

# 基于文字识别视角分析人工智能机器学习中的文字识别方法

文/冯琬婷

人工智能技术的不断发展,使得人们对其研究也变得更加深入,机器学习已经成为了人工智能中新的研究核心问题,当前已经在网络搜索、语言识别和机器视觉等方面得到广泛应用,但是对于字符集比较庞大的汉字识别,一直是亟待解决的难点问题。本文笔者结合人工智能机器学习中文字识别产生的背景和主要识别方法进行分析,旨在为今后的具体应用奠定基础。

【关键词】文字识别 人工智能 机器学习

科技和互联网技术的发展为人们向智能化方向的迈进提供了基础,而机器学习则是人工智能领域中一个非常重要的构成部分,但是其应用还需要进一步的扩展。现如今人工智能中的机器人、网络搜索和语音识别等已经取得了一定的进展,但是文字识别尚未得到有效应用,需要进行更加深入全面的研究。

## 1 人工智能机器学习中的文字识别背景阐述

在人工智能技术尚未得到广泛应用之前,文字的自动化识别录入是一项十分艰巨却需要急切解决的问题。我们在生活中所采用的传统文字识别方法往往是以其基本外形为基础,在对文字字符之间的差别进行统计和分析,再找到一组最优的可以代表文字之间差异的统计学参数,实现对文字的筛选和识别,进而实现对文字的识别和自动录入。自从计算机技术产生以来,人们就已经对相关方法进行优化研究,但是最终的文字识别结果却不尽如人意。比如对一些十分简单的英文识别还存在不少问题,无法得到较好的识别率,对于其他字符集更加庞大的语言来说,此种方法的识别率更低,还无法满足人们对文字识别更高的要求。

通过对传统文字识别特点进行分析,发现其缺点表现在以下几个方面:

(1) 文字像素抖动等噪声大,文字呈现出微弱形变,这些对形态统计参数的影响都很大;

(2) 文字种类多样化,当前世界上现存的文字种类大概有 2500 多种,而且随着时间的不断推进,文字也会发生相应的演变,想要再进行文字识别,就需要对文字进行重新

(3) 识别花费的时间相对更长。人工智能技术在上世纪五十年代产生,经过六十多年的发展,已经成为涵盖语言学、计算机、信息控制、生物行为学等多学科的技术,最主要的目的就是要实现机器向人那样思考。历史上比较典型的人工智能应用案例有,IBM 公司在 1997 年研发的深蓝计算机打败了当时的国际象棋大师卡斯帕罗夫;2012Google 研制的无人驾驶系统正式获得车辆许可证。学习能力一般被认为是人类所特有的一种行为,根据人工智能大师 simon 的观点来看,机器学习就是要在不断的工作过程中增强自身处理任务的能力,比如在解决下一个问题时可以比之前做的更好。所以文字识别方法也应该具备良好的学习能力,只具备较好的识别率并不足够。

## 2 人工智能机器学习中的文字识别主要方法分析

### 2.1 对识别分类器的分析

对人工智能机器学习中文字识别方法的分析必须要对文字识别的分类器进行研究,而 BP 神经网络人类器则是众多文字识别分类器中的一种。BP 分类器的使用首先需要构建神经网络,然后再分别设计 net1,2,3,这三个数字分别代表着不同特征的数据,这些数据中包含输入和输出层,另外也包含两个隐层。每组数据中的节点也表现出较大的差异性,比如 net1 中的节点数量分别为 64/128 和 1;net2 中的节点数量分别为 24/48/1,Unicode 为输出范围,编码范围为 x4;net3 中的节点数量分别为 60/128/1,输出范围也是 Unicode,编码范围也为 x4。

初始化网络时,initwb 和 initnw 都是比较常见的初始化方式。Initwb 在进行初始化全部网络偏置量和权值时需要自定义矩阵,而 initnw 是一种默认的初始化方式,在具体运行过程中需要应用到 widrow 和 nguyen 两种算法。这样就能够做到在不同的层次中时权重和偏置值得到有效的初始化,而且在进行控件输入的过程中,所有的神经元活动都可以被有效的输入其中,从而减少神经元当中的浪费现象,也使训练速率变得更快。

### 2.2 对文字识别性能的分析

就汉字而言,我们在日常生活中经常接触到的汉字数量一般是 6000 多个,以医药领域为标准,常用的汉字数量为 2000 多个,在这种情况下,就需要对六百个字符的数字进行图像构建,字符图像数量共计 10 幅,将所有

构建的图像分为六组,每一组中选择一幅图像作为测试数据,其余的九幅图像作为训练数据,在应用到 BP 神经网络文字识别系统之后可以发现以下现象:

文字识别系统在交叉验证过程中就会对部分文字进行相应的判断,也会检测到错误的字符数,错误字符数与总字符数之间的比值也就是最终所得到的错误自检率。在具体应用这个系统的过程中,识别率预计达到 2000 字符左右,但是仍然需要对该系统进行相应的完善,方法如下:对于每组中的字符,单个分类器需要在这些字符特征的基础上进行分辨,在整个文字识别的过程中想要有效的识别所有的字符,那么具体使用到的分类器也应该包含这些字符。这就需要提前划分好小组划分 2000 多个汉字,对于分类器的训练,需要结合每一个小组的具体特征。最终在对文字进行识别时,不同字符图像所代表的小组特征必须要被整个系统所接纳,所以说分类器需要具备各个小组所拥有的共同特征。对于分类器中出现的不同结果,统计出不同最多的那些特点,将这些特征代表的数据进行筛选,而最终的识别结果就是票数最多的那个。如果将其具体应用在交叉验证方法当中,将会在验证了结果的有效性之后再输出最终的结果。

## 3 总结

人工智能的出现为我们的生活和生产方式带来了巨大的改变,加之科技和网络技术的发展,人们已经成功的步入了信息时代,在人们生活中的各个方面都离不开信息技术。人工智能机器在众多的智能化产物中由于自身具备的独特特点得到了人们的关注,这也是自动驾驶技术中不可或缺的关键技术,对其进行研究具有显著的现实意义。基于此种发展背景,笔者从文字识别角度出发,对人工智能机器学习中文字识别发展的背景和主要方法进行分析,旨在促进其发展过程中问题的解决。

## 参考文献

- [1] 黄鑫. 基于人工智能机器学习的文字识别方法研究 [J]. 通讯世界, 2016 (13): 234.
- [2] 李文轩. 基于深度学习的路牌文字识别 [D]. 华南理工大学, 2016.
- [3] 李雷. 基于人工智能机器学习的文字识别方法研究 [D]. 电子科技大学, 2013.

## 作者单位

辽宁师范大学 辽宁省大连市 116000