7Q-04

MoRAM：CNN回帰モデルに対する予測根拠の視覚的説明*

下村真生† 中村和幸‡
明治大学††

1 はじめに

人工知能を用いた判別や回帰予測は高精度で，医療や農業など幅広い分野での応用が期待される一方，その予測過程が不明瞭という課題がある．これを解釈することで説明可能なAIを目指すことがxAIの目的である．Guidottiら [1] らによってxAIは大きく4つに分類され，得た出力に対して予測根拠を視覚的に説明する”Outcome Explanation”は専門家以外でも解釈が容易であるとの利点がある．畳み込みニューラルネットワーク (CNN) を用いた分類問題への解釈手法はGrad-CAM[2] 等が存在するが，回帰問題への解釈手法に関する研究は数が少ない．代表例となるRAMは，回帰問題への解釈が可能な一方，適用可能なモデルが限られ，予測値が負の場合に誤った可視化結果を出す傾向がある．本稿ではRAMのモデル制約を緩和し，予測値の正負に関わらず適切な解釈を可能にするModulus reweighted Grad-RAM(MoRAM) を提案し，その有効性を計算機実験により示す．

2 特徴量マップによる説明

CNN分類モデルに対する説明において，A^k を特徴量マップ (CNNの最後の畳み込み層における k 番目のチャンネル)，w_k^c を k 番目のチャンネルのクラス c に対する重要度 とする時，顕著量マップ L^c は次式で定義される：

$$L^c = \sum_k w_k^c A^k. \tag{1}$$

この L^c を元画像サイズまでアップサンプリングした $\hat{L^c}$ を元画像に重ねることで，あるクラス c の判別根拠となった領域を表すヒートマップによる視覚的説明を得る．分類モデルに対する手法群 (CAMs) は w_k の定義によりGrad-CAM++[3] 等が発表されている．

回帰モデルに対する視覚的説明は，CAMsから派生した手法である．回帰モデルの出力サイズは1であるため，分類モデルの A^k と w_k^c の定義を回帰モデルに適用して回帰モデルの顕著量マップの定義は次となる：

$$L = \sum_k w_k A^k. \tag{2}$$

先行手法であるRAMでの w_k の定義は，A^k をGlobal Average Pooling (GAP) に通して得られる k 個のノードと出力 Y の間の重みである．この定義によりRAMの適用可能なCNNモデルは，出力層付近が畳み込み層，GAP，出力層の順に直列に結合している必要があるという制約がある．この制約を緩和したものがGrad-RAM[5] である．これは w_k を以下のように定義する：

$$w_k = \frac{1}{Z}\sum_i\sum_j \frac{\partial Y}{\partial A_{ij}^k}. \tag{3}$$

出力 Y から A^k への勾配を取る形へ変更することでRAMのように出力層付近での制約がなくなる．RAMのモデル制約内で，Grad-RAMはRAMと同義になる．

3 提案手法

回帰モデルに対する手法群は，Y や w の値域が正の場合にのみ有効である．活性化関数としてReLU関数を通した後の特徴量マップで位置 (x, y) が正にも負にも重要な特徴である場合，w が正負ともに存在するため式 (2) から打ち消され $L_{xy} \simeq 0$ となり，特徴が反映されない．回帰モデルでは，出力へ影響を与えた特徴を可視化することが重要であるため，正負どちらへ効く情報も重要である．したがってMoRAMでは重み w に式 (3) を用いて顕著量マップを次のように定義する：

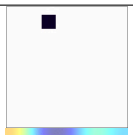
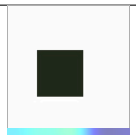
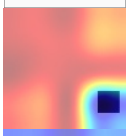
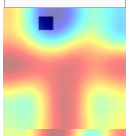
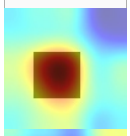
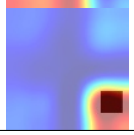
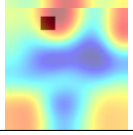
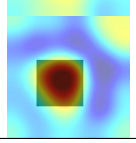
$$L = \sum_k |w_k| A^k. \tag{4}$$

4 実験と考察

シミュレーションデータを作成し，CNN回帰モデルを学習する．次にCNN回帰モデルに対してRAM, MoRAMで可視化実験を実施する．最後に可視化結果と特徴領域のピクセル精度による評価を行う．

*MoRAM: Visual Explanation for CNN Regression Model
†Masaki Shimomura・Meiji University
‡Kazuyuki Nakamura・Meiji University

