人の注目領域を用いたProtoPFormerによる詳細画像識別の精度向上

落合 祐馬[†1,a)]　平川 翼[†1,b)]　山下 隆義[†1,c)]　藤吉 弘亘[†1,d)]

概要：プロトタイプベースのモデルはプロトタイプを用いて構築されるモデルであり，モデルの判断根拠の可視化などに活用されている．プロトタイプはクラスごとに割り当てられるクラスの特徴的な領域を学習により獲得するものであり，入力画像と各クラスのプロトタイプの類似度を評価することにより画像の分類を行う．Convolurional Neural Networks (CNN) に対してプロトタイプを適用したモデルにProtoPNet がある．ProtoPNet は背景に注目することや複数あるプロトタイプが同じ領域に注目してしまう問題点がある．そこで，Vision Transformer (ViT) に ProtoPNet を適用した ProtoPFormer が作成された．ProtoPFormer はプロトタイプを用いた判断根拠の可視化が可能である．本研究では，詳細画像分類の精度向上を目的として ProtoPFormer に人の知見を導入する Branch を提案する．詳細画像分類ではクラス特有の領域に注目することが重要である．そこで，追加する Branch は人の知見に基づいてクラス特有の領域に注目するように設計された．本論文では，ProtoPFormer と本手法の精度を比較し，判断根拠を可視化したアテンションマップの注目領域を比較することで，本手法の有効性を評価する．

1. はじめに

Vision Transformer (ViT) [1] は Transformer [2] を画像認識の分野に適用したモデルであり，ViT が公表されるまでに画像認識や物体検出などの分野で使用されていた Convolutional Neural Network (CNN) [3] と比べて高い精度を獲得している．これは入力された画像内の関係性を求める Self Attention を複数並列に処理する Multi-head Self Attention を用いることで，アンサンブル学習のように異なる特徴を得ることが可能である．一方で，これらの特徴がどのようにモデルの判断に影響を与えているのかを理解するためには，何らかの可視化手法が必要となる．その一つとして，ViT をはじめとしたモデルの判断根拠を可視化する手法にプロトタイプベースのモデルが存在する．プロトタイプはクラスごとに割り当てられるクラスの特徴的な領域を学習により獲得するものである．入力画像と各クラスのプロトタイプの類似度を評価することにより，画像の分類を行う．CNN に対してプロトタイプを適用したモデルに ProtoPNet [4] がある．ProtoPNet は入力画像に畳み込み処理を適用することで得た特徴マップとプロトタイプを比較しながら画像分類を行う．この手法はプロトタイプごとに可視化することができるため，エキスパートな人材が一般的な人に対して判断根拠を提示する際にどこに注目したのかを細かい領域ごとに説明することが可能になる．しかし，ProtoPNet は背景に注目することや複数あるプロトタイプが同じ領域に注目してしまう問題点がある．対して ViT は離れた特徴を捉えることができる．この強みを利用するために，ProtoPNet に ViT を適用したモデルが ProtoPFormer である．ProtoPFormer は入力された画像の特徴を捉えながらプロトタイプを使用することで判断根拠の可視化が可能である．また，別の研究で Attention Branch Network (ABN) [5] が存在する．この研究では CNN に人の知見を導入することで精度を向上させている．そこで，本研究では，ProtoPFormer に人の知見を導入することで詳細画像分類の精度を向上させながら，細かい領域ごとに注目した判断根拠を可視化し，適切な注目を提示するための構造を提案する．人の知見は Human Knowledge Branch を追加することで導入する．本論文では，ProtoPFormer と本手法の精度を比較し，判断根拠を可視化したアテンションマップの注目領域を比較することで，本手法の有効性を評価する．学習に CUB-200-2010 を，人の知見として Bubble 情報を用いる．精度の比較には，CUB-200-2010 に対する正解率を用いる．

[†1] 現在，中部大学
Presently with Chubu University
a) ochiai20031@mprg.chubu.ac.jp
b) hirakawa@isc.chubu.ac.jp
c) takayoshi@isc.chubu.ac.jp
d) fujiyoshi@isc.chubu.ac.jp

