十大最受欢迎机器学习Python库

2018-03-26 Python网络爬虫与数据挖掘

AI的快速发展，让机器学习走向了巅峰，今天我们盘点一下最受欢迎的机器学习库（ML），希望你能够在这里寻找到你未来一段时间内的“利器”。

Pipenv

Pipenv是今年初开源的用于管理依赖项的官方推荐工具。Pipenv最初是由Kenneth Reitz创立的一个项目，旨在将其他包管理器（如NPM或yarn）的创意整合至Python中。安装virtualenv和virtualenvwrapper，并确保依赖项的依赖项版本的可重复性（在这里阅读更多关于这方面的信息）。使用Pipenv，你可以指定所有的依赖关系，通常使用命令添加，删除或更新依赖项。该工具可以生成一个文件，使得你的构建是确定性的，它可以帮助你避免那些难以捉住的BUG。

PyTorch

今年Facebook推出的DLT框架PyTorch，在深度学习社区中很受欢迎。PyTorch是构建在流行的Torch框架之上，尤其是它是基于Python的。考虑到过去几年人们一直在使用Python进行数据科学研究，这也是深度学习库大部分是使用Python的原因。

最值得注意的是，PyTorch已经成为了众多研究人员的首选框架之一，因为它实现了新颖的动态计算图范例（Dynamic Computational Graph paradigm）。当使用TensorFlow，CNTK或MXNet等框架编写代码时，必须首先定义一个称为计算图的东西。该图指定了我们的代码将运行的所有操作，这些操作稍后会被编译并被框架优化，以便能够在GPU上并行运行得更快。这个范例被称为静态计算图，因为你可以利用各种优化，而且这个图形一旦建成，就可以运行在不同的设备上。然而，在诸如自然语言处理之类的任务中，工作量通常是可变的。在将图像提供给算法之前，把图像调整为固定的分辨率，但不能对可变长度的句子进行相同的处理。这恰恰能体现PyTorch和动态图表的优势，通过让你在代码中使用标准的Python控制指令，图形将在执行时定义，给你更多自由空间，这对于几个任务来说是必不可少的。

当然，PyTorch也会自动计算梯度，并且速度非常快，而且是可扩展的。

Caffe2

这听起来可能不太现实，Facebook今年也发布了另一个的DL框架——caffe2。原来的Caffe框架已被广泛使用多年，并以非常不错的性能和经过测试的代码库而闻名。然而，最近DL的趋势使这个框架在某些方面显得有些out。于是Caffe2就成为了它的替代品。

Caffe2支持分布式训练、部署，支持最新的CPU和CUDA的硬件。虽然PyTorch可能更适合研究，但Caffe2更适合大规模部署。其实，你可以在PyTorch中构建和训练模型，同时使用Caffe2进行部署！这不是很好吗？

Pendulum

去年，Arrow是一个旨在使你更轻松，同时使用Python date time类进入了榜单，而今年是Pendulum。

Pendulum的优点之一是它是Python标准datetime类直接替代品，因此你可以轻松地将其与现有代码集成，并且只有在需

要时才能使用其功能。作者特别注意确保时区能够正确处理，默认情况下使每个实例时区感知自己的时区。你也将得到一个扩展timedelta，这样日期时间算术更容易。

与其他的库不同，它努力使API具有可预测的行为。如果你正在做一些涉及日期的小事，请查看更多的文档。

Dash

如果你正在做数据科学，你可能会使用Python生态系统中的Pandas和scikit-learn等优秀的工具。还可以使用JupyterNotebook管理你的工作流程。但是，当你和那些不知道如何使用这些工具的人一起做一项工作的时候，你该怎么办？你如何建立一个界面，使人们可以轻松地玩转数据，并在整个过程中对其进行可视化？过去，你或许需要一个专业的JavaScript前端团队来构建这些GUI。

Dash近几年发布的一个用于构建Web应用程序的开源库，尤其是在纯Python语言中利用数据可视化的Web应用程序。它建立在Flask，Plotly.js和React 之上，并提供了接口，所以你不必学习这些框架也能进行高效的开发。如果你想了解更多关于Dash的有趣应用，点击这个地方。

PyFlux

Python中有许多库用于研究数据科学和ML，但是当你的数据是随着时间的推移而变化的度量（例如股票价格，仪器的测量值等等）时，这对于大部分库来说是一个比较棘手的问题。

PyFlux是一个专门为时间序列而开发的 Python开源库。时间序列研究是统计学和计量经济学的一个子领域，目标可以描述时间序列如何表现（以潜在的因素或兴趣的特征来表示），也可以借此预测未来的行为。

