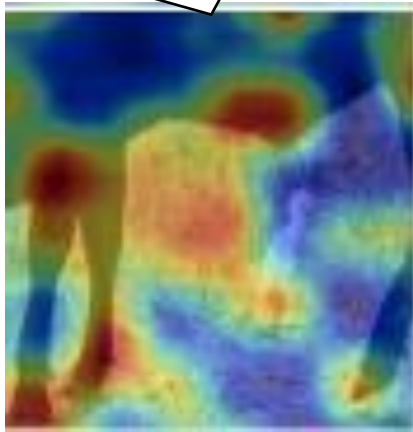


機械学習による動画像分類器の 判断根拠を可視化する手法

元画像



従来手法
(RISE)



本手法
(BORexs)



京都大学

CREST

京都大学 大学院情報学研究科 通信情報システム専攻

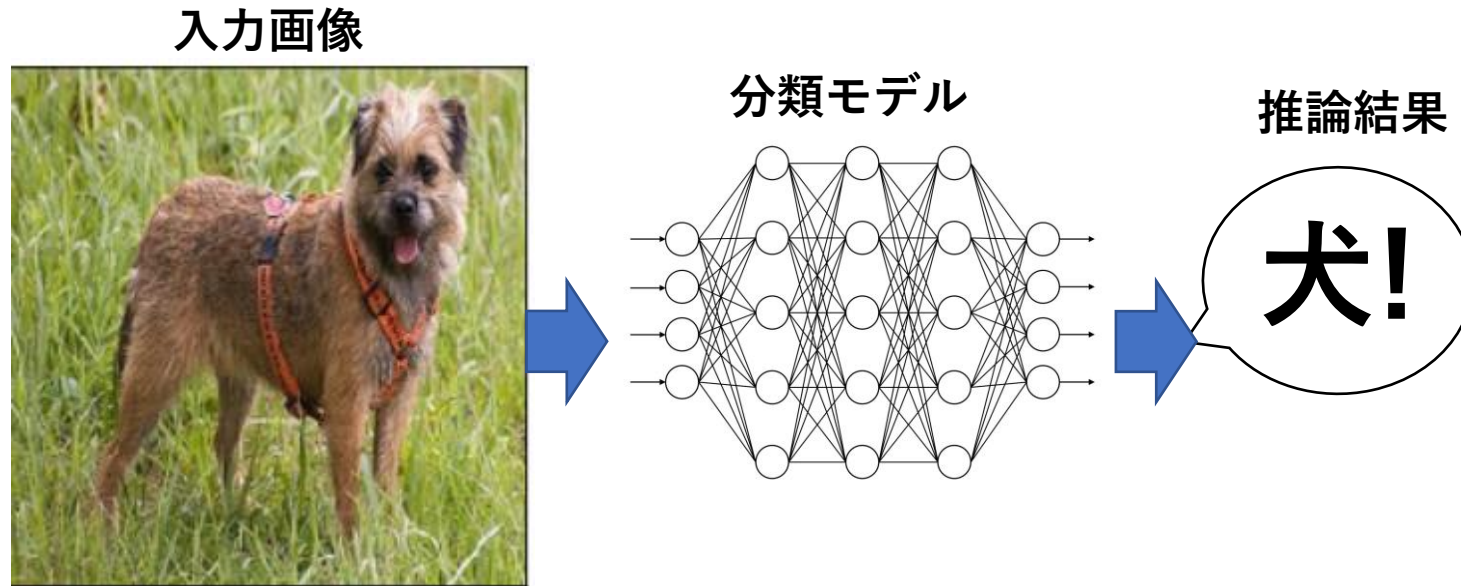
准教授 末永幸平

2022年6月28日 京都大学 新技術説明会

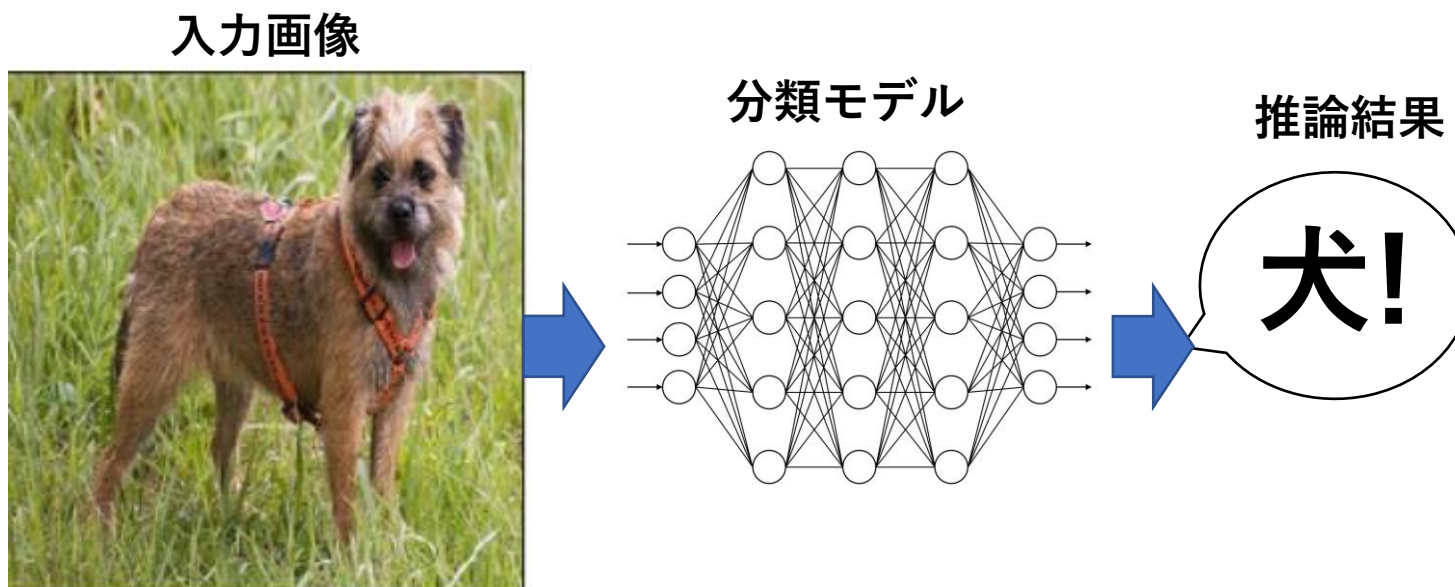
京都大学 大学院情報学研究科 末永幸平



背景: 動画像分類への機械学習の適用

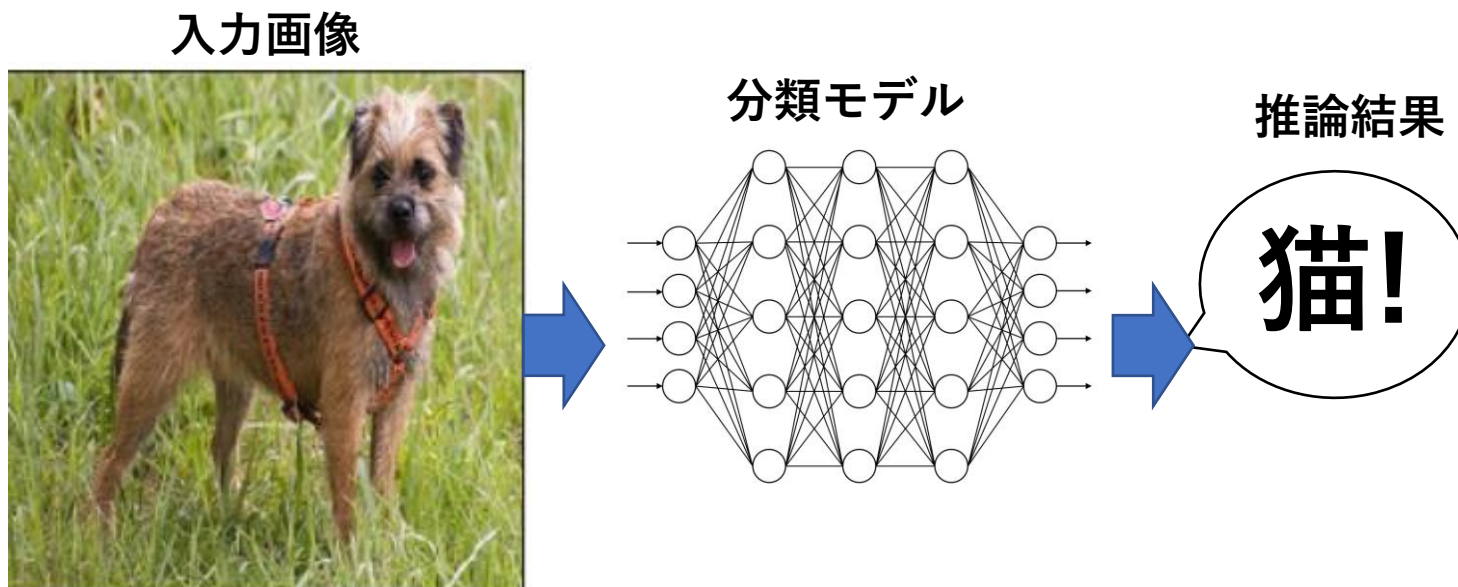


課題: 分類結果の解釈困難性



- 「正しい」根拠に基づいて分類しているかどうか不明
 - 本当に犬に着目して「犬!」と言ったのか?
 - (例えば訓練データ中に緑の背景の犬が多かったせいで) 背景が緑色の画像を「犬!」と判断していないか?

課題: 分類結果の解釈困難性



- 誤分類の原因をつきとめにくい
 - 何に着目して「猫!」と判断したのか?

研究分野: ヒートマップによる分類根拠の説明手法

分類結果を得るために重視した箇所を視覚化したヒートマップ（**顕著性マップ; saliency map**）を生成する手法

分類結果

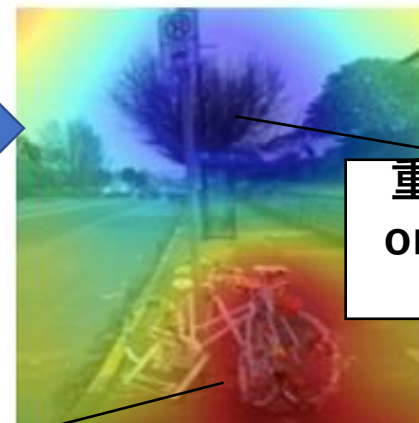
自転車!



