

進捗報告

1 今週やったこと

- GA の実験

2 実験の具体案

2.1 探索空間の縮小

- ラベル付きデータによる予測のうち top5 のラベルに対して探索
- 事前学習における予測ラベルについてラベル付きデータとして扱うか否かの探索

2.2 適応度の工夫

- validation data に対しての data augmentation
- validation loss を適応度として使用
- 探索データ以外に誤データを train data として組み込む

3 実験

3.1 実験 1: 誤データの組み込み

個体の評価時に探索データにランダムなラベルを持つ誤データを混ぜて学習させた。また validation data にも data augmentation をした。表 3, 4 に示す。また、選択は 2 個体をエリート選択、残りをトーナメント選択 ($k=2$) とし、交叉は一様交叉、突然変異はランダム遷移を用いた。

3.1.1 結果

図 1, 3 と表 1 に結果を示す。

図 1 について適応度と正答率が上がっていることが分かる。一方で、世代ごとに適応度のばらつきが激しい。これは探索データとは別で入れている誤データのラベルが世代ごとにランダムであるからだと考えられる。

また、図 3 について相関係数は 0.9 ほどであり、より相関関係が強いといえる。また、正答率ごとの適

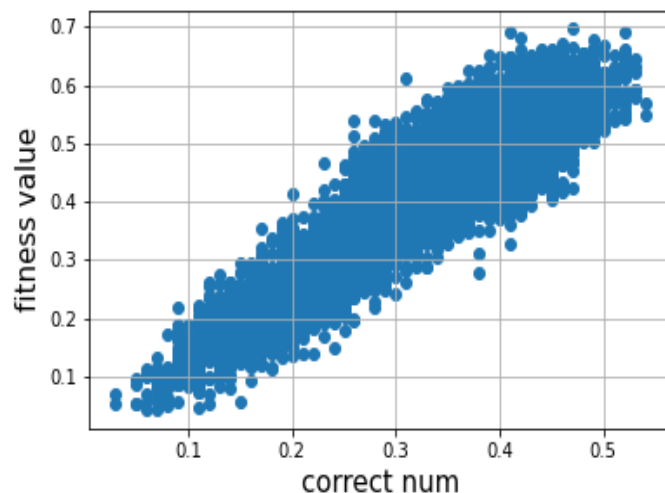


図 1: 実験 1 の散布図

表 1: 実験 1 の結果

閾値	採択数	正答数	識別率
なし	100	54	0.774
40/50	59	40	0.762
45/50	24	18	0.768
baseline(ラベル付きデータのみ)			0.776
baseline モデルによる疑似ラベル 正答率 (76/100)			0.778

応度について最小値及び最大値がより直線的である。これにより正答率ごと平均的な適応度もより直線的であり、前回よりも正答率が上がった理由であり、誤データを入れたほうがより学習できるといえる。

一方表 1 より事前学習から得られる疑似ラベルや baseline も超えることができていないためハイパーパラメータなどの改善をする必要性がある。

改善案としてより多くの誤データを含ませることが一点、また今回誤データとして選ばれたラベルなしデータについて全て統一し、それに対するラベルは各世代ごとに統一している。これについて全世代でラベルを統一すること、また各世代ごとにラベルなしデータについても入れ替えることについても試す必要があると考えられる。

3.2 実験2:事前学習の top5 に対する探索

遺伝子についてこれまでとりうる値が10個あったが top5 のみに探索範囲を絞ることで5個に減らすことで探索空間の削減を狙った。また validation data にも data augmentation をした。表 5, 6 に示す。また、選択は2個体をエリート選択, 残りをトーナメント選択 ($k=2$) とし, 交叉は一様交叉, 突然変異はランダム遷移を用いた。

個体について, 各遺伝子は $\{0,4\}$ の5つのいずれかを持つ。各遺伝子座に対応するデータの top5-accuracy と遺伝子 $\{0,4\}$ とがそれぞれ対応しており, 遺伝子の整数値をラベルに変換することで表現型を得る。

