機械学習モデルの説明性

Explainable AI の紹介

栗原 理央

株式会社ブレインパッド
リードデータサイエンティスト

本日お話する内容

1. 機械学習の説明性とは
    - 背景
    - 必要なケース
    - 手法概要
2. Explainable AI とは
    - 概要
    - 実装済みの手法
    - food101 データセットを用いた実験
3. まとめ

機械学習の説明性とは

01

背景

機械学習は研究から**社会実装**のフェーズへ

- **ブラックボックスだと業務適用しにくい**

総務省がAI利活用ガイドラインを公表（10個の原則のうち下記2つ）

- **公平性の原則**
  - 想定外のバイアスがないか
- **透明性の原則**
  - サービスに必要十分な説明性があるか

なぜ説明性が必要か

サービスを提供する事業者としての**説明責任**

モデル開発における**精度改善**、**デバッグ**

ユーザーの信頼や**納得感**の獲得、関係者への**説得材料**

人間の**知的好奇心**

実際のプロジェクトで説明性が必要だったケース

「機械学習モデルの根拠がないと信用してもらえない！」

「この需要の増減ってなんで？」

「ユーザー離脱を防ぐマーケティング施策を打ちたい！」

「精度改善したいけど、何を変更したらいいんだろう？」

4つのタイプの説明可能性フレームワーク

説明可能なモデル　or　事後説明モデル

グローバルな説明　or　ローカルな説明

参考: Challenges Towards Production-Ready Explainable Machine Learning,
https://www.usenix.org/system/files/opml20-paper-veiber_0.pdf

説明可能なモデル：解釈性の高いモデルを設計

- 決定木や線形モデルなど、多くの場面で使われている

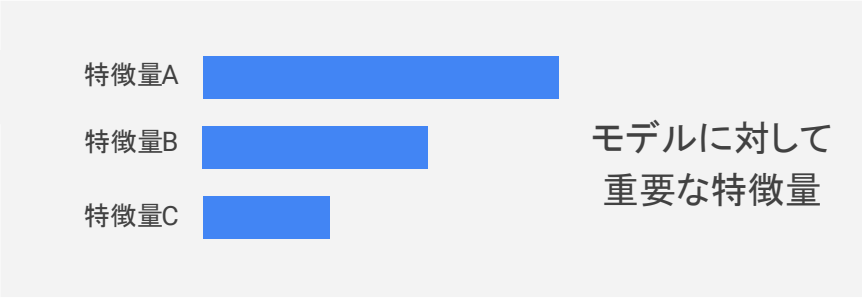
事後説明モデル：複雑なモデルを近似して学習済みモデルに説明性を付与

- 深層学習などより複雑なアルゴリズムの発展とともに研究が進んでいる

モデルの精度と**説明可能性**はトレードオフ

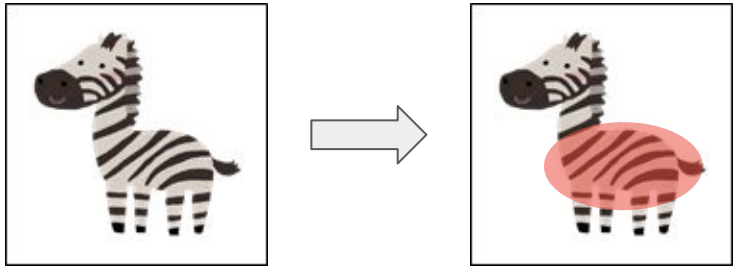
グローバルな説明：モデル全体に対して重要な特徴は何か

機械学習モデル



モデルに対して
重要な特徴量

ローカルな説明：特定のサンプルに対して各特徴量が寄与しているか



入力サンプルに対して
重要な特徴量

Explainable AI とは

02

Explainable AI の概要

- 機能
  - 各データに対して、予測結果に及ぼした特徴量の影響を説明
  - 画像分類タスクや表データの分類/回帰タスクに対応
- 手法：4つのフレームワークに当てはめると・・・
  - 説明可能なモデル or **事後説明モデル**
  - グローバルな説明 or **ローカルな説明**
- ユースケース
  - モデルが期待どおりに動作しているか確認
  - モデルのバイアスの確認
  - モデルや学習データの改善方法を確認

実装済みの手法

- Sampled Shapley
- Integrated gradients
- XRAI (eXplanation with Ranked Area Integrals)

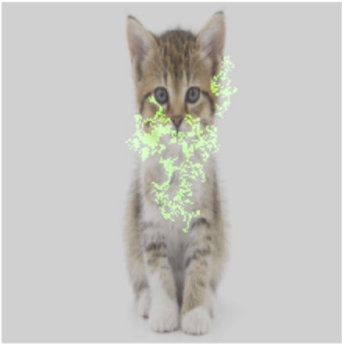
参考: https://cloud.google.com/ai-platform/prediction/docs/ai-explanations/overview

