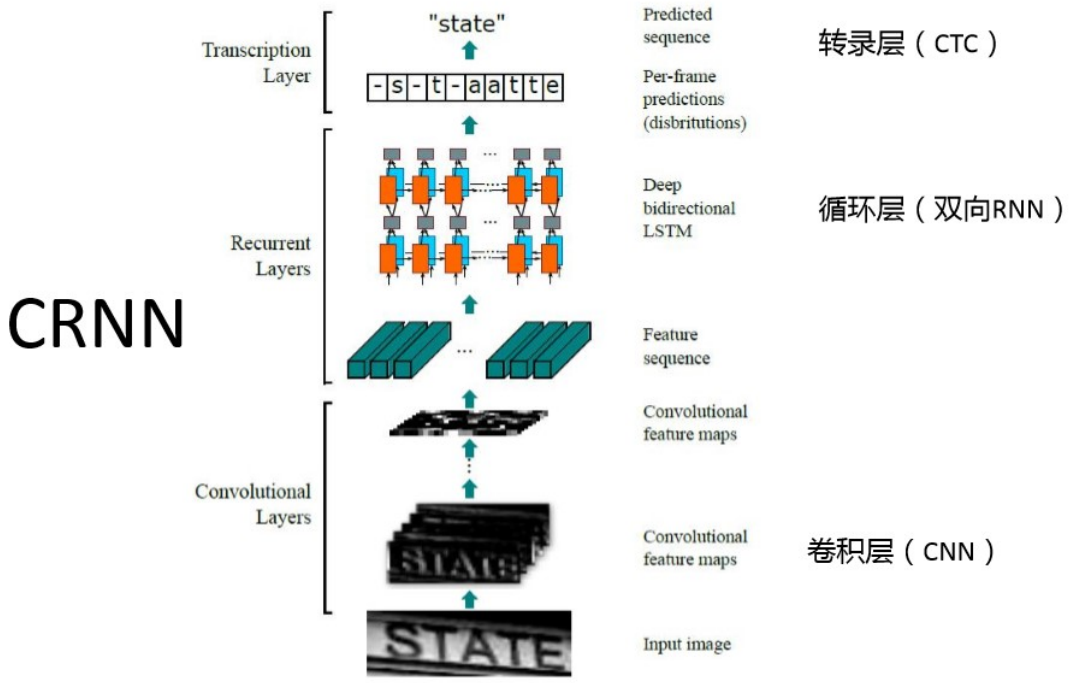
公认最优秀、最精确的开源 OCR 系统，除了极高的精确度，Tesseract 也具有很高的灵活性。它可以通过训练识别出任何字体，也可以识别出任何 Unicode 字符。

项目设计

由于变长、不规则倾斜、大小不一的情况，对齐不易，先切割然后识别的策略不会优先考虑。总体上会主要参考 CRNN 的结构[2]，即卷积层、循环层、转录层的三层结构：1.卷积层提取图像特征；2.循环层将特征处理为连续帧的序列结果；3.转录层将序列结果处理合并为最终识别结果。

采取这种结构主要有 2 点好处：1.现有的数据集可以直接训练学习，不需要详细的标注（例如字符），模型训练对文本长度也无约束；2.在场景识别中表现具竞争力[2]。

具体网络设计中，会根据识别结果好坏，考虑调整为 densenet+blstm 或其他结构[3]。



引用

[1] https://github.com/hwalsuklee/awesome-deep-text-detection-recognition

[2] https://arxiv.org/abs/1507.05717

[3] https://github.com/senlinuc/caffe_ocr