2. 関連研究

本章では代表的なプロトタイプベースのモデルである ProtoPNet，ProtoPFormer について述べる．

2.1 ProtoPNet

Prototypical part network (ProtoPNet) [4] は CNN を用いたプロトタイプベースのモデルである．ProtoPNet は，畳み込み層は入力画像から特徴マップを抽出し，プロトタイプ層で抽出した特徴マップとプロトタイプとの類似度を求め，全結合層で類似度を入力として画像の分類を行うモデルである．

プロトタイプ層で使用するプロトタイプは，勾配降下法を用いて学習される．学習では，式 (1) で表される損失関数を最小化するようプロトタイプと畳み込み層の重みを更新する．

$$\min_{P;w_{conv}} \frac{1}{n}\sum_{i=1}^{n} \mathrm{CrsEnt}(h\Gamma g_p \Gamma f(x_i), y_i) + \lambda_1 \mathrm{Clst} + \lambda_2 \mathrm{Sep} \tag{1}$$

ここで P は最適化の対象となるプロトタイプ，w_{conv} は畳み込み層の重み，n はサンプル数，x_i はデータセット内 i 番目のサンプル，y_i は真のラベル，f は畳み込み処理を表す関数，g_p はプロトタイプを用いて特徴を表す関数，h は最終的な予測を行うための関数である．また，λ_1 と λ_2 はそれぞれ損失関数にかけられる係数，式 (1) はクロスエントロピーコスト，Clst コスト，Sep コストの 3 つの項から構成される．クロスエントロピーコストは全結合層の出力と真のラベルと比較し，誤差を最小化するように学習する．Clst コストは同じクラスのプロトタイプを近づけるために使用する．Sep コストは異なるクラスのプロトタイプを離すために使用する．式 (2) と式 (3) は，それぞれ Clst コストと Sep コストを示す．

$$\mathrm{Clst} = \frac{1}{n}\sum_{i=1}^{n} \min_{j:p_j \in P_{y_i}} \min_{z \in \mathrm{patches}(f(x_i))} |z - p_j|_2^2 \tag{2}$$

$$\mathrm{Sep} = -\frac{1}{n}\sum_{i=1}^{n} \min_{j:p_j \notin P_{y_i}} \min_{z \in \mathrm{patches}(f(x_i))} |z - p_j|_2^2 \tag{3}$$

この時 z は畳み込み層から出力された特徴マップを表し，P_{y_i} はクラス y のプロトタイプを表し，p_j は最適化の対象となるプロトタイプを表す．式 (2) では各サンプルの特徴マップと真のラベルに対応するプロトタイプとの距離を最小化する．式 (3) では各サンプルの特徴マップと真のラベルに対応しないプロトタイプとの距離を最大化する．これらの式から，ProtoPNet は全結合層とプロトタイプを同時に学習し，画像分類を行うモデルであることがわかる．

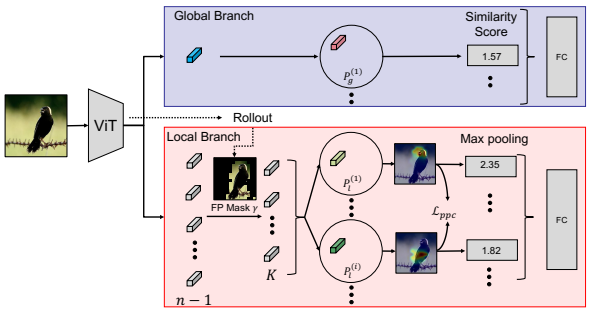

図 1 ProtoPFormer のモデル構造

2.2 ProtoPFormer

ProtoPFormer[6] 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．モデルを構造を図 1 に示す．ProtoPFormer は最初に ViT に画像を入力することでクラストークンと画像トークンが出力される．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．

ProtoPFormer は ProtoPNet と同様に各クラスの特徴を表すベクトルであるプロトタイプを勾配降下法を用いて学習する．

ProtoPFormer は式 (4) で表される損失関数を最小化するよう学習する．

$$\mathcal{L}_{PPC} = \lambda_\mu \mathcal{L}^{\mu}_{PPC} + \lambda_\sigma \mathcal{L}^{\sigma}_{PPC} \tag{4}$$

ここで λ_μ と λ_σ はそれぞれの係数である．$\mathcal{L}^{\mu}_{PPC}$ は同じクラスのプロトタイプが完全に同じにならないように使用する．$\mathcal{L}^{\sigma}_{PPC}$ はプロトタイプが注目する領域を小さくするように使用する．