3.2.1 結果

図 2, 4 と表 2 に結果を示す。

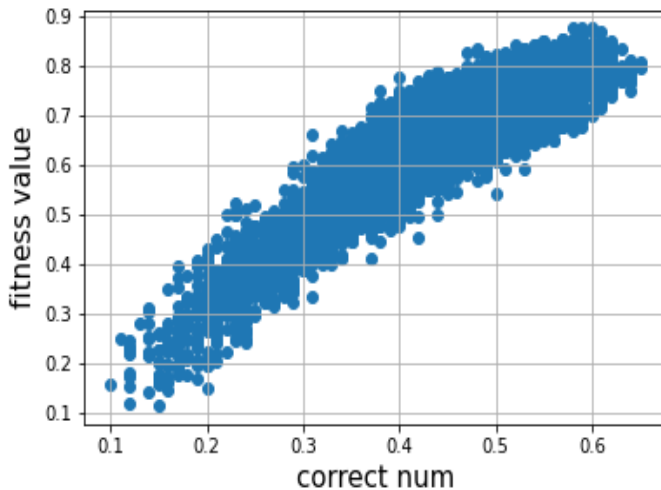


図 2: 実験 2 の散布図

表 2: 実験 2 の結果

閾値	採択数	正答数	識別率
なし	100	55	0.737
40/50	52	37	0.759
45/50	22	15	0.752
baseline(ラベル付きデータのみ)			0.761
baseline モデルによる疑似ラベル 正答率 (77/100)			0.789

結果として表 2 から識別率や正答率はましになったといえる。また図 2 について相関係数は 0.9 ほど

であったが正答率ごとの適応度については少し曲線的であり, 実験 1 の誤データをいれることについて検討すべきであると考えられる。

また, この実験について探索するのは 10 クラス識別のうち 5 クラスであり, 探索データについて top5-accuracy は 93% であった。先ではあるが今後より難しいタスクに適応するには厳しいと考えられる。