入力画像

説明手法

顕著性マップ

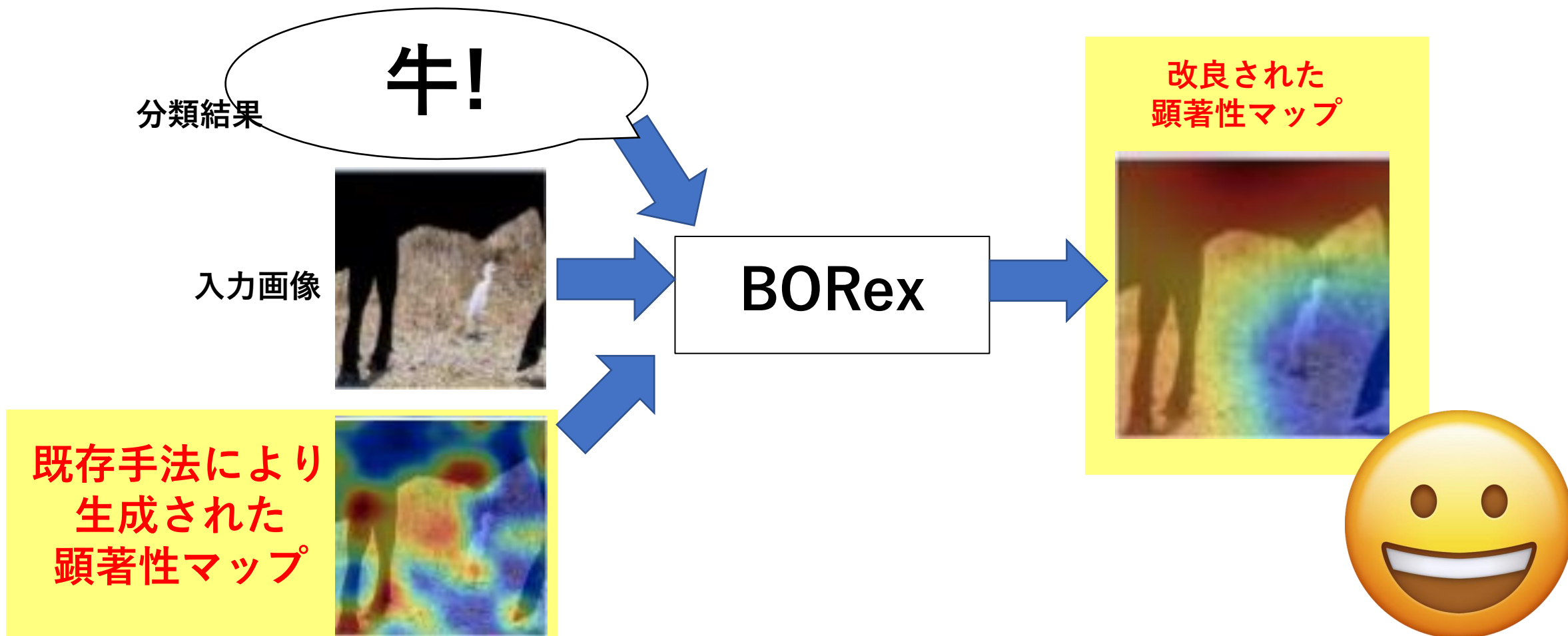


重視されなかった
or ネガティブだっ
た箇所は青

重視された箇所は赤

新技術: BORex

既存手法により得られた顕著性マップを改良する手法



アウトライン

- 従来技術とその問題点
- 新技術: BORex
- 実験結果
- 産学連携について

従来技術: ホワイトボックス型

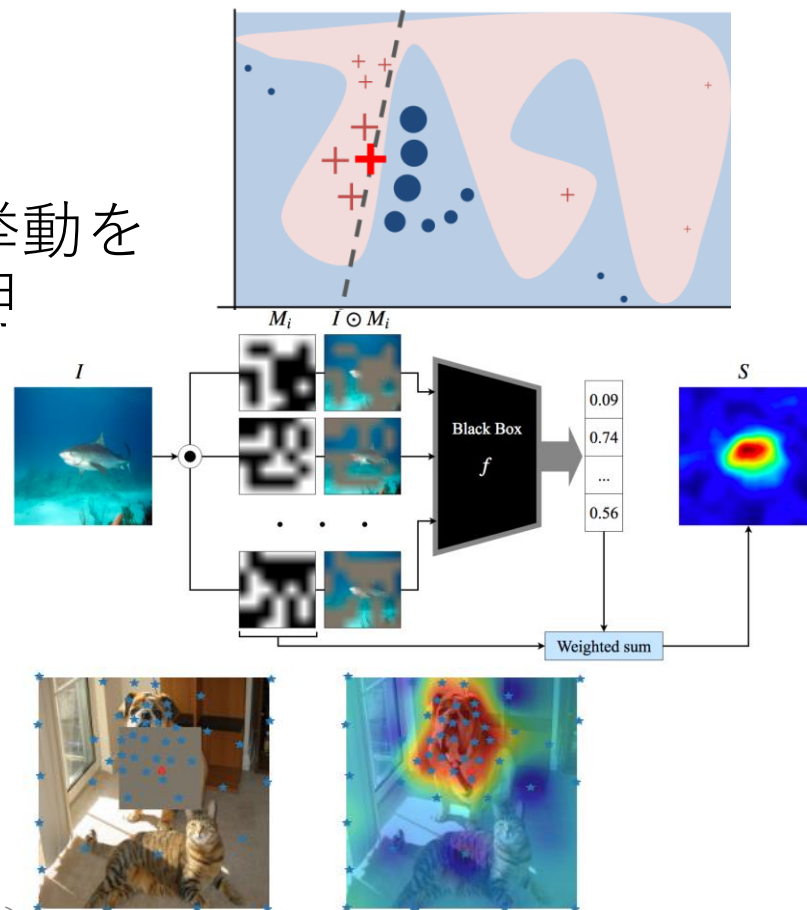
推論時のモデル内部の情報 (DNN における勾配等) を利用して顕著性マップを生成する手法

- Grad-CAM [Salvaraju+ ICCV'17]
 - DNN の計算した指定したラベルの確信度に対する feature map の勾配から顕著性マップを生成
- Grad-CAM++ [Chattopadhyay+ WACV'18]
 - Grad-CAM を一般化することにより高品質な顕著性マップを生成するように改良した手法

