AI 从业者都应该知道的实验数据集

本文作者：黄善清 2018-11-03 17:27

数据集对于深度学习模型的重要性不言而喻，然而根据性质、类型、领域的不同，数据集往往散落在不同的资源平台里，急需人们做出整理。 fast.ai 近期将这些重要的数据集汇总到了一篇文章里， AI 科技评论把文章编译如下。

少了数据，我们的机器学习和深度学习模型什么也干不了。这么说吧，那些创建了数据集、让我们可以训练模型的人，都是我们的英雄，虽然这些人常常并没有得到足够的感谢。让人庆幸的是，那批最有价值的数据集后来成了「学术基准线」——被研究人员广泛引用，尤其在算法变化的对比上；不少名字则成为圈内外都耳熟能详的名称，如 MNIST、CIFAR 10 以及 Imagenet 等。

身为 fast.ai 的一员，我们自觉欠这些数据集的创建者一句真挚的感谢，所以我们决定，通过与 AWS 合作，把一些最重要的数据集集中整理在一处，数据集自身采用标准格式，存储服务器也是快速的、可靠的（请参阅下方的完整列表与链接）。如果您在研究中使用了这些数据集，我们希望您记得引用原始论文（我们已经在表单中提供引用链接）；如果您将它们用作商业或教育项目的一部分，请考虑添加致谢文及数据集原链接。

我们之所以经常在教学中引用这些数据集，是因为它们就是学生们很有可能遇到的数据类型的绝佳例子，此外，学生可以将自己的工作与引用这些数据集的学术成果进行对比，从而取得进步。此外，我们也会使用 Kaggle Competitions 数据集，Kaggle 的 public leaderboards 允许学生在世界最好的数据集里测试自己的模型，不过 Kaggle 数据集并不会在本次表单中出现。

图像分类领域

1）MNIST

经典的小型（28x28 像素）灰度手写数字数据集，开发于 20 世纪 90 年代，主要用于测试当时最复杂的模型；到了今日，MNIST 数据集更多被视作深度学习的基础教材。fast.ai 版本的数据集舍弃了原始的特殊二进制格式，转而采用标准的 PNG 格式，以便在目前大多数代码库中作为正常的工作流使用；如果您只想使用与原始同样的单输入通道，只需在通道轴中选取单个切片即可。

引文：http://yann.lecun.com/exdb/publis/index.html#lecun-98

下载地址：https://s3.amazonaws.com/fast-ai-imageclas/mnist_png.tgz

2）CIFAR10

10 个类别，多达 60000 张的 32x32 像素彩色图像（50000 张训练图像和 10000 张测试图像），平均每种类别拥有 6000 张图像。广泛用于测试新算法的性能。fast.ai 版本的数据集舍弃了原始的特殊二进制格式，转而采用

标准的 PNG 格式，以便在目前大多数代码库中作为正常的工作流使用。

引文：https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

下载地址：https://s3.amazonaws.com/fast-ai-imageclas/cifar10.tgz

3) CIFAR100

与 CIFAR-10 类似，区别在于 CIFAR-100 拥有 100 种类别，每个类别包含 600 张图像（500 张训练图像和 100 张测试图像），然后这 100 个类别又被划分为 20 个超类。因此，数据集里的每张图像自带一个「精细」标签（所属的类）和一个「粗略」标签（所属的超类）。

引文：https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

下载地址：https://s3.amazonaws.com/fast-ai-imageclas/cifar100.tgz

4) Caltech-UCSD Birds-200-2011

包含 200 种鸟类（主要为北美洲鸟类）照片的图像数据集，可用于图像识别工作。分类数量：200；图片数量：11,788；平均每张图片含有的标注数量：15 个局部位置，312 个二进制属性，1 个边框框。

引文：http://vis-www.cs.umass.edu/bcnn/

下载地址：https://s3.amazonaws.com/fast-ai-imageclas/CUB_200_2011.tgz

5) Caltech 101

包含 101 种物品类别的图像数据集，平均每个类别拥有 40—800 张图像，其中很大一部分类别的图像数量固为 50 张左右。每张图像的大小约为 300 x 200 像素。本数据集也可以用于目标检测定位。

