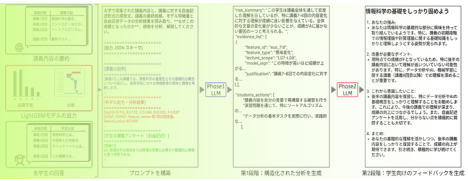
機械知覚&ロボティクスグループ
中部大学　工学研究科　情報工学専攻

MPRG
MACHINE PERCEPTION AND ROBOTICS GROUP

2025年度　山下研究室　修士論文発表　アブストラクト

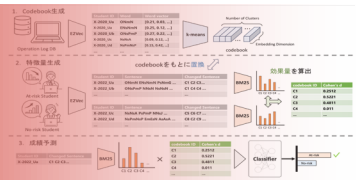
Natural Language Processing, LLM, Learning Analytics

自由記述文データを用いた講義改善フィードバックの構築に関する研究
小池 正基



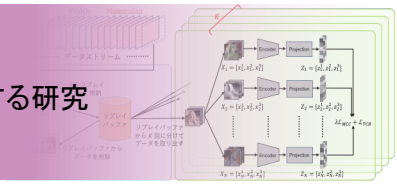
At-risk Prediction, Feature Weighting, Learning Behavior Patterns, E2Vec

学習行動ログデータを用いた成績予測における説明性向上に関する研究
舘 良太



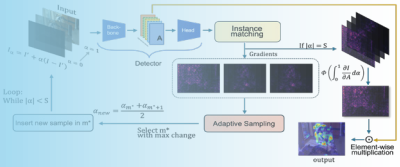
Self Supervised Online Continual Learning

自己教師ありオンライン継続学習における収束速度と勾配相関の改善に関する研究
今井 孝洋



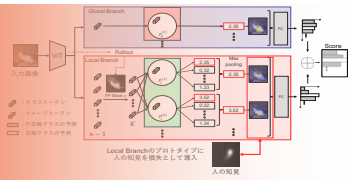
Explainable AI, Integrated Gradients

物体検出モデルの判断根拠可視化における解釈性向上に関する研究
仲井 悠真



Prototype, Human-in-the-loop

プロトタイプ法における人の知見の組み込みによる精度向上に関する研究
落合 祐馬



自由記述文データを用いた講義改善フィードバックの構築に関する研究

TP24006 小池 正基　　指導教授：山下 隆義

1. はじめに

デジタル教科書の操作を記録した学習行動ログデータから，講義を十分に理解できていない学生を早期発見し，学習に対する改善等を提供する試みが行われている [1]．しかし学習行動ログデータは，学生の講義に対するモチベーションや，内容についての解釈といった抽象的な情報を考慮できない．そのため，具体的な学習支援を提供することが難しい．本研究では，学生の主観が含まれる自由記述形式の学習日誌に注目する．予備調査より，講義に対する日誌の変化と成績の相関が高いことが判明した．そこで，学習日誌の変化量を用いた成績予測モデルを構築し，成績予測時の分析結果を用いた LLM 推論によって，より学生にとってわかりやすく，具体的な根拠と改善策を提供可能な学生向けフィードバックシステムの実現を図る．

2. 研究背景

教育分野において学生の成績の要因を調査する研究が活発である．Stephanie ら [2] は，工学系の大学生に対して自由記述アンケートを実施し，回答内容と GPA に相関があるか調査した．分析の結果，成績の高い学生はエンジニアの実務内容や目的を表現する単語や，数学，科学に対するポジティブな発言が見られ，成績の低い学生が用いる単語と明確な差が確認された．これらから，学生が記録した学習日誌を分析することで，将来の成績を予測するとともに，講義へのモチベーション低下など，成績変動に関与する要因を根拠として具体化できると考えられる．

3. 予備調査

本研究で用いる学習日誌データセットに対して分析を行い，成績との関連性を調査する．本調査では，九州大学で収集した，情報科目の講義における学習日誌と，受講者の成績で構成されるデータセットを用いる．学習日誌は講義直後に，①今回の講義内容の説明，②講義内容や講義についてわかったこと，③講義内容についてわからなかったこと，④講義への質問，⑤講義の感想の 5 項目に回答する形式で行われた．学習日誌は 2020 年から 2022 年までの期間，合計 377 名の学生に対して実施された．各受講者の成績は A，B，C，D，F の 5 段階評価である．

本データセットは講義回ごとに同一の学生から学習日誌を収集しているため，学生 s の講義 i 回目の回答 $x_{s,i}$ は $x_{s,1}, x_{s,2}, \ldots, x_{s,15}$ のような時系列データとして表現できる．このとき，各回答は講義順に時系列が進み，講義 i 回目の回答が講義 $i+1$ 回目の回答に影響しうるため，連続性や変化（改善・停滞・悪化）といった時系列的な情報が包含される．この時系列情報を埋め込み間距離を用いて定量化することで，各学生の継続的な講義への取り組み方を調査する．講義間における学習日誌の埋め込み間距離 Dist の算出方法を式 (1) に示す．

$$\mathrm{Dist}(i,g) = \mathrm{median}\Big(\sum_{n \in U_g} 1 - \frac{\mathbf{A}_i^\top \mathbf{A}_{i+1}}{\|\mathbf{A}_i\|_2 \, \|\mathbf{A}_{i+1}\|_2} \Big) \quad (1)$$

ここで，$\mathbf{A}_i$ は Word2Vec[3] によって得られた講義 i 回目の学生回答の埋め込み表現，U_g は成績 g の学生群である．式 (1) で示すように，隣接する講義回の文章埋め込みに対してコサイン距離を求めることで，各講義間の意味的な変化量を抽出する．また，コサイン距離は学生ごとに変化が大きいため，全学生から算出したコサイン距離の中央値を利用する．

予備実験の結果を図 1 に示す．図 1 より，成績が高い学生ほど講義間コサイン距離が大きく，講義 14 から 15 回目における学習日誌の回答内容が大きく変化していることが確認できる．この結果から，学習日誌の変化量は，成績予測に有用な指標であるといえる．

4. 提案手法

予備実験より，学習日誌における内容の変化量が成績に大きく影響していることが示された．本研究ではこの結果

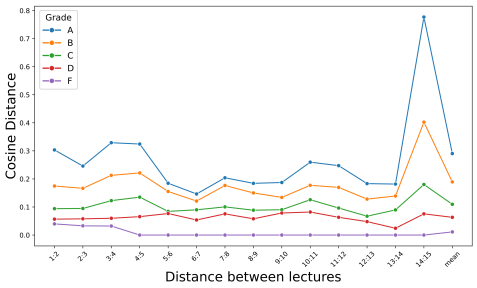


図 1: 講義間のコサイン距離

を踏まえ，以下の 2 つから構成されるアプローチによって，より効果的なフィードバックを提供可能な学習支援を実現する．

1. 学習日誌の取り組み差分情報を用いた成績予測
2. LLM による学生向けフィードバックの生成

4.1. 学習日誌の取り組み差分情報を用いた成績予測

各学生の全 15 回分の学習日誌を文章埋め込みに変換し，予備実験と同様に変化量を算出することで，成績予測モデルの精度向上を図る．成績予測モデルには，予測結果の判断根拠を容易に算出可能である点と系列データへの解釈性の高さから，Light Gradient Boosting Machine（LightGBM）を利用する．予測には，予備実験で用いた講義間コサイン距離だけでなく，講義間ユークリッド距離，学習日誌の各講義回における回答文字数，講義全体の欠席回数を特徴量として利用する．講義間コサイン距離はベクトル間の類似度から文脈や意味的な変化量を算出し，講義間ユークリッド距離はベクトル間の直接的な距離関係から使用単語の変化や文章構成の変化を定量化するものである．

4.2. LLM による学生向けフィードバックの生成

LightGBM モデルによって得られた成績予測結果とモデルの分析を利用することで，より説得力のある，学生向けの学習方法フィードバックを生成する．図 2 に本手法の概要を示す．本手法は LLM による 2 段階推論によって生成文の品質向上を目指す．第 1 段階では，モデルの SHAP 分析結果から寄与度が高い講義回を抽出し，寄与度が高くなった根拠を講義内容と学習日誌の回答から予測する．このとき，寄与度の高い要素が複数ある場合に備え，出力を構造化することで網羅性の欠落を防止する．第 2 段階では，第 1 段階で得られた分析結果と講義資料を入力することで，根拠と改善策を分かりやすく提示した学生向けのフィードバック文を生成する．各 LLM は GPT-4o の生成文を真値として訓練する．

5. 評価実験

本章では，提案手法の有効性を検証するための評価実験を行う．

5.1. 実験条件

学習日誌を用いた際の成績予測精度を比較し，提案手法の有効性を検証する．ベースラインとして，文章埋め込みを直接入力する LightGBM モデルを用いる．各講義回ごとにモデルを訓練し，その出力の平均値を最終的な予測確率とする．評価には分割交差検証法 (k=5) を行い，Accuracy，F1-score を用いて提案手法の有無における精度の変化を検証する．

