Notes on the Overview

D. NSYM

2022 年 5 月 5 日

説明法とは、ユーザが欲しい追加情報を機械学習モデルから抽出する技術。『どんな追加情報が欲しいか』はデータ/応用によって異なる。仮説に基づいて⊠明法の研究開発を進めていることが多い。

1　CNN 向けの説明手法・フレームワーク

1.1　Grad-CAM[Selvaraju et al., 2017]

最後の convolutional layer からの勾配のヒートマップを生成することによって視覚的な説明を提供。

Grad-CAM では、最終的な畳み込み特徴マップに対する分類スコアの勾配を利用して、分類スコアに最も影響を与える入力イメージの部分を識別する。この勾配が大きくなる場所は、最終的なスコアがデータに最も依存する場所を示す。

1.2　LIME[Ribeiro et al., 2016]

解釈可能なモデルを局所的に学習することで、任意の分類器や回帰の局所的に忠実な説明を生成する

- どの特徴が予測に重要だったかを提示する。
- モデルを説明対象データの周辺で線形モデルで近似する。
- 線形モデルの係数の大小で, 各特徴の重要度合いを測る。

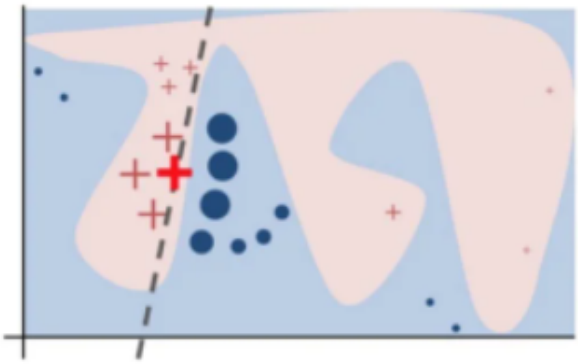

図 1: LIME の概要

1.3 ProtoPNet[Chen et al., 2019]

モデルの性能を犠牲にすることなく、画像中のプロトタイプの部品を識別できる prototypical part network（ProtoPNet）を提案

1. 深層特徴抽出モジュール f で、入力画像の一部の領域を表現する特徴ベクトル x を得る
2. f で訓練画像の一部の領域を表現する特徴ベクトル x' を得て、プロトタイプ：訓練データの代表点とする
3. 入力画像の各領域 x と訓練画像の代表領域 x' との類似度を評価し、各領域間の類似度 $\boldsymbol{S}$ を得る
4. $\boldsymbol{S}$ を入力に取る線形予測モジュール L で識別。L の重み w_{mk} の大きさは、入力画像の m 番目の領域と訓練画像の k 番目の領域が似ていることが予測に影響していることを意味する。

図 2: ProtPNet の概要

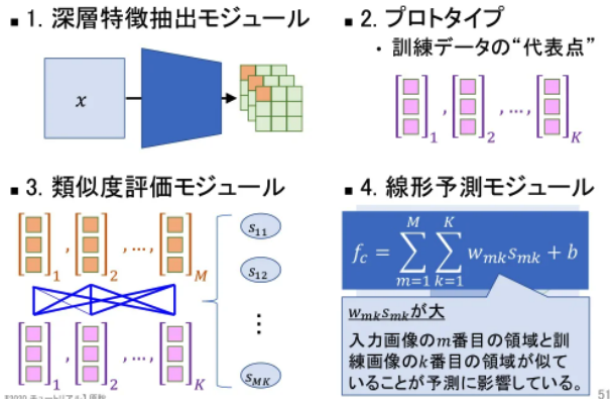

図 3: ProtPNet の仕組み

参考文献

[Chen et al., 2019] Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., and Su, J. K. (2019). This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32.

[Ribeiro et al., 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

[Selvaraju et al., 2017] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.