4 来週の課題

- 実験設定やハイパーパラメータの改良

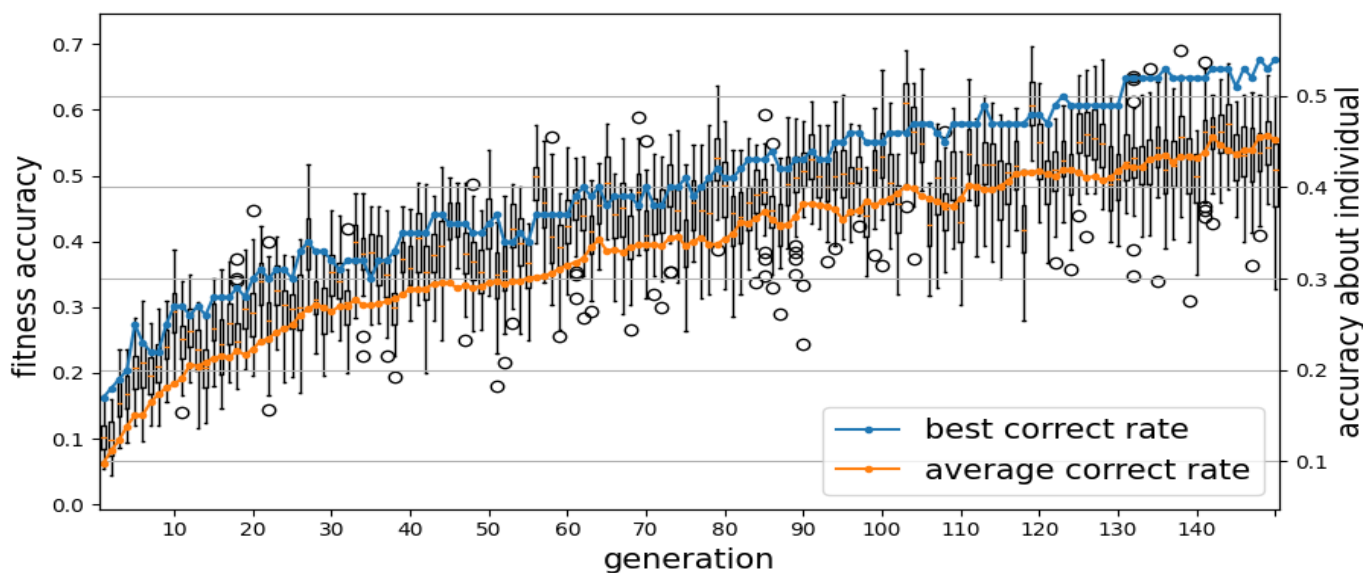


図 3: 実験 1 : 適応度の推移

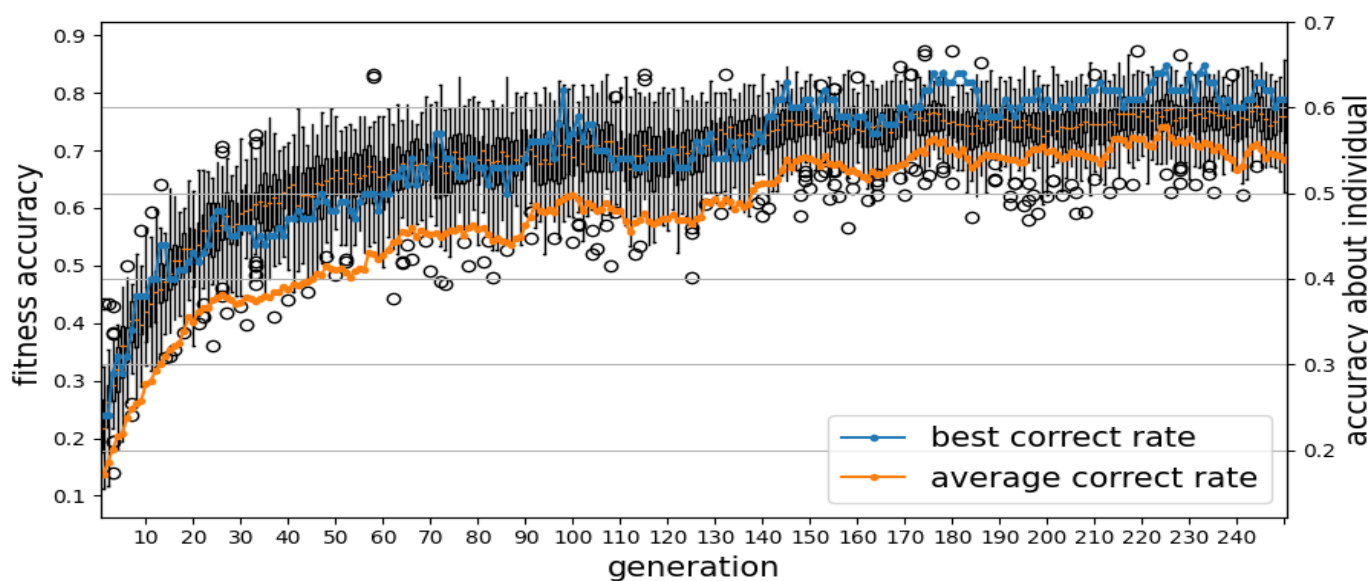


図 4: 実験 2 : 適応度の推移

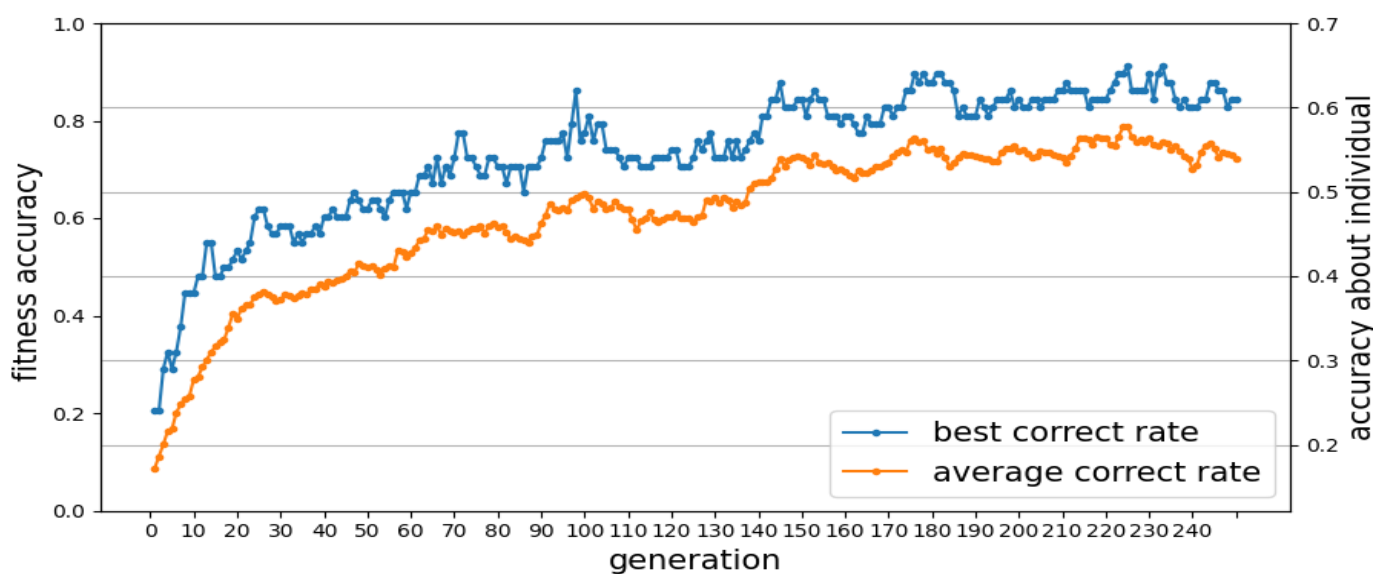


図 5: 実験 2 : 適応度の推移

表 3: 実験 1 における GA の設定

ラベル付きデータ	50
探索データ	100
誤データ (ラベルなし)	100
交叉率	1.0
突然変異率	
遺伝子ごと	0.5
遺伝子座ごと	0.1

表 4: 実験 1 における SimCLR の設定

model	Encoder	ResNet18
	Projection head	2 層 MLP(shape:2048to512)
	classifier	MLP(shape:2048to10)
Encoder 学習		
train_data(unlabeled)		50000
batch size	1024	
epochs	500	
optimizer	Adam(lr=1.0 * 10 ⁻³)	
GA 中の個体評価		
train_data(labeled)		100+100
batch size	32	
epochs	10	
loss	Cross Entropy Loss	
optimizer	Adam(lr=5.0 * 10 ⁻²)	
validation data	50 * 3	
探索された個体の評価		
train_data(labeled)		50+100
batch size	64	
epochs	100	
loss	Cross Entropy Loss	
optimizer	Adam(lr=5.0 * 10 ⁻³)	
test data	10000	

表 5: 実験 2 における GA の設定

ラベル付きデータ	50
探索データ	100
交叉率	1.0
突然変異率	
遺伝子ごと	0.5
遺伝子座ごと	0.1

表 6: 実験 2 における SimCLR の設定

model	Encoder	ResNet18
	Projection head	2 層 MLP(shape:2048to512)
	classifier	MLP(shape:2048to10)
Encoder 学習		
train_data(unlabeled)		50000
batch size	1024	
epochs	500	
optimizer	Adam(lr=1.0 * 10 ⁻³)	
GA 中の個体評価		
train_data(labeled)		100
batch size	32	
epochs	10	
loss	Cross Entropy Loss	
optimizer	Adam(lr=5.0 * 10 ⁻²)	
validation data	50 * 3	
探索された個体の評価		
train_data(labeled)		50+100
batch size	64	
epochs	100	
loss	Cross Entropy Loss	
optimizer	Adam(lr=5.0 * 10 ⁻³)	
test data	10000	