変分情報ボトルネックを用いた画像分類による視覚的説明性に関する研究

○小松 翔太，　平川 翼，　山下 隆義，藤吉 弘亘

○ Shota KOMATSU，Tsubasa HIRAKAWA,
Takayoshi YAMASHITA, and Hironobu FUJIYOSHI

中部大学
{komatsusyouta, hirakawa}@mprg.cs.chubu.ac.jp, {takayoshi, fujiyoshi}@isc.chubu.ac.jp

＜要約＞ Class Activation Mapping における説明性向上を目的として，変分情報ボトルネックを導入した画像分類法を提案する．提案手法により，分類クラスに関連する情報のみを保持するように特徴抽出を行うことで，より明瞭な Attention Map の獲得を目指す．SUN397 データセットを用いた評価実験により，重要な領域により注目できることを示す．
＜キーワード＞画像分類，視覚的説明性，変分情報ボトルネック

1　はじめに

機械学習は，学習したモデルの推論を行い結果を出力するが，その出力結果の根拠を人間が理解することは困難である．モデルの判断根拠を示すことは，モデルの信頼性向上に繋がる．判断根拠の視覚的な可視化手法として，Class Activation Mapping (CAM) [1] や Grad-CAM [3] など，様々な手法が提案されている．しかし，CAM は本来の注目クラス以外のクラスにも同時に反応するという問題がある [2]．ただし，問題点を指摘した ReCAM [2] は，入力画像に対して CAM を用いて生成した Attention Map を用いて算出した特徴マップに，再度 CAM を用いて Attention Map の可視化を行うことで高品質な Attention Map を獲得している．そのため，根本となる CAM は改善されておらず，CAM において発生している問題点の改善には至っていない．

そこで，本研究では変分情報ボトルネック (VIB) [4] を導入した画像分類法を提案する．VIB で定義される目的関数を用いることで，入力データの情報を圧縮しつつ，データに含まれる分類クラスに関する固有の特徴を捉える効果が得られる．これにより，分類クラスに関連した固有の特徴のみ捉える Attention Map を獲得することで，CAM における説明性の改善を目指す．

2　先行研究

2.1　判断根拠の可視化手法

判断根拠の可視化手法として，学習済みのモデルに対して画像を入力して推論を行い，推論時に貢献度が高い領域を Attention Map として可視化を行う方法がある．Class Activation Mapping (CAM) [1] とは，ネットワーク出力に対する全結合層の結合重みと特徴マップを用いることで推論時に貢献度が高い領域を Attention Map として可視化する手法である．CAM の可視化構造を図 1 に示す．CAM において，特徴マップの値を計算する際，Global Average Pooling(GAP) という方法が用いられている．これにより，各特徴マップを平均値として出力することにより特徴マップの次元数を削減でき，パラメータ数の削減に貢献している．

しかし，CAM では本来の注目クラス以外の複数クラ

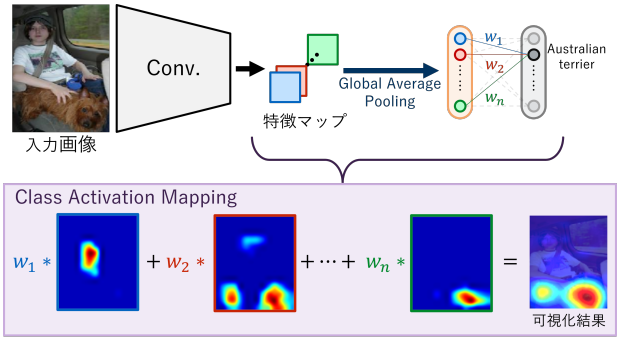

図 1　CAM のネットワーク構造

スにも同時に反応する可能性がある [2]．ただし，Re-CAM [2] は，入力画像に対して CAM を用いて生成した Attention Map を用いて算出した特徴マップに，再度 CAM を用いて Attention Map の可視化を行うことで高品質な Attention Map を獲得するアプローチである．そのため，根本となる CAM は改善されておらず，CAM において発生している問題点の改善には至っていない．

