4Q-03

異常度算出手法の変更によるGANを用いた異常検知性能の評価

森仁美† 中野美由紀‡ 丸千尋† 小口正人†
†お茶の水女子大学 ‡津田塾大学

1 はじめに

異常検知の技術は医療診断，製造業における部品検査など様々な分野において実用化され，盛んに研究が行われている．さらに，昨今の機械学習の発展，深層学習の利用により高精度な異常検知が可能となった．

深層学習のモデルの１種である敵対的生成ネットワーク (GAN)[1] を用いた異常検知についても，既に多くの手法が提案されている．GAN を用いる場合，正常性からの逸脱度合いを示す異常度の算出には，異常検知対象のデータと GAN の再構築データとの差分である再構築誤差と，識別器の判定結果である識別誤差の２つの誤差が指標に用いられる．２つの指標の比率が異常検知精度に影響を与える可能性があるが，既存論文の多くは異常検知の精度を研究対象にしており，各指標が精度に与える影響について詳細な分析を行っているものはあまり見られない．

そこで本稿では，GAN を用いた異常画像検知における異常度算出指標の比率を変化させることで，各指標が異常検知の精度に与える影響について考察を行った．さらに，異常検知モデルの訓練に用いるデータや検知対象の異常データの種類を変えて比較し，データの特性に合わせた分析を行った．

2 関連研究

GAN を用いた異常検知手法には，様々なものが存在する．Schlegl らの AnoGAN[2] は，正常データのみで事前学習した GAN を用いて異常データを検知する．この手法は，未知のデータに最も近しいデータを生成するノイズを求める際に更新学習を行うため，多くの時間を要するという特徴をもつ．本稿で使用した Efficient GAN[3] は BiGAN[4] をベースとした異常検知手法である．BiGAN では，通常の GAN の構造にデータからノイズを求めるエンコーダを導入し，効率的に学習や生成を行うことを可能にしている．Efficient GAN の詳細については３節にて記述する．

これら多くの研究では，異常検知の精度向上を目的としており，例えば，先の研究では異常度の算出指標である再構築誤差と識別誤差の比率を９：１に固定して実験が行われている．よって本研究においては，異常度の算出部分に焦点を当て，精度との関係を調べた．

Evaluating the Performance of Anomaly Detection with GAN by Varying Anomaly-scores
†Hitomi Mori
‡Miyuki Nakano
†Chihiro Maru
†Masato Oguchi
†Ochanomizu University
‡Tsuda University

3 GANとGANによる異常検知について

GAN は，生成器 G と識別器 D で構成される生成モデルである．(図１）識別器は，訓練データから抽出した本物データ x，生成器がノイズを元に生成した偽物データ G(z)，の２つの入力に対し本物かを示す推定確率を出力し，それを元に誤差を計算して訓練を行う．

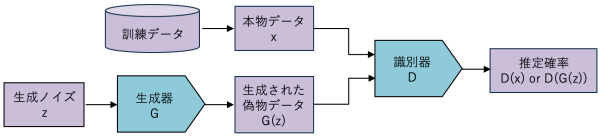

図 1: GAN のアーキテクチャ

異常検知においては，正常データのみを用いて GAN を事前学習することで異常データの検知が可能となる．Efficient GAN では，訓練時，生成器はノイズ z を入力として本物に近い生成データ G(z) を出力，エンコーダ E は本物データ x を入力とし，x に対応するノイズ E(x) を出力，識別器は画像とノイズのペアを入力とし，いずれのペアが入力されたかを示す推定確率を出力，の流れで事前に学習が行われる．(図２）

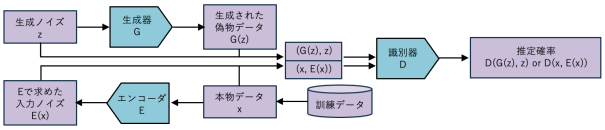

図 2: Efficient GAN 訓練時のアーキテクチャ

異常検知時，未知のデータ x からエンコーダが求めたノイズ E(x) を元に，生成器が再構築したデータ G(E(x)) と x の差分である再構築誤差 (residual loss) と，識別器に入力したペアの判定結果である識別誤差 (discrimination loss) の２つを出力する．(図３）その後，それらを元に次式で異常度を計算する．

$$\text{loss} = (1-\lambda)\times\text{residual loss}+\lambda\times\text{discrimination loss}$$

λ は２つの誤差の比率を制御する係数で，λ が小さいほど再構築誤差の影響が，λ が大きいほど識別誤差の影響が大きくなる．上の式で求めた異常度が事前に設定した閾値より大きければ，異常と判定されたと言える．

