Examples are not Enough, Learn to Criticize! Criticism for Interpretability

Been Kim [1] Rajiv Khanna [2] Oluwasanmi Koyejo [3]

[1]Allen Institute for AI

[2]UT AUSTIN

[3]UIUC

NIPS, 2016/ Presenter: Ji Gao

Outline

1. Motivation
2. Method
   - Maximum Mean Discrepancy (MMD)
   - Use MMD for Prototype Selection
   - Criticism
3. Experiment

Motivation: Interpretablity

- Deep learning becomes popular in decision making.
- However, lack of transparency and interpretability in deep learning models is problematic
- An example: 10/2 Google algorithm fail puts 4chan's wrongly named Las Vegas gunman on top of search. At the same time, Facebook and YouTube put forged news on the first page.
- In some studies, interpretable models also outperforms complex models.

Previous solution: Example-based explanation

- **Example-based explanation**: Use prototypes to develop rules for decision making
- One popular approach: Case-based Reasoning. *Aamodt, Agnar, and Enric Plaza. "Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches"*
- Problem of examples:
  - Over-generalization
  - Complicated operations, i.e. regularization, might conflict with single examples.

Criticism

- Criticism samples: Those data points that don't fit the model well
- Criticism samples could be viewed as a complementary to the prototype samples.
- Bayesian model criticism (BMC): Study bayesian statistics to evaluate fitted bayesian models
- **Motivation:** Use statistical idea to generate criticism samples

Maximum Mean Discrepancy

Definition: Maximum Mean Discrepancy

Suppose $\mathcal{F}$ is a function space, P,Q are probability distributions, then the MMD of the two distributions is defined as

$$\mathrm{MMD}(\mathcal{F}, P, Q) = \sup_{f \in \mathcal{F}} (E_{X \sim P}[f(X)] - E_{Y \sim Q}[f(Y)]) \tag{1}$$

When $\mathcal{F}$ is a reproducing kernel Hilbert space (RKHS) with kernel k, the suprenum is achieved at

Witness function

$$f(x) = E_{X \sim P}[k(x, X)] - E_{Y \sim Q}[k(x, Y)] \tag{2}$$

Maximum Mean Discrepancy(Contd.)

Square of MMD:

$$\mathrm{MMD}^2(\mathcal{F}, P, Q) = E_{X,X'\sim P}[k(X, X')] - 2E_{X\sim P, Y\sim Q}[k(X, Y)] + E_{Y,Y'\sim P}[k(Y, Y')] \tag{3}$$

Sample Approximation

Given samples $X = x_i \sim P, i = 1..n$ and $Z = z_j \sim q, j = 1..m$:

$$\mathrm{MMD}_b^2(\mathcal{F}, X, Z) = \frac{1}{n^2}\sum_{i,j\in[n]} k(x_i, x_j) - \frac{2}{nm}\sum_{i\in[n], j\in[m]} k(x_i, z_j) + \frac{1}{m^2}\sum_{i,j\in[m]} k(z_i, z_j) \tag{4}$$

$$f(x) = \frac{1}{n}\sum_{i=1..n} k(x, x_i) - \frac{1}{m}\sum_{j=1..m} k(x, z_j) \tag{5}$$

Use MMD for Prototype Selection

Problem formulation:

Given n samples $X = \{x_i, i = 1..n\}$,suppose S is a subset of $\{1, 2..n\}$. Minimize the discrepancy $\mathrm{MMD}^2(\mathcal{F}, X, X_S)$ between X and X_S, where $X_S = \{x_i, \forall i \in S\}$.

Let

$$
\begin{aligned}
J_b(S) &= \frac{1}{n^2} \sum_{i,j=1}^{n} k(x_i, x_j) - MMD^2(\mathcal{F}, X, X_S) \\
&= \frac{2}{n|S|} \sum_{i \in [n], j \in S} k(x_i, x_j) - \frac{1}{|S|^2} \sum_{i,j \in S} k(x_i, x_j)
\end{aligned}
$$

Note that as $\frac{1}{n^2} \sum_{i,j=1}^{n} k(x_i, x_j)$ is a constant, Maximize $J_b(S)$ is equivalent to minimize $MMD^2(\mathcal{F}, X, X_S)$
$J_b(S)$ is a linear combination of $k(x_i, x_j)$.

