机器学习工程师纳米学位毕业项目（算式识别）
开题报告

杨学冉

17 November 2018

研究领域背景

深度学习是机器学习领域中的重要分支，近几年深度神经网络在图像识别中取得重大突破。神经网络的发展给计算机视觉领域带来了新的解决方案，在众多神经网络中，CNN是应用最广泛的一种。随着计算机算力的提升，利用大数据训练的CNN网络在计算机视觉领域发挥了越来越重要的作用。特别是自2010年以来，MNIST、CIFAR、ImageNet等规范化的数据集被越来越多的用于图像分类和各种竞赛，催生了像ResNet、VGG16等众多被广泛运用的算法模型。相对于传统的计算机视觉算法，采用深度神经网络特别是卷积神经网络的算法实现了更好的成绩。

硬件方面，GPU的运算能力越来越强。Nvidia公司主推用GPU芯片做通用计算，并提供了CUDA工具包进行深度学习软件的开发，为深度神经网络相关的算法模型提供算力基础。

问题声明

算式识别项目要求使用深度学习算法来识别图像中的算式。算式中的字符种类是有限的，本质上，算式识别是对每张图片中可能出现的字符进行分类。所以该问题可以视为多类别的分类问题。由于每个算式中的字符是储存在一张图片中，我们需要对每个字符进行分类识别，而不是对整张图片中信息的分类识别。

字符序列s = s1,s2, s3,...sn，这里s1,s2,s3,sn均为算式字符集中的一个字符，n是有限的，约为10。我们的目标是通过算法识别每个字符sn，并最终得到完整的算式s。

数据集和输入

数据集是已标注算式内容的10万张计算机生成的RGB彩色图片，每张图片包含一个公式。公式中可能出现的字符有“0～9”10个数字、“+、-、*”3个运算符和一对括号，共计15种字符。即对于单个字符的多类分类，分类的数目是15。每张图

片的大小都是300x64像素。图像中的字符存在噪声、旋转和缩放。
在训练前需将数据集按一定比例随机分成训练集、验证集和测试集。最终使用测试集对模型进行评估。

方案声明

这是一个对图像文件中的字符序列进行分类识别的问题。需要提取字符特征，拟采用CNN实现算式中字符特征的提取。由于图像中的算式字符是连续的序列，而我们需要提取其中的每个字符并拼接成算式。有两种潜在的方案：一种是依据字符位置对图像进行分割，分割后的每个片段包含一个字符，并对每个片段用CNN算法实现字符识别。另一种是借助RNN对连续字符序列进行分类识别。
Baoguang Shi团队提出了CRNN方法，这一方法是CNN和RNN的结合：使用CNN获取图像的特征图，并将特征序列作为RNN的输入，得到图像中字符的预测分类，并最终转换成字符序列。

基准模型

提出CRNN模型的Baoguang Shi团队在IIIT 5K-word数据集中对连续文本实现了97.6%的准确率（词典数为50），在SVT（Street View Text）数据集中实现了96.4的准确率（词典数为50），而在IC03和IC13数据集中也取得了类似的准确率。考虑到这几个数据集中的字符及场景的复杂程度要比本项目使用的数据集要大很多，本项目的最终准确率结果应比上述结果要好。本项目要求实现至少99%的准确率。

图1. SVT数据集中的图像示例

图2. 算式识别中的图像示例

评估矩阵

使用准确率作为分类准确程度的指标，即在测试集中，正确分类的算式数目占测试集中算式总数目的比例。这里正确分类的算式应当是算式内所有字符都需正确识别，否则会使算式内容不成立从而导致识别无意义。

项目设计

本项目将借鉴CRNN算法训练模型提取算式字符序列。项目实施主要包含图像预处理、算式字符串提取和算式生成。

数据集中所有图片大小一致，因此对图片尺寸拟不做处理，RGB文件中的色彩信息对字符识别的作用不大，因此需将整个数据集进行灰度处理。使用灰度图像训练、验证和测试模型。

字符提取是本项目的重点内容，将借助CNN实现算式字符串的特征提取，使用RNN处理特征图中的连续字符特征。并最终得到连续的算式字符串。

项目实施过程中还将借助可视化工具对模型中的图像处理步骤进行可视化处理，以明确图像在模型中处理的变化。