5.2. 実験結果

各手法における講義理解度予測精度の比較を表 1 に示す．表 1 より，ベースラインよりも Accuracy が 6.43pt，F1-score が 8.48pt 向上し，提案手法の有効性を確認した．次に，定性評価として SHapley Additive exPlanations (SHAP) による特徴量寄与度を可視化し，モデルの判断根拠を分析

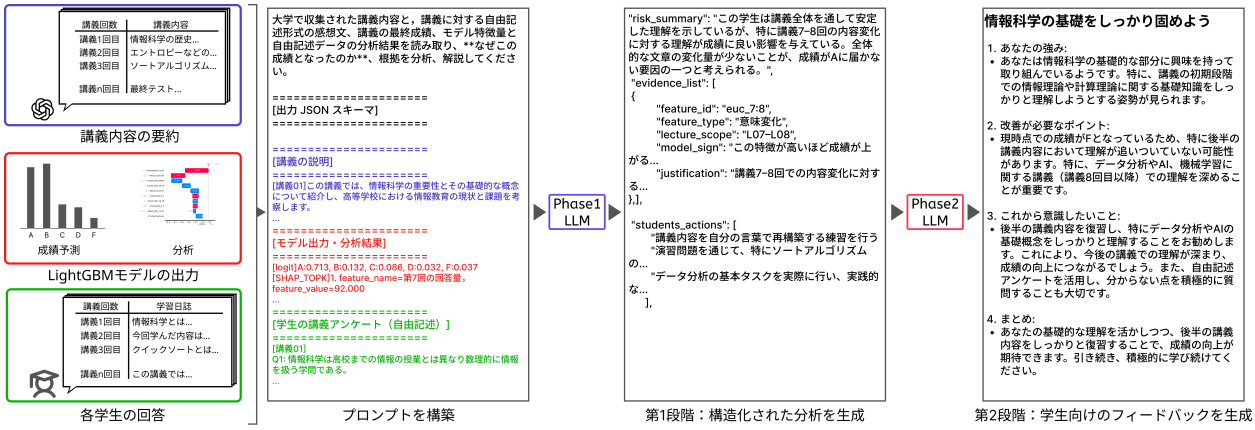
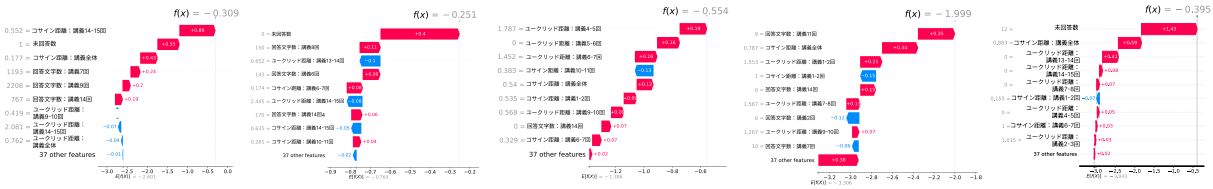


図 2: LLM によるフィードバックシステムの概要



(a) 成績 A (b) 成績 B (c) 成績 C (d) 成績 D (e) 成績 F

図 3: 各成績の学生における判断根拠の可視化

表 1: 成績予測精度の比較

指標	ベースライン	提案手法
Accuracy	45.70	**52.13**
F1-score	33.20	**41.68**

する．図 3 に各成績の学生に対する寄与度分析結果を示す．図 3 より，成績 A，B の学生に対してモデルは講義 14，15 回間のコサイン距離や未回答数，全体的なコサイン距離の平均に注目していることがわかる．この結果から，成績が高い学生ほど，講義最終時点の回答内容の変化や，講義全体を通して一貫性のある回答がされているかといった要素に注目していることがわかる．一方で，成績 C，D の学生に対して，モデルは講義 4〜7 回目のユークリッド距離や，講義 11 回時点の回答など，中盤〜後半の学習日誌に注目している．この結果から，成績が低下する傾向にある学生は成績が高い学生と比較し，講義中間以降の取り組みに問題があると解釈できる．成績 F の学生については，学習日誌の未提出回数と全体のコサイン距離に注目していることから，学生の講義不参加といった傾向に強く注目した予測を行っていることが確認できる．以上より，成績予測モデルは各成績の学生が示す学習日誌の変化から，講義全体における取り組みの継続性に注目して予測を行なっていると考えられる．

5.3. LLM による学生向けフィードバックの評価

LLM によって生成されるフィードバック内容を定性的に評価する．図 4 に，成績 C の学生に対するフィードバックの生成結果を示す．図 4 より，予測時の判断根拠をもとに，初期の講義内容である符号化やエントロピーなどの単元に対する取り組みが活発であるといった強みの分析と，後半の講義内容であるデータ分析や AI の単元における理解不足といった改善点を明確に提示している．この傾向は図 3 に示す成績 C の学生の寄与度分析と一致しており，モデルの分析内容をフィードバックに適用できていることを示唆している．さらに，「非構造化データの処理方法について理解を深めるべきである」といった具体的な改善点が言及されていることから，各学生に対して講義理解を促す具体的なフィードバックを提示できていることが示唆される．

情報科学の基礎をしっかり固めよう

1. あなたの強み:
• あなたは情報科学の基礎的な部分において、特に初期の講義での理解がしっかりしています。情報の符号化やエントロピーの概念に関する基礎知識がしっかりしていることが、講義の進行に役立っています。
2. 改善が必要なポイント:
• 講義の後半、特にアルゴリズムやデータ分析、AIの内容に関する理解がやや不足している可能性があります。これらの分野の具体的な手法や計算量の違いを復習することで、より高い成績を目指せるでしょう。
3. これから意識したいこと:
• アルゴリズムやデータ分析、AIの応用について、実践的な演習を通じて知識を定着させましょう。また、後半の講義で扱う非構造データの処理方法やパターン認識の基本概念についても、理解を深めることをお勧めします。
4. まとめ:
• あなたの基礎的な理解を活かしつつ、特に後半の講義内容を補強することで、さらに高い成績を目指せるでしょう。引き続き頑張ってください。

図 4: LLM によるフィードバック結果

6. おわりに

本研究では，学生の理解度や解釈を読み取りやすい自由記述形式の学習日誌に注目し，学生の成績予測を活用するフィードバックシステムを開発した．各講義における学習日誌の変化が成績と強く相関していることが示されたため，これを決定木モデルの入力に利用することで，Accuracy が 6.43pt，F1-score が 8.48pt 向上することを確認した．さらに，成績予測モデルの SHAP 分析において，講義中間時点における学習日誌の内容変化に対して寄与度が高いと示された．この結果を用いた多段階 LLM の構築により，学生に対して具体的なフィードバックを提示可能となった．今後は，フィードバックで示す成績改善案のさらなる具体化と，実環境における有効性の調査を行う予定である．

参考文献

[1] Yağcı, “Educational data mining: prediction of students’ academic performance using machine learning algorithms”, Smart Learn. Environ., 2022.

[2] Gratiano *et al.*, “Can a five minute, three question survey foretell first-year engineering student performance and retention?”, ASEE, 2016.

[3] Mikolov *et al.*, “Distributed Representations of Words and Phrases and their Compositionality”, NeurIPS, 2013.

研究業績

[1] Koike *et al.*, “An Investigation of the Relationship between Open-Ended Questionnaires and Lectures”, LAK, 2025.
(他 学会発表 2 件)

学習行動ログデータを用いた成績予測における説明性向上に関する研究

TP24009 舘良太　　指導教授：山下 隆義

1. はじめに

教育現場における学習管理システム（LMS）やデジタル教材配信システムの普及に伴い，教育・学習活動に関するデータを収集・解析し，教育改善に活用する Learning Analytics の研究が活発に行われている．これにより，学習者の進捗状況，学習傾向などを把握することで，各学習者の学修状況に合わせた支援を提供することが可能となる．こうした背景から，操作ログのような大規模なデータを扱い，学習者ごとの特徴を捉える手法として，教育分野における機械学習モデルの活用が期待されており，早期退学者の検出や学習行動の改善を目的とした成績予測の研究が多く行われている．

宮崎らは，自然言語処理を用いて操作間の前後関係や時間間隔を保持した分散表現を生成する手法を提案している．しかし，学生特徴の生成には学習量を直感的に表現できる Bag-of-Words（BoW）によるヒストグラム特徴を用いており，操作頻度分布の偏りを十分に考慮できないため，希少だが重要な操作が特徴として反映されにくいという問題がある．そこで本研究では，BoW と同様に頻度情報に基づきつつ，文書長正規化と逆文書頻度を考慮可能な BM25 と，クラス間差の大きさを定量化できる効果量を組み合わせた重み付けにより，操作頻度分布の偏りとクラス間差の双方を考慮した成績予測手法を提案する．

2. 先行研究

宮崎らは，学習操作間の前後関係や時間間隔といった情報を保持した分散表現の生成手法である E2Vec[1] を提案している．E2Vec の概要を図 1 に示す．E2Vec は，前処理，埋め込み，集約の 3 つのモジュールで構成される．