$\mathcal{L}^{\mu}_{PPC}$ を式 (5) に示す．

$$\mathcal{L}^{\mu}_{PPC} = \frac{1}{(m_l^c)^2}\sum_{i \neq j} \max(t_\mu - \left|\hat{\mu}_i^c - \hat{\mu}_j^c\right|^2, 0) \tag{5}$$

ここで，m_l は Local Branch のプロトタイプの数，c はクラス，$\hat{\mu}$ はプロトタイプ，t_μ が閾値である．この損失関数は同じクラスのプロトタイプ同士の距離が閾値 t_μ よりも小さくなるように学習する．

$\mathcal{L}^{\sigma}_{PPC}$ を式 (6) に示す．

$$\mathcal{L}^{\sigma}_{PPC} = \mathrm{tr}\left(\max\left(0, \hat{\sum} - t_\sigma\right)\right) \tag{6}$$

$\hat{\sum}$ はプロトタイプの共分散行列の対角成分の平均，t_σ は閾値である．この損失関数はプロトタイプが注目する領域を小さくするよう学習する．

3. 提案手法

本手法では ProtoPFormer に人の知見を組み込むために Human Knowledge Branch を追加した構造提案する．

3.1 人の知見を用いた ProtoPFormer

提案手法の構造は図 2 に示すように ViT と Global Branch，Local Branch，Human Knowledge Branch の 3 つの Branch から構成される．ViT に画像を入力することでクラストークンと画像トークンが出力される．出力されたクラストークンは Global Branch に入力され，画像トークンは Local Branch と Human Knowledge Branch に入力される．各 Branch では入力されたトークンとプロトタイプとの類似度を求め，全結合層に入力することでクラススコアを出力する．そのため，クラススコアは Branch の数と同じ 3 つ出力される．出力された 3 つのクラススコアの平均値をモデルの出力とする．

3.1.1 Bubble Mask

Human Knowledge Branch に入力された画像トークンの中から使用するトークンを選択する際に使用する人の注目領域を用いた Mask である．最初に出力される ViT[1] の画像トークンのサイズに注目領域を合わせるために，単一の値にハイキュービック補間を用いて縮小する．次に，単一の値が高い，領域を使用する Bubble Mask を作成する．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．このように Human Knowledge Branch で使用する画像トークンを選択する．

3.1.2 Human Knowledge Branch

Human Knowledge Branch は、Bubble Mask で抽出された各画像トークンとプロトタイプとの類似度を求め、クラス確率を計算する．各プロトタイプは異なる領域を学習しているため、同じクラスを学習しているプロトタイプでも、プロトタイプごとに異なる領域に注目する．入力される画像トークンとプロトタイプの類似度は、類似度コサインを使用する．そのため、画像トークンとプロトタイプの次元数は同じ必要がある．類似度は、一つのプロトタイプにつき抽出された画像トークンの数だけ生成される．その中から最も大きい値を、プロトタイプの類似度スコアとして出力する．

4. 評価実験

本章では，提案手法である ProtoPFormer への人の知見の導入による有効性を確認するために ProtoPFomer との精度と注目領域の比較を行う．

4.1 データセット

本実験では，データセットとして CUB-200-2010[7] と Bubble 情報 [7] を使用する．CUB-200-2010 は，200 種類の鳥の画像であり，学習データは 2000 枚，テストデータは 1000 枚の画像が含まれる．画像サイズはさまざまである．Bubble 情報は CUB-200-2010 の画像に対して複数人が注目する画像上の座標を入力したデータセットである．Bubble 情報の値は 0 から 1 の連続値で表現される．Bubble 情報の作成方法は画像に対して特徴的な部分に人が点を打つ．この作業を複数人に行ってもらいより多くの人が注目する領域ほど 1 に近い値となる．