2.2　変分情報ボトルネック

変分情報ボトルネック (VIB) [4] とは，データ $x \in X$ とラベル $y \in Y$ を与えた際にラベル Y に関する情報を保持しつつ小さな潜在空間 $z \in Z$ に圧縮する手法である．具体的には，式 (1) の目的関数を最大化するネットワークパラメータ θ を推定する．

$$R_{IB}(\theta) = I(Z, Y; \theta) - \beta I(Z, X; \theta) \tag{1}$$

ここで，β はラベル Y の情報保持とデータ X の情報圧縮のバランスを調整するハイパーパラメータである．式 (1) の空間 Z に対する周辺分布を変分近似して整理した時，式 (2) の目的関数の最適化問題に帰着する．

$$\begin{aligned} J_{VIB}(\theta) = \mathbb{E}_{\epsilon - p(\epsilon)}[-log\, q(y_n | f(x_n, \epsilon))] \\ + \beta\, KL[p(z|x_n), r(z)] \end{aligned} \tag{2}$$

3　提案手法

本研究は，CAM を用いた Attention map の説明性の改善を目的としている．目的実現のため，変分情報ボトルネック (VIB) を導入した画像分類法を提案する．提案手法では，VIB の目的関数を用いることで，入力データの情報を可能な限り圧縮しつつ，分類クラスに関する情報を保持する．これにより，分類クラスに関連する特徴のみを抽出し，各クラスを表現する重要な特徴を捉える Attention Map を獲得することで説明性の改善を図る．

3.1　ネットワーク構造

提案手法のモデル構造を図 2 に示す．具体的に，まず入力画像を畳み込み層に入力し，特徴マップを獲得する．次に，特徴マップから Global Average Pooling (GAP) により平均 μ を算出する．これは，特徴マップごとの個別の特徴を表現するベクトルとして定義する．

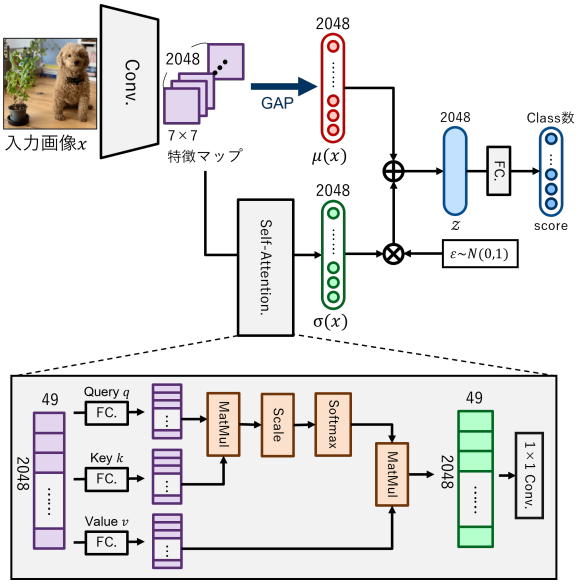

図 2　提案手法のネットワーク構造

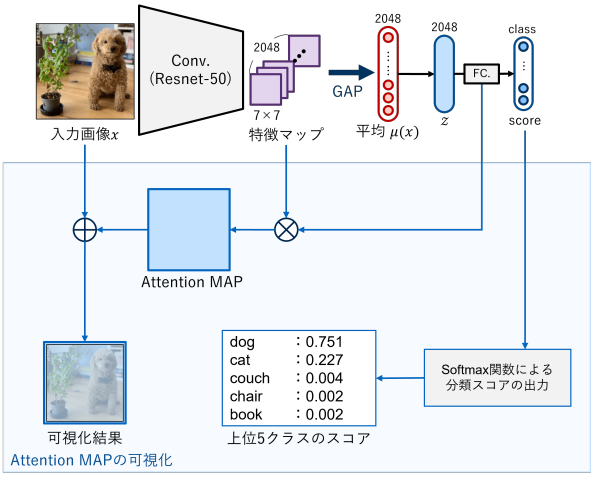

図 3　推論・可視化時のネットワーク構造

また，特徴マップ間の Self-Attention を求めて分散 σ を推定する．特徴マップ間の関連度を分散と仮定することで，各特徴マップが自己を含めたどの特徴マップに注目しているか計算できる．ここで，分散 σ に対して正規分布に従った乱数を乗算することで微分可能な定数にする Reparameterization trick [5] を適用する．これにより，ネットワーク全体の学習を可能にしている．次に，平均 μ と分散 σ に従う正規分布によって圧縮表現 z を獲得した後，出力層にて計算を行い各クラスに対するスコアを獲得する．ここまでで算出した各パラメータを用いて，VIB の目的関数である式 (2) により誤差を求めることで学習を行う．