従来技術: ブラックボックス型

モデル内部の情報は使用せず，モデルの入出力のみから
顕著性マップを生成する手法

- LIME [Ribeiro+ KDD'16],
SHAP [Lundberg+ NeurIPS'17]
 - 入力データに近いデータでのモデルの挙動を
近似する線形モデルを学習し説明に使用
- RISE [Petsiuk+ BMVC'18]
 - 画像にランダムなノイズを乗せた画像
に対するモデルの出力から
顕著性マップを計算
- [Mokuwe+ IJCNN'20]
 - ガウス過程回帰 (GPR) によって画像の
重要な場所を少ない推論回数で判断



従来技術の問題点

しばしば低品質な顕著性マップが出力され，それを改良する手段がない

分類結果

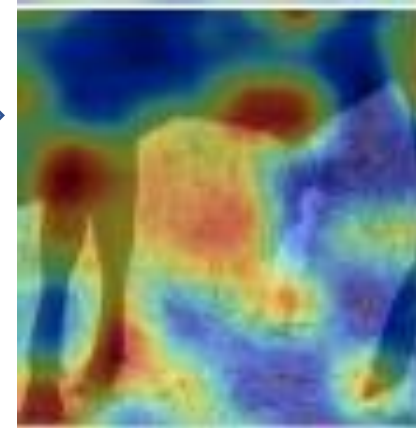
牛!



入力画像

既存手法

顕著性マップ
(RISEにより生成)

