Progress

Mizuno Yasuaki

September 20, 2022

ファミリー数削減

前回はファミリー数を百前後で学習を行ったが、ファミリー数を十に減らして、様々なニューラルネットワークのモデルを試す。ファミリー数を減らしたときのデータの構成を表 1 に示す。

Table 1: 削減データ（訓練データ）の構成 [1]

family	count	family	count
0	846	5	136
1	864	6	116
2	163	7	273
3	136	8	1152
4	1538	9	99

[1] sum=5370, kind=83

削減データを用いた学習結果①

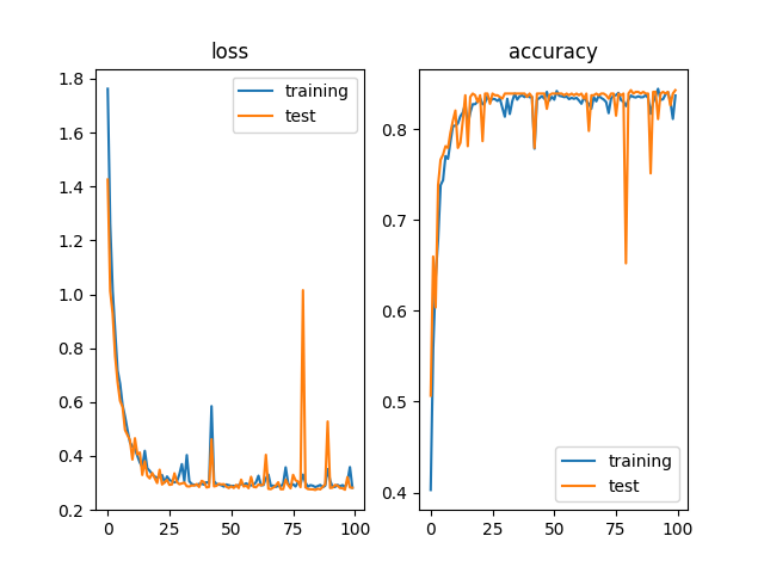
前回使用したニューラルネットワークのモデル (Simple_FNN と Simple_CNN) を利用して学習を行い、学習結果を表 2 に示す。

Table 2: 学習結果

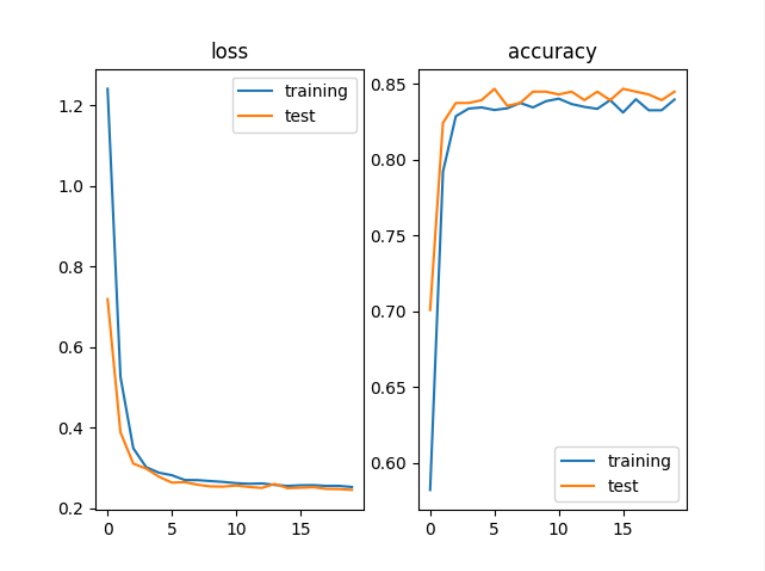
モデル	精度
Simple_FNN	0.8429906368255615
Simple_Cov	0.8448598384857178

削減データを用いた学習結果②

また、それぞれのモデルの学習過程を図と図にそれぞれ示す。



(a) Simple FNN



(b) Simple CNN

他のデータベースの評価①

使用していたデータベース (GPCR) はデータ数が 71442 であり、ファミリー数は 86 である。よって、単位ファミリーにおける平均データ数は

$$\frac{71442}{86} = 830.72$$

ある。それに対して、より大きなデータベース (COG-100-2892) においてはデータ数が 3131952 であり、ファミリー数は 2892 である。よって、同様に単位ファミリーにおける平均データ数は

$$\frac{3131952}{2892} = 1082.97$$

となる。単位ファミリーあたりにおけるデータ数はの差は二百程度であるが、データに偏りが大きければあまり意味がない。

他のデータベースの評価②

他のデータベース (COG-100-2892) について調べた結果を表 3 に示す。簡単のため、10 より小さいファミリーを示した。

Table 3: データベースの構成 [2]

family	count	family	count
0	1414	5	3070
1	1256	6	842
2	880	7	2483
3	1648	8	2018
4	1244	9	1772

