ノイズに対する頑健性と視覚的説明性を獲得する Attention Branch Transformer の提案

ER19043　鈴木雅司　　指導教授：藤吉弘亘

1. はじめに

Attention Branch Network (ABN) [1] は，Attention 機構による精度の向上とアテンションマップの可視化による視覚的説明を両立する手法である．ABN のベースモデルである畳み込みニューラルネットワーク (CNN) は，局所的な帰納的バイアスを持つため物体のテクスチャを重視して学習する．そのため，画像の意味的な構造やスタイルの変化等に対応できない．これらの問題を解決するために，本研究では局所的な帰納的バイアスの影響を受けにくいモデルである Vision Transformer (ViT) に Attention Branch を導入した Attention Branch Transformer (ABT) を提案する．

2. Vision Transformer

ViT は自然言語処理の Transformer を画像処理に応用した手法である．ViT は画像を均等なパッチに分割し，埋め込み処理で各パッチをベクトルの系列である Patch Token に変換する．そして，学習可能な Token である Class Token を追加し，これらの Token を Transformer Encoder (TE) へ入力する．TE では Multi-Head Attention で全ての Token 間の関係性を捉え，Token 間で相互に値を更新し特徴抽出を行う．ViT は Token 間の広域的な特徴を捉えるため，CNN と比べ局所的な帰納的バイアスの影響が少ない．そのため，ViT は画像の意味的な構造やスタイルの変化およびテクスチャノイズ等のノイズに対して頑健となる．

3. 提案手法

本研究では，より頑健であり説明性の高い説明可能 AI (XAI) の実現を目指し，ViT にアテンションマップを出力する Attention Branch を導入した手法 ABT を提案する．図 1 に提案手法のネットワーク構造を示す．提案手法の TE では Patch Token のみで特徴抽出を行い，各 Branch の最終層で Class Attention (CA) および Class Token を用いて推論を行う．提案手法の損失関数 $L(\mathbf{x}_i)$ として，式 (1) のように各 Branch の学習損失の和を用いる．ここで，L_{att}, L_{per} は各 Branch の学習損失，x_i は入力サンプルを表す．また，各 Branch の学習損失は，各 Branch より出力されるクラス確率と教師ラベルとのクロスエントロピー誤差として算出される．

$$L(\mathbf{x}_i) = L_{att}(\mathbf{x}_i) + L_{per}(\mathbf{x}_i) \quad (1)$$

Attention Branch では，式 (2) に示すように CA に基づきアテンションマップを生成する．ここで，$\mathbf{Q}_a, \mathbf{K}_a$ は Attention Branch における CA の Query，Key，d, h は各パッチの埋め込み次元数および head 数，$\mathbf{M}_{att}$ はアテンションマップを表す．

$$\mathbf{M}_{att} = \mathrm{Sigmoid}(\mathbf{Q}_a\mathbf{K}_a^\top/\sqrt{d/h}) \quad (2)$$

また，Perception Branch では，式 (3) に示すように 1 層目の TE の Attention weight とアテンションマップの要素積をとることで特徴マップの強調を行う．ここで，$\mathbf{A}, \mathbf{Q}_p, \mathbf{K}_p$ は Perception Branch における 1 層目の Multi-Head Self Attention の Attention weight および Query，Key を表す．

$$\mathbf{A} = \mathrm{Softmax}(\mathbf{Q}_p\mathbf{K}_p^\top/\sqrt{d/h})\mathbf{M}_{att} \quad (3)$$

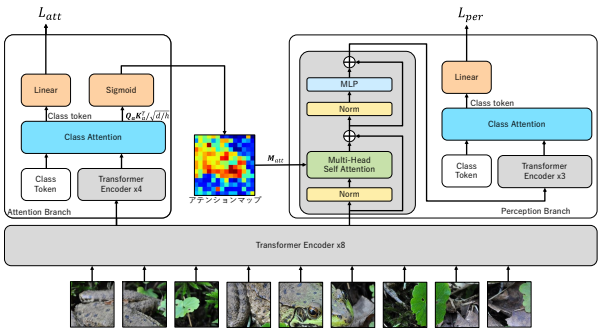

図 1: 提案手法のネットワーク構造

4. 評価実験

本章では，提案手法と ABN，ViT の精度，耐ノイズ性およびアテンションマップによる視覚的説明性の比較結果について述べる．

4.1. 実験条件

学習用データセットには ImageNet を用いる．評価用データセットには，ImageNet と耐ノイズ性の評価に ImageNet-A，ImageNet-R，ImageNet-C を用いる．ここで，ImageNet-A は ImageNet から誤識別の多い画像，ImageNet-R は創作物の画像を集めたデータセットであり，ImageNet-C は ImageNet に 19 種類のノイズを付加したデータセットである．ImageNet-A および ImageNet-R では Top 1 accuracy，ImageNet-C では各ノイズに対する Corruption Error (CE) の平均である mean Corruption Error (mCE) で評価する．ABT-B は埋め込み次元数を 786，head 数は 12 であり，ABT-S と ABT-T は，埋め込み次元数と head 数をそれぞれ 1/2,1/4 に変更したモデルとする．

4.2. 実験結果

提案手法と ABN，ViT の精度と頑健性の比較結果を表 1 に示す．表 1 より，精度，頑健性共に ABT-B が最も優れている事が確認できる．

表 1: 提案手法と従来手法の精度と頑健性の比較

Model	Param	INet Top-1 ↑	INet-A Top-1 ↑	INet-R Top-1 ↑	INet-C mCE ↓
ABN-34	36.4 M	71.4	3.3	34.2	77.9
ABN-50	43.6 M	79.9	11.3	41.2	65.4
ABN-101	62.6 M	81.5	18.4	44.3	58.8
ViT-T	5.7 M	72.2	7.0	33.2	71.3
ViT-S	22.1 M	79.9	19.2	42.5	55.7
ViT-B	86.6 M	82.0	27.9	45.3	49.4
ABT-T	8.6 M	71.3	7.5	32.6	71.1
ABT-S	33.1 M	80.1	23.6	43.5	53.6
ABT-B	129.9 M	**82.8**	**32.5**	**46.2**	**47.9**

ABN と提案手法の出力したアテンションマップを図 2 に示す．図 2 は左から入力，ABN および提案手法の Attention Branch から出力したアテンションマップである．図 2 より，提案手法のアテンションマップは推定物体の形状に沿った Attention が得られている事が確認できる．この結果より，提案手法は視覚的説明において形状の表現に有効であると確認できる．

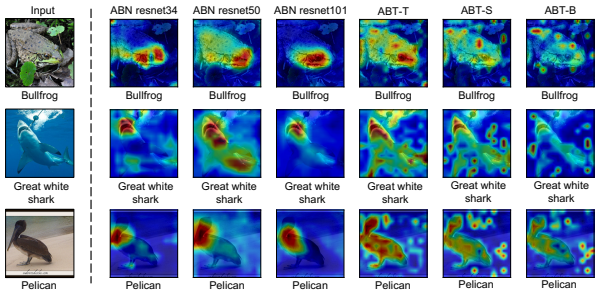

図 2: アテンションマップの可視化

5. おわりに

本研究では，より頑健であり説明性の高い XAI の実現を目指し，ViT に Attention Branch を導入した手法 ABT を提案した．本研究では ABT が精度，耐ノイズ性，視覚的説明性に有効であることを確認した．今後は人から AI への働きかけが可能となるよう，提案手法へ人の知見を導入する手法を検討する．

参考文献

[1] H. Fukui, *et al.*, “Attention Branch network: Learning of attention mechanism for visual explanation”, CVPR, 2019.