前処理では，操作ログを自然言語処理の “文字”，“単語”，“文章” に対応付け，文字列表現に変換する．文字は 1 文字 1 操作に対応し，操作名を 1 文字に変換して表現する．各操作名と文字の対応表を表 1 に示す．単語は最大 1 分間かつ 15 文字以内で構成され，隣接する文字間に操作の時間間隔に対応した文字を挿入して表現される．文章は複数の単語から構成され，操作間隔が 5 分以上となるまでを 1 つの文章とする．

埋め込みでは，前処理により文字列表現に変換した操作ログを用いて fastText を学習し，学習済み fastText により各単語を 100 次元ベクトルに埋め込む．さらに，単語ベクトルを平均化して文章埋め込みベクトルを生成する．

集約では，文章埋め込みベクトルに対して k-means によるクラスタリングを行い，codebook を生成する．生成した codebook を用いて BoW アプローチにより学生の特徴ベクトルを生成する．

E2Vec は BoW アプローチに基づく特徴量生成を行うため，操作頻度分布の不均衡を考慮できず，モデルが “希少だが重要な操作” を捉えることが困難となる．

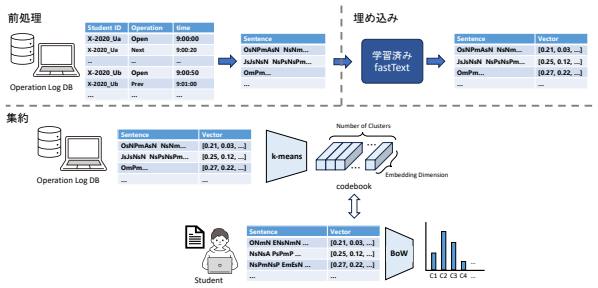


図 1：E2Vec の概要

3. 提案手法

本研究では，効果量に基づく重み付けを適用した，E2Vec ベースの BM25 特徴による成績予測手法を提案する．本手法の概要を図 2 に示す．本手法は，codebook 生成モジュール，特徴量生成モジュール，成績予測モジュールの 3 つで構成される．

表 1：操作名と文字の対応表

操作名	説明	文字
NEXT	次のページへ移動	N
PREV	前のページへ移動	P
OPEN	教材を開く	O
ADD MARKER	マーカーを引く	A
CLOSE	教材を閉じる	C
PAGE JUMP	指定したページへ移動	J
GET IT	ページ内容について理解した	G
OTHERS	低頻出の操作	E
short interval	1 から 10 秒の時間間隔	s
medium interval	10 から 300 秒の時間間隔	m
long interval	300 秒以上の時間間隔	l

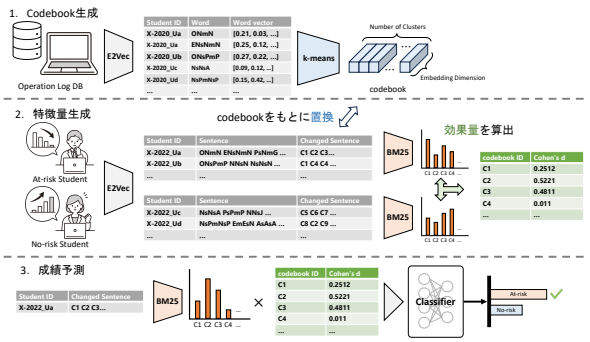


図 2：提案手法の概要

3.1. Codebook 生成モジュール

操作ログデータに E2Vec の前処理を適用し，操作単語の分散表現を獲得する．得られた操作単語ベクトルを k-means++でクラスタリングし，codebook を作成する．各操作単語を最も近いクラスタに置換し，学生ごとのクラスタ列を生成する．

3.2. 特徴量生成モジュール

クラスタ列に BM25[2] を適用し，操作の頻度と希少度を考慮した特徴量を算出する．BM25 とは，文書集合における，ある単語の重要度を測るための尺度である．BM25 特徴 $BM25_{i,j}$ は式 (1) で定義される．

$$\mathrm{BM25}_{i,j} = \mathrm{IDF}_j \cdot \frac{f_{i,j}\,(k_1+1)}{f_{i,j} + k_1\left(1 - b + b\frac{L_i}{L}\right)} \tag{1}$$

$$\mathrm{IDF}_j = \log \frac{N - n_j + 0.5}{n_j + 0.5} \tag{2}$$

ここで，N は学生数，n_j はクラスタ j を含む学生数，$f_{i,j}$ は学生 i のクラスタ j の出現回数，L_i は学生 i のクラスタ総出現回数，k_1 と b は BM25 のハイパーパラメータである．

BM25 特徴を用いてクラス間の行動差を定量化するために，式 (3) のように効果量 [3] を算出する．

$$d_j = \frac{\bar{x}_{\text{at-risk},\,j} - \bar{x}_{\text{no-risk},\,j}}{s_p} \tag{3}$$

ここで，$\bar{x}_{\mathrm{at-risk},j}$ は低成績クラスにおけるクラスタ j の BM25 平均値，$\bar{x}_{\mathrm{no-risk},j}$ は高成績クラスにおけるクラスタ j の BM25 平均値，s_p はプールされた標準偏差である．特徴量 $x_{i,j}$ は式 (4) で求める．

$$x_{i,j} = \mathrm{BM25}_{i,j} \times |d_j| \tag{4}$$

3.3. 成績予測モジュール

式 (4) により算出された特徴量 $x_{i,j}$ をクラスタごとに集約することで，各学生を K 次元の特徴量ベクトルとして表現する．この特徴量ベクトルを機械学習モデルに入力し，At-risk と No-risk の二値分類を行う．本研究では，分類モデルとして Random Forest を用いる．

4. 評価実験

本実験では，定量的評価として，提案手法を用いて成績予測を行い，E2Vec と成績予測精度を比較する．定性的評価として，BM25 値と効果量の分析を行い，各成績において重要な操作を調査する．また，SHAP を用いて成績予測に寄与する特徴量の分析を行う．

4.1. 実験条件

本実験では，九州大学で収集された LMS の操作ログデータを使用する．A，D はコースの種類を表し，2020 年，2021 年，2022 年はコースが開講された年を表す．A-2020 と D-2020 は，fastText の学習および codebook の生成にのみ使用し，A-2021，A-2022，D-2021，D-2022 の 4 コースを成績予測に用いる．1 コースを訓練，別の 1 コースを評価とした組合せを全通り実施し，計 12 通りの実験を行う．成績は A，B，C，D，F の 5 段階評価であり，F は単位不合格を意味する．本研究では，A，B を No-risk，C，D，F を At-risk として扱う．分類モデルには Random Forest Classifier を使用し，Grid Search によりハイパーパラメータ探索を行う．評価指標には F1-Score を用いる．

4.2. 定量的評価

表 2 に，4 コースに対する E2Vec と提案手法の分類精度を示す．表 2 より，評価データにコース A のデータを用いた場合に最大 0.41 pt の精度向上が確認できた．これより，訓練データと評価データで異なるコースを用いた場合でも精度が向上し，汎化性能があることが確認できる．

表 2：E2Vec と提案手法による分類精度比較

train	test	E2Vec	Ours
A-2021	A-2022	0.72	**0.74**
	D-2021	**0.60**	**0.60**
	D-2022	0.53	**0.55**
A-2022	A-2021	0.71	**0.77**
	D-2021	**0.67**	0.58
	D-2022	0.51	**0.52**
D-2021	A-2021	0.53	**0.77**
	A-2022	0.24	**0.65**
	D-2022	**0.64**	0.48
D-2022	A-2021	0.59	**0.79**
	A-2022	0.38	**0.72**
	D-2021	**0.85**	0.67

4.3. 定性的評価

定量的評価において最も高い予測精度であった D-2022 を対象として，BM25 値，効果量，および SHAP に基づき，各成績クラスの特徴的な操作パターンと，クラス間を識別するために重要な操作パターン，成績の向上・低下に寄与する操作パターンを分析する．表 3 に各クラスで BM25 平均値が高い操作パターン上位 5 件を示す．表 3 より，At-risk クラスは操作間隔が短い“ s ”を含む“ Next ”の連続操作パターンであること，No-risk クラスは操作間隔が短い“ m ”を含む“ Next ”の連続操作や“ Page Jump ”や“ Add Marker ”を含む操作パターンであることが確認できる．このことから，ページ移動効率や教材内容を理解しようとする能動的な学習行動が成績に関係していると考える．

表 4 に 各クラスで効果量が高い操作パターン上位 5 件を示す．表 4 より，At-risk を特徴づける操作パターンは，操作間隔が短い ”s ”を多く含む“ Next ”と“ Prev ”で構成された操作パターンであることが確認でき，No-risk を特徴づける操作パターンは，操作間隔が長い ”m ”と“ l ”を含む“ Next ”と“ Prev ”で構成された操作パターンであることや，”Add Marker ”を含む操作パターンであることが確認できる．このことから，ページごとの滞在時間や教材内容を整理する行動がクラス間の差であると考える．

表 3：D-2022 における BM25 平均値が高い操作

At-risk Operation	No-risk Operation
CsCsCm	NNNNsNsCl
GNmGNl	JsNJJPJJNJNJPN
NNsNNsNNNsNsNs	NmNmNmNm
NsNsNsNsNsNsNs	NmNsNsNmNsNs
GsNmNsGmNsGm	AmAsAm

