[10/22 中間報告]

AIの判断根拠の可視化による信頼性向上に関する研究

ゴーチェ研究室 / 米村慶太

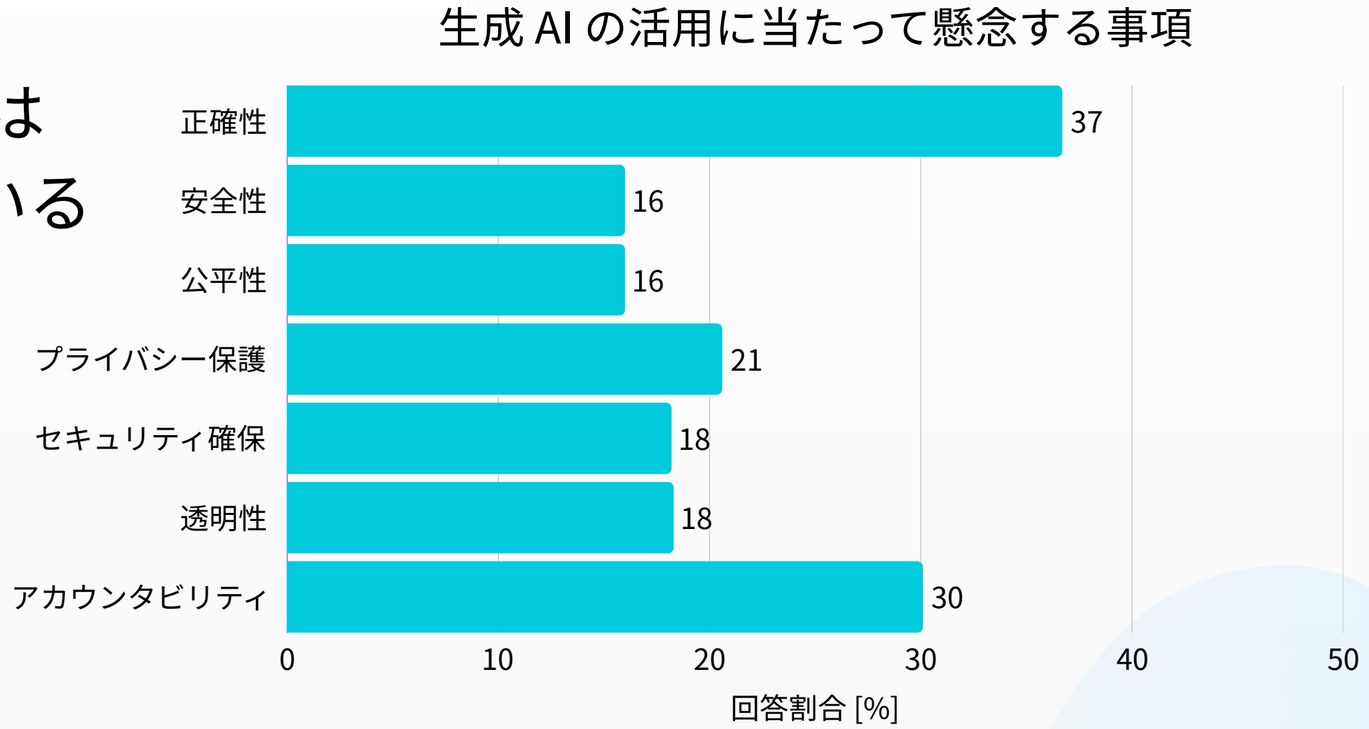
[目次]

[1. 研究背景]

近年、 画像認識をはじめとするAI は
多くの分野で活用されている

- 製造
- 医療
- 自動運転 など

一方、課題も...



三菱総合研究所作成「生成AIの信頼に関するアンケート調査」（2024年6月）

- AIは「なぜその判断をしたのか」が人間に分かりにくい（**ブラックボックス問題**）
- 特に医療や安全性が求められる分野 → **判断根拠の説明性（Explainability）**が重要

[2. 研究目的]