一瞥したが、極端にデータ数が少ないファミリーは少なかった。

[2] sum=3131952, kind=2892

他のデータベースを用いた学習結果①

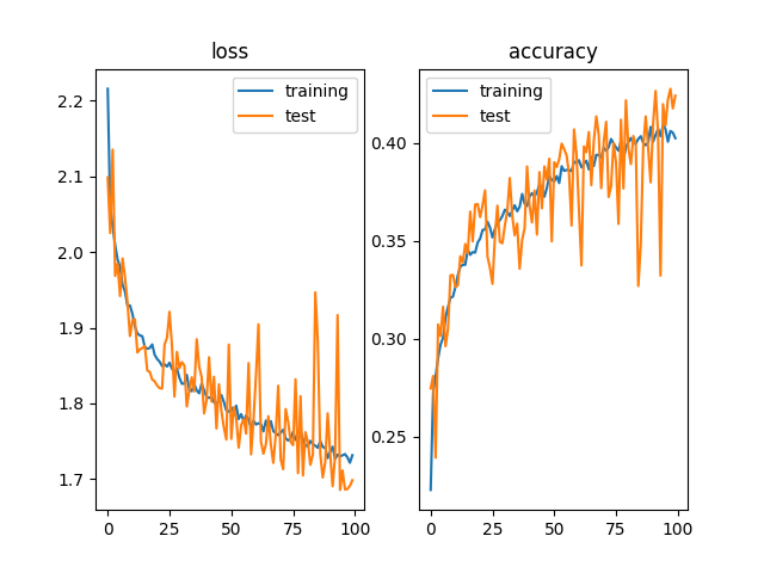
ニューラルネットワークのモデル (Simple_FNN と Simple_CNN) を利用して学習を行い、学習結果を表 4 に示す。

Table 4: 学習結果

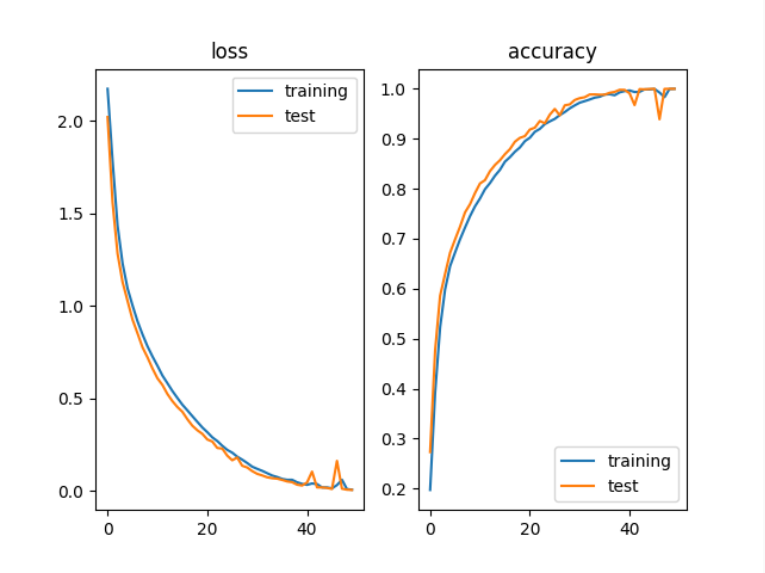
モデル	精度
Simple_FNN	0.42407429218292236
Simple_Cov	1.0

削減データを用いた学習結果②

また、それぞれのモデルの学習過程を図 2a と図 2b にそれぞれ示す。



(a) Simple FNN



(b) Simple CNN

まとめ

- データの偏りがあったため、他のデータベースを用いる
- それぞれのファミリーのデータ数に偏りをなくす?
- 精度が高すぎる原因やテストのほうが精度が高い原因を探る
- より複雑なニューラルネットワークモデルを組む
- 『ゼロから作る Deep Learning』のニューラルネットワークでモデルを組む