「説明可能AIによるAI高信頼化」と「信頼できる説明可能AI」について

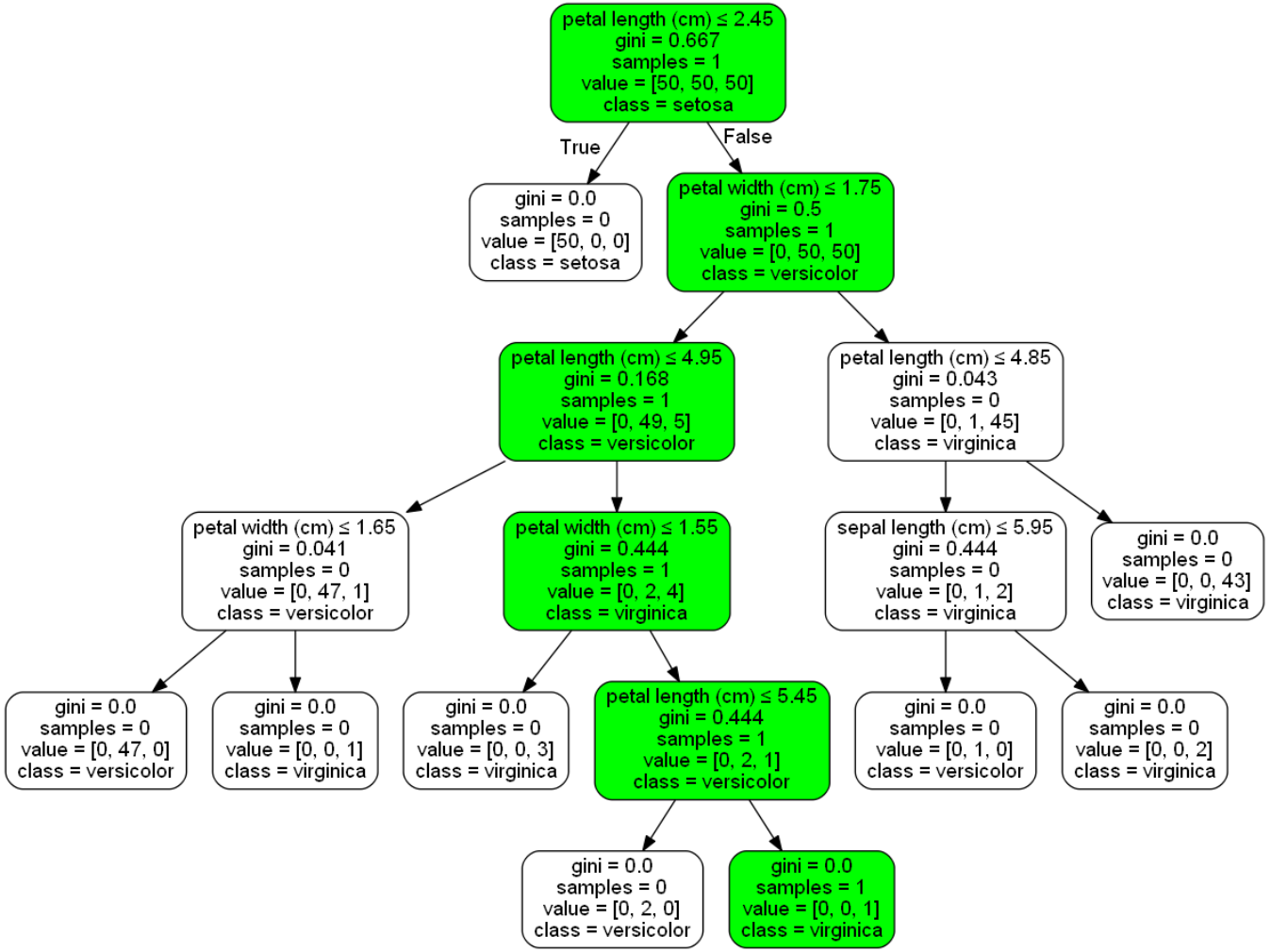
理研AIP / 東工大 情報理工学院
佐久間 淳

分類モデルにおける説明とは

- 説明付き分類器　$f : \mathbb{R}^D \to \{0,1\} \times E$（分類結果：$\{0,1\}$、説明：$E$）
- 説明とは
  - 入力(データ)と出力(分類結果)の関係を明らかにする補助情報
  - 分類結果のみを受け取るより, 判断の信頼性が増す

  → 「説明可能AIによるAI高信頼化」
- 分類モデルの計算過程が単純なら…
  - 計算過程全体が人間に解釈可能(線形モデル、決定木)
  - 計算過程自体が説明となる
- 分類モデルの計算過程が複雑なら…
  - 計算過程が解釈不可能(NN、GBDT)
  - 解釈可能な一側面を説明として切り出す

