2020/12/25 이유진 (youjin lee)

Paper Review

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

Been Kim, Rajiv Khanna, Oluwasanmi O. Koyejo

NIPS, 2016

Abstract

예제 기반 설명은 매우 복잡한 분포의 해석가능성을 개선시키기 위한 노력의 일환으로 널리 사용되었다. 그러나 프로토타입만으로 복잡도의 요지를 표현하기에는 거의 충분하지 않다. 사용자가 더 나은 mental 모델을 구성하고 복잡한 데이터 분포를 이해하기 위해서, 또한 프로토타입에서는 포착되지 않았던 무언가를 설명하기 위해서 criticism(비평)이 필요하다. 베이지안 모델의 형태에서 영감을 받아 우리는 프로토타입과 비평을 효과적으로 학습하는 해석가능성을 목적으로 설계된 MMD-critic을 개발했다. 사람 피험자로 진행된 파일럿 연구는 MMD-critic이 사람들이 이해하고 추론하는데 유용한 프로토타입과 비평을 선택한다는 것을 보여주었다. 또한 MMD-critic이 선택한 프로토타입을 평가했고 기본 모델들과 비교하여 경쟁력 있는 성능을 보여주었다.

Introduction and Related Work

ML이 사람의 의사결정에 있어 광범위하게 사용되면서 투명성과 해석가능성의 중요성은 점점 커지고 있다. 해석가능성은 심각한 결과를 초래하는 결정이 가능한 분야에서 특히 더 중요하다. 예를 들어 폐렴 위험 예측 사례는 해석가능한 모델이 복잡한 모델이 간과한 데이터 내의 중요하지만 놀라운 패턴들을 더 나타낼 수 있음을 보여주었다.

사람 추론 연구는 예제(프로토타입)의 사용은 전략적 의사 결정을 위한 효율적인 전략 발전의 기본임을 보여주었다. 이 맥락을 이은 인기있는 연구는 실제 문제에 성공적으로 적용시킨 사례 기반 추론(CBR)이었다. 최근에는 비지도학습에서 베이지안 프레임워크와 CBR 기반 접근법을 결합하여 사용자 해석가능성을 더욱 개선했다. 지도학습의 경우 예제 기반 분류 모델이 데이터의 요약된 관점을 제시하면서 해석이 불가능한 방법에서도 비교할 만한 성능을 이룩했다. 그러나 모델의 행동을 설명하기 위해 예제에만 의존하면 지나치게 일반화되고 오해가 생길 수 있다. 물론 데이터 분포가 깨끗할 경우, 데이터를 충분히 표현하는 프로토타입 예제들의 집합이 존재한다는 점에서 예제만으로 충분할 수도 있다. 하지만 실제 데이터에서 이런 경우는 거의 없다. 복잡한 데이터에 대한 모델 피팅 시 성능을 높이기 위해 모델에 bias를 추가하는데(regularization) 이 bias는 데이터 분포와 충돌할 수 있다. 따라서 해석 가능성을 유지하기 위해서는 좋은 설명을 제공하지 않는 프로토타입 예제들이 있는 입력 데이터 영역의 부분을 나타내는 통찰력을 전달하는 것이 프로토타입 예제와 함께 중요하다. 우리는 적절치 않은 데이터들을 모델 비평 샘플이라 부른다. 프로토타입과 함께 비평은 사람들이 복잡한 데이터로 이루어진 모델이 더 나아지는데 도움을 준다.

Bayesian model criticism(BMC)는 베이지안 모델을 기반으로 특정 모델이 데이터를 설명하지 못할 수 있는 영역과 방법들을 식별하도록 도와 모델 발전과 선택을 목적으로 개발된 평가용 프레임워크이다. 모델 설계의 중요한 영역으로 빠르게 발전되었고, 베이지안 통계학자들은 모델 비평을 모델 구성, 추론, 비평 사이클에서 중요한 요소로 간주한다. Lloyd와 Ghahramani는 통계적 모델 비평을 위해 최대 평균 불일치(MMD)를 이용한 탐색적

접근법을 제시했고, 모델이 가장 잘못 표현한 데이터들의 입력 영역 부분을 식별하기 위해 목적 함수를 사용하여 탐색했다. 고전적인 2개의 샘플 테스트 목적이나 베이지안 모델 비평 목적 대신에, 우리는 MMD의 새로운 적용을 깊게 생각했고 그와 관련된 목적 함수를 프로토타입과 비평 샘플을 선택하는 주요 접근법으로 간주했다.