表 4：D-2022 における効果量が高い操作

At-risk Operation	No-risk Operation
NsPmNmNm	PPPsPPPPPNNNNN
NsNsNsNl	NsPPNsNPsNmNNN
NNNNNNNsPsPPs	AsCm
PsPsPNNsNm	NNmCl
PPPsPsPPsPm	NNsNPsPl

さらに，予測結果の説明を目的として SHAP により特徴量の寄与度を分析した．図 3 に SHAP 値の分布を示す．その結果，操作パターン “OsJsNPPPPNNNNN ”，“AsCm”，“PPPPPPPPPsPmPP”，“NsNsNmOmNNNNNN” は特徴量値が高いほど負の寄与を示す傾向が見られ，一方で “NsNm”，“NNNNNsNNNNNNNN” は特徴量値が高いほど正の寄与を示す傾向が見られた．これより，成績に良い影響を与えるクラスタは “Page Jump” や “Add Marker” を含むパターンであり，成績に悪い影響を与えるクラスタは “Next” や “s” で構成されたパターンであることが分かる．以上より，希少な操作を含む学習行動が成績向上に寄与する一方で，短時間の連続ページ遷移が成績低下に寄与する可能性が示唆された．

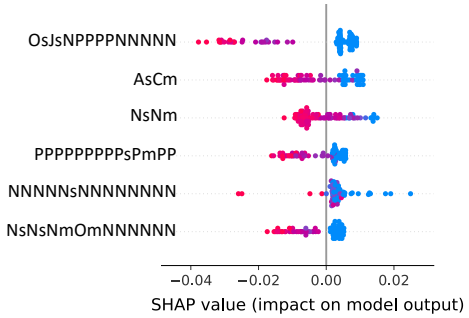


図 3：SHAP による寄与度の可視化結果

5. おわりに

本研究では，効果量に基づく重み付けを適用した E2Vec ベース BM25 特徴による成績予測を行った．結果から，従来手法と比べ，訓練と評価で異なるデータを使用した場合でも精度向上が確認できた．また，BM25 値と効果量，および SHAP による分析により，操作間隔の違いや “Page Jump”，“Add Marker” といった希少操作の有無がクラス間の差として現れており，教材の熟読やページ移動効率が成績に関係していると考えられる．今後は，重み付けを特徴量ではなく損失関数に適用した成績予測モデルの構築を行う予定である．

参考文献

[1] Y. Miyazaki *et al.*, “E2Vec: Feature Embedding with Temporal Information for Analyzing Student Actions in E-Book Systems”, EDM, 2024.

[2] S. E. Robertson *et al.*, “Okapi at TREC-3”, TREC-3, 1995.

[3] J. Cohen, “Statistical Power Analysis for the Behavioral Sciences (2nd ed.)”, Lawrence Erlbaum Associates, 1988.

研究業績

[1] R.Tachi *et al.*, “Grade Prediction Using fastText Features Weighted Through Differential Pattern Mining”, LAK, 2025.

(他 2 件)

自己教師ありオンライン継続学習における収束速度と勾配相関の改善に関する研究

TP24001 今井 孝洋　　指導教授：山下 隆義

1. はじめに

深層学習モデルを学習する場合，データを事前に収集・蓄積して学習するオフライン学習と，蓄積せずに逐次データを入力して学習するオンライン学習の 2 つのアプローチがある．オフライン学習は，蓄積したデータセットを繰り返し用いて学習するため高い精度を達成しやすい一方で，データの蓄積に伴うストレージコストが課題となる．オンライン学習は，収集したデータを即座に学習し，学習後はそのデータを破棄するため，ストレージコストを大幅に削減できる．しかし，収集したデータに対してリアルタイムでラベル付けして学習するのは困難である．

これに対して，自己教師ありオンライン継続学習は，逐次入力するデータに対して自己教師あり学習を行うことで，ラベルを付与するコストを低減できる．しかし，従来の自己教師ありオンライン継続学習手法は，2 つの課題がある．(i) パラメータ更新時の勾配に相関が発生し，モデルが特定のデータに過度に適合することで汎化性能が低下する．(ii) 自己教師あり学習の収束が遅く，収集したデータを破棄するまでの短い時間で十分に学習することが困難である．

そこで本研究では，コサイン類似度を使用した学習データの選択によって勾配の相関を抑制し，マルチクロップ対照損失によって自己教師あり学習の収束速度を改善する自己教師ありオンライン継続学習手法を提案する．実験により，提案手法が従来手法と比較して分類精度を改善することを示す．

2. 自己教師ありオンライン継続学習

自己教師ありオンライン継続学習 (Self-Supervised Online Continual Learning: SSOCL) は，逐次入力されるラベルのないデータであるデータストリームを用いて継続的に学習するアプローチである．データストリームは，連続するデータ間に強い相関を持つと共に，時間の経過に伴いデータ分布が変化する非定常性という 2 つの特性がある．このような特性を持つデータストリームで学習する SSOCL には 2 つの課題がある．

課題 1：パラメータ更新時の勾配の相関．図 1 に t 回目のイタレーションと $t+1$ 回目のイタレーションにおけるパラメータ更新時の勾配のコサイン類似度を示す．従来の深層学習モデルは，サンプル間に相関がないデータで学習を行うことを仮定しており，勾配の類似度は 0 に近い値となる．一方で，サンプル間に相関のあるデータで学習を行うと，勾配の類似度が 1.0 に近い値になる．勾配の類似度が高いと，モデルのパラメータを特定のデータ分布に過度に適合する方向へ更新するため，汎化性能が低下する．

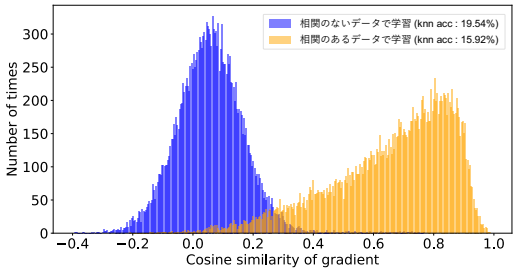


図 1: パラメータ更新時の勾配の類似度

課題 2：自己教師あり学習の収束の遅さ．図 2 に教師あり学習と自己教師あり学習の収束速度を示す．図 2 より，自己教師あり学習は，教師あり学習と比較すると学習収束が遅いことがわかる．これは，データ分布が時間と共に変化する実世界において，自己教師あり学習の学習が不足する可能性を示している．

3. 提案手法

本研究では，SSOCL における 2 つの課題に対処する手法を提案する．提案手法は，勾配の相関に対処するためコサイン類似度を用いたデータ選択を行い，自己教師あり学習の学習収束の遅さに対処するため Multi-Crop Contrastive Loss (MCC Loss) を導入する．

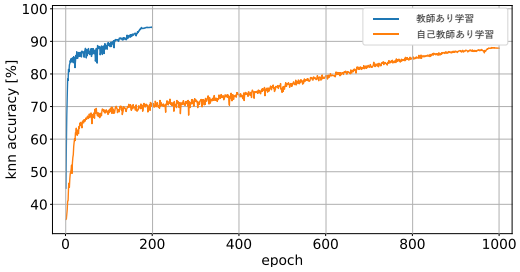


図 2: 教師あり学習と自己教師あり学習の収束速度

3.1. 学習プロセス

提案手法の学習プロセスを図 3 に示す．図 3 に示すように，提案手法は，データストリームで観測されたサンプルを固定サイズのバッファに追加する．その後，バッファからランダムにサンプリングしたデータで K 個のミニバッチを作成し，学習に使用する．このとき，自己教師あり学習の収束の遅さに対処するため，損失関数に MCC Loss を導入する．従来の Contrastive Loss は，1 枚の画像に異なる 2 種類のデータ拡張を加えて得た 2 つのクロップに対し，それらの特徴量を 1 対 1 で近づけ，異なる画像から得たクロップは，遠ざけるように学習を行う．これに対して，提案手法で導入する MCC Loss は，3 種類以上の異なるデータ拡張を適用し，得られた各クロップの特徴量をそれらの平均特徴量へと同時に近づける．クロップ数を増加させることで，1 回のパラメータ更新においてより多くの情報を効率的に学習できるため，自己教師あり学習の収束の遅さを改善することが可能である．

K 個のミニバッチで学習後，バッファ内のサンプル数が一定数を超えている場合，コサイン類似度に基づいて多様なサンプルのみをバッファに保持し，冗長なサンプルを削除する．これにより，提案手法は，データストリームが持つ相関のある冗長なデータでの学習を防止し，勾配の相関に対処する．その後，データストリームで観測したサンプルに対して同様の処理を繰り返す．以下では，MCC Loss とコサイン類似度によるデータ選択について詳細に説明する．