決定木は自明な説明を与える



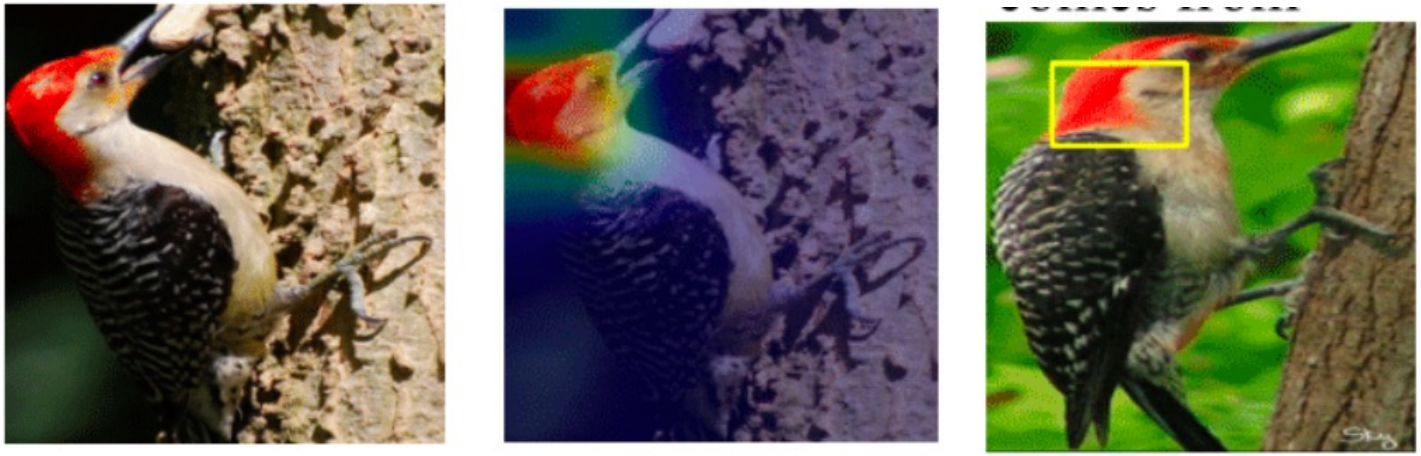
https://stackoverflow.com/questions/55878247/how-to-display-the-path-of-a-decision-tree-for-test-samples

分類モデルにおける説明とは

- 説明付き分類器　$f : \mathbb{R}^D \rightarrow \{0,1\} \times E$（分類結果、説明）
- 説明とは
  - 入力(データ)と出力(分類結果)の関係を明らかにする補助情報
  - 分類結果のみを受け取るより, 判断の信頼性が増す

  → 「説明可能AIによるAI高信頼化」
- 分類モデルの計算過程が単純なら…
  - 計算過程全体が人間に解釈可能(線形モデル、決定木)
  - 計算過程自体が説明となる
- 分類モデルの計算過程が複雑なら…
  - 計算過程がブラックボックス (NN、 GBDT)
  - 解釈可能な一側面を説明として切り出す

代表的な説明手法



- 特徴(画素)帰属「分類結果に影響を与えた特徴は頭の部分」
- 見本による説明「頭の部分がこの見本に似てるから」
- 言語化された説明「頭が赤いから」
- 反事実仮想による説明「もし頭が黒かったら…」
- 解釈可能コンセプトによる説明「頭の色を表す潜在変数」

Chen et al., This Looks Like That: Deep Learning for Interpretable Image Recognition. NeurIPS'19

特徴帰属 (GradCAM)

(a) Original Image　　(c) Grad-CAM ‘Cat’

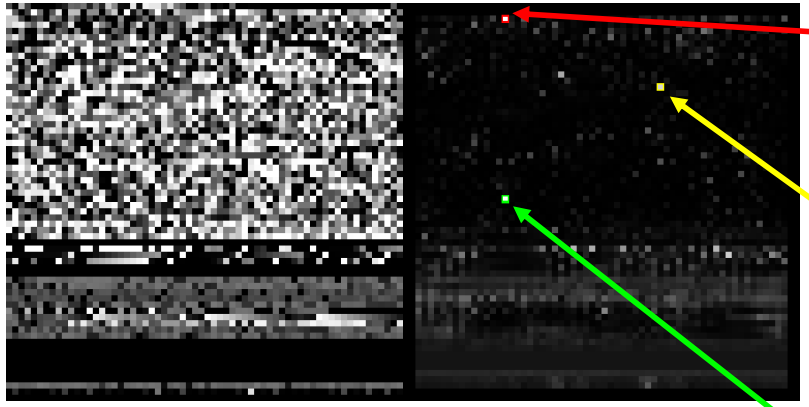
- 猫と判断するために重視した領域の提示
- 「この部分を見て猫と判断しました」

