TSUBAME 共同利用　令和7年度　学術利用　成果報告書

利用課題名　シンセティック テキスト: AI 生成テキストの検出
英文：SYNTHETIQ TEXT: AI-generated text detection/localization

利用課題責任者
越前 功（Isao Echizen）

所属: 国立情報学研究所
Affiliation: National Institute of Informatics
URL: https://research.nii.ac.jp/~iechizen/official/

邦文抄録（300 字程度）

生成 AI の急速な発展により、AI が生成したテキスト・画像・映像と人間が作成したコンテンツの判別が困難になっている。本プロジェクトでは、TSUBAME4 の GPU 計算資源を活用し、AI 生成コンテンツの検出に関する研究を 4 つの方向で実施した。(1) AI 生成テキスト検出のための文単位分類モデルの学習とハイパーパラメータ最適化、(2) AI 論文査読システムの脆弱性分析のための retrieval-augmented classification フレームワークの構築、(3) ブラックボックス敵対的攻撃手法の開発と画像偽造検出モデルの脆弱性評価、(4) ディープフェイク検出器のショートカット学習と人口統計バイアスの因果分析を行い、それぞれ定量的な成果を得た。これらの成果は 4 本の学術論文として国際会議・論文誌に投稿中または準備中である。

英文抄録（100 words 程度）

Using TSUBAME4 GPU resources, we conducted research on AI-generated content detection as part of the SYNTHETIQ TEXT project. Four research directions were pursued: (1) sentence-level AI-generated text classification with hyperparameter optimization, achieving high detection accuracy; (2) interpretability analysis of AI paper review systems using a retrieval-augmented classification framework; (3) development of black-box adversarial attack methods achieving near-100% success rates and evaluation of image forgery detection model vulnerabilities; and (4) causal auditing of deepfake detectors, revealing shortcut learning and demographic bias pathways. Four papers are under review or in preparation at major venues.

Keywords: 5つ程度
AI-Generated Text Detection; Adversarial Attack; Image Forgery Detection; Deepfake Detection; Causal Analysis

背景と目的

生成 AI の進化に伴い、高品質なテキスト・画像・映像の自動生成が容易になり、偽情報の拡散や学術不正といった社会的脅威が深刻化している。特に大規模言語モデル（LLM）により生成されたテキストは人間が書いたものとの判別が極めて困難であり、また拡散モデル等により生成された画像・映像もリアリティが飛躍的に向上している。これらの AI 生成コンテンツを正確に検出する技術の確立は喫緊の課題である。

本プロジェクトでは、TSUBAME4 の GPU 計算資源を活用し、AI 生成コンテンツの検出技術について、テキスト・画像・映像の各モダリティにわたる包括的な研究を実施した。深層学習モデルの学習・評価に加え、検出システムの脆弱性分析や因果分析による信頼性評価も行い、実用的かつ頑健な検出技術の構築に向けた知見を得た。

概要

本プロジェクトでは以下の 4 つの研究を実施した。

1. AI 生成テキスト検出モデルの学習・最適化

TSUBAME4 の GPU ノード上で、AI 生成テキストを文単位で検出する分類モデルの学習を行った。複数のエンコーダアーキテクチャの比較検討、自動ハイパーパラメータ探索の導入、学習データセットの整備・改善を実

施した。小型モデルが大型モデルと同等の精度かつ大幅に高速な推論を実現できることを実証し、検出アプリケーション SYNTHETIQ TEXT に採用した。

2. AI 論文査読システムの脆弱性分析

IRIS フレームワーク（retrieval-augmented classification model）を実装し、AI による論文査読システムが論文品質を判別する際に依拠する特徴の解析を行った。1,448 本の学術論文をチャンク分割し FAISS vector DB に格納、learnable query vectors によるセマンティック検索と linear attention mechanism による長文書分類を実施した。Cohen's d 効果量、query overlap 分析、retrieval pattern プロファイリング等の解釈性分析も行った。

3. ブラックボックス敵対的攻撃と画像偽造検出の脆弱性評価

オートエンコーダによる特徴量分離の学習、Stable Diffusion モデルの推論・最適化、CLIP モデルの活用を行った。既存の敵対的攻撃手法（ADBA、RayS 等）と画像偽造検出モデル（FakeShield、TruFor、FOCAL、IFOSN）のテストを実施した。ImageNet、Cifar-10 等の画像分類データセットと CASIA-V1、Columbia 等の画像偽造検出データセットを用いた大規模実験を行った。

4. ディープフェイク検出モデルの因果分析

EfficientNet-B0 によるディープフェイク検出モデルの評価と因果分析を実施した。Celeb-DF v2 で学習し、v1 および DFD データセットでクロスドメイン評価を行った。16 の解釈可能な特徴量（周波数エネルギー、顔ランドマーク、人口統計情報等）を抽出し、FCI 因果構造学習アルゴリズムと介入実験を組み合わせた分析を行った。

結果および考察

1. AI 生成テキスト検出

モデルアーキテクチャの改善とハイパーパラメータ最適化により、高い分類精度を達成した。小型モデルが大型モデルと同等の精度で推論時間を大幅に短縮できることを実証し、実用的な検出アプリケーションとして外部提供を実現した（2026 年 3 月、日本物理学会向け提供）。

2. AI 論文査読システムの脆弱性分析

IRIS モデルが論文品質を区別する 3 つのメカニズムを発見した：(a) content absence － 高品質論文にのみ存在する詳細な手法記述（Query 2 で 34.7 倍の検索ギャップ）、(b) quality difference － 同種コンテンツにおける記述の質の差、(c) preprocessing asymmetry － 低品質文書の 55.8% に未フィルタの LaTeX artifacts が存在。8 つの query vectors のうち 5 つが Cohen's d > 1.0 を達成し、Queries 3–7 が超相関クラスター（pairwise r = 0.92–0.99）を形成することから、モデルが冗長な特徴に依存していることが示唆された。

3. 敵対的攻撃と画像偽造検出

オートエンコーダによる特徴量分離を活用した敵対的攻撃手法により、ブラックボックス設定で ImageNet テストデータセットにおいてほぼ 100%の攻撃成功率かつクエリ数中央値 1 を達成した。また、Stable Diffusion を用いた敵対的偽造画像の生成により、画像偽造検出モデルの脆弱性を評価し、低い F1/IOU スコアを達成しつつ視覚的品質を維持することに成功した。

4. ディープフェイク検出の因果分析

ディープフェイク検出器がドメイン固有のショートカット（圧縮方式やデータセット由来情報）に依存し、真の改ざんアーティファクトを学習していないことを発見した。クロスデータセット精度が 95.2%（ドメイン内）から

52.3%（DFD）に低下し、汎化性能の欠如を確認した。因果分析により、直接的なショートカット依存（Domain → Prediction）と隠れた人口統計バイアス経路（Race → Domain → Prediction）を同定した。

まとめ、今後の課題

本プロジェクトでは、TSUBAME4 の GPU 計算資源を活用し、AI 生成テキスト検出、AI 査読システムの脆弱性分析、敵対的攻撃による画像偽造検出の評価、ディープフェイク検出の因果分析という 4 つの研究を実施した。各研究において定量的な成果を得るとともに、AI 生成テキスト検出のアプリケーションの外部提供も実現した。今後の課題として、(1) 新たな生成モデルへの対応と検出手法の継続的更新、(2) テキスト・画像・映像を横断するマルチモーダル検出の統合、(3) 実環境における検出モデルの頑健性評価と改善を進める予定である。