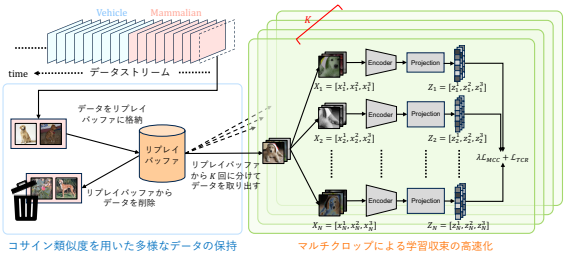


図 3: 提案手法の概要

3.2. Multi-Crop Contrastive Loss

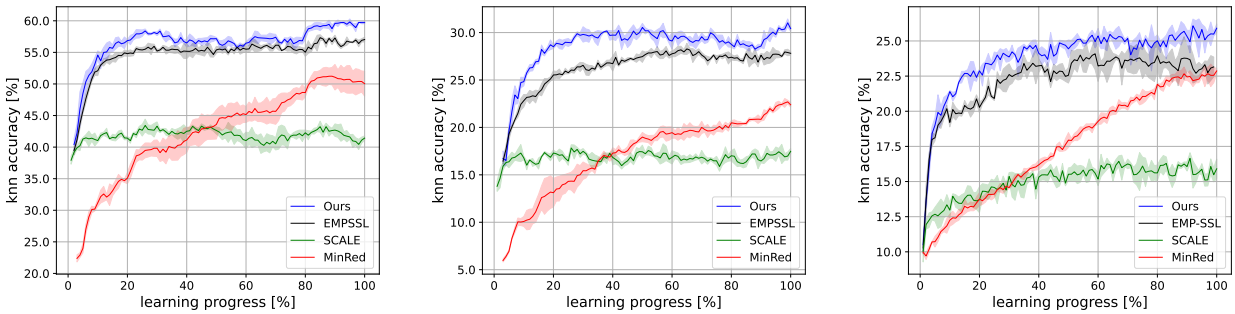
自己教師あり学習の収束速度は，クロップ数の増加によって高速化可能であることが知られている [3]．提案手法は，従来の 2 クロップのみを対象とした Contrastive Loss を 3 クロップ以上に拡張した Multi-Crop Contrastive Loss (MCC Loss) を導入する．MCC Loss を式 (1) に示す．

$$\mathcal{L}_{MCC} = \frac{1}{Nb}\sum_{i=1}^{N}\sum_{j=1}^{b}\left(-\log\frac{\exp(\bar{\mathbf{z}}^j\cdot\mathbf{z}_i^j/\tau)}{\sum_{k=1}^{N}\sum_{l=1}^{b}\exp(\bar{\mathbf{z}}^j\cdot\mathbf{z}_k^l/\tau)}\right) \tag{1}$$

ここで，N はクロップ数，b はバッチサイズ，$\mathbf{z}$ は各サンプルの特徴量，τ は温度パラメータ，$\bar{\mathbf{z}}^j$ はサンプル $\mathbf{x}^j$ の

表 1: 学習終了時の kNN 分類精度 [%]

	CIFAR10			CIFAR100			ImageNet100		
	Seq	Seq-bl	Seq-im	Seq	Seq-bl	Seq-im	Seq	Seq-bl	Seq-im
MinRed[2]	50.04	51.18	46.41	22.38	23.20	21.26	22.87	22.71	20.46
SCALE[1]	41.41	40.85	41.31	17.49	16.93	17.03	15.46	15.41	15.77
EMP-SSL[3]	57.02	57.32	57.31	27.81	28.40	27.90	22.79	22.01	22.99
Ours	**59.71**	**59.96**	**58.67**	**30.41**	**30.32**	**30.00**	**25.81**	**25.64**	**25.24**



(a) Seq-CIFAR10 (b) Seq-CIFAR100 (c) Seq-ImageNet100

図 4: 学習過程における kNN 分類精度 [%]

平均特徴量を示し，$\bar{\mathbf{z}}^j$ は式 (2) で求める．

$$\bar{\mathbf{z}}^j = \frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i^j. \tag{2}$$

MCC Loss は，1 枚の画像に対して N 種類の異なるデータ拡張を加え，各サンプルの平均特徴 $\bar{\mathbf{z}}^j$ を計算する．そして，各データの特徴量 $\mathbf{z}_i^j$ をその平均特徴量 $\bar{\mathbf{z}}^j$ に同時に近づける．また，異なるデータの平均特徴 $\bar{\mathbf{z}}^j$ から各データの特徴量 $\mathbf{z}^l$ を遠ざけるように学習する．

3.3. コサイン類似度による多様なデータの選択

コサイン類似度を用いたデータ選択の目的は，データストリームから多様なデータを選択して学習に利用することである．コサイン類似度の計算は，バッファ内サンプル $\mathbf{x}_i$ と同時に保存した代表的な特徴量 $\bar{\mathbf{z}}_i^*$ を用いて計算する．提案手法のデータ選択は，式 (3) で定式化できる．

$$\mathbf{x}_i^* = \arg\min_{\mathbf{x}_i\in\mathcal{M}}\min_{\mathbf{x}_j\in\mathcal{M}}\mathrm{Sim}\left(\bar{\mathbf{z}}_i^*, \bar{\mathbf{z}}_j^*\right) \tag{3}$$

$$\bar{\mathbf{z}}_i^* \leftarrow \alpha\bar{\mathbf{z}}_i^* + (1-\alpha)\,\bar{\mathbf{z}}_i \tag{4}$$

ここで，$\mathcal{M}$ はバッファ，$\mathbf{x}_i^*$ はバッファに保存するデータ，$\bar{\mathbf{z}}_i^*$ は平均特徴 $\bar{\mathbf{z}}_i$ の指数移動平均であり，$\mathbf{x}_i^*$ と $\bar{\mathbf{z}}_i^*$ をバッファに保存する．リプレイバッファ内で類似度が高いデータを削除し，類似度が低く多様なデータを優先してリプレイバッファに保持する．

4. 評価実験

データストリームで学習した各手法のクラス分類における分類精度を評価する．

4.1. 実験条件

評価には CIFAR10/100，ImageNet100 の 3 つのデータセットを用いて，データストリームの構築を行う．各データセットを従来研究 [1] に従って，Seq，Seq-bl，Seq-im データストリームを構築する．Seq は，クラスごとのデータ数を統一し，データ分布が一定のタイミングで変化する模擬的なデータストリームである．また，より現実的なデータストリームで評価を行うため，Seq-bl はデータ分布の変化境界を曖昧にし，Seq-im はクラス毎のデータ数を不均衡にする．これにより，Seq-bl は現実世界で発生する環境の滑らかな変化を，Seq-im はクラス毎の出現頻度の偏りという現実的な不均一性を再現する．

4.2. 実験結果

各手法の学習終了時の分類精度を表 1 に示す．表 1 より，提案手法は，CIFAR10 で最大 19.11pt，CIFAR100 で最大 13.39pt，ImageNet100 で最大 10.35pt の精度向上を確認できる．また，MinRed [2] は，Seq-im での精度低下が確認できる．これは，クラス毎の出現頻度を不均一にすることで，観測回数の少ないクラスに対して学習が収束しなかったためだと考えられる．一方で，提案手法は，Seq-im の精度は他のデータストリームと同程度であり，これはより現実的なデータストリームにおいても学習が可能であることを示している．

次に，学習過程における分類精度の比較を行う．各手法の学習過程における分類精度の推移を図 4 に示す．図 4 より，提案手法の kNN 分類精度は，学習進捗が 20% の時点において，Seq-CIFAR10 で約 56.0%，Seq-CIFAR100 で約 28.0%，Seq-ImageNet100 で約 23.0% である．これは，他手法の学習終了時点での精度と同等かそれ以上である．これは，提案手法の Multi-Crop Contrastive Loss によって学習収束を高速化したことに起因すると考えられる．

5. おわりに

本研究では，自己教師ありオンライン継続学習における勾配の相関による性能劣化と学習収束の遅さに対処する手法を提案した．提案手法は，コサイン類似度によるデータ選択によって，バッファ内に多様なデータを保持し学習に用いることで勾配の相関に対処し，MCC Loss を導入することで自己教師あり学習の収束の遅さに対処した．実験結果より，複数のデータセットにおいて提案手法は，従来手法よりも高い分類精度を達成することを確認した．今後は，ImageNet21K などより大規模なデータセットを使用して，より実世界に近いデータストリームを構築し，その有効性を評価する予定である．

参考文献

[1] Yu *et al.*, “SCALE: Online Self-Supervised Lifelong Learning Without Prior Knowledge”, CVPRW, 2023.

[2] Purushwalkam *et al.*, “The challenges of continuous self-supervised learning”, ECCV, 2022.

[3] Tong *et al.*, “Emp-ssl: Towards self-supervised learning in one training epoch”, arXiv, 2023.