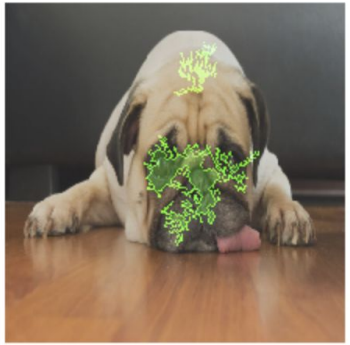
Sampled Shapley

- テーブルデータの分類／回帰モデルに対応
- 各特徴量の貢献具合だけでなく、予測値に対してプラスに寄与したのかマイナスに寄与したのかを出力
- Shapely 値を利用して変数の寄与を説明
  - 特徴量があるときないときの予測値の差をみる
  - 特徴量の追加順も考慮する（全組み合わせ・全順番）
- ただし、全パターンは計算せず、一定回数ランダムに計算された寄与のサンプル平均を Shapley 値とする

Integrated gradients

- テーブルデータの分類／回帰モデル、画像データの分類モデルに対応
- 下記画像は入力画像を分類するために有効だと計算されたピクセルを表す
- Grad-CAM, SmoothGrad などと同様に、勾配情報を用いる手法

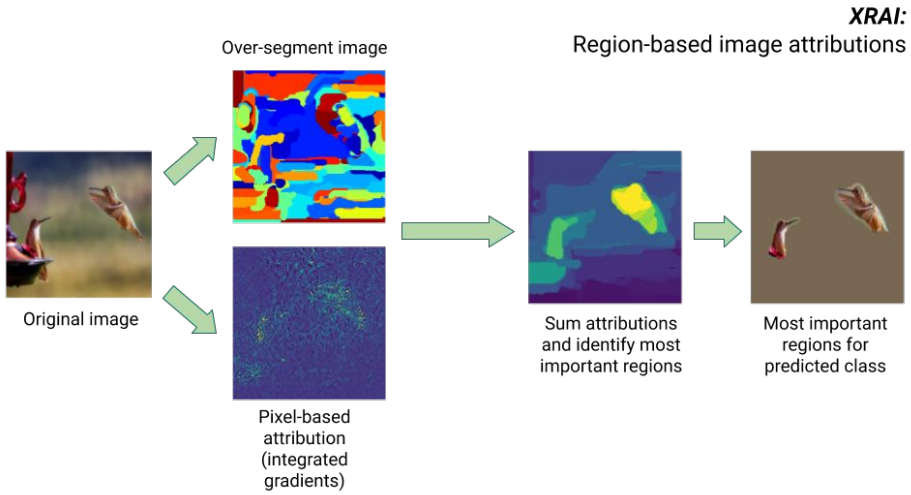




参考: Axiomatic Attribution for Deep Networks, https://arxiv.org/abs/1703.01365

XRAI (e**X**planation with **R**anked **A**rea **I**ntegrals)

- 画像データの分類モデルに対応
- 予測に重要だった箇所を、ピクセル単位ではなくリージョン単位で出力
- Integrated gradients と追加の処理を組み合わせた手法



参考: https://cloud.google.com/ai-platform/prediction/docs/ai-explanations/overview
http://cs.brown.edu/people/pfelzens/papers/seg-ijcv.pdf

どんなアウトプットが出てくるのか？

結果に納得感があるか？

モデルの精度改善のヒントにできそうか？

画像分類タスクで実験

データセット：food101

- 101個の食品クラス
- 各クラスごとに、750枚の学習用画像と250枚のテスト画像
  - 学習時は750枚をtrain: 600, validation: 150に分割
- 最大辺の長さが512ピクセル
  - 学習時は (256, 256)にリサイズ

XRAIを使用

参考: https://www.tensorflow.org/datasets/catalog/food101
https://github.com/GoogleCloudPlatform/ml-on-gcp/blob/master/tutorials/explanations/ai-explanations-image.ipynb

分類難易度別に2つのモデルを構築

※独断と偏見により目視で振り分け（各 6クラスずつ）

①分類しやすそうなクラス群

(val_acc: 0.87)

ワッフル

枝豆

マカロン

など

②分類が難しそうなクラス群

(val_acc: 0.60)

ステーキ

ポークチョップ

フィレミニョン

など

正しく分類できた例

ドーナツ (99%)

