1．データセットの作成と学習方法のノウハウ，学習後の評価方法

寺本 篤司 藤田医科大学医療科学部放射線学科

ディープラーニングは大量のデータを利用して学習することで，高度な処理を行うことができる。そのため，処理に利用するデータ（データセット）の構築は最も重要な作業である。本誌の読者は，主に医用画像を対象としたディープラーニングの研究に興味があると思われるが，医用画像は一般的な自然画像とは画像のフォーマットが異なり，データを多く集めることが困難なことも多い。本稿では，医用画像を対象としたディープラーニング処理を行う際に必要なデータセット作成技術や，学習，評価方法に関するノウハウについて解説する。

データセットの作成

1．良いデータセットとは

ディープラーニングの真価を発揮できるデータについて述べる。ディープラーニングの学習時には，入力と出力を結ぶネットワークの内部パラメータが，大量のデータを用いて計算される。処理結果の良しあしはデータが決め手になることが多く，以下の点について留意しながらデータを準備するとよい。

① 十分な数量のデータ

対象や学習させるネットワーク規模にもよるが，100パターン程度はないと，学習が不安定になることや未知のデータに対する処理性能が低くなることが多い。少数しか集められない場合には，後述のデータ拡張を利用して水増しする方法を用いれば，性能低下を軽減することができる。

② バランスの良いデータ配分

例えば，データを2つのクラスに分類する場合には，2つのクラスに属するデータ数に偏りが生じない方がよい。極端な例として，肺結節の良悪性鑑別を行う場合に，良性結節のデータ数が全体の90%，悪性結節のデータ数が10%の学習データで分類処理を行わせた場合，すべてのデータに対して良性と出力すれば正解率90%となり良好な結果のように見えるが，それは実際には役に立たないことは明らかである。そのため，できるかぎりクラス間に偏りがないようにデータを収集する必要がある。ただし，アンバランスになる問題は疫学的な問題などから解決できないこともあり，後述のデータ拡張を利用して水増しすることで，バランスを取ることもできる。

③ 質の高い正解データ

画像分類処理を行う場合には，画像ごとにその画像がどのクラスに属しているのか最初に分類しておく。また，画像から臓器などの領域を抽出する場合などに用いられる領域抽出処理を行う場合には，入力画像に対応する領域画像（ラベル画像）を準備する。このように，入力するデータとペアになるデータをセットで用意する必要があり，それをデータセットと呼ぶ（**図1**）。これらは，正解データとして学習に用いられる。その質は処理性能に大きな影響を及ぼすため，一定の基準で正解データを作成する必要がある。注意すべき点として，正解データを数人で分担して作成した場合，正解データに判断基準の個人差によるバラツキが含まれることである。多量のデータに対して正解データを作成する作業は非常に負担が大きいが，複数人で作業した場合にも，最初にしっかりと基準を決めておくこと，最後に1名が確認・修正するなど，データが均質になるような工夫が必要である。

2．画像形式

一般的な自然画像を対象としたディープラーニングの研究では，階調数が256

〈0913-8919/19/￥300/ 論文 /JCOPY〉