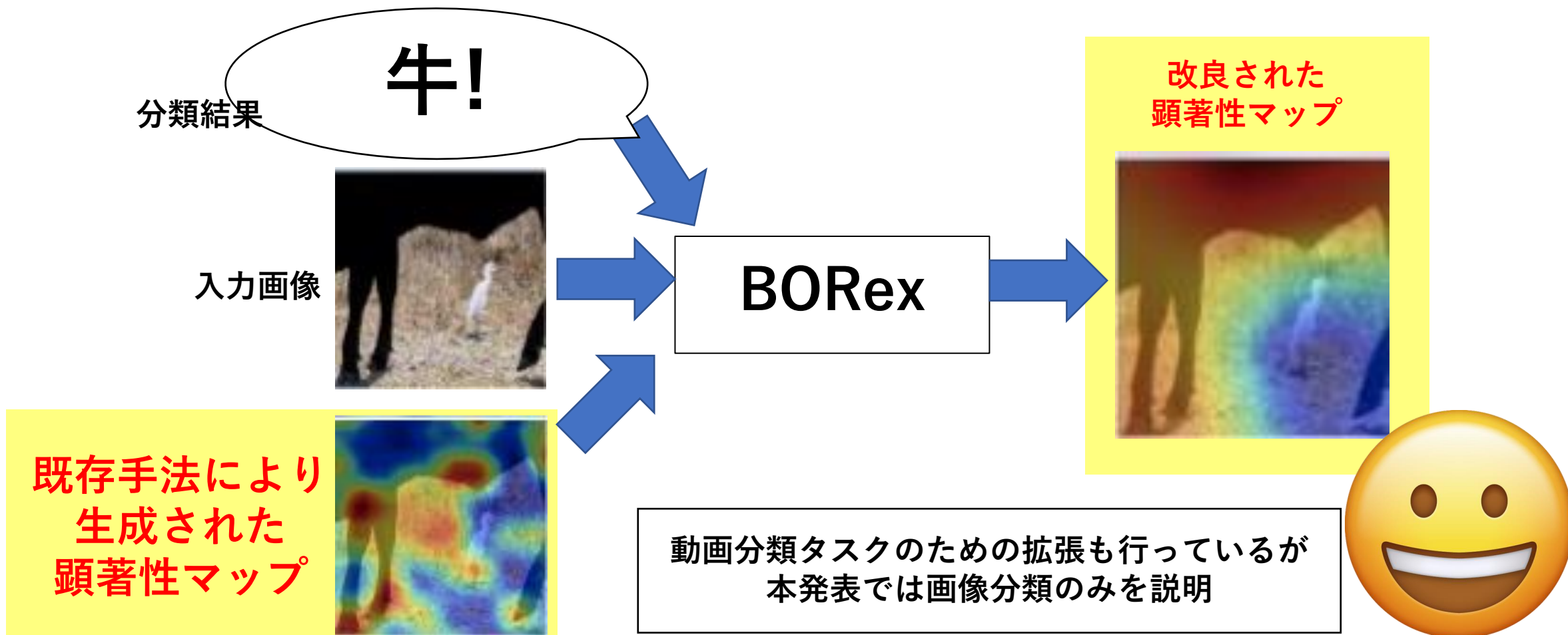


アウトライン

- 従来技術とその問題点
- 新技術: BORex
- 実験結果
- 産学連携について

新技術: BORex

既存手法により得られた顕著性マップを改良する手法



BORex の実行例

既存手法で得られた顕著性マップに比べて
よりわかりやすいマップが得られている

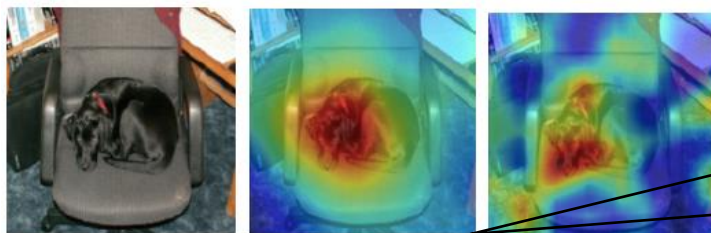
ラベル

dining
table



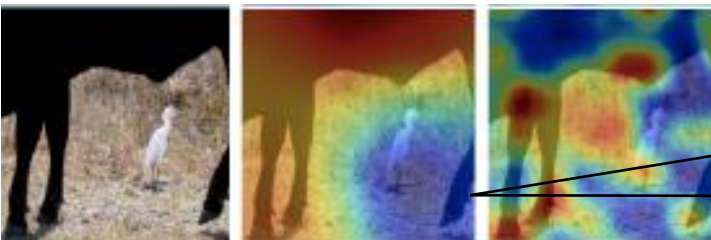
ダイニングテーブルが
正しく指摘されている

labrador
retriever



ラブラドルレトリバーが
正しく指摘されている

COW



画像の縁に存在する牛が
正しく指摘できている

入力画像

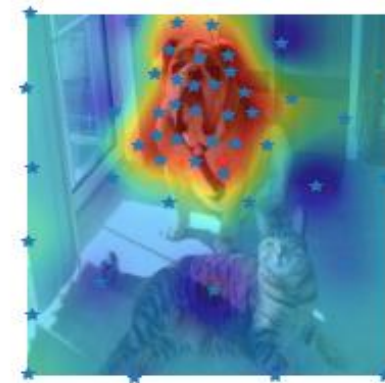
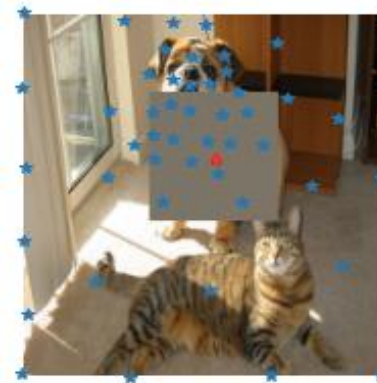