4.2 実験条件

本実験は特徴抽出を行うに Data-efficient Image Transformer (DeiT) の tiny モデルを使用する．学習条件は ProtoPFormer の論文と同じ 6 つ seed 値 (1028, 2678, 3566, 4686, 5328, 6186) の 6 つで実験を行う．また，最適化手法は AdamW [8], 初期学習率は 1e-5, スケジューラは Cosine Learning Rate Scheduler, エポック数は 200, バッチサイズは 64 とする．また，バックボーンネットワークである ViT からは画像トークンが 196 個出力される．FP Mask と Bubble 情報を使用して作成する Bubble Mask では出力された 196 個の中から 81 個の画像トークンを使用する．また，使用するプロトタイプの数は各クラス 10 個使用する．

4.3 人の知見導入による比較

はじめに ProtoPFormer と提案手法の認識精度の比較を表 1 に示す．表 1 から明らかなように，ProtoPFormer に人の知見を用いた Human Knowledge Branch を組み込むことで，認識精度が 1.25pt 向上，平均で 1.34 pt 向上，していることを確認できる．これはマスクを人の知見に沿った形になるように変更したことにより，プロトタイプが重要な画像トークンに注目したからだと考える．

表 1 認識精度 [%]

	最大値	平均値
ProtoPFormer	42.4	40.7
HK Branch の追加	**45.3**	**40.8**

次に ProtoPFormer と提案手法における可視化結果を図 3 に示す．左側 2 枚の可視化結果では分類対象により注目しており，分類結果は従来手法と同じように正しく行うことができている．3 枚目の画像では注目町域が分類対象により注目し，分類が正しく行われていることが確認できる．4 枚目の画像では提案手法でも正しく分類が行うことはできなかったが，従来手法と比べて注目領域が分類対象により注目していることが確認できる．