Selvaraju et al., Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. ICCV’17

特徴帰属によるマルウェア機能分析
[YS, ACM Computers & Security 2019]

- マルウェアはその機能が簡単にはわからないように作成されている
- 作成は簡単だが, 機能解析には手間がかかる
- マルウェアバイナリを画像化した上で, マルウェア分類器を作成
- 「AIがどこを見てマルウェアを分類したか」を手がかりに,マルウェアの機能分析を効率化

Win32.Gaobotの解析例



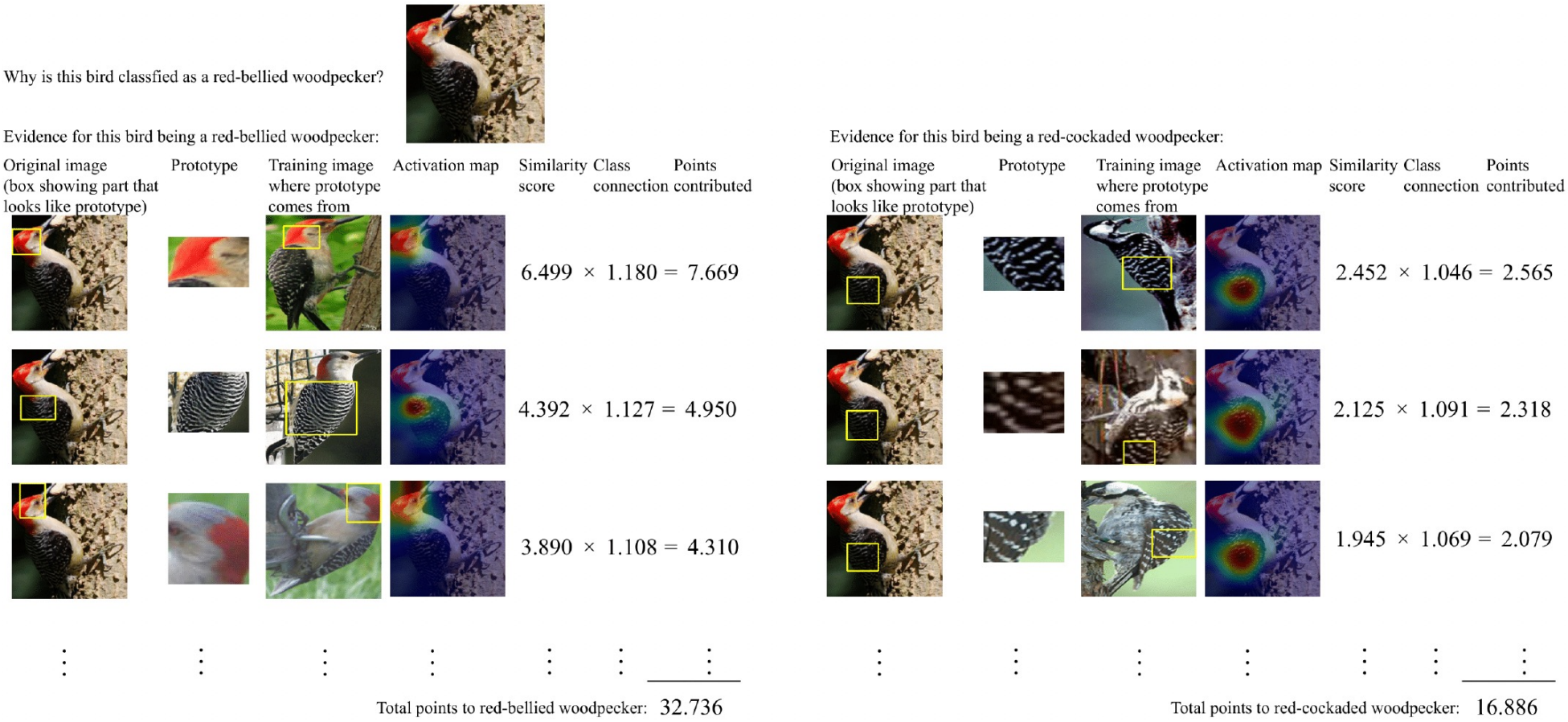
IRCサーバーに接続してチャットルームに入る機能

DDoS攻撃を行うために、パケットを指定された宛先にリダイレクトする機能

傍受したHTTP通信の内容に 文字列 "PAYPAL "が含まれているかどうかを確認する機能。

Neural Malware Analysis with Attention Mechanism, ACM Computers & Security, Issue C, Vol. 87, 37 pages, 2019.