研究業績

[1] Imai *et al.*, “Faster convergence and Uncorrelated gradients in Self-Supervised Online Continual Learning”, ACCV, 2024. (他 3 件)

物体検出モデルの判断根拠可視化における解釈性向上に関する研究

TP24010　仲井悠真　　指導教授：山下　隆義

1. はじめに

自動運転や医療画像解析において，深層学習による物体検出モデルは，高い性能だけでなく，高い信頼性が要求される．しかし，深層学習モデルは，その判断根拠がブラックボックスである．このような背景から，物体検出モデルの判断根拠を人間に理解可能な形で示す説明可能な AI（XAI）が注目されている．物体検出に特化した手法として ODAM[1] が提案されている．ODAM は勾配情報に基づいて可視化を行うため，入力画像に対する勾配消失や局所的な勾配ノイズの影響を受けやすいという課題がある．

そこで本研究では，ODAM が抱える勾配依存による課題を軽減するため，入力画像に関する情報を持たないベースライン画像から入力画像に至るまでの過程を考慮できる勾配計算法である Integrated Gradients[2] を導入する．さらに Integrated Gradients における補間画像の生成方法に起因する積分近似誤差および勾配ノイズの問題に着目する．これらを低減するため，勾配変動と空間的変化に基づいてサンプリング位置を適応的に制御する機構を導入した Adaptive IG-ODAM を提案する．

2. Integrated Gradients

XAI において，可視化結果が入力と予測の関係を適切に反映していることが求められる．既存の勾配ベース手法は，勾配消失により重要な特徴を捉えられないという感度の欠如が課題である．この課題に対し，ベースラインから入力までの直線経路に沿って勾配を積分し，各特徴量の寄与度を算出する Integrated Gradients が提案されている．画像タスクでは，ベースラインとして全画素がゼロの画像が用いられることが多く，経路全体の勾配情報を用いることで，単一の入力画像の勾配に依存しない忠実な寄与度推定が可能となる．

一方で，実装上は経路積分を有限個の補間点により近似するため，図 1 に示すように，補間経路を一様にサンプリングした場合には，勾配が急激に変化する区間に十分な補間点が割り当てられず，積分近似誤差や勾配ノイズが生じやすいという課題がある．特に，深層モデルにおいて勾配が非線形に変化する場合，この近似誤差は可視化結果の忠実性に影響を及ぼす可能性がある．

また，Integrated Gradients は主に単一の予測出力を対象とした設定を想定しており，物体検出のように複数のインスタンスや出力を同時に扱うマルチインスタンス環境への直接的な適用には，依然として課題が残されている．

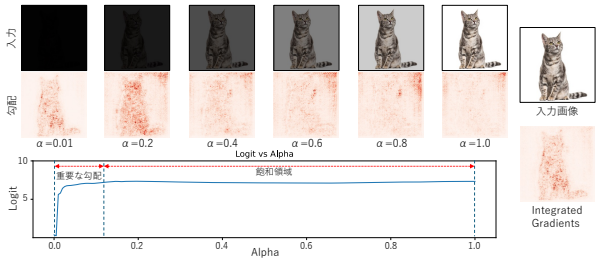


図 1: Integrated Gradients による寄与推定と補間経路上での勾配飽和の例

3. 提案手法: Adaptive IG-ODAM

本研究では，Integrated Gradients を物体検出の判断根拠可視化手法である ODAM に導入した IG-ODAM を提案する．さらに，一様サンプリングに起因する積分近似誤差や勾配ノイズを低減するため，補間経路上の重要区間にサンプリング点を適応的に配置する Adaptive IG-ODAM を提案する．

3.1. ODAM への Integrated Gradients の導入

IG-ODAM は，入力画像とベースライン画像を結ぶ補間経路全体の勾配情報を用いることで，単一の画像における勾配に依存する従来手法に見られる局所的な偏りの影響を低減する．さらに，物体検出特有のマルチインスタンス環境へ適用するために，IoU に基づく位置的類似度とクラススコアの類似度を統合したインスタンスマッチングを導入する．これにより，補間経路上において同一インスタンスを一貫して追跡しながら寄与度推定する．

物体 p に対する予測クラススコア $s^{(p)}(I)$ を寄与度推定の対象とし，ベースライン画像 I' から入力画像 I への補間経路 $I_\alpha = I' + \alpha(I - I')$（$\alpha \in [0,1]$）に沿って経路積分を行う．物体検出では，補間画像ごとに検出結果の数や順序が変化するため，単純な対応付けでは同一インスタンスを追跡できないという問題がある．そこで IG-ODAM では，入力画像 I における物体 p の BBox 座標と予測クラススコアから構成される検出結果 D_t を基準とし，m 番目の補間画像 $X_m = I_{\alpha_m}$ から得られる検出結果集合 $\phi(X_m)$ 内の各検出 $D_{j,m}$ との間で，位置類似度 s_{loc} およびクラススコア類似度 s_{cls} を用いた類似度を定義する．

$$\text{Sim}(D_t, D_{j,m}) = s_{\text{loc}}(D_t, D_{j,m}) \cdot s_{\text{cls}}(D_t, D_{j,m}) \quad (1)$$

各補間画像においては，$\text{Sim}(D_t, D_{j,m})$ が最大となる検出結果を対応インスタンス $\hat{d}_m$ として選択する．この対応付けにより，補間経路全体にわたって同一インスタンスを一貫して追跡しながら，寄与度推定を行うことが可能となる．

特徴マップ A_k に対するチャネル重み $w_k^{(p)}$ は，補間経路上の勾配を積分することで式 (2) のように定義される．

$$w_k^{(p)} = \int_0^1 \frac{\partial s^{(p)}(I_\alpha)}{\partial A_k}\, d\alpha \quad (2)$$

実装上は，補間経路を一様に分割し，有限個の補間点に基づく数値積分によって近似することで，インスタンス固有のヒートマップを生成する．

3.2. Spatial-Guided Adaptive Sampling

Integrated Gradients における一様サンプリングに起因する課題に対して，補間経路上のサンプリング点を動的に再配置する Spatial-Guided Adaptive Sampling を導入した Adaptive IG-ODAM を提案する．本手法は，物体検出モデルの出力が急激に変化する補間区間に重点的にサンプルを配置することで，積分近似誤差および勾配ノイズの低減を目的とする．図 3 に提案手法のモデル図を示す．

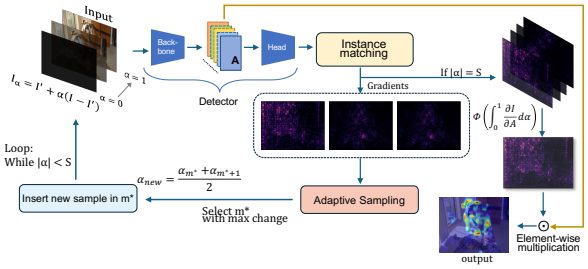


図 3: Adaptive IG-ODAM のモデル構造

Adaptive IG-ODAM では，補間経路上の連続するサンプリング点 α_m と α_{m+1} の間における重要度を評価し，重要度の高い区間を逐次的に細分化する．重要度評価には，勾配の変動量と予測 BBox の空間変動の両方を用いる．

まず，勾配変動 g_m を式 (3) により定義する．

$$g_m = \|G(\alpha_{m+1}) - G(\alpha_m)\|_1 \quad (3)$$

ここで，$G(\alpha_m)$ は補間画像 α_m における対象物体の検出スコアに対する特徴マップの勾配を表す．次に，連続する補間画像間における予測 BBox の空間変動 s_m を，IoU に基づいて式 (4) のように定義する．ここで，$B(\alpha_m)$ は補間画像 α_m に対する予測 BBox を表す．

$$s_m = 1 - \text{IoU}\big(B(\alpha_m), B(\alpha_{m+1})\big) \quad (4)$$

これらを重み係数 λ を用いて統合し，各補間区間の優先度スコア R_m を式 (5) により算出する．

$$R_m = \lambda g_m + (1 - \lambda) s_m \quad (5)$$

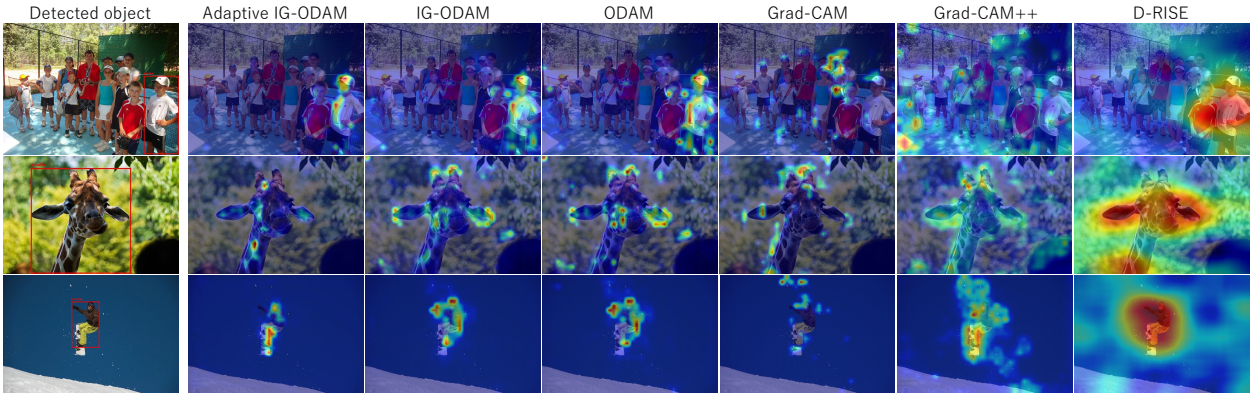


図 2: DETR による物体検出結果に対する判断根拠の可視化結果

優先度スコアの高い区間を逐次的に細分化することで，重要な補間区間にサンプルを集中的に割り当てる．Spatial-Guided Adaptive Sampling により得られた非一様な M 個の補間点に基づき，物体 p に対するチャネル k の重み $w_k^{(p)}$ を台形則に基づき式 (6) のように近似する．ここで，$G_k(\alpha_m)$ は補間画像 α_m におけるチャネル k に対応する勾配を表す．

$$w_k^{(p)} \approx \sum_{m=1}^{M-1} \frac{1}{2} \left[G_k(\alpha_m) + G_k(\alpha_{m+1}) \right] (\alpha_{m+1} - \alpha_m) \quad (6)$$