BORex
(提案手法)

RISE
(既存手法)

BORex のアイデア

改良対象の顕著性マップをガイドとして用いるように
[Mokuwe+ IJCNN'20] を拡張

- [Mokuwe+ IJCNN'20]: ガウス過程回帰 (GPR) を用いて画像中の領域の重要度を推定しつつ次にノイズを乗せる場所を選択
- BORex は顕著性マップを入力として取れるように [Mokuwe+] を拡張
 - その他に重要度の計算方法や GPR によって得られた推定値から顕著性マップを生成する手法に工夫



BORex と従来技術全般との比較

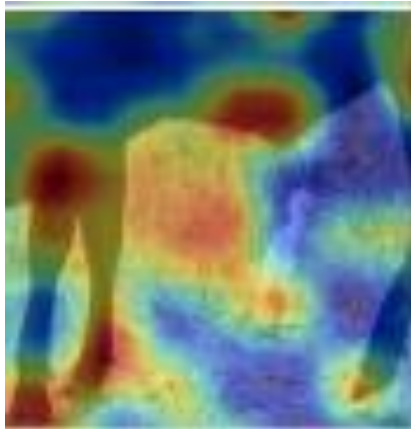
BORex は顕著性マップの**改良**に用いることが可能

既存手法では低品質かもしれない
顕著性マップが得られても
それをさらに改良することが困難

入力画像

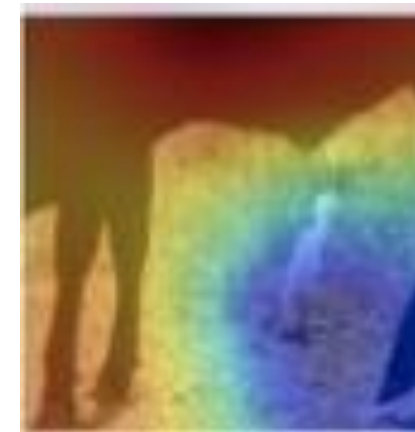


既存手法 (RISE)