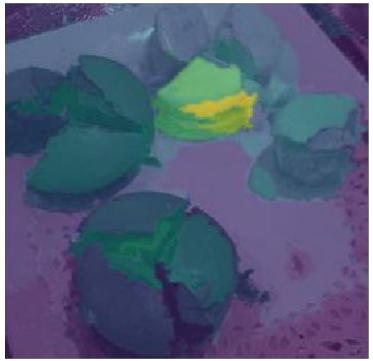
枝豆 (99%)

ワッフル (99%)

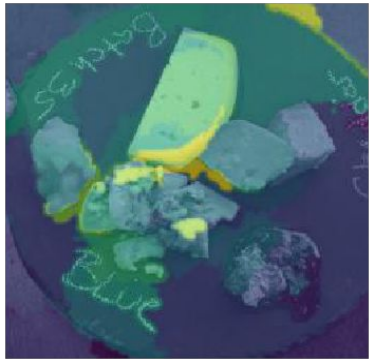
誤分類の例

正解：マカロン
予測：**ドーナツ (76%)**

正解：マカロン
予測：**ワッフル (47%)**



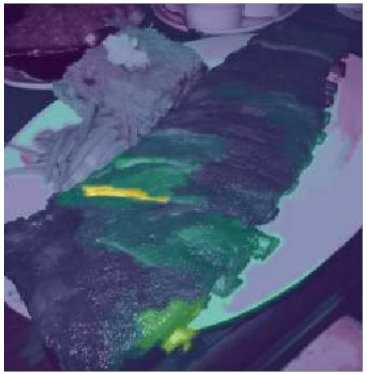
正解：チーズプレート
予測：**ビビンバ (48%)**



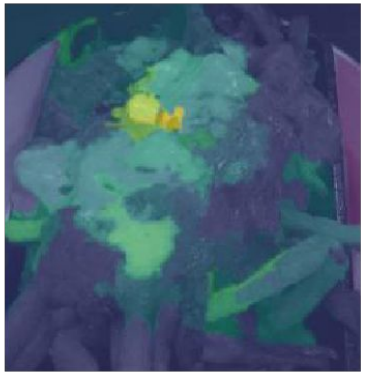
②分類が難しそうなクラスのモデル

正しく分類できた例

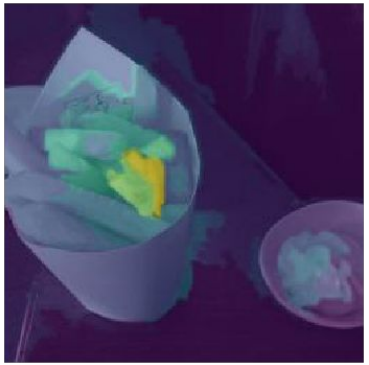
ベイビーバックリブ (99%)



プーティン (93%)



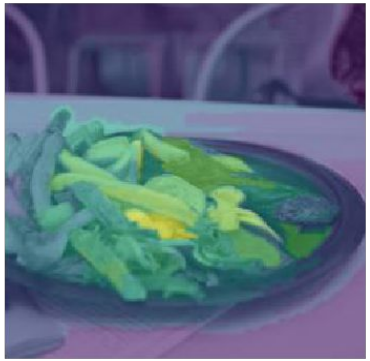
フライドポテト (94%)



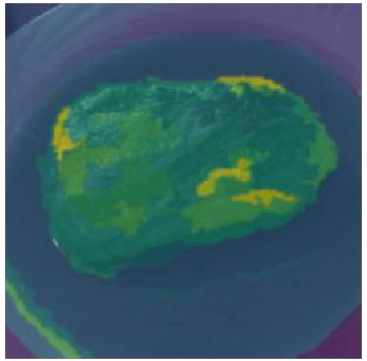
②分類が難しそうなクラスのモデル

誤分類の例

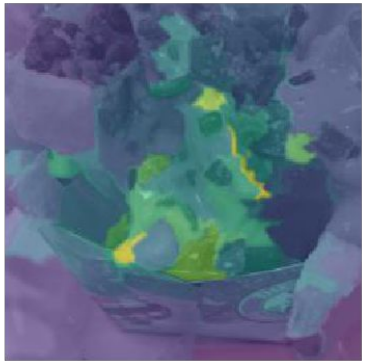
正解：ベイビーバックリブ
予測TOP2：**プーティン (85%)**
ベイビーバックリブ (11%)



正解：フィレミニョン
予測TOP2：**ステーキ (45%)**
ポークチョップ (41%)



正解：フライドポテト
予測：**プーティン (98%)**



まとめ

03

まとめ

- Explainable AI
  - 特定のサンプルに対する各特徴量の貢献度を出力
  - 画像分類タスクや表データの分類/回帰タスクに対応
- 使ってみた結果
  - 納得感は得られた
  - 取り組んでいるタスクの課題についてヒントが得られた
    - データセットのバイアス、アノテーションの信頼性など

Thank you