進捗報告

1　今週やったこと

GA を用いた DataAugmentaion

2　実験

前回に引き続きＧＡを用いたアンサンブル学習のための DataAugmentation の実験を行った．
今回は個体の精度を適応度とした集団 1 と，後述の fitness2 を適応度とした集団 2 とを，並行して学習を進めた．共に個体数は 15 とした．今回計算量削減のためアンサンブルを行う個体を集団 1 から上位 3 つ，集団 2 から 5 つを選抜し，集団 2 ではこの 5 つの個体を求めて，精度をよくすることを目標とする．

2.1　実験データ

実験データは cifar10 を用いて，事前学習では epoch 数 300，train_data を各ラベル 5000 枚の計 50000 枚使用し，ＧＡで学習する際は epoch 数 100，train_data は各ラベル 200 枚のオリジナルとそれらすべてを DataAugmentaion したものとを合わせ計 4000 枚とし，test_data は共に 10000 枚とした．また事前学習での accuracy は 0.8475 である．

2.2　遺伝的アルゴリズム

2.2.1　探索空間

探索する水増し操作として画素値操作 (Sharpness，Posterize，Brightness，Autoconstrast，Equalize，Solarize，Invert，Contrast，ColorBalance)，変形操作 (Mirror，Flip,Translate X/Y，Shear X/Y，Rotate) の 16 種類の操作であり，今回はそれらすべてを個別にどの程度強くかけるかおよびどの順序でかけるかということを探索する．各操作についての強度の最大最小を設定し，それを-100%から 100%まで 25%ずつ分 11 段階の度合いとする．ただし，Autocontrast，Equalize，Invert，Mirror については適用するか否かであるためパラメータが 0 以上で適用するとした．強度は 0 から 5 の整数値を持つ 15 個の遺伝子を実数値コーディングによって表現する．また，適用順序に関しては同様に 15 個の遺伝子を持つ順列コーディングによって表現する．確率は 10%ごと 11 段階の実数地コーディングによって表現する．つまり，探索空間は $2^5 * 11^{11} * 15! * 11^{16}$ となる．

2.2.2　選択

選択について，エリート選出によって最も適応度の高い 2 つの個体を選択する．なお，この二つは後述する交叉，突然変異は受けずに次の世代に追加する．残りの選出にはトーナメント選出を用した．トーナメント選出は集団の中から任意の数 (トーナメントサイズ) の個体のうち最も適応度の高い個体を選出し次の世代に追加する．今回トーナメントサイズは 2 とした．

2.2.3　交叉

強度，確率を表す染色体については 2 点交叉，順序を表す染色体については部分写像交叉を用いた．2 点交叉は一対の親染色体をそれぞれ同じ場所で三分割し中央の染色体を入れ替えて交叉を行う．部分写像交叉は親遺伝子を二分割し入れ替える際重複をなくす交叉法で，重複のあった遺伝子について，それに該当した重複する遺伝子座を見つけ，それに対となっているもう一方の親の遺伝子を参照する．

2.2.4　突然変異

強度，確率を表す染色体について，対象となる遺伝子の値を各 50%の確率に 1 増減させ，　順序を表す染色体について，染色体の一部を逆順にする操作か，染色体を二つに分け前後を入れ替える操作のいずれかを行うものとした．

2.2.5　多様性維持

多様性を維持するために，上記 3 つの操作 (選択，交叉，突然変異) を行った集団に対し，適用順序を表す染色体について一致するものが 3 つ以上あれば，それが 2 つになるように一部の個体を突然変異させたうえで次の世代の集団とした．これを集団 2 にのみ適用した．

2.2.6 適応度

集団 2 において個体 i の test_data10000 枚の予測値の集合を $\mathbf{pred}(i)$，accuracy を $f_{acc}(\mathbf{pred}(i))$ とし，予測値の集合の集合 A に対するアンサンブルによる accuracy を $f_{ens\_acc}(A)$ とする，また集団 1 のうち上位 3 個体の予測値の集合の集合を B とする．

$$f_{ens\_acc}(X) = f_{acc}\left(\frac{1}{\#X}\sum_{\mathbf{a}\in X}\mathbf{a}\right)$$

$$fitness2_i = \frac{1}{{}_{15-1}\mathrm{C}_{5-1}}\sum_{A} f_{ens\_acc}(A+B)$$

$$U = \{\mathbf{pred}(1), \mathbf{pred}(2), ..., \mathbf{pred}(n)\}$$

$$A = A \subset U \mid A \text{ have } \mathbf{pred}(i) \wedge \#A = 5$$

とした．

2.3 実験

2.3.1 パラメータ

表 1 に学習パラメータを示す．表 2 に GA の設定

表 1: 学習パラメータ

optimizer	Adam
learning rate	0.001
loss function	categorical_crossentropy
batch size	128
epoch size	30