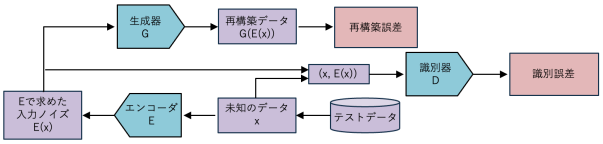

図 3: Efficient GAN 異常検知時のアーキテクチャ

4　実験

4.1　実験概要

0から9までの手書き数字の画像データセットである MNIST データを対象に，Efficient GAN を用いた異常画像検知を行った．その際，データに関する条件を変えた上で，異常度の式の λ を0から1まで0.1ずつ変化させ，2つの誤差の比率を変えたときの精度の変化を調べた．訓練データの種類を変えて行った実験では，異常データを「2」，各画像の訓練枚数を1500に設定し，訓練データに「4」と「6」,「3」と「9」,「7」と「8」を用いた．また，異常データの種類を変えて行った実験では，訓練データを「4」と「6」，訓練枚数を1500に設定し，異常データに「4」と「6」以外の8つの数字を用いた．また，全ての実験において，閾値はテストデータの異常度の高い方から50番目の値に設定し，異常度が閾値を超えたデータについて訓練済みモデルが異常と予測したと判断した．その上で，精度を示す値として，モデルの予測結果と実際の結果を元に F1-score を算出した．

4.2　実験結果

訓練データの種類を変化させた結果を図4に示す．これより，λ の値が小さいほど F1-score が大きくなることが読み取れ，訓練データによらず同様の傾向が見られる．この結果より，再構築誤差比率が高いほど精度が良くなることがわかった．これは訓練データが「4」と「6」の場合における，λ が 0.1，0.9 のときのテストデータの異常度を表した散布図からも読み取れる．(図5，図6)

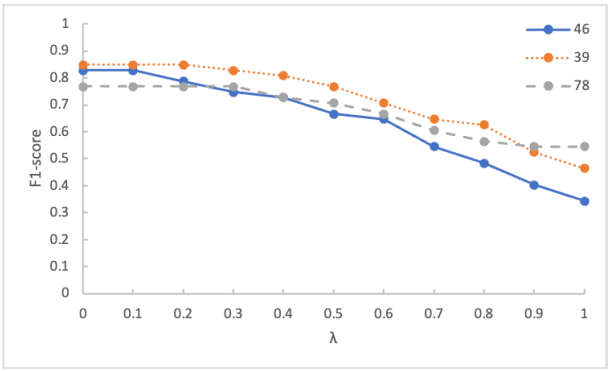

図 4: 各訓練データにおける誤差比率を変えたときの F1-score の変化

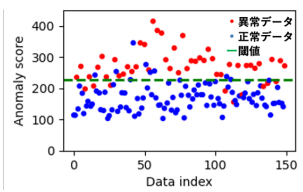

図 5: λ=0.1 の異常度

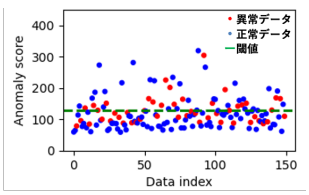

図 6: λ=0.9 の異常度

異常データの種類を変化させた結果を図7に示す．これより，ほとんどの数字において λ の値が小さいほど F1-score が高い傾向にあることが読み取れる．唯一，異常データが「1」の場合は誤差比率を変えてもほとんど精度の違いが見られなかった．これは，4の縦棒を斜めに書いた場合，6の縦線がまっすぐな場合など,「1」が訓練データに用いた2つの数字に比較的似ているため，異常検知の精度が他の数字より低いことが要因と考えられる．しかし，全ての異常データについて，λ が 0.5 未満のとき F1-score の最高値であることから，異常データの種類によらず，識別誤差より再構築誤差の比率が高い方が精度が良いとわかった．

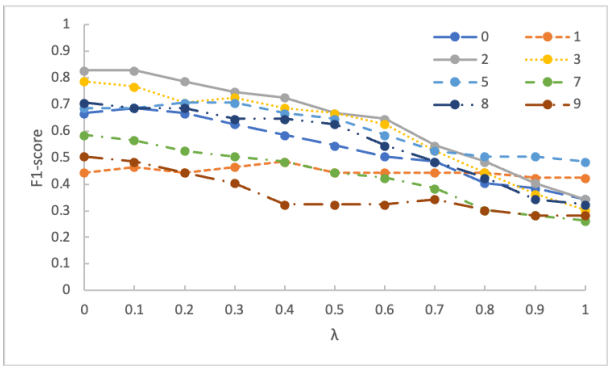

図 7: 各異常データにおける誤差比率を変えたときの F1-score の変化

5　まとめと今後の課題

画像データに対して GAN を用いた異常検知を実施し，データに関する条件を変えた上で再構築誤差と識別誤差の比率と精度の関係を調査した．訓練データの種類によらず再構築誤差比率が高いほど精度が高くなること，異常データの種類によらず識別誤差より再構築誤差の比率が高い方が精度も良くなることがわかった．今後は，条件を更に変化させて実験するとともに，画像以外のデータに対しても同様の研究を行いたい．

参考文献

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde- Farley, S. Ozair, A. Courville, and Y. Bengio. ”Generative Adversarial Nets,” arXiv:1406.2661, 2014

[2] T. Schlegl, P. Seebock, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, ”Unsupervised anomaly detection with generative adversarial networks to guide marker discovery,” In International Conference on Information Processing in Medical Imaging, pp. 146–157, 2017.

[3] H. Zenati, C. S. Foo, B. Lecouat, G. Manek, and V. R. Chandrasekhar, ”Efficient GAN-based anomaly detection,” arXiv:1802.06222, 2018.

[4] J. Donahue, P. Krahenbuhl, and T. Darrell, ”Adversarial feature learning,”arXiv:1605.09782, 2016.