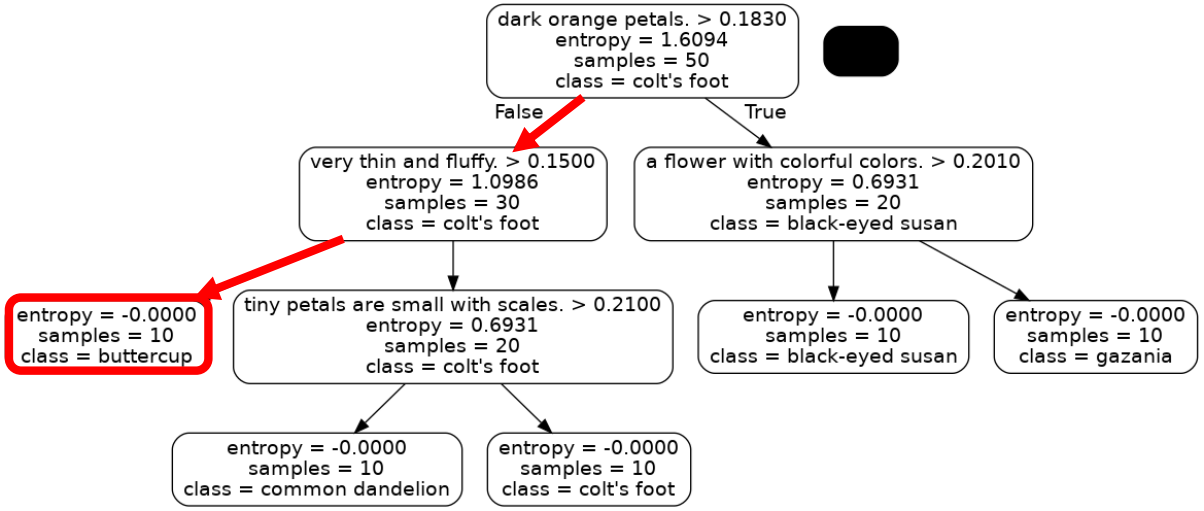
見本による説明 (ProtoPNet)



- 別の画像の類似した部分を列挙することによる説明
- 「ここが別のキツツキ画像のこの領域と似ているからキツツキと判断しました」

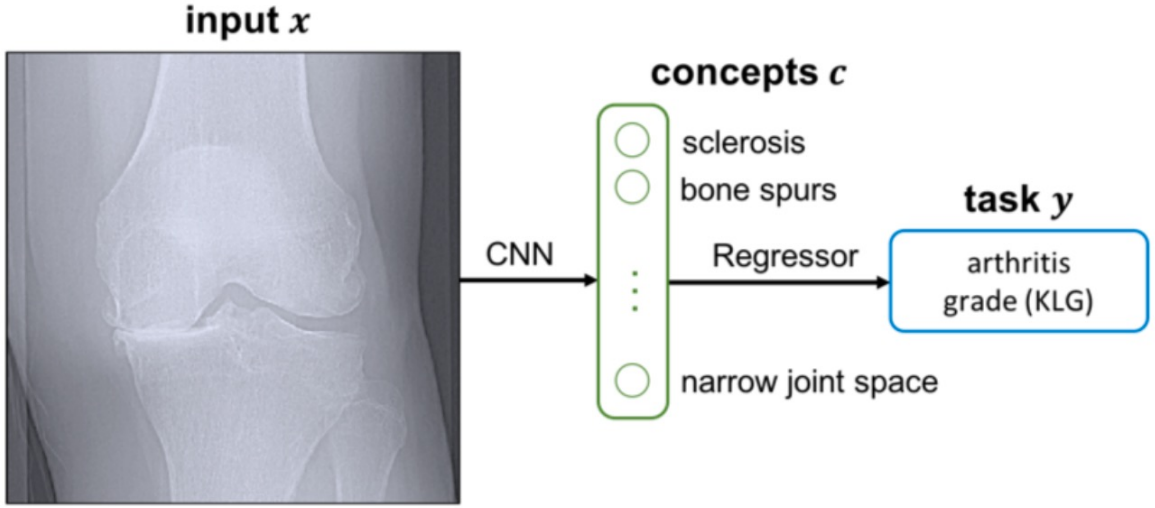
Chen et al., This Looks Like That: Deep Learning for Interpretable Image Recognition. NeurIPS'19

自然言語による説明 [SFAS'22, under review]



- 異なるクラスを識別できるフレーズを言語生成モデルで発見的に探索
- 発見フレーズと画像の類似度で決定木を構成
- 「濃いオレンジ花弁では<u>ない</u>」かつ「薄くふわふわしている」→butter cup