3.2 推論処理と Attention Map の可視化

推論時には Self-Attention を用いずに推論を行う．これは，推論時は特徴マップ間の関連度を再度計算する必要がないためである．Attention Map は，各クラスに対するスコアから分類予測の結果を出力し，推論時の特徴マップと全結合層の結合重みを用いて出力する．

4 実験

本実験では，提案手法と変分情報ボトルネック (VIB) を導入していないモデルによる比較を行う．具体的には，分類精度の比較と獲得される Attention Map の可視化結果に対する比較を行う．

4.1 実験条件

データセットには，SUN397 データセット [6] を使用する．SUN397 は 397 種類のシーンカテゴリから推論を行うデータセットである．このデータでは，シーンを構成する様々なオブジェクトから推論を行う必要があるため，可視化時に分類クラスに関係した情報のみを抽出できているか調査するのに適したデータであると考える．全データのうち，18,050 枚を train 用データ，18,050 枚を validation 用データとして用いる．

実験条件として，学習回数は 100 Epochs，BatchSize は 256，学習率は 0.1 と設定し誤差関数には Cross Entropy Loss，ネットワークには ResNet-50 を用いる．

変分情報ボトルネックのハイパーパラメータである β は，$10^0 \sim 10^{-7}$ に設定する．β の数値が大きいほど入力画像のデータをより小さい潜在空間の情報として圧縮して獲得することが可能である．ここで β に一定の値の範囲を設けて設定する理由は，変分情報ボトルネックの β には適切なパラメータサイズが存在しており，適切なパラメータを選択することでデータに対する頑健性を獲得する効果が得られるためである [4]．

4.2 実験結果

各モデルの分類精度と従来手法の Resnet-50 を 100% とした場合の精度の性能差を表 1 に示す．従来手法に比べ提案手法は 1.0 ∼ 3.0pt ほど精度が低下しており，性能では −2.0 ∼ −7.0pt ほどの精度差が確認できる．ただし，本研究では可視化時の視覚的説明性の改善を目的としているため，改良に伴い低下した精度の改善は今後の課題とする．

ここで，提案手法のモデルに適切であると考えられるパラメータ β のモデルを 1 つ選択して Attention Map の可視化実験を行うが，今回は精度比較実験において精度が比較的高く，可能な限り大きい β である $\beta = 10^{-1}$ のモデルを用いる．

表 1　精度比較

モデル	β	分類精度（性能差）[%]
ResNet-50	–	53.37 (100.00)
VIB-ResNet-50	10^0	50.60 (94.81)
	10^{-1}	52.14 (97.70)
	10^{-2}	50.28 (94.21)
	10^{-3}	51.92 (97.28)
	10^{-4}	52.15 (97.71)
	10^{-5}	49.79 (93.29)
	10^{-6}	52.31 (98.01)
	10^{-7}	51.67 (96.81)

次に，waiting_room と television_studio の 2 クラス各 3 枚を入力画像として用いた Attention Map の可視化結果を図 4 に示す．クラスごとの各画像で共通する判別に重要な定性的特徴について，次のような特徴が挙げられる．

・**waiting_room** 大量の椅子が並んだ特徴

・**television_studio** テレビカメラ，コメンテーター

図 4 の可視化結果について，提案手法では比較的重要な特徴領域に注目できていることが確認できる．また，従来手法では“部屋の角”や“カーテン”といった，屋内で共通して存在しており分類クラスの判別に効果的でない特徴領域にも部分的に注目していることが確認できるが，提案手法ではそれらの特徴への注目が少ないことが確認できる．よって，提案手法では従来手法に比べ分類クラスに関する情報のみを捉えやすくなり，重要な領域に注目できると考えられる．