最後に，得られたチャネル重みを用いて，インスタンス固有のヒートマップ $H^{(p)}$ を生成する．ここで，A_k はチャネル k の特徴マップを表す．

$$H^{(p)} = \mathrm{ReLU}\left(\sum_k w_k^{(p)} \circ A_k \right) \quad (7)$$

4. 評価実験

提案手法の忠実性と空間識別能力を評価するため，比較実験を行う．判断根拠の忠実性は Deletion / Insertion テストの AUC により評価し，空間識別能力は Visual Explanation Accuracy（VEA）と Energy-based Pointing Game（EBPG）を用いて測定する．

物体検出モデルには Backbone に ResNet-50 を用いた DETR を使用し，MS COCO データセット上で Grad-CAM，Grad-CAM++，D-RISE，ODAM と比較する．なお，本実験では，勾配変動と空間変動の寄与を等しく考慮するため，重み係数 λ を 0.5 に設定する．

4.1. Deletion, Insertion

Deletion / Insertion テストは，可視化手法がモデル予測に重要な領域をどの程度正確に特定できるか，忠実度を評価する指標である．Deletion では，ヒートマップに基づき画素を重要度順にランダム値で置換し，予測スコアの低下を測定する．Insertion では，ベースライン画像に重要画素を順次追加し，予測スコアの上昇を測定する．本実験では，両テストを 100 ステップで実施し，信頼度推移から AUC を算出する．

実験結果を表 1 に示す．IG-ODAM は，従来の物体検出向け可視化手法である ODAM と比較して，Deletion スコアを 55.25 から 51.48 に低減し，Insertion スコアを 15.37 から 18.14 に向上させることで，忠実度の向上を示した．さらに，Adaptive IG-ODAM は，Deletion スコア 46.48，Insertion スコア 25.88 と最良の性能を示した．これは，Spatial-Guided Adaptive Sampling により経路積分に使用する補間画像が最適化され，積分近似誤差が低減されたためと考えられる．

4.2. VEA, EBPG

VEA は物体形状との一貫性を，EBPG は物体領域への局在精度をそれぞれ評価する指標である．実験結果を表 2 に示す．Adaptive IG-ODAM は，IG-ODAM と比較して，VEA を +0.0528 ポイント，EBPG を +0.1133 ポイント向上させ，空間的一貫性および局在精度の双方で性能向上を示した．これは，Spatial-Guided Adaptive Sampling により，補間経路上でモデル出力が大きく変化する区間に重点的なサンプリングが行われたためである．

表 1: 各手法の Deletion/Insertion 評価結果

Method	Deletion↓	Insertion↑
Grad-CAM	72.82	11.23
Grad-CAM++	72.60	11.04
D-RISE	57.57	13.23
ODAM	55.25	15.37
IG-ODAM	51.48	18.14
Adaptive IG-ODAM	**46.48**	**25.88**

表 2: VEA と EBPG の評価結果

Method	VEA ↑	EBPG ↑
Adaptive IG-ODAM	**0.1492**	**0.3934**
IG-ODAM	0.0964	0.2801

4.3. 定性的評価

図 2 に，各手法による可視化結果を示す．IG-ODAM は，従来手法と比較してノイズが低減され，物体境界をより正確に捉えている．一方，Grad-CAM および Grad-CAM++ では背景や他物体への注目が生じやすく，ODAM ではマルチインスタンス環境において注目領域の分散が見られる．さらに，Adaptive IG-ODAM はインスタンス固有の注目領域を明確に分離することで，従来手法で見られた注目の分散を最も効果的に抑制していることがわかる．

5. おわりに

本研究では，物体検出における説明可能性の向上を目的として，IG-ODAM および Adaptive IG-ODAM を提案した．IG-ODAM は，Integrated Gradients を ODAM に統合し，補間経路全体の勾配情報とインスタンスマッチングにより，マルチインスタンス環境におけるインスタンス単位の判断根拠可視化を実現した．さらに Adaptive IG-ODAM では，勾配変動と予測 BBox の空間変動に基づく Spatial-Guided Adaptive Sampling を導入することで，積分近似誤差およびノイズの削減を達成した．評価実験の結果，忠実度および空間識別能力の両面で既存手法を上回る性能を確認した．

今後は，得られたヒートマップを知識蒸留における教師信号として活用する手法を検討する．

参考文献

[1] Chenyang ZHAO, Antoni B. Chan, “ODAM: Gradient-based Instance-Specific Visual Explanations for Object Detection”, ICLR. 2023.

[2] Sundararajan, *et al.*, “Axiomatic Attribution for Deep Networks.” arXiv. 2017.

研究業績

[1] Nakai, *et al.*, “IG-ODAM: Instance-Aware Visual Explanations for Object Detection with Integrated Gradients,” MVA. 2025.

（他 3 件）

プロトタイプ法における人の知見の組み込みによる精度向上に関する研究

TP24003 落合祐馬　　指導教授：山下隆義

1. 概要

プロトタイプ学習は，認識に有効な画像中の局所領域をプロトタイプとして選択し，特徴ベクトルを学習する手法である．推論時は入力画像とプロトタイプとの認識クラスを特定するとともに認識の判断根拠となる局所領域を同時に得ることができる．しかし，データ駆動で獲得したプロトタイプは，背景やロゴマークなどの不適切な領域に注目してしまう問題がある．この問題は医療や自動運転など判断根拠が重要な分野では深刻である．

本研究では，モデルの信頼性向上を目的として，人の知見を認識モデルに組み込む Human-in-the-Loop（HITL）のアプローチを取り入れた手法を提案する．具体的には，ProtoPFormer[1] をベースに，人間の注目領域や対象物体の欠損箇所を人の知見として学習時に損失関数として Human Knowledge Loss（HKLoss）を導入する．これにより，適切なプロトタイプの選択を誘導し，不適切な領域への注目を抑制する．

2. ProtoPFormer

ProtoPFormer は，Vision Transformer (ViT) をベースにしたプロトタイプ学習を行うモデルである．画像全体に注目する Global Branch と，局所領域に注目する Local Branch から構成される．推論時は Local Branch では ViT の注目領域をもとに FP Mask を作成し，前景にある対象物のイメージトークンのみを抽出する．その後，Global Branch ではクラストークンとプロトタイプ Local Branch では抽出されたイメージトークン とプロトタイプとの類似度を計算する．次にプロトタイプごとに出力された最大の類似度を全結合層に入力し，各 Branch ごとにクラス確率を出力する．最後にそれぞれのクラス確率の平均をモデルが出力するクラス確率とする．これらの推論時の処理に加えて，学習時にはモデル全体を学習するための Cross Entropy Loss と，Local Branch 内のプロトタイプを学習するために式 (1) と式 (2) の損失関数を用いて学習する．

$$\mathcal{L}^{\mu}_{PPC} = \frac{1}{(m_l^c)^2} \sum_{i \neq j} \max\left(t_\mu - \left\|\hat{\mu}_i^c - \hat{\mu}_j^c\right\|^2, 0\right) \quad (1)$$

$$\mathcal{L}^{\sigma}_{PPC} = tr\left(\max\left(0, \sum - t_\sigma\right)\right) \quad (2)$$

$\mathcal{L}^{\mu}_{PPC}$ は同じクラスのプロトタイプが類似しているほど大きな損失を与える．ここで，m_l^c は各クラスで使用するプロトタイプの数であり，μ はプロトタイプである．t_μ は閾値である．$\mathcal{L}^{\sigma}_{PPC}$ はプロトタイプが注目する領域を小さくする損失である．$\sum$ はプロトタイプの共分散行列の対角成分の平均であり，t_σ は閾値である．

これにより，従来の CNN をベースとした手法と比較して高い認識精度を持つ一方で，背景やロゴマークなどへ過剰に注目してしまう問題が指摘されている．

3. 提案手法

本研究では，ProtoPFormer の局所領域に注目をする Local Branch 内のプロトタイプに対し，人の知見を損失として導入する．損失を導入する概要図を図 1 に示す．

3.1. プロトタイプの選択的誘導

全てのプロトタイプを一律に「人の知見」に近づけると，全てのプロトタイプが同じ領域に注目し，多様性がなくなる．これを防ぐため，本手法では画像ごとに，人の知見に最も近い注目領域を持つプロトタイプの組み合わせを動的に探索する．探索の方法として，入力画像のラベルに基づき，対応するクラスに属するプロトタイプを対象とし，対象となるプロトタイプから得られる全ての組み合わせから人の知見との類似度が最大となった組み合わせに含まれるプロトタイプに対してのみ損失を適用する．これにより，多様性を維持しつつ，モデルの出力を人の知見へと整合させる．