新手法によって顕著性マップを改良可能
→ **信頼性の高い説明が可能に**

新手法 (BORex)



BORex とホワイトボックス型従来技術 (Grad-CAM++等) との比較

- BORex はブラックボックス型手法に属する
 - モデルの内部の情報にアクセスすることなく
顕著性マップの改良が可能

アウトライン

- 従来技術とその問題点
- 新技術: BORex
- 実験結果
- 産学連携について

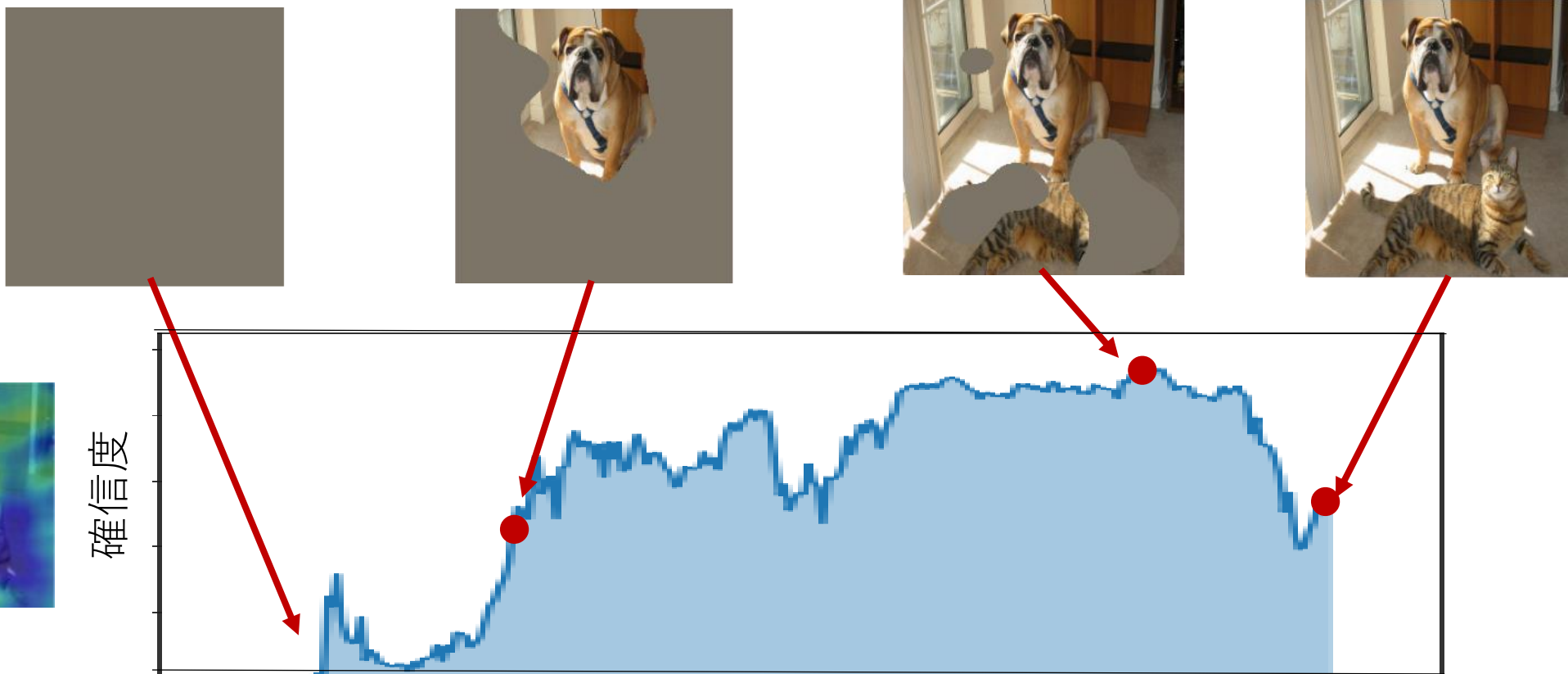
実験: セッティング

BORex が既存手法により生成された顕著性マップを改善できているかどうかを検証

- Pascal VOC データセット中の各画像について以下の顕著性マップを生成
 - 既存手法により生成された顕著性マップ
 - (1): RISE により生成
 - (2): Grad-CAM++ により生成
 - BORex により生成された顕著性マップ
 - (3): (1) を入力として生成
 - (4): (2) を入力として生成
- (1) と (3), (2) と (4) の品質を比較
- 使用したモデル: ResNet 152

