人の順位付けによる深層学習モデルの視覚的説明の傾向調査

ER19090 **若山浩之**　　**指導教授：藤吉弘亘**

1. はじめに

深層学習モデルの推論結果に対する注視領域をアテンションマップとして可視化することで，モデルが入力画像のどこに注目して判断したかを人が理解できるようになる．このような視覚的説明性の良し悪しを表すアテンションマップの評価基準はないため，どのようなアテンションマップを獲得すると有効であるかは不明である．そこで本研究では，人の順位付けによるアテンションマップの評価法を提案し，様々な深層学習モデルのアテンション特性を調査する．

2. Grad-CAM

Grad-CAM [1] は，ネットワークの勾配を利用して出力クラスに対するアテンションマップを可視化する手法である．はじめに，画像をモデルに入力して計算した特徴マップから，クラス出力に対する勾配を計算する．次に，特徴マップの各チャネルごとに勾配の平均を求め，特徴マップに重み付けする．重み付けした特徴マップに ReLU 関数を適用し，その出力をアテンションマップとする．

3. アテンションマップの評価方法

提案するアテンションマップの評価方法を図 1 に示す．本研究では以下の手順により，人の判断根拠と類似しているアテンションマップを示すモデルの順位付けを行う．

Step 1: Grad-CAM でアテンションマップを可視化

Step 2: 被験者が分類時に注目すると考える順で順位付け

Step 3: 順位付け結果からスコアを集計

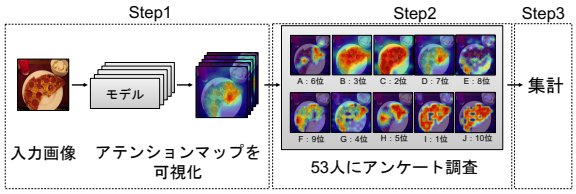

図 1: 評価方法

評価指標として，図 2 に示す順位スコアを用いる．被験者が順位付けしたアテンションマップの順位に対応した点数を小計スコアとして，被験者数分の小計スコアを加算し，合計スコアとする．合計スコアの順位を合計順位として，合計順位に対応した点数を順位スコアとする．順位スコアは合計スコアを正規化するため，合計スコアと比べて外れ値の影響を受けにくい．

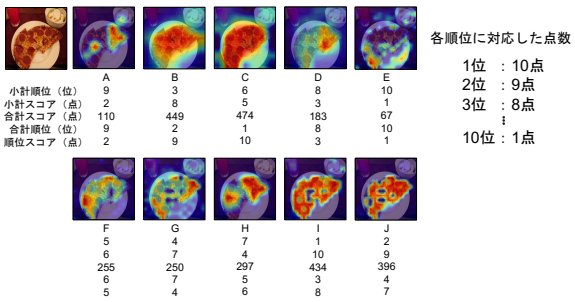

図 2: 順位スコア

4. 評価実験

本実験では，人の順位付けにより深層学習モデルのアテンションマップの傾向を調査する．

4.1. 実験条件

10 種の深層学習モデルのアテンションマップを Grad-CAM で可視化し，53 人の被験者にモデル情報を伏せて公平な評価の下，調査を行う．比較するモデルを表 1 に示す．SA は Self-Attention，Conv は畳み込み，精度は ImageNet-1k の評価示す．それぞれのモデルは ImageNet-1k で学習済みであり，パラメータ数が互いに近いモデルを用いて実験する．評価画像は，10 クラス 5 枚ずつの計 50 画像を使用する．

表 1: 比較するモデル

モデル	学習方法		特徴抽出		パラメータ数	精度 [%]
	教師あり学習	自己教師あり学習	Conv	SA		
ResNet	√	-	√	-	23M	75.8
DINOResNet	-	√	√	-	23M	75.3
ConvNeXT	√	-	√	-	28M	82.0
Swin	√	-	-	√	29M	81.1
CaiT	√	-	-	√	17M	80.7
DeiT	√	-	-	√	22M	79.9
DINODeiT	-	√	-	√	22M	77.0
CoAT	√	-	√	√	20M	81.9
XCiT	√	-	√	√	26M	81.8
DINOXCiT	-	√	√	√	26M	77.8

4.2. 実験結果

各評価画像で獲得した順位スコアを各クラス別に加算して比較する．評価結果例を表 2 に示す．全データは，50 枚の評価画像で獲得した順位スコアの合計である．合計データの括弧内に分散を示す．全データの結果から，XCiT が最もスコアが良いモデルとなることがわかる．次に CoAT，DINOResNet，ResNet と続いている．畳み込み構造の ResNet と Self-Attention 構造の DeiT を比較すると ResNet の方が，順位スコアの順位が高い．Swin と，ResNet の構造から Swin に近づけるよう設計した ConvNeXT を比較しても，畳み込みベースである ConvNeXT の順位が高い．1 位と 2 位である XCiT と CoAT は，畳み込み処理と Self-Attention の構造の両方を取り込んだモデル設計であり，最も人の注視領域と類似した傾向となる．

表 2: 評価結果 [位]

モデル	ピザ	電車	封筒	金魚	チワワ	アトリ	ホルン	全データ
ResNet	2	2	5	4	4	4	5	4 (5.64)
DINOResNet	5	2	5	3	2	2	5	3 (5.61)
ConvNeXT	7	7	7	5	8	3	4	7 (7.93)
CoAT	1	1	2	6	1	6	5	**2** (8.14)
Swin	9	8	10	9	6	8	9	10 (5.13)
DeiT	6	5	3	2	5	4	3	5 (4.9)
DINODeiT	10	10	8	8	10	10	8	10 (5.55)
CaiT	3	6	1	6	7	8	2	6 (7.36)
XCiT	4	4	3	1	3	1	1	**1** (3.72)
DINOXCiT	8	9	9	10	9	7	10	8 (6.29)

ラベルありで学習した XCiT とラベルなしの自己教師あり学習（DINO）で学習した DINOXCiT のアテンションマップの可視化例を図 3 に示す．括弧内に表 2 の全データの順位を示す．DINO を使用した場合，教師ありの XCiT と比較して，認識対象とアテンションの領域が類似していないことがわかる．また，DINOXCiT は XCiT と比べて順位が低いことから，人の注視領域と異なるといえる．

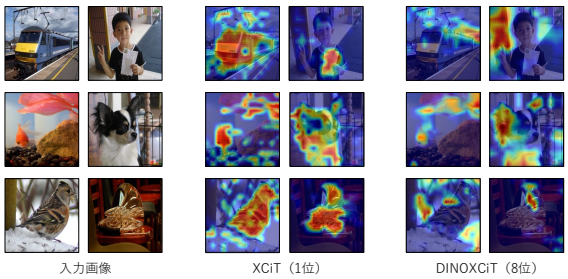

図 3: XCiT と DINOXCiT のアテンションマップ

5. おわりに

本研究では，深層学習モデルの認識特性の傾向調査した．実験結果から，人の注視領域と類似したモデルは XCiT，CoAT になることが分かった．今後は，アテンションの定量的な評価を行う．

参考文献

[1] R. Selvaraju, *et al.*,“Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization”，ICCV，2019.