コンセプトに基づく説明



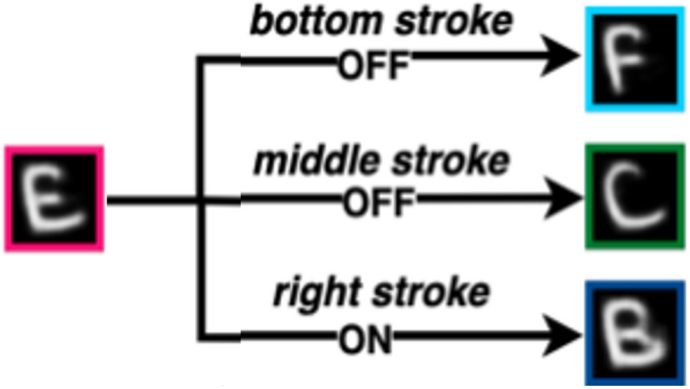
過程は解釈できないが
結果は解釈できる特徴抽出

過程が解釈できる線形分類器

- 「膝関節の間が狭く」、「骨棘がある」から変形性膝関節症の重症度が高い
- 各画像に各コンセプト値を教師ラベルとして割り当てる必要がある

Pang Wei Koh, et al. Concept bottleneck models. In ICML, 2020.

反事実による原因の説明 [TFAS, AAAI'22]



$$\mathcal{L}(X) = \frac{1}{|X|} \sum_{\mathbf{x} \in X} [\mathcal{L}_{\text{VAE}}(\mathbf{x}) + \lambda_{\text{R}} \mathcal{L}_{\text{R}}(\mathbf{x})] + \lambda_{\text{CE}} \mathcal{L}_{\text{CE}}(X).$$

潜在変数を求めるVAE

潜在変数を単純化する正則化

潜在変数の変化と識別ラベルの相互情報量を最大化

- クラスの違いを特徴づけるコンセプト(e.g., 下に線がある/ない)を生成モデルで学習 (コンセプトの提示不要)
- 反事実仮想による説明例: "E"は
  - 下に線がある (もしなかったらFになる)
  - 真ん中に線がある (もしなかったらCになる)
  - 右に線がない (もしなかったらBになる)

から、"E"である

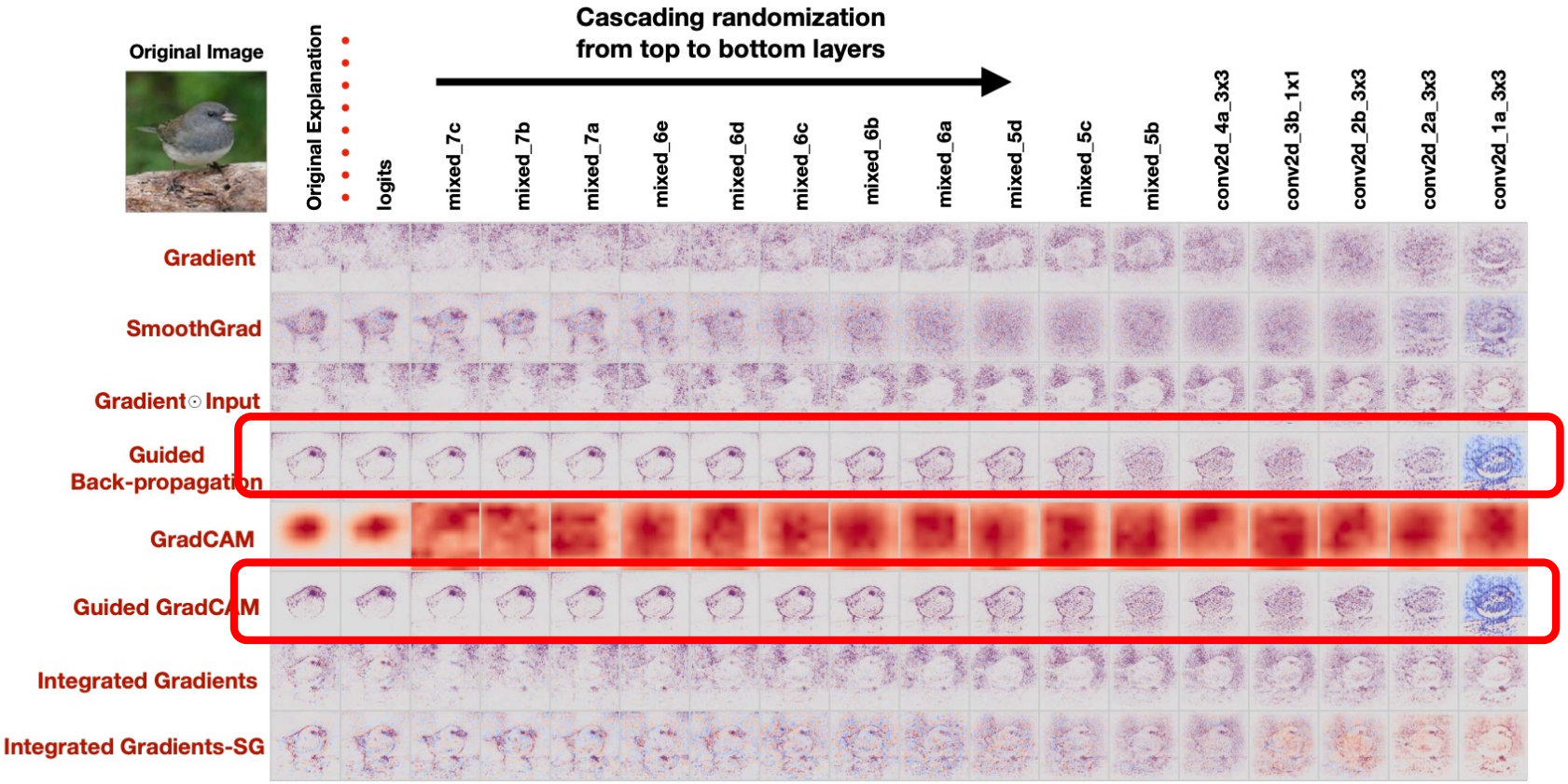
Unsupervised Causal Binary Concepts Discovery with VAE for Black-box Model Explanation, Thien Q Tran, Kazuto Fukuchi, Youhei Akimoto, Jun Sakuma, AAAI2022.