を示す．

表 2: 実験パラメータ

世代数	32
交叉率	0.9
突然変異率	
強度，確率 (遺伝子ごと)	0.06
順序 (染色体ごと)	0.1

2.3.2 結果

図 1，に accuracy の最良値及び平均値の推移を示す．

また，今回の最良値が 0.8812 であり，前回の最良値は 0.8846 であった．

2.4 まとめ

今回の実験で記載はしていないが，集団 1 よりも集団 2 のほうが個体ごとの accuracy が高いものとなっていたが，個体数や初期値における誤差だと考えられる．しかしその結果，集団 2 では accuracy が高いものほど fitness も高くなってしまっており，アンサンブル学習としての意味は少ないように感じた．この結果からアンサンブル学習用の個体を探索するためには先に良個体の探索を済ませたのちに探索したほうが良いと考えられる．また，今まで愚直に平均して fitness をとっていたが，個体同士の関係をもっとよくみて fitness を与えるべきだと思った．

3 来週の課題

- 前期発表の資料の仕上げ
- 絵本についてのリサーチを進める

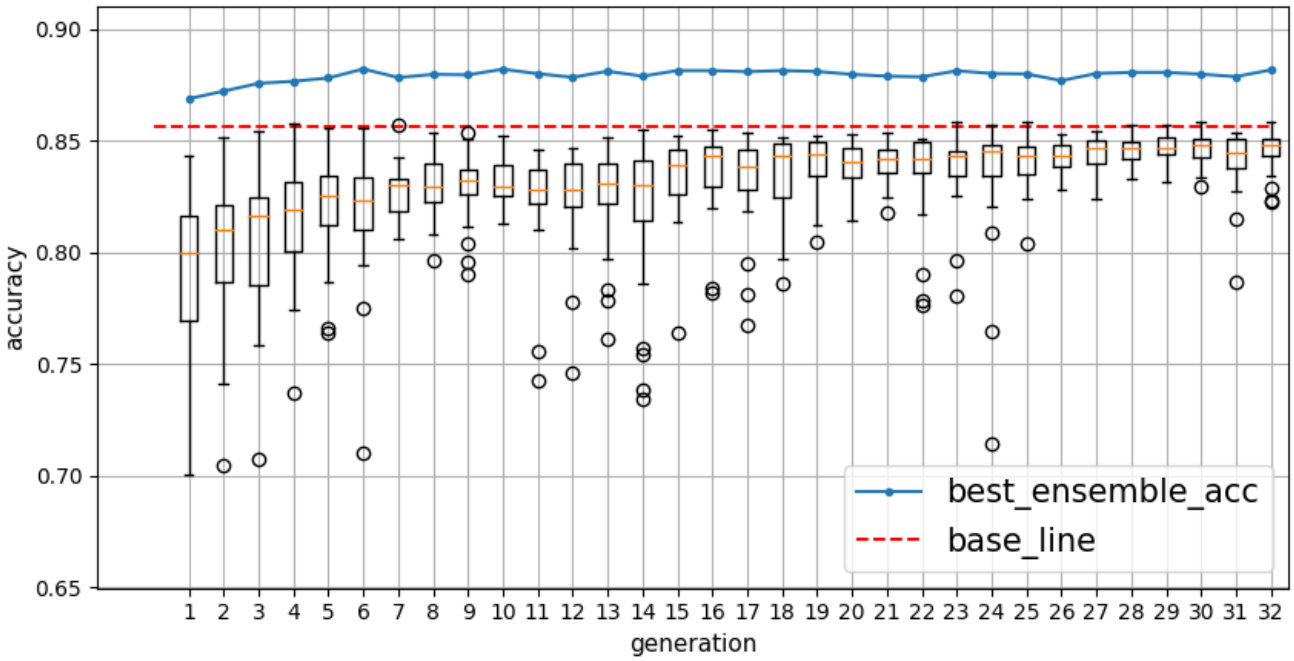

図 1: 集団 2 の accuracy の推移