人の注目領域を用いた ProtoPFormer による詳細画像識別の高精度化

EP20031 落合祐馬　　指導教授：山下隆義

1. はじめに

ProtoPFormer [1] は，入力画像の前景に注目するようにマスクを適用しつつ，複数のプロトタイプを用いること画像分類を行う．しかし，ProtoPFormer は不適切な領域を注目することがある．本研究では ProtoPFormer に対して，人の知見として注目領域をモデルに与えることで，精度向上を図る．

2. ProtoPFormer

ProtoPFormer は複数のプロトタイプを用いて画像分類を行うことに加えて注目領域を可視化することができる手法である．ProtoPFormer のモデル構造を図 1 に示す．プロトタイプ層は Global Branch と Local Branch から構成される．Global Branch では画像全体の情報をもつクラストークンを入力として大域的な特徴とプロトタイプの類似度を算出する．Local Branch では画像の各パッチの情報をもつ画像トークンを入力として各局所的な特徴とプロトタイプの類似度を算出する．その際，バックボーンネットワークの注目領域から作成された FP Mask を用いて前景の画像トークンのみを対象とする．そして，プロトタイプごと求めた類似度について Max pooling を行う．その後，最大値を求める．最後に，全結合層を適用し，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立の平均を求めて出力する．

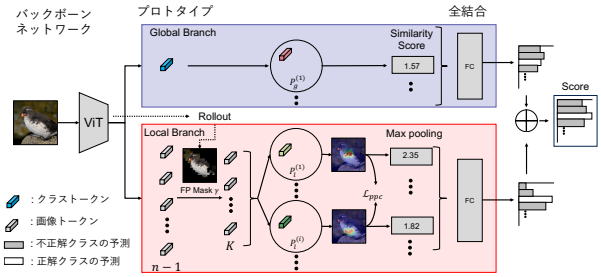

図 1：ProtoPFormer のモデル図

学習時に使用する損失関数を式 (1) と式 (2) に示す．

$$\mathcal{L}^{\mu}_{PPC} = \frac{1}{(m^c_l)^2} \sum_{i \neq j} \max\left(t_\mu - \left\|\hat{\mu}^c_i - \hat{\mu}^c_j\right\|^2, 0\right) \quad (1)$$

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

$\mathcal{L}^{\mu}_{PPC}$ は同じクラスのプロトタイプが似ているほど大きな損失を与える．ここで，m_l 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．

3. 提案手法

ProtoPFormer に人の知見を組み込むために Human Knowledge Branch を追加した構造を提案する．提案手法の構造を図 2 に示す．追加した Human Knowledge Branch では人が認識する際の注目領域をもとに Mask を作成する．Mask を作成するには，まず最初に，ViT[2] が出力する画像トークンのサイズになるように注目領域を縮小する．次に，縮小された注目領域の値に順位をつける．最後に，順位が上位の注目領域に対応する画像トークンのみを残す．その後，ProtoPFormer の Local Branch と同様にプロトタイプとの類似度を計算し，全結合層を適用する．各 Branch から出力された 3 つのクラス確率の平均をモデルの出力とする．学習の際に使用する損失関数は ProtoPFormer と同様である．

4. 評価実験

本実験では提案手法と ProtoPFormer の精度を比較して有効性を検証する．

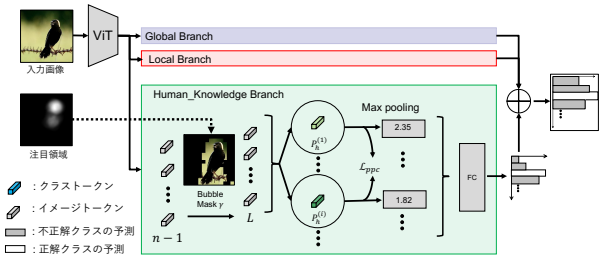

図 2：提案手法

4.1. 実験条件

データセットとして CUB-200-2010 と Bubble 情報を用いる．Bubble 情報は人の知見をもとに作成したデータセットである．バックボーンネットワークには DeiT の tiny モデルを使用する．学習条件はエポック数を 200, バッチサイズは 64 とする．各クラスに対するプロトタイプ数は 10 とする．Local Branch と Human Knowledge Branch で使用する画像トークンの数はそれぞれ 81 個とする．

4.2. 実験結果

ProtoPFormer と提案手法を 6 つのシード値で実行した結果の最大値，平均値，標準偏差を表 1 に示す．表 1 より ProtoPFormer に人の知見を導入することで認識精度の最大値が約 1.54pt 向上していることが確認できる．

表 1：認識精度の比較

	最大値	平均値	標準偏差
ProtoPFormer	42.40	40.68	1.07
提案手法	**43.65**	**42.22**	**0.96**

ProtoPFormer と提案手法のアテンションマップの可視化結果を図 3 に示す．図 3 上部の Parakeet Auklet では ProtoPFormer は背景を含む領域に注目しているが，提案手法は頭部と腹部付近に注目し，正しい予測結果となった．しかし，図 3 下部の Brewer Blackbird では頭部から腹部にかけて鳥の領域を注目したが，正しくない予測結果となった．このことから，提案手法は人の注目領域を用いて分類することが可能となるが，アピアランスが似ている場合は画像分類を間違える場合がある．

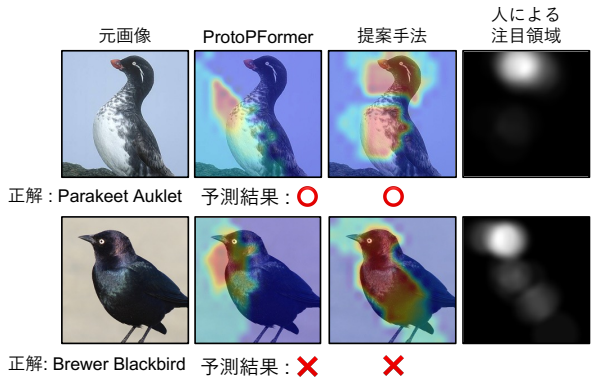

図 3：アテンションマップの可視化

5. おわりに

本研究では ProtoPFormer に人の知見を導入する手法を提案し，認識精度の向上とより適切な注目領域の獲得を確認した．今後は各 Branch の寄与率を調査することにより手法の有効性を確認する．

参考文献

[1] Mengqi Xue et al. ”ProtoPFormer: Concentrating on Prototypical Parts in Vision Transformers for Interpretable Image Recognition.” arXiv, 2022.

[2] Alexey Dosovitskiy et al. ”An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.” arXiv, 2020.