引文：http://www.vision.caltech.edu/feifeili/Fei-Fei_GMBV04.pdf

下载地址：https://s3.amazonaws.com/fast-ai-imageclas/caltech_101.tar.gz

6) Oxford-IIIT Pet

包含 37 种宠物类别的图像数据集，每个类别约有 200 张图像。这些图像在比例、姿势以及光照方面有着丰富的变化。本数据集也可以用于目标检测定位。

引文：http://www.robots.ox.ac.uk/~vgg/publications/2012/parkhi12a/parkhi12a.pdf

下载地址：https://s3.amazonaws.com/fast-ai-imageclas/oxford-iiit-pet.tgz

7) Oxford 102 Flowers

包含 102 种花类的图像数据集（主要是一些英国常见的花类），每个类别包含 40—258 张图像。这些图像在比例、姿势以及光照方面有着丰富的变化。

引文：http://www.robots.ox.ac.uk/~vgg/publications/papers/nilsback08.pdf

下载地址：https://s3.amazonaws.com/fast-ai-imageclas/oxford-102-flowers.tgz

8) Food-101

包含 101 种食品类别的图像数据集，共有 101,000 张图像，平均每个类别拥有 250 张测试图像和 750 张训练图像。训练图像未经过数据清洗。所有图像都已经重新进行了尺寸缩放，最大边长达到了 512 像素。

引文：https://pdfs.semanticscholar.org/8e3f/12804882b60ad5f59aad92755c5edb34860e.pdf

下载地址：https://s3.amazonaws.com/fast-ai-imageclas/food-101.tgz

9) Stanford cars

包含 196 种汽车类别的图像数据集，共有 16,185 张图像，分别为 8,144 张训练图像和 8,041 张测试图像，每个类别的图像类型比例基本上都是五五开。本数据集的类别主要基于汽车的牌子、车型以及年份进行划分。

引文：https://ai.stanford.edu/~jkrause/papers/3drr13.pdf

下载地址：https://s3.amazonaws.com/fast-ai-imageclas/stanford-cars.tgz

自然语言处理领域

1) IMDb Large Movie Review Dataset

用于情感二元分类的数据集，其中包含 25,000 条用于训练的电影评论和 25,000 条用于测试的电影评论，这些电影评论的特点是两极分化特别明显。另外数据集里也包含未标记的数据可供使用。

引文：http://ai.stanford.edu/~amaas/papers/wvSent_acl2011.pdf

下载地址：https://s3.amazonaws.com/fast-ai-nlp/imdb.tgz

2) Wikitext-103

超过 1 亿个语句的数据合集，全部从维基百科的 Good 与 Featured 文章中提炼出来。广泛用于语言建模，当中包括 fastai 库和 ULMFiT 算法中经常用到的预训练模型。

引文：https://arxiv.org/abs/1609.07843

下载地址：https://s3.amazonaws.com/fast-ai-nlp/wikitext-103.tgz

3) Wikitext-2

Wikitext-103 的子集，主要用于测试小型数据集的语言模型训练效果。

引文：https://arxiv.org/abs/1609.07843

下载地址：https://s3.amazonaws.com/fast-ai-nlp/wikitext-2.tgz

4) WMT 2015 French/English parallel texts

用于训练翻译模型的法语/英语平行文本，拥有超过 2000 万句法语与英语句子。本数据集由 Chris Callison-Burch 创建，他抓取了上百万个网页，然后通过一组简单的启发式算法将法语网址转换为英文网址，并默认这些文档之间互为译文。

引文：https://www.cis.upenn.edu/~ccb/publications/findings-of-the-wmt09-shared-tasks.pdf

下载地址：https://s3.amazonaws.com/fast-ai-nlp/giga-fren.tgz

5) AG News

496,835 条来自 AG 新闻语料库 4 大类别超过 2000 个新闻源的新闻文章，数据集仅仅援用了标题和描述字段。每个类别分别拥有 30,000 个训练样本及 1900 个测试样本。

引文：https://arxiv.org/abs/1509.01626

下载地址：https://s3.amazonaws.com/fast-ai-nlp/ag_news_csv.tgz

6) Amazon reviews - Full

34,686,770 条来自 6,643,669 名亚马逊用户针对 2,441,053 款产品的评论，数据集主要来源于斯坦福网络分析项目（SNAP）。数据集的每个类别分别包含 600,000 个训练样本和 130,000 个测试样本。

引文：https://arxiv.org/abs/1509.01626

下载地址：https://s3.amazonaws.com/fast-ai-nlp/amazon_review_full_csv.tgz

7) Amazon reviews - Polarity

34,686,770 条来自 6,643,669 名亚马逊用户针对 2,441,053 款产品的评论，数据集主要来源于斯坦福网络分析项目（SNAP）。该子集的每个情绪极性数据集分别包含 1,800,000 个训练样本和 200,000 个测试样本。