表 1: RAM と MoRAM による視覚的説明

正解値	-0.78	-0.32	0.95
元画像			
RAM			
MoRAM			

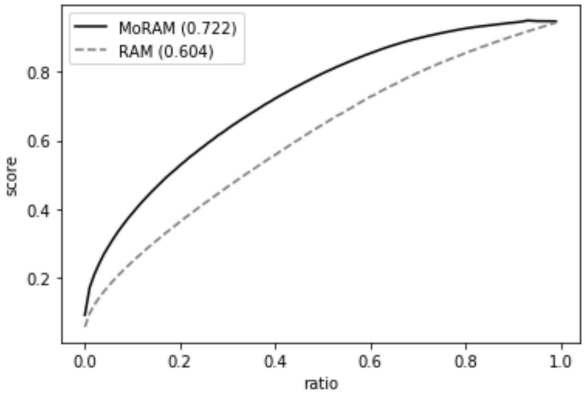


図 1: ピクセル精度 (括弧内は評価曲線下側面積)

今回のデータは矩形を，位置と面積をランダムに描画した 3 万枚の画像を使用する．目的変数は，3 万データの矩形面積を $[-\pi, \pi)$ にリスケーリングし，正弦関数に入れた値域 $[-1, 1]$ の値とする．CNN は VGG16 を基に，最後の畳み込み層の後に GAP，出力層 (サイズ:1，出力関数:恒等関数) を直列に繋いだモデルを使用する．学習により MAE が 0.01，決定係数が 0.99 の高精度モデルを作成した．このモデルに対して RAM と MoRAM を適用した結果は図 1 である．なおアップサンプリング手法は LANCZOS 法を採用した．正解値が負の場合，RAM では特徴となる矩形付近の L が低くなったが，MoRAM では矩形周辺の L が高く，適切に特徴を捉えた．一方で正解値が正の場合も RAM に比べ MoRAM では矩形領域内で L が高い領域が増え，より適切な可視化を得た．以上の定性的評価により MoRAM は適切に可視化できる手法である示唆を得た．

最後に，下記手順で定量的評価を行う．1) L をアップサンプリングしたものを $\hat{L}$ とする．2) $R = \frac{\hat{L}_{ij}}{\sum_x \sum_y \hat{L}_{xy}}$ を算出する．3) R を昇順に並べ，累積和をとり累積重要度とする．4) $r \in [0, 1]$ に対し累積重要度が r 以上のピクセルを黒，r 未満を白とした画像を作成する．5) これと矩形領域のピクセル精度を算出する．これによりある重要度以上の領域と，真の矩形領域との重複度を測定する．図 1 は評価結果である．本実験から $^\forall r \in [0, 1]$ に対して MoRAM が RAM に比べピクセル精度が高く，真の領域と予測領域の一致度を測る評価曲線の下側面積も MoRAM は RAM を上回った．以上の定量的評価から MoRAM は適切な可視化をしたと言える．

5 おわりに

本稿では CNN 回帰モデルに対する予測根拠の視覚的説明手法として MoRAM を提案し，定性的評価と定量的評価の両面から既存手法 RAM と比較をし，MoRAM は適切な視覚的説明を得ることができることを示した．本稿では正解値に負を含む場合の検証を行なったが，正解値に負を含まない場合には MoRAM と RAM の可視化結果が同一となるため，MoRAM は RAM に比べより汎用性の高い手法である．本研究ではシミュレーションデータを用いた検証であったため，今後は実現象のデータを用いた追加検証が今後の課題である．

参考文献

[1] R. Guidotti *et al.*, “A Survey of Methods for Explaining Black Box Models,” ACM Comput. Surv., 51, 5, Article 93 (January 2019), 42 pages, 2019.

[2] R.R. Selvaraju *et al.*, “Grad-cam: Visual explanations from deep networks via gradient-based localization”, In Proceedings of the IEEE International Conference on Computer Vision, pp. 618-626, 2017.

[3] A. Chattopadhay *et al.*, Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks, In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, pp. 839-847, 2018.

[4] Z. Wang *et al.*, “Diabetic retinopathy detection via deep convolutional networks for discriminative localization and visual explanation,” arXiv preprint arXiv:1703.10757, 2017.

[5] 下村真生, 中村和幸, “農作物収量予測に向けた可視化手法の適用分析事例,” 人工知能学会全国大会論文集，2020.