ここで，可視化結果は改善されたものの分類には失敗した例について，lake のクラスから 1 枚を入力画像として与えた際の結果を図 5 に示す．ここで，lake における判別に重要な定性的特徴について，次のような特徴が挙げられる．

・**lake** 広大な水域と山や森などの自然の背景

可視化結果では，提案手法において湖や湖の水が反射した背景に注目されており，湖でなく山を注目してい

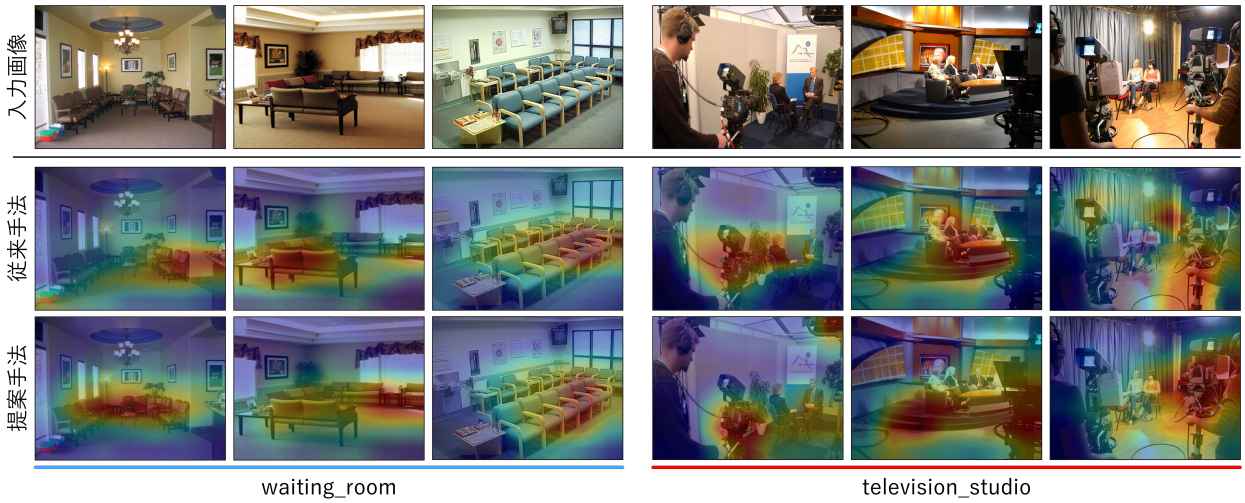

図 4　Attention Map の可視化結果

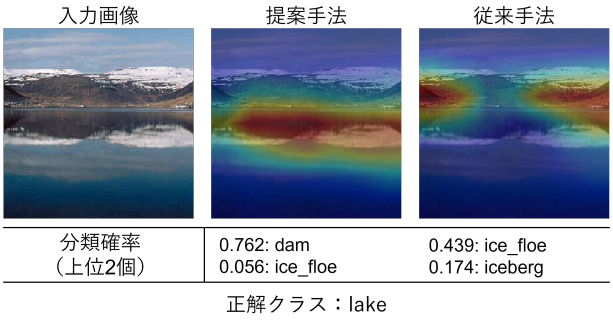

図 5　可視化結果のみ改善された結果

図 6　lake と ice_floe の類似例

る従来手法より重要な特徴に注目していることが確認できる．しかし，分類結果では dam や ice_floe が出力されており，分類に失敗したことがわかる．両者で出力されている ice_floe について画像を図 6 に示すが，図 5 における提案手法や従来手法の注目領域から，水に反射した背景に映った雪山の特徴を誤って認識したと考えられる．

また，比較的類似したクラスに分類したことで失敗した例について，forest_path, forest_road, bamboo_forest の 3 クラス各 1 枚を入力画像として与えた際の可視化