説明と信頼

- 説明可能AIはAIの高信頼化に役に立つ
- しかし、AIによる説明自体が信頼できないケースがある
  - 説明手法自体に問題がある場合
  - 説明可能AIが攻撃を受けている場合
  - 本質的要因(対象と因果関係あり)と非本質的要因(対象と偶然に相関)の取り違え
  - 「人間が求める」説明と「AIが合理的と思う」説明が異なる

説明手法自体に問題がある場合

モデルパラメータをランダム化しても説明が変化しない



モデルに依存しない説明は単なるフィルタ

→説明がついている=信頼性が高い、ではない

→「説明手法」が信頼できない場合がある

Adebayo et al., Sanity Checks for Saliency Maps, NeurIPS'18

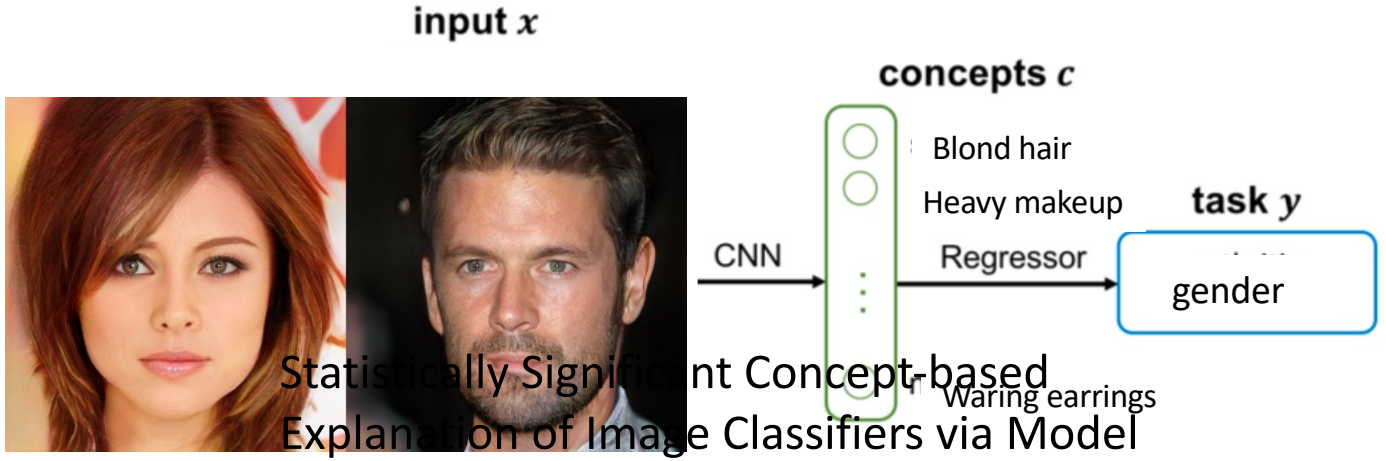
説明可能AIが攻撃を受けている場合

あらかじめ指定した特徴帰属による説明が表れるように、画像に微小な敵対的ノイズを混入

A. K. Dombrowski et al. “Explanations can be manipulated and geometry is to blame” NeurIPS 2019

本質的要因と非本質的要因の取り違え

- 顔画像からの性別識別におけるコンセプトに基づく説明



Statistically Significant Concept-based Explanation of Image Classifiers via Model Knockoffs

本来性別と関係ない要因を重要な要因と説明
「太っている」,,「目の下の涙袋」
「白髪」「波打った髪」「(画像が)ボケている」など

Xu et al., Statistically Significant Concept-based Explanation of Image Classifiers via Model Knockoffs. IJCAI'23

Statistically Significant Concept-based Explanation of Image Classifiers via Model Knockoffs [XFAS, IJCAI'23]