ヒートマップの品質の評価尺度 (insertion metric) について

- ・ 顕著性マップの重要度が高い値からピクセルを挿入し、どれだけ早く確信度が上がるか
- ・ **高いほど高品質**



実験結果

**BORex によって顕著性マップの品質が
有意に改善していることを確認**

Compared with	Metric	Statistics	p-value
RISE	Insertion	65484.0	1.016050e-23**
	Deletion	44608.0	0.887350
Grad-CAM++	Insertion	54406.0	5.089895e-08**
	Deletion	34083.0	0.000679**

Insertion metric において
RISE の顕著性マップを有意に改善

Insertion metric において
Grad-CAM++ の顕著性マップを
有意に改善

片側Wilcoxon 符号順位検定による検定結果

帰無仮説: BORex の各指標が元の顕著性マップの各指標よりも優れていない

* は $p < 0.05$, ** は $p < 0.001$ を表す

アウトライン

- 従来技術とその問題点
- 新技術: BORex
- 実験結果
- 産学連携について

想定される分野・用途

- 機械学習モデルの挙動の高品質な説明が求められる safety critical な分野での使用
 - 医療や自動運転等?
- 既存手法では十分な顕著性マップが得られない際の適用
 - 推論に時間のかかるモデルを使用したり画像が巨大なために RISE 等の手法が収束するまでに時間がかかりすぎる場合に
 - モデル内部の情報を使用することができず, Grad-CAM++ 等のホワイトボックス型手法が使用できない場合に

実用化に向けた状況

- 現状
 - PyTorch モデルに特化した研究用プロトタイプを実装
 - ResNet 152 モデルによる PascalVOC 中の画像で Grad-CAM++ と RISE の顕著性マップを改良できることを確認
 - Kinetics-500 中の動画で, Grad-CAM++ と RISE の動画拡張で生成した顕著性マップをある程度改良できることを確認
 - やや改良の余地あり
- 有用性を向上させるために必要なタスク
 - 他のモデル, 画像, 顕著性マップ生成手法での有用性検討
 - ガウス過程回帰のパラメタのチューニング手法
 - 動画のための改良の研究

希望する協働の形態

- 本発明の特許のライセンスを受けて自社開発をしていただける企業を希望
 - 有用と思われる産業分野
 - 判断根拠の説明が求められる医療、金融、自動運転などの分野
 - 機械学習モデルのデバッグを継続的に行う必要のある分野
 - 希望する連携形態
 - ライセンス契約を結んだ上で当方の研究用プロトタイプを企業側に提供して、それを元に自社開発をしていただく
- 共同研究や技術指導については応相談

本技術に関する産業財産権

発明の名称 ： 情報処理装置、情報処理方法および情報処理プログラム

出願番号 ： 特願2021-180647

出願人 ： 国立大学法人京都大学

発明者 ： 末永幸平、菊池淳、和賀正樹

産学連携の実績

スタートアップを含む複数の企業との 産学連携の経歴あり

- 実績
 - 自動車系企業との形式検証技術に関する共同研究
 - スタートアップとの形式検証技術に関する共同研究・技術指導
 - リーガルテック
 - 特許検索サービス
 - ブロックチェーン
 - 画像処理系企業との解釈可能性向上手法に関する共同研究・技術指導
- 末永の研究経歴は以下を参照
 - researchmap: <https://researchmap.jp/ksuenaga>
 - ORCID: <https://orcid.org/0000-0002-7466-8789>
 - LinkedIn: <https://www.linkedin.com/in/ksuenaga/>

お問合せ先

国立大学法人京都大学内
株式会社TLO京都
京大技術移転部

TEL 075 - 753 - 9150

FAX 075 - 753 - 9169

e-mail event@tlo-kyoto.co.jp