輕鬆識別文字，這款Python OCR庫支持超過80種語言

程序員教授 今天

以下文章來源於Python大數據分析 ， 作者朱衛軍

Python大數據分析

分享python編程、可視化設計、大數據分析、機器學習等技術以及數據分析案例，包...

OCR是什麼？

有一款軟件叫掃描全能王，想必一些小伙伴聽過，這是一個OCR集成軟件，可以將圖像內容掃描成文字。

所以說，OCR作用是對文本資料的圖像文件進行分析識別處理，獲取文字及版面信息。

OCR的全稱叫作“Optical Character Recognition”，即光學字符識別。

這算是生活裡最常見、最有用的AI應用技術之一。

細心觀察便可發現，身邊到處都是OCR的身影，文檔掃描、車牌識別、證件識別、銀行卡識別、票據識別等等。

OCR本質是圖像識別，其包含兩大關鍵技術：文本檢測和文字識別。

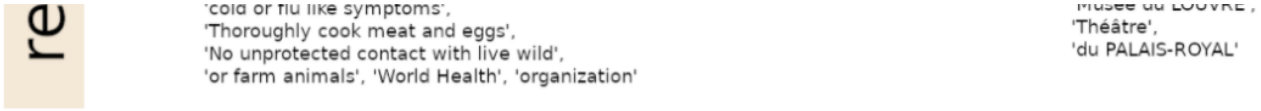
先將圖像中的特徵的提取並檢測目標區域，之後對目標區域的的字符進行分割和分類。


图像输入
前期处理
二值化
图像降噪
倾斜校正
中期处理
版面分析
字符切割
字符识别
版面还原
后期处理

關於EasyOCR

Python中有一個不錯的OCR庫-EasyOCR，在GitHub已有9700star。它可以在python中調用，用來識別圖像中的文字，並輸出為文本。

> https://github.com/JaidedAI/EasyOCR



EasyOCR支持超過80種語言的識別，包括英語、中文（簡繁）、阿拉伯文、日文等，並且該庫在不斷更新中，未來會支持更多的語言。

支持语言	代号
Abaza	abq
Adyghe	ady
Afrikaans	af
Angika	ang
Arabic	ar
Assamese	as
Avar	ava
Azerbaijani	az
Belarusian	be
Bulgarian	bg
Bihari	bh
Bhojpuri	bho
Bengali	bn
Bosnian	bs
Simplified Chinese（简体中文）	ch_sim
Traditional Chinese（繁体中文）	ch_tra
Chechen	che
Czech	cs
Welsh	cy
Danish	da
Dargwa	dar
German	de
English	en
Spanish	es
Estonian	et
Persian (Farsi)	fa
French	fr
Irish	ga
Goan Konkani	gom

Goan Konkani	gom
Hindi	hi
Croatian	hr
Hungarian	hu
Indonesian	id
Ingush	inh
Icelandic	is
Italian	it
Japanese	ja
Kabardian	kbd
Korean	ko
Kurdish	ku
Latin	la
Lak	lbe
Lezghian	lez
Lithuanian	lt
Latvian	lv
Magahi	mah
Maithili	mai
Maori	mi
Mongolian	mn
Marathi	mr
Malay	ms
Maltese	mt
Nepali	ne
Newari	new
Dutch	nl
Norwegian	no
Occitan	oc
Polish	pl

Portuguese	pt
Romanian	ro
Russian	ru
Serbian (cyrillic)	rs_cyrillic
Serbian (latin)	rs_latin
Nagpuri	sck
Slovak (need revisit)	sk
Slovenian	sl
Albanian	sq
Swedish	sv
Swahili	sw
Tamil	ta
Tabassaran	tab
Thai	th
Tagalog	tl
Turkish	tr
Uyghur	ug
Ukranian	uk
Urdu	ur
Uzbek	uz
Vietnamese (need revisit)	vi

安裝EasyOCR

安裝過程比較簡單，使用pip或者conda安裝。

```
pip install easyocr
```

如果用的PyPI源，安装起来可能会耽误些时间，建议大家用清华源安装，几十秒就能安装好。

使用方法

EasyOCR的用法非常简单，分为三步:

- 1.创建识别对象；
- 2.读取并识别图像；
- 3.导出文本。

我们先来举个简单的例子。

找一张路标图片，保存到电脑：

接着撸代码：

```
# 导入easyocr
import easyocr
# 创建reader对象
reader = easyocr.Reader(['ch_sim','en'])
# 读取图像
result = reader.readtext('test.jpg')
```

```
# 结果
result
```

输出结果：

```
[1]: [([[360, 78], [784, 78], [784, 226], [360, 226]], '石石路', 0.9229596853256226),
      ([[414, 218], [709, 218], [709, 292], [414, 292]],
       'shishi Rd.',
       0.22739121317863464),
      ([[84, 312], [642, 312], [642, 488], [84, 488]], '西城大道', 0.23274241387844086),
      ([[129, 457], [379, 457], [379, 522], [129, 522]],
       'xicheng Ave.',
       0.6641761660575867),
      ([[366, 530], [741, 530], [741, 662], [366, 662]],
       '竹园支路',
       0.9881148338317871),
      ([[426.1210468075817, 631.0778621224576],
        [608.9805783261377, 655.5597278827723],
        [600.8789531924183, 696.9221378775424],
        [418.0194216738624, 672.4402721172277]],
       'zhuyuan',
       0.6594952940940857),
      ([[604, 646], [677, 646], [677, 695], [604, 695]], 'Rd.', 0.9779377579689026)]
```