- コンセプトによる説明の二つの問題点
  1. どのコンセプトも多少は識別対象に関係する
  2. 線形モデルは不必要に多くの要因を重要と判断する

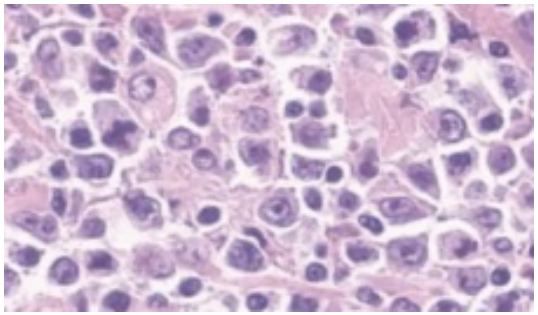
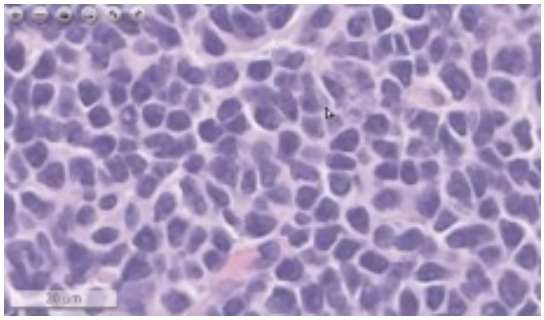
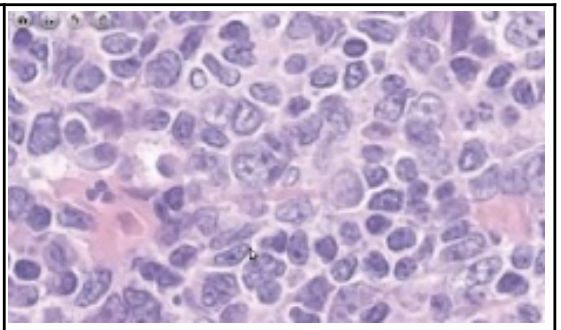
  → 説明がスパースにならない
- 提案法
  - 識別器での利用前提とするスパース特徴抽出
  - Knock-offによる「不必要な要因」の採用率(False discovery rate)の理論的な抑制

Xu et al., Statistically Significant Concept-based Explanation of Image Classifiers via Model Knockoffs. IJCAI'23

「人間が求める」説明と「AIが合理的と思う」説明が異なる

- 人間のよる判断の根拠
  - 分野の背景知識と整合した判断根拠
- AIによる判断の説明
  - 与えられたデータの範囲内で合理的な説明

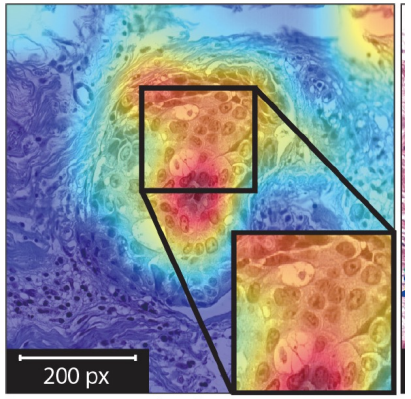
悪性リンパ腫の３病型の特徴

病型	 DLBCL	 FL	 Reactive
細胞の特徴	肥大化した細胞が多い	角ばった細胞が多い 染色による色が濃い	形や大きさは様々
細胞分布の特徴	散らばっている	集まっている	集まっている～中程度

