Grad-CAM に基づく視線一貫性指標の定義と深層学習モデルへの適用

米村 慶太[†]
[†] 有明工業高等専門学校 創造工学科 情報システムコース
Gauthier Lovic 研究室

1 背景と目的

深層学習モデルの判断根拠を可視化する手法として Grad-CAM [1] は広く用いられるが, 可視化結果の安定性を定量評価する標準指標は少ない．本研究は，同一クラス内で Grad-CAM がどれだけ一致するかを「**視線一貫性**」として定義し，

1. モデルの一貫性指標と性能の関係性の分析
2. 学習時正則化項としての利用

を通じて，その有用性を検証する．

2 視線一貫性指標の定義

2.1 Grad-CAM の概要

Selvaraju らによって提案された Grad-CAM は, 画像分類モデルの予測に寄与する領域を可視化し，**モデルの「視線」**を理解するための手法である．畳み込みニューラルネットワーク（CNN）において, 対象とするマップ $A^k (k = 1, ..., M)$ を持つ畳み込み層の出力と，クラス c に対するスコア y^c の勾配を用いて，

$$\begin{aligned} \alpha_k^c &= \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{i,j}^k} \\ \mathrm{CAM}^{(c)} &= \mathrm{ReLU}\left(\sum_k \alpha_k^c A^k \right) \end{aligned} \tag{1}$$

のように計算される．ここで，Z は特徴マップの空間サイズに基づく正規化定数であり, M は対象層における特徴マップ数を表す．

本研究では, 空間分布の形状差に着目するため, CAM を $[0, 1]$ に正規化して扱う．

2.2 指標の定義

データセット $D = \{(x_i, y_i)\}_{i=1}^{|D|}$ に対し，クラス c に属し，かつモデルが正しく分類したサンプル集合を

$$D_c^+ := \{(x_i, y_i) \in D \mid \hat{y}_i = y_i = c\} \tag{2}$$

と定義する．D_c^+ に含まれる各サンプル x_i について，クラス c に対する Grad-CAM を計算し，その平均マップを

$$\overline{\mathrm{CAM}}_D^{(c)} := \frac{1}{|D_c^+|} \sum_{x_i \in D_c^+} \mathrm{CAM}^{(c)}(x_i) \tag{3}$$

と定める．

この平均マップは，**当該クラスにおいてモデルが共通して参照していると解釈可能な「代表的視線」**を表す．各サンプルに対する視線一貫性損失は，

$$\ell_{\mathrm{CAM}}(x_i, c) := \frac{1}{H'W'} \left\| \mathrm{CAM}^{(c)}(x_i) - \overline{\mathrm{CAM}}_D^{(c)} \right\|_2^2 \tag{4}$$

で与えられ，H', W' は Grad-CAM マップの空間解像度である．

データセットに対する視線一貫性指標 $\mathcal{L}_{\mathrm{CAM}}$ は，これらの損失のサンプルごと・クラスごとの平均として定義される．

本指標は，**同一クラス内における説明のばらつき**を直接的に測定する点に特徴がある．

2.3 学習時正則化への応用

提案した視線一貫性指標を，学習過程における制御項として用いる．分類損失 L_{cls} に対し，一貫性正則化項 $\mathcal{L}_{\mathrm{CAM}}^{\mathrm{train}}$ を加えた総損失関数を

$$\mathcal{L} := L_{\mathrm{cls}} + \lambda \mathcal{L}_{\mathrm{CAM}}^{\mathrm{train}} \tag{5}$$

と定義する．ここで $\mathcal{L}_{\mathrm{CAM}}^{\mathrm{train}}$ は，学習用に最適化された視線一貫性指標であり，基本的には**データセット全体ではなく，一つのミニバッチに対して定義されるもの**とする．

平均マップ $\overline{\mathrm{CAM}}_D^{(c)}$ は，ミニバッチ間のばらつきを抑制するため**指数移動平均（EMA）**により更新する．

また，学習初期における不安定な Grad-CAM に過度に拘束されることを避けるため，エポックの進行に応じて正則化係数 λ を段階的に増加させる**カリキュラム学習**を導入した．

3 実験

Chest X-Ray および Brain MRI データセットを用い，CNN の深さを段階的に変化させた複数のモデルに対して評価を行った．各モデルについて分類精度と視線一貫性指標を算出し，両者の関係性を分析した．さらに，提案した正則化項を導入した場合の Grad-CAM の空間分布および性能への影響を検証した．[1]

実験の結果，

- 分類性能と視線一貫性の間には，**必ずしも相関は存在しない**
- 正則化により Grad-CAM の空間的一貫性は向上する一方で，分類性能の改善は限定的である
- いずれも，**データセットに依存した挙動を示す**

ことが確認された．

4 結論

本研究では, Grad-CAM を単なる可視化手法としてではなく，**定量的に評価・制御可能な対象**として扱う枠組みを提案した．クラス内視線一貫性指標により，説明の安定性を客観的に分析できることを示した．

実験結果から，一貫性の向上が必ずしも性能向上に直結しないことが明らかとなり, XAI における評価指標の多面的な検討の必要性が示唆された．提案手法は，データセットに依存して有効性が大きく異なる手法であり，「**同一クラスにおいて注目領域が比較的一貫して定義可能なタスク**」において有効である可能性が示唆された．

参考文献

[1] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, と D. Batra, 「Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization」, *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 10 月 2019, doi: 10.1007/s11263-019-01228-7.

[1]本研究で使用した実装および関連する資料は，GitHub 上で公開している．
URL: https://github.com/59GauthierLab/Yonemura_Research_Public