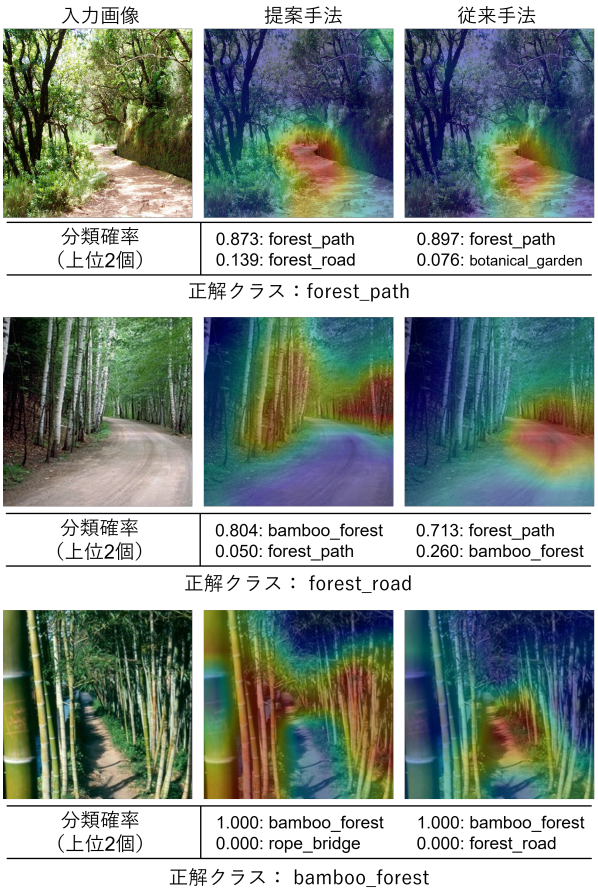

図 7　比較的類似したクラスによる可視化結果

結果を図 7 に示す．各クラスの判別に重要な定性的特徴について，次のような特徴が挙げられる．

- **forest_path**　森や林などの自然の背景と小道
- **forest_road**　森や林などの自然の背景と道路
- **bamboo_forest**　竹林

図 7 から，特に forest_road で誤分類が発生しており，同じ”道”や”森”が特徴となりやすい forest_path や，森のような竹林が特徴となる bamboo_forest に分類されてしまったと考えられる．このように背景画像には類似した場合も多く，SUN397 では細かくクラスが分割されているために分類タスクが高難易度となり，表 1 において全体的な精度へ影響したと考えられる．

4.2.1　アンケート実験

ここで，提案手法により Attention Map が改善されたか定量的に比較する．具体的に，30 名を対象に図 4 や図 5，図 7 を含む 15 個のデータに対する可視化結果を示し，提案手法か従来手法についてどちらの Attention Map が良い結果であるかアンケート実験を行った．実験結果を表 2 に示す．結果より，提案手法の Attention Map の方が良好であるという結果が得られた．

表 2　15 個の可視化結果によるアンケート実験

提案手法が良好と判別されたデータ数	割合 [%]
11 / 15	73.3

5　まとめ

本研究では，CAM における Attention Map の視覚的な説明性改善のため，変分情報ボトルネックを導入した画像分類法を提案した．提案手法では，変分情報ボトルネックの目的関数を用いることで，入力データの情報を可能な限り圧縮しつつ，分類クラスに関する情報を保持する．これにより，分類クラスに関連する特徴のみを抽出し，各クラスを表現する重要な特徴を捉える Attention Map を獲得する．Attention Map の可視化実験から，提案手法では変分情報ボトルネックを導入していない場合に比べて，画像中から分類クラスの判別に重要な特徴により注目した結果が得られていることを確認した．また，27 名を対象にしたアンケート実験より，提案手法を用いることで Attention Map の可視化結果が視覚的に改善されたことがわかった．

今後の予定としては，分類精度の向上やより詳細な特徴の分析などが挙げられる．

参考文献

[1] B. Zhou *et al.*, “Learning Deep Features for Discriminative Localization”, CVPR, 2016.

[2] Z. Chen *et al.*, “Class Re-Activation Maps for Weakly-Supervised Semantic Segmentation”, CVPR, 2022.

[3] C. Wang *et al.*, “Grad-cam: Visual explanations from deep networks via gradient-based localization”, ICCV, 2017.

[4] A. Alemi *et al.*, “Deep Variational Information Bottleneck”, ICLR, 2017.

[5] D.P. Kingma *et al.*, “Auto-Encoding Variational Bayes”, ICLR, 2014.

[6] J. Xiao *et al.*, “SUN Database: Exploring a Large Collection of Scene Categories.” IJCV, Vol.119, pp.3–22, 2016.