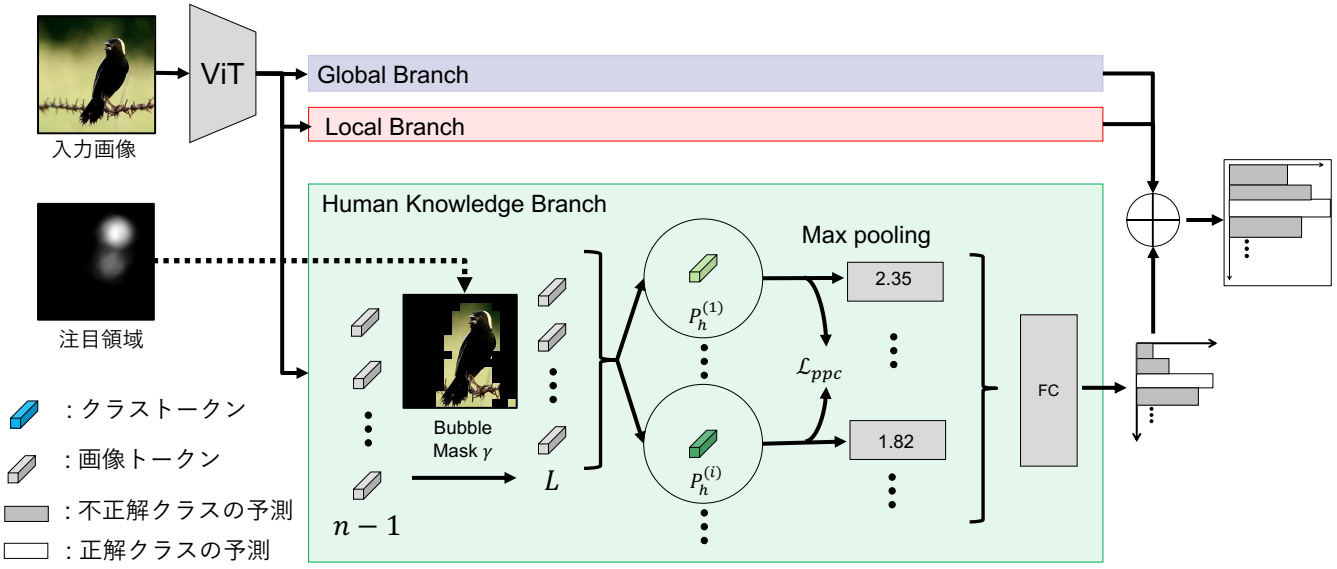

図 2 提案手法のモデル構造

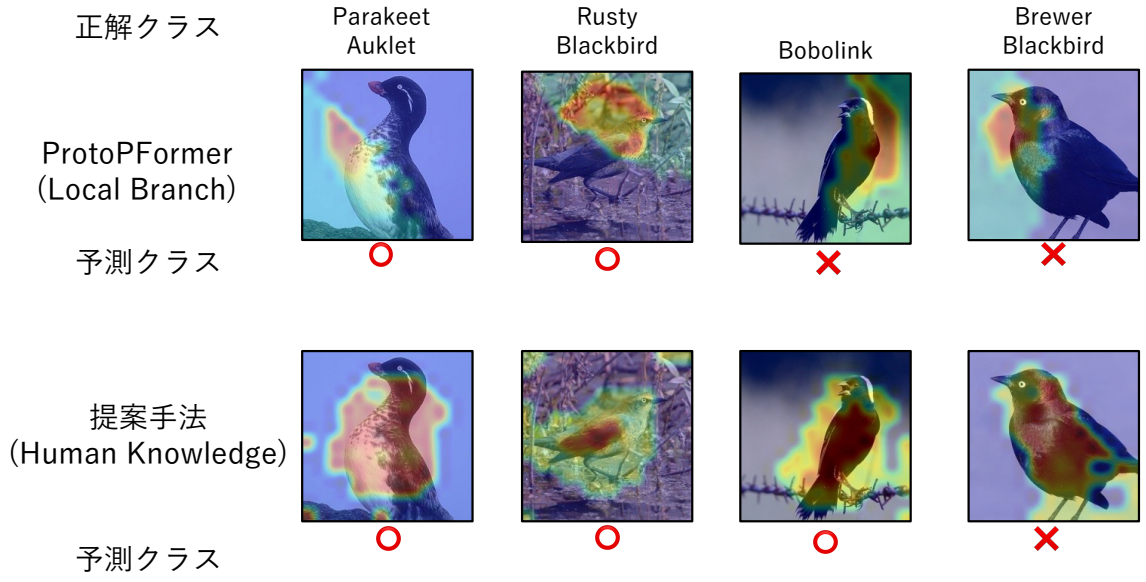

図 3 類似度の可視化結果

4.4 Ablation study

各 Branch の精度への寄与率を確認するために， Local Branch と Global Branch が存在しない場合の精度を確認する．3 つの Branch が存在する場合，その精度が低下するほど重要な Branch であると考えることができる．

表 2 Ablation styde の認識精度 [%]

削除した Branch	最大値	平均値
Local Branch	44.5	42.7
Global Branch	40.2	34.7

Local Branch が削除された際では最大値が 0.8pt 減少，平均値は 1.9pt 増加した．このことから Local Branch は精度にあまり関係せず，追加した Human Knowledge Branch が Local Branch の役割を担っていると考えられる．Global Branch を削除した際は最大値が 5.1pt 減少，平均値は 6.1pt 減少した．大きく精度が減少した原因として Global Branch は画像の全体からクラス分類をするのに対して残りの二つの Branch は画像の局所的な部分からクラス分類をするため複数の Branch を使用する際に得られる相乗効果がなくなってしまったことにより，精度が低下したのではないかと考える．

5. おわりに

本研究では，人の知見を導入するための Branch を提案した．提案手法の精度が向上したことから，人の知見を導入することは詳細画像識別の精度向上に有効であることが確認できた．加えて，注目領域を確認したところ分類対象により注目し，分類を正しく行えるようになることがある

と確認できた．しかし，画像によっては提案手法でも正しく分類が行うことはできなかった．これは分類対象に注目するだけではなく，詳細な変化を捉えることができるとさらに精度が向上すると考えられる．今後は，AI から人に対して知見を共有ができる仕組みの構築を目指す．

参考文献

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[2] Ashish Vaswani, Shazeer Ahmed, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012.

[4] Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. This looks like that: Deep learning for interpretable image recognition. In *NeurIPS*, 2019.

[5] Hiroshi Fukui, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Attention branch network: Learning of attention mechanism for visual explanation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[6] Mengqi Xue, Qihan Huang, Haofei Zhang, Lechao Cheng, Jie Song, Minghui Wu, and Mingli Song. Protopformer: Concentrating on prototypical parts in vision transformers for interpretable image recognition, 2022.

[7] J. Deng, J. Krause, and L. Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. In *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 580–587, Los Alamitos, CA, USA, jun 2013. IEEE Computer Society.

[8] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.