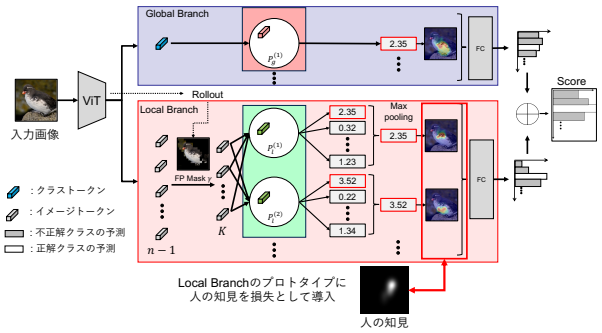


図 1：提案手法の概要図

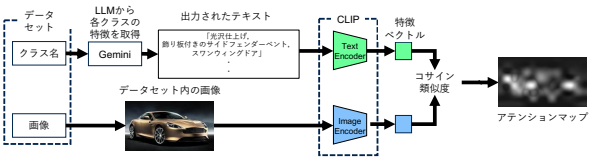


図 2：擬似知見生成手順

3.2. Human Knowledge Loss (HKLoss)

HKLoss を式（3）に示す．選択されたプロトタイプの注目領域 y_{pred} と，人の知見 y_{true} との間の二乗誤差を損失関数とする．

$$L_{HK} = \left(\frac{1}{m}\sum_{i=1}^{m} y^i_{pred} - y_{true}\right)^2 \quad (3)$$

ここで m は選択されたプロトタイプの数である．この損失により，プロトタイプの活性化領域が，人間が重要と考える領域（鳥の頭部，車のライト，製品の欠陥部など）に収束する．

3.3. 擬似的な人の知見の生成

人の知見データが存在しないデータセットに対して，人の知見を用いたい場合は，図 2 に示す二段階の手順を用いて，擬似的な人の知見を生成する．第一段階として，LLM である Gemini に対し，データセット内のクラス名を入力として与え，そのクラスの特徴を 3〜5 個のテキストとして出力する．第二段階では，出力されたテキストと画像との対応付けを行うために CLIP を用い，テキスト記述に対応する画像内の注目領域を特定する．これらの手順によって得られた注目領域を「擬似的な人の知見」として学習に導入する．

4. データセット

本研究では CUB-200-2011，MVTec，Stanford-Cars，Stanford-Dogs の 4 つのデータセットを用いた．また，人の知見として，CUB-GHA，欠陥部分のセグメンテーションマスクを利用，人の知見がないデータセットにおいては擬似的な人の知見の生成手法を用いて，人の知見を導入した．

5. 評価実験

提案手法の適用による認識精度の変化，および可視化結果に基づく定性的な評価を行った．

5.1. 定量的評価

各データセットにおける精度比較結果を表 1 に示す．CUB および MVTec において，8 ポイント以上の大幅な精度向上を確認した．これは，識別において重要な「局所的な特徴」への誘導が成功したことを示している．また，Stanford-Cars においては，既存手法と比較して 12 ポイント以上の大幅な精度向上を確認した．対して，Stanford-Dogs にお

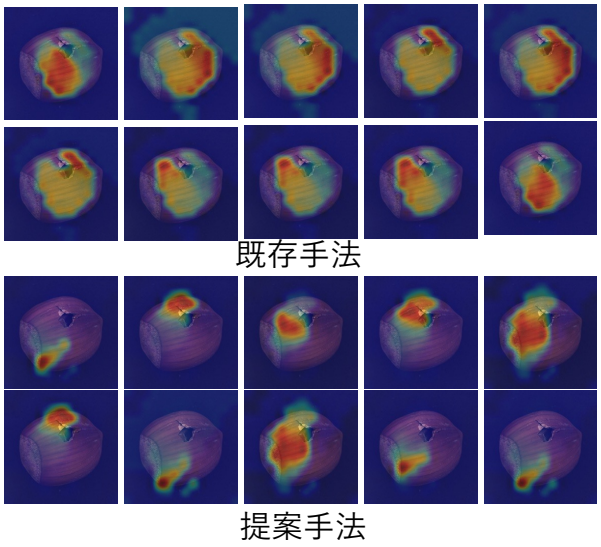


図 3：MVTec における注目領域の比較

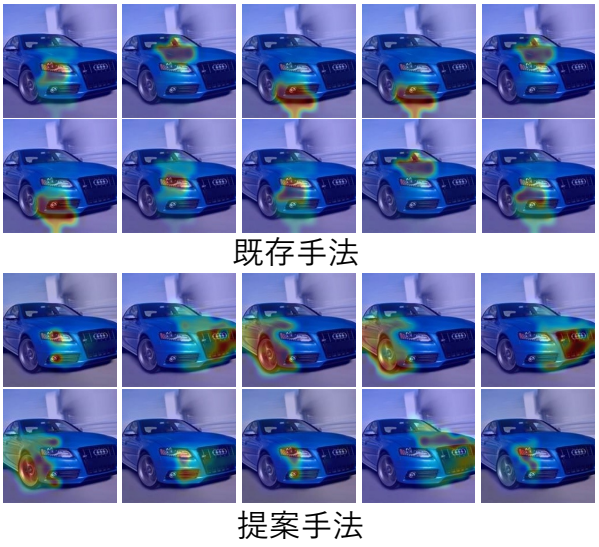


図 4：Stanford-Cars における注目領域の比較

いては，既存手法と比較して 0.18 ポイントの僅かな精度向上が見られた．

表 1：従来手法と人の知見を加えた際の精度比較 [%]

データセット	既存手法	提案手法	追加実験
CUB-200	81.19	**89.78**	81.52
MVTec	89.28	**97.42**	95.83
Stanford Cars	87.94	**99.89**	99.86
Stanford Dogs	80.71	**80.89**	78.48

5.2. 定性的評価と可視化

CUB-200，MVTec，Stanford-Cars，Stanford-Dogs における注目領域の比較結果を図**??**，3，4，5 に示す．実験に使用したモデルのプロトタイプは各クラス 10 個であるため，1 つの入力画像に対して 10 個の可視化結果を生成した．図**??**では，提案手法の注目領域が既存手法と比べ局所的になり，人の知見に近づいたことが確認できた．図 3 では，物体全体に注目する既存手法に比べ，提案手法は局所的な注目をした．図 4 では，既存手法の注目領域に多様性がなく，提案手法では様々なパーツに注目した．図 5 では，提案手法と既存手法で差異が見られなかった．

6. 考察

実験結果から，データセットによって提案手法の効果が異なることがわかった．CUB-200，MVTec，Stanford-Cars においては，注目領域が局所的になり，さらに多様性を持ったことで，提案手法が精度向上をもたらした．一方で，Stanford-Dogs においては，大きな精度向上は見られなかった．これは提案手法と既存手法の注目領域に差異が見られなかったため，生成された擬似的な人の知見が局所的な領域に集中せず，全体に注目したことで，提案した損失が機能しなかったためと考えられる．

7. 結論

本研究では，LLM と CLIP を活用した低コストな擬似的な人の知見の生成手法と人の知見を損失として導入する HKLoss を ProtoPFormer に導入する手法を提案した．実験により，提案手法が車種識別等において強力な正則化として機能し，大幅な精度向上をもたらすことを確認した．今後の展望として，Stanford-Dogs をはじめとした様々なデータセットに対応するため，LLM と CLIP を用いた擬似的な人の知見の生成手法の精度向上が考えられる．

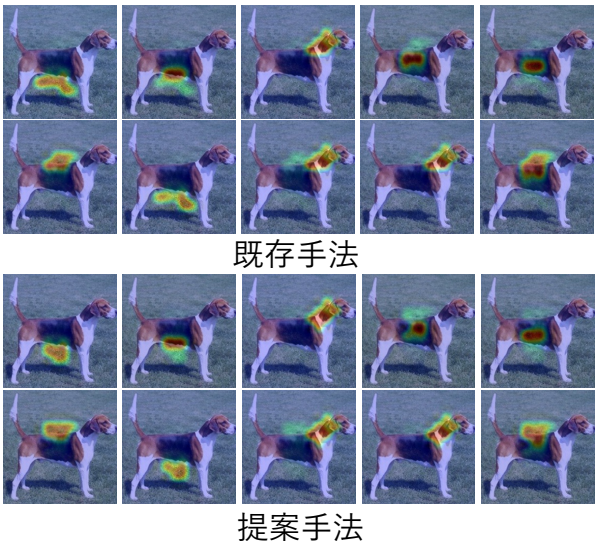


図 5：Stanford-Dogs における注目領域の比較

参考文献

[1] Mengqi Xue et al. ”ProtoPFormer: Concentrating on Prototypical Parts in Vision Transformers for Interpretable Image Recognition.” IJCAI, 2024.

[2] Rong, Yao et al. ”Human Attention in Fine-grained Classification” arXiv, 2021

[3] P. Welinder et al. ”Caltech-ucsd birds 200,” California Institute of Technology, 2011.

研究業績

[1] 落合祐馬 等，“ プロトタイプ法 ProtoPFormer への人の知見の組み込みによる精度向上 ”，画像センシングシンポジウム，2025
（他 1 件）