裸地認識モデル作成における不均衡問題解消のためのデータ抽出手法

八千代エンジニヤリング株式会社　正会員　○吉田　龍人　色川　瑞希　藤井　純一郎
非会員　岩村　尚人

１．はじめに

機械学習において不均衡なデータで学習したモデルは，少数データに対する性能が低下することがある．ニューラルネットワークを用いた画像認識モデルを不均衡データで学習させる場合，データ量に合わせた Loss の重み付けや，少量データを水増しするオーバーサンプリング，多量データを減らすアンダーサンプリングなどが対策として実施される [1]．一方でこれらの手法がかえって学習に悪影響を及ぼすことがある．例えば極めて不均衡なデータで重み付けを変えた場合，最適化計算時の少数データに対する誤差関数項が過大となり，Weight の更新が不安定となる．オーバーサンプリングでは，似た拡張画像を繰り返し学習することで過学習を誘引する．単に多数データを減らすだけのアンダーサンプリングでは，多数データが持つ多様性が損なわれ，汎化性が低下することが懸念される．よって学習データの不均衡問題は適切な手法によって解決する必要がある．

本研究は不均衡データを使った CNN の分類モデルの学習にて，ミニバッチ作成時のデータ抽出手法を改善することで不均衡問題の解消を行うものである．従来手法と提案手法の 2 手法でモデル作成時の学習曲線を比較し，未学習画像を推論することで各モデルの性能の違いを評価する．

２．実験内容

本研究は空撮画像から切り出した 1 辺 224px の画像において，裸地化の有無を 2 クラスで分類するタスクにて検証を行う．**図-1** に教師となる Train 画像例を示す．正常画像は主に全体が草で覆われた画像で構成されるが，一部にアスファルトなどの人工構造物を含んだ画像も含まれる．一方，異常画像は植生の一部が枯れて，地面が露出した画像で構成される．データセットの内訳を**表-1** に示す．表より正常と異常の枚数に偏りあることが分かる．Train 画像はいずれも 2021 年 11 月に撮影された画像で構成されている．Test 画像は 2020 年に撮影された画像であるが撮影日が 7 月，8 月，10 月と様々で，学習画像とは特徴量が大きく異なる**図-2** のような画像も含まれる．

不均衡問題解決のアプローチでは Yan[1]らの手法を参考にした．Yan らは学習時にモデルに入力するミニバッチの中身をランダムに抽出するのではなく，少量データと多量データの比率を揃えて抽出することで，不均衡問題が解消されることを示した．ただし論文にて示された手法では，多量データの数から 1Epoch のミニバッチ数を決定し，ミニバッチ数に合わせて少量データを複数回サンプリングするアルゴリズムであったため，各 Epoch で同一データを学習するデメリットがあった．そこで 1Epoch のミニバッチ数を少量データの数によって決定し，ミニバッチ数に合わせて多量データをアンダーサンプリングするアルゴリズムに変更することで，Epoch ごとに同一データを複数回学習することがないよう手法の改善をした．

(a) 正常画像　(b) 異常画像

図-1 Train 画像例

表-1 データセット内訳

	正常	異常
Train	5354 枚	219 枚
Test	2103 枚	551 枚

(a) 7 月画像例　(b) 8 月画像例

図-2 Test データ例

キーワード　深層学習，不均衡データ，画像分類，アンダーサンプリング
連絡先　〒111-8648　東京都台東区浅草橋 5-20-8 CS タワー 八千代エンジニヤリング(株)　ＴＥＬ03-5822-6843

3．実験条件および結果

学習を行うにあたって，モデルは ResNet18，Loss は Cross Entropy，Optimizer は学習率を 0.001，重み減衰を 0.0001 とする Adam，Epoch は 100，Batch size は 16 に固定した．この条件に対して，正常と異常を合わせた 5573 枚からランダムに 16 枚ずつ取得する従来のミニバッチの抽出手法と，必ずバッチ内に正常画像が 12 枚，異常画像が 4 枚含まれるようにした提案手法の 2 手法で得られたモデルを比較した．従来手法の場合エポックごとに全ての画像から 5568 枚の画像がランダム抽出されるが，提案手法の場合，5354 枚の正常画像から 648 枚，219 枚の異常から 216 枚がエポックごとに抽出される．Validation には，Test 画像のうち正常と異常のそれぞれから 2 割ずつランダム抽出した 530 枚を使用し，各 Epoch 終了時点で Validation Loss が最も低くなったモデルをベストモデルとした．学習時は全ての画像に対してランダムに反転と回転を加えるデータ拡張を行った．

