2P-02

モーフィング画像生成による説明可能性を有した画像分類モデルの提案

佐藤 汰樹† 小池 敦†
一関工業高等専門学校†

1 はじめに

近年，機械学習の社会利用が進んでいる．その上で，機械学習モデルのブラックボックス性は，推論結果の社会利用にあたり障壁となりうる要素である．そのため，機械学習モデルの説明可能性についての研究が盛んに行われている．

画像分類モデルに着目すると，広く使われている先行研究の手法として， LIME[1]， SHAP[2]， Grad-CAM[3] などが挙げられる．これらの手法は，出力に大きな影響を与えた特徴量の提示を説明とし，図1のようなヒートマップで可視化される．一方で，これらの手法による説明は，入力画像が分類先クラスに分類されることについて，ユーザへの説明になっていない．そのため，特に誤分類時において，ユーザのモデル出力に対する納得が得られにくいと考える．

例えば図1のように，ラベルが「1」である画像（図1a）を，分類モデルが「2」と分類した場合を考える．先行研究の手法による説明（図1b,1c,1d）では，ユーザはなぜこの入力が「2」と分類されたのか理解しにくいため，モデル出力に納得できないと考える．ユーザがモデル出力に納得するためには，「入力画像が分類先クラスに分類される可能性」を理解しやすくなる説明が必要だと考える．

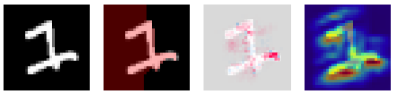
そのため本稿では，ユーザが分類モデルの出力に納得しやすくなるような説明を出力する，画像分類モデル兼説明モデルを提案する．提案手法の説明は，図2のように入力画像から分類先クラスの典型的な画像へのモーフィングにより行う．この説明により，入力画像が分類先クラスに分類される可能性をユーザが理解しやすくなり，特に誤分類時において，ユーザが分類モデルの出力に納得しやすくなると考える．

実験の結果，一定程度の分類精度を確保しつつ，他のモデル非依存な説明手法と比較して，特に誤分類時において主観的に納得しやすい説明が出力された．

2 ID-CVAE

提案手法では， Intrusion Detection CVAE (ID-CVAE)[4] を画像分類モデルとして用いる． ID-CVAE は VAE[5] や CVAE[6] に基づいており，図3に示す構造を持つ．

ID-CVAE の学習において，損失関数は KL ダイバージェンスで表されるモデル誤差 $Loss_{KL}$ と任意の再構成損失 $Loss_{Rec}$ の2種類を用いる． $Loss_{KL} + Loss_{Rec}$ を最小化するようなモデルのパラメータを，学習によって求める．



(a) (b) (c) (d)

図1: 既存手法による，一般的な分類モデルの説明の例
（a）入力とした「1」の画像，推論結果は「2」
（b）LIME （c）SHAP （d）Grad-CAM

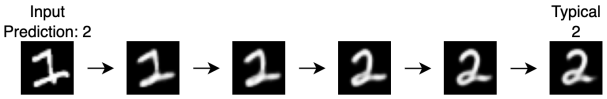


図2: 提案手法のモデルが出力する説明の例
入力は図1aと同一．推論結果は「2」.

入力 x に対するラベル l^* の推論は式(1)で行われる．ただし， p を入力次元数， f を特徴量次元数， L をラベル集合，エンコードを関数 $E(x) : \mathbb{R}^p \to \mathbb{R}^f$，デコードを関数 $D(z, l) : (\mathbb{R}^f, L) \to \mathbb{R}^p$ として表す．

$$l^* = \underset{l \in L}{\operatorname{argmin}}\, Loss_{Rec}(x, D(E(x), l)) \tag{1}$$

3 提案手法

提案手法は画像分類モデル兼説明モデルであり，入力画像が分類結果のクラスに分類される可能性を，入力画像をモーフィング（図2）して説明する．

説明生成のため， ID-CVAE を画像分類モデルとして用いる．説明に用いる連続的な画像の生成を，モデルの特徴空間内で特徴量をモーフィングすることにより行う．

入力画像のモーフィング先であるクラスの典型画像は，次の手順で生成することにした．まず，対象クラスに属するトレーニング画像それぞれをエンコードし，特徴量ベクトル集合を得る．続いて，その特徴量ベクトル全体の重心となる特徴量ベクトルを求める．最後に，重心となる特徴量ベクトルをデコードして，クラスの典型画像を得る．

入力画像 x がラベル l と分類されたときの説明生成処理を，図4 のように行う． $X_l \subset \mathbb{R}^p$ をラベル l であるトレーニン

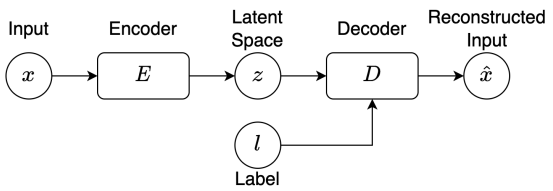


図3: ID-CVAEの構造

An Image Classification Model Featuring Explainability by Image Morphing
†Taiki Sato, †Atsushi Koike
†National Institute of Technology, Ichinoseki College

