細川　岳大

引き継ぎ資料

1　はじめに

研究背景として，分類問題におけるラベル付けのコスト削減のため一部にのみラベルが付いたデータである半教師あり学習 (Semi Supervised Learning: SSL) が提案されており，高精度の結果が出ているものの，ラベル数が少なくなると精度が不安定になってしまう報告があり，ラベル付きデータを遺伝的アルゴリズム (Genetic Algorithm: GA) で自動的に少し増やしてあげることで安定させることが目標である．

2　手法および要素技術について

手法について，探索したいラベルなしデータを入力に，それに対応したラベルの集合を出力として学習する．このラベルの集合を GA によって探索する．この時適応度は学習されたモデルに対しラベル付きデータの識別率によって決定する．

探索によって得られたある程度正しいラベル付けがなされたとされるデータに元々のラベル付きデータを加えて SSL で学習して最終的なモデルを得て，それにおけるテストデータの識別率を指標とする．

要素技術は以下のものがあり調べてたり私の卒論に簡単に書いてあるものを参考にするなどしてください．

- SSL
- Contrastive Learning
- SimCLR
- Pseudo Label
- Genetic Algorithm
- (FixMatch)

また， SSL については研究が盛んであるため新たな手法などの調査はたまに行ってください．

しかし，GA を使う性質上一個体の適応度計算の時間が少なくて済むように様々な　SSL　の手法の中でも SimCLR のように事前に Encoder を作ってから分類器を学習する方が効率的かと思われる．その一方で，最終的な識別率の検証の際には FixMatch や他の時間のかかる手法を取り入れるのは一つの手だと思う．

3　現時点での成果と今後の展望

CIFAR10 という 10 クラス識別のデータで各ラベル 5 枚ずつ分かっている状態での実験で，新たに 100 枚のデータに GA でラベル付けをして再学習させたものが baseline に対し識別率を同程度まであげることが出来た．(baseline はモデルを小さくしているので SoTA よりはかなり低い精度である)

展望として，これまでは 50 枚と多めだったので，10 枚の場合でも同程度出せるのか，また出せたとして安定した精度となるのかという点，より付与されるラベルの精度を改善し識別率の向上，また CIFAR10 は均衡なデータのため一部データをマスクするなどしてより現実に近い不均衡データについてできるかどうかなども考えられる．

また，そのために別の手法や工夫を取り入れる必要があり，特に valid 用のラベル付きデータが非常に少なくなるので適応度の工夫が必須である．一応個人的に考えていることとして本研究から少しずれるが deep clustering などの教師なし学習を組み込めないかと考えている．

4　最後に

もしソースコードや内容について分からない点があれば chibilion1take@gmail.com に直接連絡してください．この研究自体良い成果が残せるかは非常に難しいところですが，頑張ってください．