可以看到路标上的三个路名以及拼音都识别出来了！

识别的结果包含在元组里，元组由三部分组成：边框坐标、文本、识别概率。

「关于语言：」

这段代码有一段参数['ch_sim','en']，这是要识别的语言列表，因为路牌里有中文和英文，所以列表里添加了ch_sim（简体中文）、en（英文）。

可以一次传递多种语言，但并非所有语言都可以一起使用。英语与每种语言兼容，共享公共字符的语言通常相互兼容。

前文我们给出了EasyOCR支持的语言列表，并附有参数代号。

「关于图像文件：」

上面传入了相对路径'test.jpg'，还可以传递OpenCV图像对象（numpy数组）、图像字节文件、图像URL。

再读取一张文字较多的新闻稿图片：

多地新冠疫苗已经开打！

2021年春节即将到来，冬春季疫情防控成为最受关注的话题。

虽然多地出现本土确诊病例，但是，近期伴随北京、上海、深圳等多地新冠疫苗正式开打，防控又多了一道安全屏障。

早在元旦之前，国内已有部分地区的重点人群疫苗接种开启。

例如，陕西省重点人群接种工作已从2020年12月25日全面启动。首批参加新冠病毒疫苗接种的是海关检验检疫、口岸边检、口岸进口冷链物品装运及医疗卫生机构的一线工作人员等重点人群。

此外，深圳最近也已经为重点人群启动接种工作，主要面向9类高风险人员。其中，进口冷链物流相关人员、隔离场所工作人员、海关边检人员等8类重点人群都是通过工作单位统一预约接种。

```
# 导入easyocr
import easyocr
# 创建reader对象
reader = easyocr.Reader(['ch_sim','en'])
# 读取图像
result = reader.readtext('test1.jpg')
# 结果
result
```

```
[2]: [([[57, 5], [339, 5], [339, 41], [57, 41]], '多地新冠疫苗已经开打!', 0.7373780608177185),
      ([[57, 73], [319, 73], [319, 109], [57, 109]],
       '2021年春节即将到来,',
       0.2614888846874237),
      ([[329, 73], [763, 73], [763, 111], [329, 111]],
       '冬春季疫情防控成为最受关注的话题.',
       0.238058403134346),
      ([[59, 143], [385, 143], [385, 179], [59, 179]],
       '虽然多地出现本土确诊病例,',
       0.8407993316650391),
      ([[397, 143], [463, 143], [463, 179], [397, 179]], '但是,', 0.9547151923179626),
      ([[475, 143], [645, 143], [645, 179], [475, 179]],
       '近期伴随北京',
       0.7977046370506287),
      ([[663, 143], [725, 143], [725, 179], [663, 179]], '上海', 0.8590742349624634),
      ([[739, 143], [1063, 143], [1063, 179], [739, 179]],
       '深圳等多地新冠疫苗正式开',
       0.815974771976471),
      ([[57, 193], [365, 193], [365, 229], [57, 229]],
       '防控又多了一道安全屏障',
       0.35195714235305786),
      ([[4, 196], [44, 196], [44, 228], [4, 228]], '打,', 0.996860146522522),
      ([[241, 260], [758, 260], [758, 298], [241, 298]],
       '国内已有部分地区的重点人群疫苗接种开启',
```

识别文字的准确率还是很高的，接下来对文字部分进行抽取。

```
for i in result:
    word = i[1]
    print(word)
```

输出：

多地新冠疫苗已经开打!
2021年春节即将到来,
冬春季疫情防控成为最受关注的话题.
虽然多地出现本土确诊病例,
但是,
近期伴随北京
上海
深圳等多地新冠疫苗正式开
防控又多了一道安全屏障
打,
国内已有部分地区的重点人群疫苗接种开启
早在元旦之前,
陕西省重点人群接种工作已从2020年12月25日全面启动.
首批参加新冠病毒疫
例如
口岸进口冷链物品装运及医疗卫生机构的一线工作人
苗接种的是海关检验检疫.
口岸边检.
员等重点人群6
此外,
深圳最近也已经为重点人群启动接种工作,
主要面向9类高风险人员
其中,进
口冷链物流相关人员.隔离场所工作人员.海关边检人员等8类重点人群都是通过工作单位
统一预约接种.

小结

该开源库是作者研究了几篇论文，复现出来的成果，真是一位实干家。

检测部分使用了CRAFT算法，识别模型为CRNN，它由3个主要组件组成：特征提取，序列标记（LSTM）和解码（CTC）。整个深度学习过程基于Pytorch实现。

作者一直在完善EasyOCR，后续计划一方面扩展支持更多的语言，争取覆盖全球80%~90%的人口；另一方面支持手写识别，并提高处理速度。

---------- END ----------

推荐阅读：

太可怕了！差点因为一条SQL被拖出去祭天.....

90% 人会做错常见的10道Python面试题

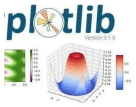
超硬核的 Python 数据可视化教程！

14个超有趣的数据分析项目，数据集都给你整理好啦

喜欢此内容的人还喜欢

Matplotlib數據可視化!

Datawhale



Python 炫技操作：推導式的五種寫法

Python編程時光

討論：對於神經網絡，不需要弄明白原理，只需要應用，是這樣嗎?

程序員大白