- 画像分類AIの判断根拠を可視化し、これを元にしたモデルの定量的な評価や精度の向上を行う

→ AIの信頼性（**「判断根拠の説明性」**）向上の1つのアプローチ

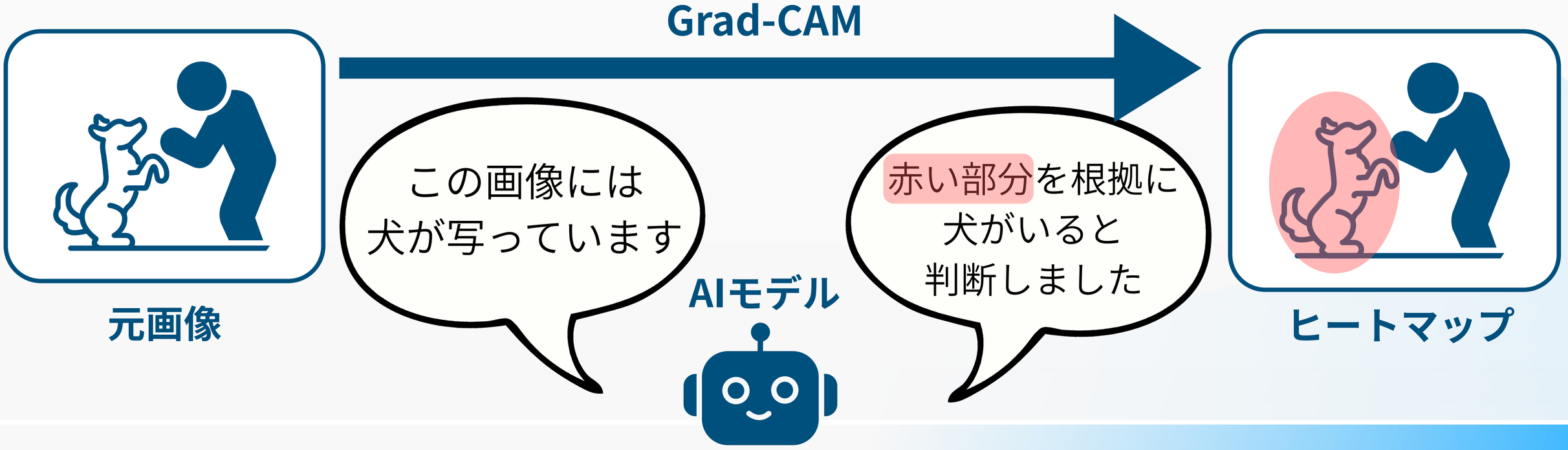
- 可視化には**Grad-CAM**を利用

→ 画像分類AIの判断根拠の可視化として利用

[3. 事前知識]

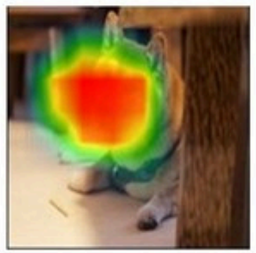
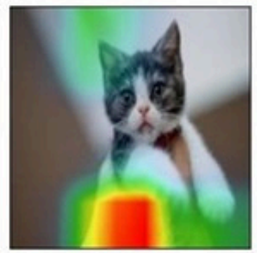
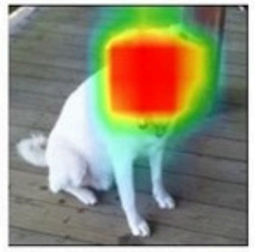
Grad-CAM: Gradient-weighted Class Activation Mapping

- CNN（画像認識タスクなどに広く使われているモデル）が予測に用いる箇所をヒートマップを使って可視化する手法
- AIの判断根拠を視覚的に理解するために利用される



[3. 事前知識]

実際の例: 正しく学習できていないモデルの Grad-CAM

No.	犬		猫	
	入力画像	+CAM画像	入力画像	+CAM画像
1				
2				

左の表を見ると...

- 正しい判断時

→ 注目領域が対象物（犬）に集中

- 誤分類時

→ 背景などの誤った箇所に注目

- Grad-CAMの出力結果からモデルの正確性を推定できる？
- 誤った判断根拠を減らすように学習することで精度向上が期待できる？

CRAFT WORK, Grad-CAMを用いたクラス分類モデルの分析 (https://craft-server.co.jp/news/179/) より引用

[4．仮説]

[仮説 1]

- 構図の類似した画像において、分類モデルが正解を出す際にはGrad-CAMの注目領域に**空間的一貫性** が見られる

→ モデルが同種の特徴に基づいて正しく判断していることを示す可能性

[仮説 2]

- 正解画像の 代表的な注目分布（平均的Grad-CAM分布）と、個々の画像のGrad-CAMの整合性を損失関数に組み込むことで
  - 分類性能
  - 視線の一貫性

  の双方が改善されると考える

[4. 仮説]

例: レントゲン写真から肺がんがあるか判定するAI

- レントゲン画像は構図がほとんど変わらない
- AさんでもBさんでも注目すべき部分は同じはずという仮定

→ 判断根拠(視線)の空間的位置には一貫性が見られるはず

→ 一貫性が見られるように学習させれば精度向上が期待できる

[5．研究手法の概要]

1. CNNによる画像分類モデルを構築（ResNetやVGGなど）

→ **Pytorch, pytorch-gradcam** を用いて構築

2. Grad-CAMを用いて、各画像に対する注目領域を可視化
3. 類似構図の画像群に対して、Grad-CAMの空間的類似度を算出

$$L_{\text{cam}} = \|\text{CAM}_c(\mathbf{x}_i) - \overline{\text{CAM}_c}\|_2^2$$

4. 損失関数に「注目領域の整合性項」を追加したモデルを検証

$$L_{\text{total}} = L_{\text{cls}} + \lambda \cdot L_{\text{cam}}$$

L_{cam} : 通常の分類誤差(クロスエントロピーなど)

L_{cls} : Grad-CAM分布の整合性損失

[6. 進捗と今後の展望]

現在の進捗:

- Grad-CAMを実装・基本的なCNN分類モデルで可視化を確認
- いくつかのデータセットで判別と視覚化
  - **CIFAR-10**：飛行機、猫、自動車など
  - **Stanford Cars**：構図一貫性が高い
- 今回の研究では構図の類似した画像であることが重要
  → 最適な画像群を探している
- 正解データでも予想以上に注目
  領域が散らばっていた
  → 領域を狭くするとどうなる？

[6. 進捗と今後の展望]

今後の展望:

- 損失関数への整合性項の導入と精度比較実験
- Grad-CAMの改良版: Grad-CAM++ などのモデルによる検証
- 代表値とは別アプローチからの損失関数の検討
  - 注目領域のばらつきを抑える方向に
  - 逆に注目領域のばらつきを増やす
  - 前述の学習方法のハイブリット型やカリキュラム学習