2 手法での学習によって得られた学習曲線を図-3 に示す．従来手法では毎 Epoch で多数の正常データを学習しているため早期から低い Train Loss を示すが，Train Loss と Validation Loss の乖離が生じており，明らかに過学習が発生している．提案手法では Train Loss と Validation Loss 間の乖離が小さく，過学習が抑制できている．2 つのモデルで Test データを推論し，異常クラスの Confidence を取得した結果を図-4 に示す．ミニバッチのデータ抽出法を改善することで，正常と異常のクラス分類精度が向上した．提案手法モデルで誤分類が発生した画像例を図-5 に示す．図-5(a)に示す人工構造物などの非草地のオブジェクトが写った正常画像は異常と誤分類される傾向にあった．図-5(b)のような Train 画像とは特徴量の異なる異常画像は正常と誤分類される傾向にあった．

4．終わりに

本研究では不均衡データを用いた画像分類モデルの学習時におけるミニバッチのデータ抽出法を改善し，結果を評価した．これにより「不均衡データでの学習の安定化」，「入力データ数削減による学習の迅速化」，「モデル性能の向上」という成果が得られた．本手法でクラス間不均衡は解消されたが，クラス内不均衡の課題は残る．今回のデータセットの正常画像には画像全体が草に覆われた画像が多いためか，モデルが非草地の非裸地化領域を含んだ画像を異常と誤分類する傾向にあった．これに対して，誤分類画像を教師に追加するだけでなく，多量データから画像を抽出するアンダーサンプリングの仕組みを改善することで問題の解決を図りたい．特に本研究ではランダムで抽出を行ったが，学習時の出力値に基づいて抽出する画像を決める仕組みへと変更することで改善が期待される．

参考文献

1) J. M. Johnson, T. M. Khoshgoftaar.: Survey on deep learning with class imbalance, Journal of Big Data, 2019.
2) Y. Yan et al.: Deep Learning for Imbalanced Multimedia Data Classification, IEEE ISM, 2015.

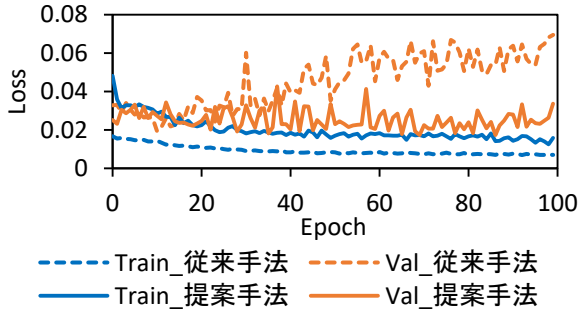
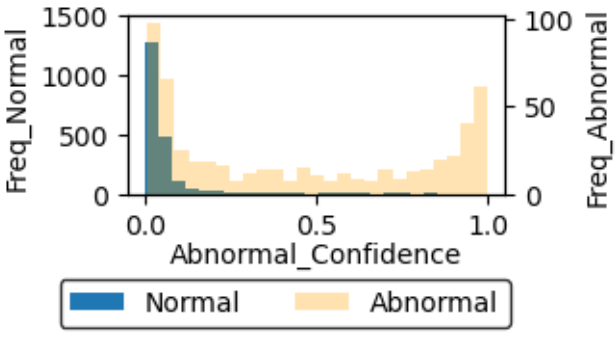


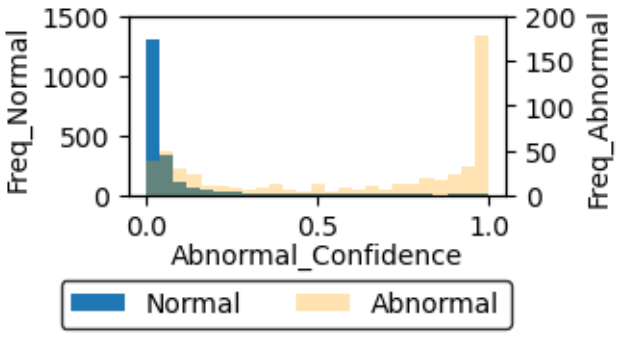
図-3 学習曲線

(a) 異常となった正常画像　(b) 正常となった異常画像

図-5 誤分類発生画像例



(a) 従来手法モデル



(b) 提案手法モデル

図-4 Test 画像に対する分類結果