

QUBO ソルバーを用いたDeep Learning Network の訓練時のミニバッチデータ組み合わせの最適化

キヤノンイメージングシステムズ株式会社

大庭 広士

開発における問題点

機械学習モデルの訓練データ収集コストが高

➡ **訓練データの不足により過学習発生**

過学習：訓練時精度は高いがテスト時の精度が低くなる現象。

手法・ツールの適用による解決

ミニバッチ学習時に訓練データ同士の組み合わせを相関が低くなるように最適化ソルバーで求める(従来手法はランダムにデータ選択)

➡ **訓練時のデータの組み合わせを工夫することで過学習を防止**

解決へのアプローチ

似たデータ(=相関が高い)同士で学習しなければ偏ったパラメータ最適化(過学習)が防止できる

相関行列の計算

- 画素値差

$$S_{ij} = -\sqrt{\sum_{n=0}^N \sum_{m=0}^M (I_i(x_n, y_m) - I_j(x_n, y_m))^2}$$

- 潜在変数間距離

$$S_{ij} = -|\vec{Z}_i - \vec{Z}_j|$$

相関が低くなる組み合わせの計算

2 値変数2 次形式最適化問題(QUBO)ソルバーによる求解 (Hを最小化)

$$H = \underbrace{\sum_i \sum_j S_{ij} \phi_i \phi_j}_{\text{相関の和}} + \lambda \underbrace{\left(\sum_{n=1}^M \phi_n - N \right)^2}_{\text{データ個数制約項}}$$

ミニバッチ 学習投入

学習条件

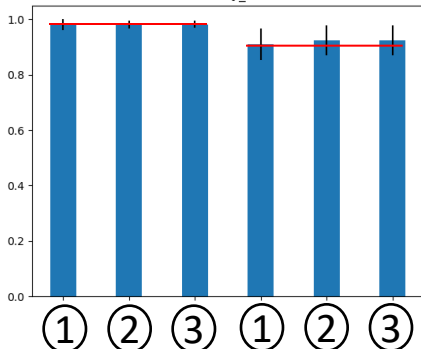
- モデル：CNN-autoencoder
- 学習データ：MNIST, Fashion-MNIST

検証結果

MNIST F-MNIST

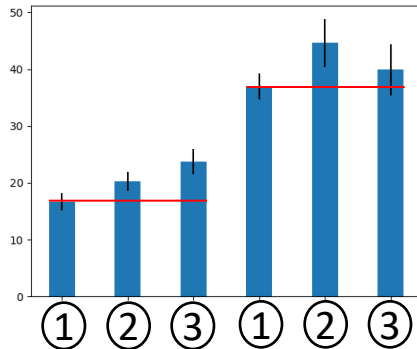
テスト時 / 訓練時精度

accuracy_ratio



クラスタリング性能

PseudeF



- (1) 無相関(ランダム選択)
 (2) 画素値差
 (3) 潜在空間距離

- 過学習防止には効果無し
- クラスタリング性能は向上

まとめ

データの組み合わせ有無・相関の種類による影響

- テスト時/訓練時精度・誤差は影響なし
- クラスタリング性能には改善がみられる

- 従来手法(ランダムなデータ選択)は誤差の最小化にはロバストな手法。
- データの組み合わせを最適化すると、誤差最小化の訓練と同時に他の性能指標を向上させられる可能性がある。

様々なデータの種類の種類・相関の種類・性能指標の種類で検証することで有効性を確認する必要がある。