우리는 ML의 해석가능성을 향상시키기 위해 프로토타입과 비평 선택 목적의 확장 가능한 프레임워크인 MMD-critic을 소개한다. 우리가 아는 선에서 위 방법은 ML에 대한 설명을 생성하기 위해 BMC 프레임워크를 활용하는 첫번째 연구이다. MMD-critic은 데이터 포인트와 잠재적인 프로토타입 사이의 유사성 측정 목적으로 MMD 통계를 사용하고, 통계를 최대화하는 프로토타입을 효율적으로 선택한다. 분석에 의하면 확장성 측면에서 특정 조건 하 프로토타입 선택을 위한 MMD는 상위 모듈의 집합 함수라는 것을 보여주었다. 우리는 근접 프로토타입 분류 모델로 MMD-critic의 성능을 정량적으로 평가하고 기존에 존재하는 방법들과 비교할 만한 성능을 달성했음을 보여주었다. 또한 사람 피험자 파일럿 연구로부터 프로토타입과 함께한 비평을 포함하는 것이 데이터 분포가 잘 설명되어야 하는 최종 작업에 도움이 된다는 결과를 제시했다.

Maximum Mean Discrepancy (MMD)

두 분포에 대한 기대치 사이의 차이를 나타내는 함수 공간 F의 상한에 대해 주어진 분산 P, Q의 차이를 측정하는 것이다.

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

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

여기서 (1), (2) 모두 MMD를 측정하는 목적 함수로 알려져 있다. Q가 P의 밀도에 underfit하면 양수, Q가 P에 overfit하면 음수임을 확인하고, 우리는 (2)를 (1)로 대체하고 그 결과를 제곱하여 아래와 같은 식으로 유도했다.

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

$MMD^2(\mathcal{F}, P, Q) \geq 0$이고 $MMD^2(\mathcal{F}, P, Q) = 0$임이 분명하고, P는 F 공간의 Q와 구별이 불가능하다. 이 모집단 정의는 표본 기대치를 사용하여 근사화 할 수 있는데, 다음은 $X = \{x_i \sim P, i \in [n]\}$인 P로부터 n개의 샘플과 $Z = \{z_i \sim Q, i \in [m]\}$인 Q로부터 m의 샘플이 주어졌을 때의 유한 표본 근사치(4)와 목적함수(5)이다. ($[n]$은 정수 집합을 의미한다)

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

$$f(x) = \frac{1}{n}\sum_{i\in[n]} k(x, x_i) - \frac{1}{m}\sum_{j\in[m]} k(x, z_j) \quad (5)$$

MMD-critic for Prototype Selection and Criticism

통계 모델 $X = \{x_i, i \in [n]\}$ 로부터 n개의 샘플이 주어질 때 $X_s = \{x_i \; \forall i \in S\}$를 만족하는 부분집합을 $S \subseteq [n]$라 하고, 커널 함수 $k(\cdot,\cdot)$와 함께 RKHS가 주어질 때 우리는 $MMD^2(\mathcal{F}, X, X_s)$를 사용하여 선택된 모든 부분집합과 샘플 사이의 MMD를 측정할 수 있다. MMD-critic은 $MMD^2(\mathcal{F}, X, X_s)$를 최소화하는 프로토타입 인덱스들 S를 선택한다. 우리의 목적을 달성하기 위해 이 문제를 정규 이산 최대화로 제시하는 것이 편리할 것이다. 이를 위해 추가적인 bias를 부여하는 $MMD^2(\mathcal{F}, X, X_s)$의 부재로 주어진 다음 비용 함수를 고려한다.

$$J_b(S) = \frac{1}{n^2}\sum_{i,j=1} k(x_i, x_j) - \ MMD^2(\mathcal{F}, X, X_S) = \frac{2}{n|S|}\sum_{i\in[n], j\in S} k(x_i, y_j) - \frac{1}{|S|^2}\sum_{i,j\in S} k(y_i, x_j) \quad (6)$$