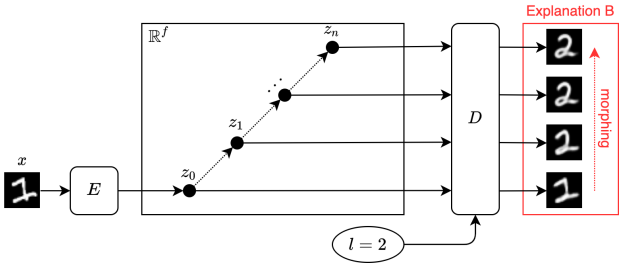


図4: 説明生成処理

グ画像集合，モーフィング画像の生成枚数を $n+1$ 枚とする．各 $i = 0, \ldots, n$ について，特徴量ベクトル z_i を式(2)で生成する．各特徴ベクトル z_i をラベル l でデコードして，モーフィング画像集合 $\{D(z_i, l)\}$ を得る．

$$z_i = \begin{cases} E(x) & \text{if } i = 0 \\ \frac{1}{|X_l|} \sum_{x_l \in X_l} E(x_l) & \text{if } i = n \\ z_0 + \frac{i}{n}(z_n - z_0) & \text{otherwise} \end{cases} \tag{2}$$

4 実験

4.1 実験設定

実験はMNISTデータセットを対象に，図5に示すモデルで行った．すべての畳み込み層において，カーネルサイズは 4×4， SamePaddingを設定した． LeakyReLUのパラメータは 0.01 とし，特徴量空間の次元数は 16 とした．また，再構成損失関数 $Loss_{Rec}$ は交差エントロピー誤差とした．

学習は最適化手法をAdam，学習係数を 1e-4 で行った．MNISTデータセットのトレーニング画像 60000枚を 9:1 に分割し，それぞれ学習用，検証用として用いた．学習済みモデルの評価に用いるテストデータセットは， MNISTデータセットのテスト画像 10000枚を用いた．

4.2 実験結果

検証用データの損失を指標とした EarlyStopping により，学習は 124 エポックで終了した．学習済みモデルによるテストデータの分類正解率は 98.13% だった．

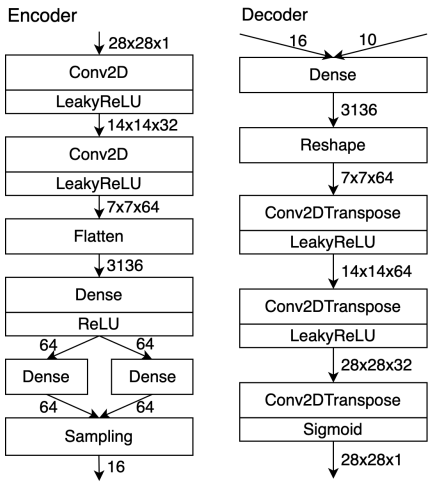


図5: 実験に使用したモデル

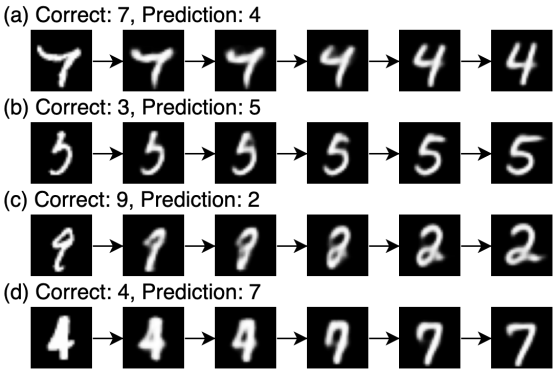


図6: 誤分類時に出力された説明の例

特に誤分類した例について，モデルが出力した説明を図6に示す．主観的に判断すると，図6a,bのように納得しやすい説明が出力された場合もあるが，図6c,dのように説明に違和感がある事例も散見された．

5 おわりに

本稿では，画像分類問題に対して，入力画像から分類先クラスの典型的な画像へのモーフィングにより，ユーザが分類モデルの出力に納得しやすくなるような説明を出力する，画像分類モデル兼説明モデルを提案した．

今後は分類精度の向上や，違和感が少ないモーフィング画像の出力，他の適用可能な問題の発見に取り組みたい．

参考文献

[1] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. pages 97–101, 02 2016.

[2] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. 12 2017.

[3] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[4] Manuel Lopez-Martin, Belén Carro, Antonio Sanchez-Esguevillas, and Jaime Lloret. Conditional variational autoencoder for prediction and feature recovery applied to intrusion detection in iot. *Sensors*, 17:1967, 08 2017.

[5] Diederik Kingma and Max Welling. Auto-encoding variational bayes. 12 2014.

[6] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.