紹介論文

Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)

Been Kim and Martin Wattenberg and Justin Gilmer and Carrie Jun Cai and James Wexler and Fernanda Viegas and Rory Abbott Sayres

In *Proceedings of the 35th International Conference on Machine Learning*, pages 2668-2677, 2018 (ICML2018)

1 Goal of this paper (TCAV)

- Accessibillity： 機械学習の専門知識をほとんど必要としない
- Customization：あらゆる概念 (性別, ...) に適応できる
- Plug-in readiness：学習モデルの再学習や変更を必要としない
- Global quantification：個別のデータに対する説明ではなく, データセット全体を単一の指標で評価することができる

2 What is TCAV (Proposal method)

- ある画像 (動物, 乗り物, ...) に対して, 概念画像 (縞模様, ドット, ...) の重要度の評価手法 (FIgire 1)

3 Method

3.1 Preparation

- (a) 関心のある概念画像 C (縞模様, ...)
- (b) ラベル付された学習データ
- (c) 訓練済みネットワーク
  - input $\boldsymbol{x} \in \mathbb{R}^n$
  - output：K class
  - network：$f : \mathbb{R}^n \to \mathbb{R}^k$ (feedforward layer l with m neurons: $f_l : \mathbb{R}^n \to \mathbb{R}^m$, logit from layer l to layer k：$f_{l,k} : \mathbb{R}^m \to \mathbb{R}$)

1. (a) を Positive set　(P_C), (b) を Negative set(N) に分ける
2. ネットワークの l 層の潜在空間上で 2 つ ($\{f_l(\boldsymbol{x}) : \boldsymbol{x} \in P_C\}$ and $\{f_l(\boldsymbol{x}) : \boldsymbol{x} \in N\}$) の線形分離を行う
   - $\boldsymbol{v}_C^l \in \gtrdot$ は concept C に対する liner CAV である

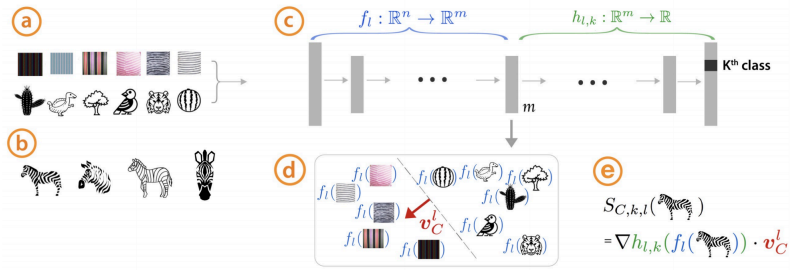

Figure1　Testing with Concept Activation Vectors

3. concept C の class k の sensitivity を以下のように定義

$$
\begin{aligned}
S_{C,k,l}(\boldsymbol{x}) &= \lim_{\epsilon \to 0} \frac{h_{l,k}(f_l(x) + \epsilon \boldsymbol{v}_C^l) - h_{l,k}(f_l(\boldsymbol{x}))}{\epsilon} \\
&= \Delta h_{l,k}(f_l(\boldsymbol{x})) \cdot \boldsymbol{v}_C^l
\end{aligned} \tag{1}
$$

4. Testing with CAVs (TCAV) を以下のように定義

$$
\begin{aligned}
&\mathrm{TCAV}_{Q_C,k,l} = \frac{|\{\boldsymbol{x} \in X_k : S_{C,k,l}(\boldsymbol{x}) > 0\}|}{|X_k|} \\
&where X_k \ is \ all \ inputs \ with \ that \ given \ label
\end{aligned} \tag{2}
$$