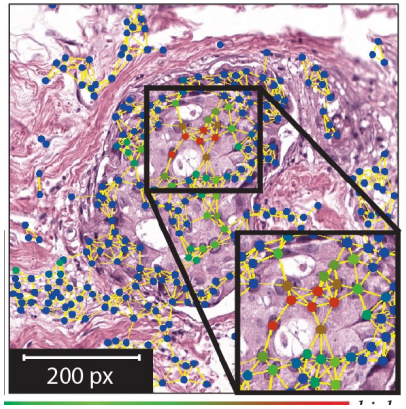
- 個々の細胞の特徴に加えて、細胞の位置関係で病型が決まる

既存のExplainable AIでは説明が不十分

画素帰属による説明
×どの細胞が重要か不鮮明
×細胞の位置や、細胞同士の関係を説明に利用できない

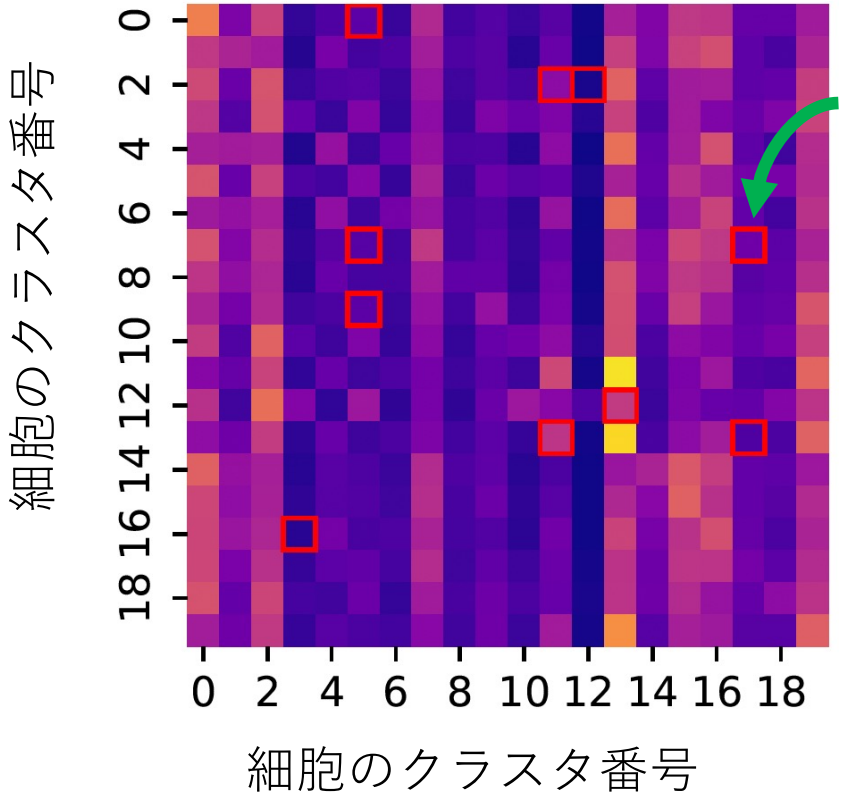


細胞グラフ分類での説明が有望？
ノード：細胞，エッジ：top-kの近傍関係
〇どの細胞(ノード)が重要か説明に利用できる
〇細胞同士の関係(エッジ)が重要か説明に利用できる



J. Guillaume, et al., "Quantifying explainers of graph neural networks in computational pathology," CVPR, 2021.

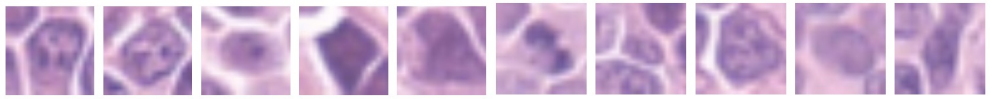
説明例



細胞クラスタ7, 17間の関係がAIが DLBCLを予測する重要な根拠と提示

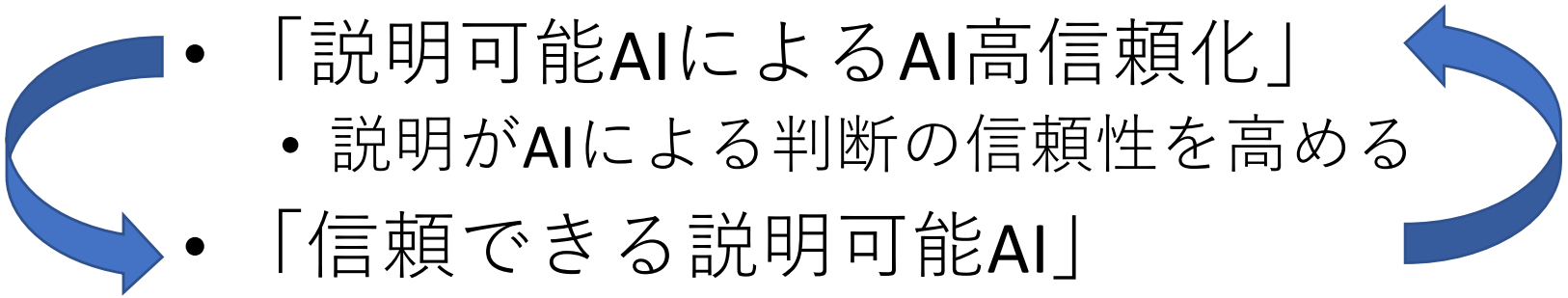
クラスタ7：

クラスタ17：



- 細胞を個別にクラスタリングし、タイプの異なる細胞クラスタ同士の関係をグラフ表現し、分類器を構成
- 病理医の印象：分類精度は82%と良好だが、なぜクラスタ7-クラスタ17の関係が識別に重要なのか、判断できない(病理学的な背景知識と整合しない)

おわりに



- 「説明可能AIによるAI高信頼化」
  - 説明がAIによる判断の信頼性を高める
- 「信頼できる説明可能AI」
  - 説明自体の信頼性が判断の信頼性には重要
    - 統計的有意性
    - 環境変化/攻撃に対する安定性
    - 背景知識との整合性