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．

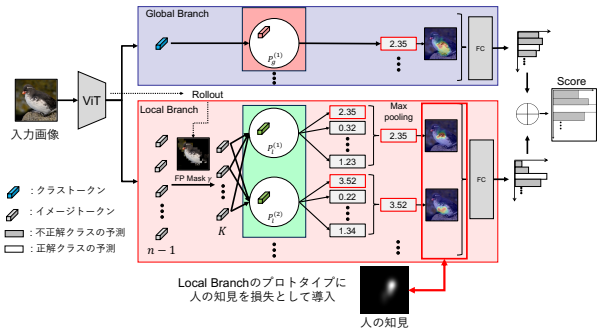

図 1：提案手法の概要図

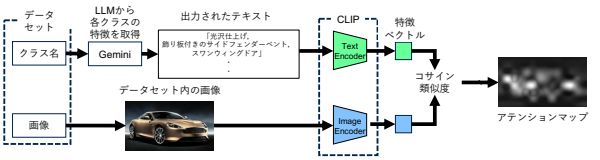

図 2：擬似知見生成手順

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

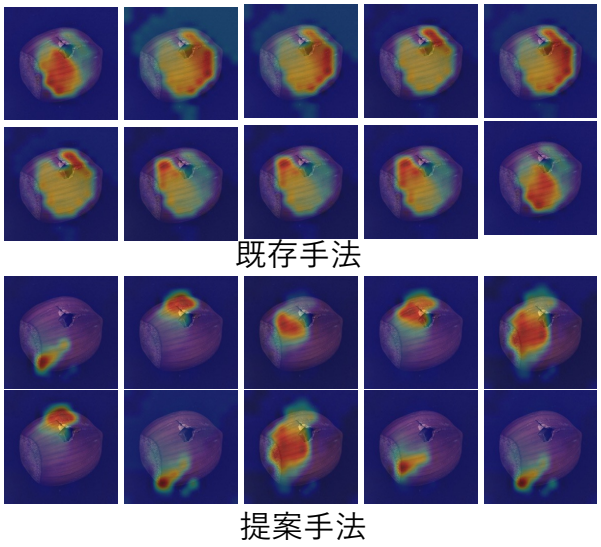

図 3：MVTec における注目領域の比較

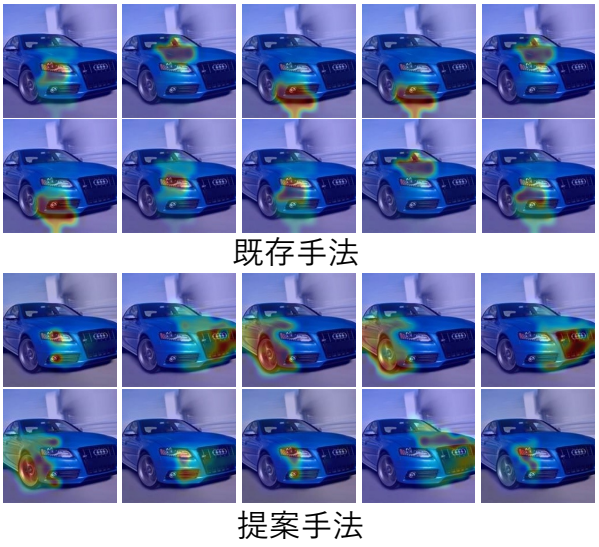

図 4：Stanford-Cars における注目領域の比較

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．

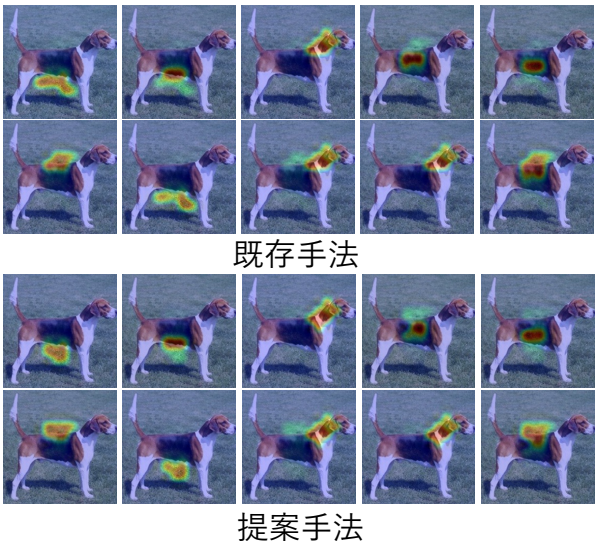

図 5：Stanford-Dogs における注目領域の比較

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）