PyFlux允许使用时间序列建模，并且已经实现了像GARCH这样的现代时间序列模型。

Fire

通常情况下，你需要为你的项目制作命令行界面（CLI）。除了传统的argparse，Python还有一些这样的工具，Clik和docopt。Fire是今年谷歌发布的软件库，在解决这个的问题上采用了不同的方法。

Fire是一个开源的库，可以为任何Python项目自动生成一个CLI，关键是自动，你几乎不需要编写任何代码或文档来构建你的CLI！你只需要调用一个Fire方法并把它所需要构建的传递给CLI。

如果你想对此有所深入了解，请阅读指南，因为这个库可以为你节省很多时间。

Imbalanced-learn

在理想的情况下，我们会有完美平衡的数据集，但不幸的是，现实世界并不是这样的，某些任务拥有非常不平衡的数据。例如，在预测信用卡交易中的欺诈行为时，你预计绝大多数交易（99.9%）是合法的。天真地训练ML算法会导致令人失望的性能，所以在处理这些类型的数据集时需要特别小心。

幸运的是，Imbalanced-learn是一个Python包，它提供了一些解决这类问题的方法，并提供一些技术的实现，它与scikit-learn兼容，是scikit-learn-contrib项目的一部分。

FlashText

如果你需要搜索某些文本并将其替换为其他内容（如大多数数据清理流程中），则通常会转为正则表达式。通常情况

下，正则表达式考研完美的解决问题。但是有时会发生这样的情况：你需要搜索的术语数量是成千上万，然后，正则表达式可能变得非常缓慢。这时FlashText是一个更好的选择，它使整个操作的运行时间大大提高了（从5天到15分钟）。FlashText的优点在于无论搜索条件有多少，运行时都是一样的，而正则表达式中运行时将随着条件数几乎呈线性增长。

FlashText证明了算法和数据结构设计的重要性，即使对于简单的问题，更好的算法也可以轻松超越最快的CPU。

Luminoth

现实生活中图像无处不在，理解其内容对于多个应用程序来说是至关重要的。值得庆幸的是，由于DL的发展，图像处理技术已经进步很多。

Luminoth是一个使用TensorFlow和Sonnet构建的用于计算机视觉的开源Python工具包。目前，它可以支持被称为Faster R-CNN的模型的形式进行对象检测。

并且Luminoth不仅是一个特定模型的实现，而是建立在模块化和可扩展的基础上的，所以定制现有的部分或用新的模型来扩展它来处理不同的问题，就可以能多地重用代码。它提供了用于轻松完成构建DL模型所需的工程工作如：将你的数据转换为用于提供数据管道（TensorFlow的记录）的格式，执行数据增强，在多个GPU训练，运行评估指标，在TensorBoard中可视化，并用简单的API或浏览器界面部署训练有素的模型，以便人们使用。

其他优秀的Python库：

PyVips

你可能从来没有听说过libvips库，首先它是一个图像处理库，如Pillow或ImageMagick，并支持多种格式。但是，与其他库相比，libvips速度更快，占用的内存也更少。PyVips是最近发布的用于libvips的Python绑定包，它与Python 2.7-3.6（甚至PyPy）兼容，易于使用pip。如果在你的应用程序中需要进行某种形式的图像处理，可以考虑一下它。

Requestium

有时，你需要自动化网络中的某些操作，如抓取网站，进行应用程序测试，填写网页表单，要想在不暴露API的网站中执行操作，自动化是必需的。Python有很好的请求库，可以让你执行一些这样的操作，但不幸的是请求获取的HTML代码可能没有表单，你可能会尝试查找表单来填充自动化任务。解决这个问题的方法是对JavaScript代码所做的请求进行反向工程，这将意味着需要花费很多时间来调试。另一个选择是转向使用Selenium这样的库，它允许你以编程方式与Web浏览器交互并运行Javascript代码。有了这个，问题就可以解决了。

Requestium库可以让你从请求开始并无缝地切换到使用Selenium，它可以作为一个请求的直接替换。它还集成了Parsel，因此编写所有用于在页面中查找元素的选择器要比其它方式更加快捷。

skorch

假如你很喜欢使用scikit-learn的API，但是遇到了需要使用PyTorch来完成工作。不要担心，skorch是一个封装，可以通过类似sklearn的接口提供PyTorch编程。如果你熟悉这些库，那么语法将很简单易懂。通过skorch，你会得到一些抽象的代码，所以你可以把更多的精力放在真正重要的事情上，比如做数据科学。

原文： https://tryolabs.com/blog/authors/alan-descoins/

近期文章：

1. 一个完整的Django入门指南 - 第1部分
2. python定期爬取GitHub上每日流行项目