IT農家のディープ・ラーニング 奮闘記

第3回　枝豆の撮影画像から学習＆評価用データセットを作る

小池 誠

枝豆の莢に含まれる粒の数を見分ける人工知能を作っています．前回（2019年4月号）は学習に用いる画像を用意するために2粒莢と3粒莢の枝豆を撮影しました．

画像を集め終わったら，次はデータセットを作ります．データセットとは，「集めた画像」と「答えとなるラベル」をセットにした後，学習用とテスト用に分けてファイルにまとめたものです．

例えばディープ・ラーニング入門でよく使われる手書き数字画像データセットMNISTは，28×28画素の手書き数字画像とその画像が0～9のどの数字かを表すラベルとがセットになっており，学習用に60000件，評価用に10000件のデータがまとめられています．

● 作る理由

データセットを作る1番の理由は，構築したニューラル・ネットワークの学習やハイパー・パラメータ・チューニングのベンチマークとして使用するためです．一般的に，ニューラル・ネットワークの評価は，あるデータセットに対して正答率が何%であるかというように行います．この評価に用いるデータが毎回異なっていたり，特定のラベルに偏っていたりしていては，純粋に結果を比較できなくなってしまいます．そこで，開発の最初に「学習に使えるデータはこれ」，「評価に使うデータはこれ」というようにデータを分けておきます．または，学習に用いるデータ，ハイパー・パラメータ・チューニングの検証に用いるデータ，評価に用いるデータと，3つに分ける場合もあります．

重要なことは，最終的に評価に用いるデータは，学習時にもチューニング時にも使っていない完全な未知のデータであること，さらに，理想を言えばラベルごとに同数のデータが含まれていることです．

そろえておきたいデータのフォーマット

● いろいろなデータ・フォーマットがある

データセットを作る際のファイル・フォーマットは複数あります．**表1**にインターネットで公開されている有名なデータセットのフォーマットをリストアップしています．

ディープ・ラーニング・ライブラリの多くはPythonで記述されていることが多いため，データセットもPythonからアクセスしやすいpickleやnpy/npzが多く使われています．

npyは，Numpy配列をシリアライズするためのフォーマットで，配列データに加え配列のshapeやdtype情報なども格納されており，アーキテクチャが異なる別のマシン上でも正しくNumpy配列を再構築できます．

npzは，複数のnpyファイルをzipファイルにまとめたものになり，どちらもNumpyパッケージを使って読み書きが可能です．ただし，Pickleやnpyなどは基本的には，データセットを読み出す際にデータセッ

表1　データセットのファイル・フォーマットにはPythonから使いやすいpickleやnpy/npzがよく使われる

データセット	ファイル・フォーマット	備　考
MNIST	バイナリ・データ，npz	バイナリ・データの詳細は，http://yann.lecun.com/exdb/mnist/を参照．kerasなどのフレーム・ワークからはnpzで取得可能
Fashion-MNIST	バイナリ・データ，npz	MNISTと同様
CIFAR-10/100	Pickle	pythonのpickleパッケージでアクセスする
The Quick,Draw!	ndjson，npy	改行区切りのJSON
Iris Dataset	CSV	数値データのみ
ImageNet	テキスト	画像URLとラベルが記されたテキスト

第1回　ブラウザで開発＆体験OKのAI開発環境Google Colaboratory（2019年3月号）
第2回　正答率UPのキモ…枝豆画像データを集める（2019年4月号）