引文：https://arxiv.org/abs/1509.01626

下载地址：https://s3.amazonaws.com/fast-ai-nlp/amazon_review_polarity_csv.tgz

8) DBPedia ontology

来自 DBpedia 2014 的 14 个不重叠的分类的 40,000 个训练样本和 5,000 个测试样本。

引文：https://arxiv.org/abs/1509.01626

下载地址：https://s3.amazonaws.com/fast-ai-nlp/dbpedia_csv.tgz

9) Sogou news

2,909,551 篇来自 SogouCA 和 SogouCS 新闻语料库 5 个类别的新闻文章。每个类别分别包含 90,000 个训练样本和 12,000 个测试样本。这些汉字都已经转换成拼音。

引文：https://arxiv.org/abs/1509.01626

下载地址：https://s3.amazonaws.com/fast-ai-nlp/sogou_news_csv.tgz

10) Yahoo! Answers

来自雅虎 Yahoo! Answers Comprehensive Questions and Answers1.0 数据集的 10 个主要分类数据。每个类别分别包含 140,000 个训练样本和 5,000 个测试样本。

引文：https://arxiv.org/abs/1509.01626

下载地址：https://s3.amazonaws.com/fast-ai-nlp/yahoo_answers_csv.tgz

11) Yelp reviews - Full

来自 2015 年 Yelp Dataset Challenge 数据集的 1,569,264 个样本。每个评级分别包含 130,000 个训练样本和 10,000 个 测试样本。

引文：https://arxiv.org/abs/1509.01626

下载地址：https://s3.amazonaws.com/fast-ai-nlp/yelp_review_full_csv.tgz

12) Yelp reviews - Polarity

来自 2015 年 Yelp Dataset Challenge 数据集的 1,569,264 个样本。该子集中的不同极性分别包含 280,000 个训练样本和 19,000 个测试样本。

引文：https://arxiv.org/abs/1509.01626

下载地址：https://s3.amazonaws.com/fast-ai-nlp/yelp_review_polarity_csv.tgz

目标检测定位

1) Camvid: Motion-based Segmentation and Recognition Dataset

700 张包含像素级别语义分割的图像分割数据集，每张图像都经过第二个人的检查和确认来确保数据的准确性。

引文：https://pdfs.semanticscholar.org/08f6/24f7ee5c3b05b1b604357fb1532241e208db.pdf

下载地址：https://s3.amazonaws.com/fast-ai-imagelocal/camvid.tgz

2) PASCAL Visual Object Classes (VOC)

用于类识别的标准图像数据集——这里同时提供了 2007 与 2012 版本。2012 年的版本拥有 20 个类别。训练数据的 11,530 张图像中包含了 27,450 个 ROI 注释对象和 6,929 个目标分割数据。

引文：http://host.robots.ox.ac.uk/pascal/VOC/pubs/everingham10.pdf

下载地址：https://s3.amazonaws.com/fast-ai-imagelocal/pascal-voc.tgz

COCO 数据集

目前最常用于图像检测定位的数据集应该要属 COCO 数据集（全称为 Common Objects in Context）。本文提供 2017 版 COCO 数据集的所有文件，另外附带由 fast.ai 创建的子集数据集。我们可以从 COCO 数据集下载页面（http://cocodataset.org/#download）获取每个 COCO 数据集的详情。fast.ai 创建的子集数据集包含五个选定类别的所有图像，这五个选定类别分别为：椅子、沙发、电视遥控、书籍和花瓶。

fast.ai 创建的子集数据集：https://s3.amazonaws.com/fast-ai-coco/coco_sample.tgz

训练图像数据集：https://s3.amazonaws.com/fast-ai-coco/train2017.zip

验证图像数据集：https://s3.amazonaws.com/fast-ai-coco/val2017.zip

测试图像数据集：https://s3.amazonaws.com/fast-ai-coco/test2017.zip

未经标注的图像数据集：https://s3.amazonaws.com/fast-ai-coco/unlabeled2017.zip

测试图像数据集详情：https://s3.amazonaws.com/fast-ai-coco/image_info_test2017.zip

未经标注的图像数据集详情：https://s3.amazonaws.com/fast-ai-coco/image_info_unlabeled2017.zip

训练/验证注释集：https://s3.amazonaws.com/fast-ai-coco/annotations_trainval2017.zip

主体训练/验证注释集：https://s3.amazonaws.com/fast-ai-coco/stuff_annotations_trainval2017.zip

全景训练/验证注释集：https://s3.amazonaws.com/fast-ai-coco/panoptic_annotations_trainval2017.zip

via fast.ai，AI 科技评论编译