Accelerate the selection of prototypes

Optimize:

$$\max_{|S| \leq m_*} J_b(S)$$

However, it's hard as there are exponential number of subsets.
Luckily (or not), we have the greedy algorithm:

Algorithm 1 Greedy algorithm, $\max F(\mathsf{S})$ s.t. $|\mathsf{S}| \leq m_*$

Input: $m_*, \mathsf{S} = \emptyset$
while $|\mathsf{S}| < m_*$ **do**
  foreach $i \in [n] \backslash \mathsf{S}$, $f_i = F(\mathsf{S} \cup i) - F(\mathsf{S})$
  $\mathsf{S} = \mathsf{S} \cup \{\arg\max f_i\}$
end while
Return: S.

Correctness of the greedy algorithm

It is proved the greedy algorithm can achieve a constant fraction of the optimal result.

Theorem

If F is any normalized, monotonic submodular function, the set S_* obtained by the greedy algorithm achieves at least a constant fraction $1 - \frac{1}{e}$ of the objective value obtained by the optimal solution i.e.

$$F(S_*) \geq (1 - \frac{1}{e}) \max_{|S| \leq m} F(S)$$

In the paper, it is proved that if $\forall i \neq j, 0 \leq k_{i,j} \leq \frac{k_*}{n^3 + 2n^2 - 2n - 3}$, where k_* is the diagonal item, the $J_b(S)$ function is monotone submodular.
Also, it is proved (not in the paper) no polynomial time algorithm can achive better approximation guarantee unless P=NP.

Criticism

We want to select those points with the largest $f_b(x)$. We have:

Criticism cost function

$$L(C) = \sum_{l \in C} \left| \frac{1}{n} \sum_{i \in |n|} k(x_i, x_l) - \frac{1}{m} \sum_{j \in S} k(x_j, x_i) \right| \tag{6}$$

Simply a summation over the witness functions.

Criticism with regularization

Use a regularizer can encourage a diverse selection of criticism points and improve the performance in practice.

Criticism with regularizer

$$\max_{|C| \leq c_*} L(C) + r(K, c) \tag{7}$$

They use the log-determinant regularizer:

log-determinant regularizer

$$r(K, c) = \log \det K_{c,c} \tag{8}$$

If the regularizer is submodular, the total objective function is submodular, and it can be approximated with the same greedy algorithm.

Experiment

Three experiments:

- Use prototypes and criticisms as a 1-NN classifier, on USPS handwritten digit dataset.
- Generating the prototypes and criticisms on Imagenet.
- Quantitative result: Human study on whether prototype and criticisms can improve interpretability

USPS handwritten digits

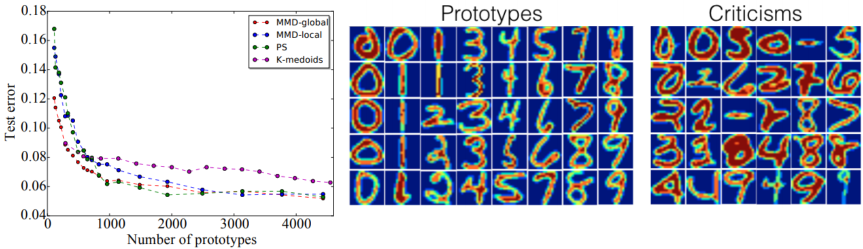

Figure 1: Classification error vs. number of prototypes $m = |S|$. MMD-critic shows comparable (or improved) performance as compared to other models (left). Random subset of prototypes and criticism from the USPS dataset (right).

Imagenet

Prototypes and criticisms:

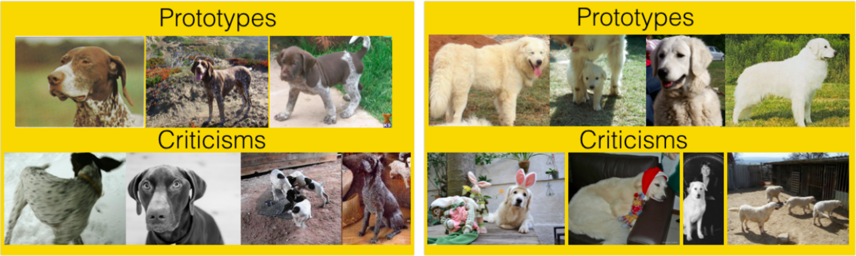

Figure 2: Learned prototypes and criticisms from Imagenet dataset (two types of dog breeds)

Study of the interpretability

- Design 4 conditions: Raw images, prototype only, uniformly sampled data and prototype with criticism.
- For each, they design 21 questions. Showing 6 different groups of a species and an image randomly sampled from one of the group. The participant is required to classify which is the group as fast as possible.
- Four conditions assign to four participants.
- Result:
  - With Proto and Criticism, participant successfully answer more questions
  - Subject think the addition of criticism "made it easier to locate defining features".