추가적인 bias $MMD^2(\mathcal{F}, X, \emptyset) = \frac{1}{n^2}\sum_{i,j=1}^{n} k(x_i, x_j)$는 S에 대하여 상수이고, 공집합에서 평가할 때 $J_b(\emptyset)$는 $J_b(\emptyset) = \min_{S \in 2^{[n]}} J_b(S) = \frac{1}{n^2}\sum_{i,j=1}^{n} k(x_i, x_j) - \frac{1}{n^2}\sum_{i,j=1}^{n} k(x_i, x_j) = 0$ 와 같기 때문에 정규화 된다. MMD-critic은 최적화 부분집합 S로써 m_*개의 프로토타입을 선택한다.

$$\max_{S \in 2^{[n]}, |S| \le m_*} J_b(S) \quad (7)$$

Submodularity and Efficient Prototype Selection

(6)을 최적화하는 것은 꽤 복잡할 수 있지만 비용 함수 $J_b(S)$가 커널 매트릭스 조건 하에 단조로운 서브 모듈이고 커널 매트릭스가 주어지면 쉽게 확인할 수 있음을 보여주었다. 이러한 결과를 바탕으로 효율적인 프로토타입 선택을 위해 탐욕 알고리즘을 설명한다. $F: 2^{[n]} \mapsto \mathbb{R}$가 집합 함수를 나타낸다고 하자. (1) 만약 $F(\emptyset) = 0$이면 F는 정규화 한다. (2) 만약 모든 부분집합 $u \subset v \subset 2^{[n]}$이 $F(U) \le F(V)$를 만족하면 F는 monotonic하다고 한다. (3) 만약 모든 부분집합 $U, V \in 2^{[n]}$ 이 $F(U \cup V) + F(U \cap V) \le F(U) + F(V)$ 를 만족하면 F는 submodular(하위 모듈)이라 한다. (F가 하위 모듈이었을 때 -F는 상위 모듈이 되고 그 반대도 역시 같다.)

우리는 하위 모듈성을 증명했고, 특별한 경우에서 (6)의 하위 모듈성을 보여주었다. 특히 다음 원리는 선형 매트릭스 함수인 일반적인 이산 최적화 문제들을 고려하고, 단조롭고 하위 모듈인 또는 단조롭거나 또는 하위 모듈이거나 하는 문제들을 위한 매트릭스 조건에서 충분함을 보여주었다.

정리2(Monotone Submodularity for Linear Forms): 대칭일 필요가 없고 $h_* = max_{i,j\in[n]} h_{i,j} > 0$을 만족하는 매트릭스 $H \in \mathbb{R}^{n\times n}$라 하고, $h_{i,j} = h_*$이면 $e_{i,j} = 1$이고 반대는 $e_{i,j} = 0$인 이진 매트릭스 $E \in [0,1]^{n\times n}$라 하며 이 때 E에 대해 orthogonal complement $E' = 1 - E$이다. $S \subseteq 2^{[n]}$가 주어졌을 때 $F(H, S) = < A(S), H > \forall S \in S$인 선형식을 고려하여 $m = |S|$일 때 아래의 함수를 정의한다.

$$\alpha(n, m) = \frac{a(S \cup \{u\}) - a(S)}{b(S)}, \qquad \beta(n, m) = \frac{a(S \cup \{u\}) + a(S \cup \{v\}) - a(S \cup \{u, v\}) - a(S)}{b(S \cup \{u, v\}) + d(S)} \quad (8)$$

1. $h_{i,j} \le h_*\alpha \ \forall \ 0 \le m \le m_*, \forall (i,j) \in E'$이면 $F(H, S)$는 monotone이다.
2. $h_{i,j} \le h_*\beta \ \forall \ 0 \le m \le m_*, \forall (i,j) \in E'$이면 $F(H, S)$는 submodular이다.

추론3(Monotone Submodularity for MMD): 커널 매트릭스 $K \in \mathbb{R}^{n\times n}$는 $k_{i,i} = k_* > 0 \ \forall i \in [n]$이며 모든 값이 양수이고 diagonally dominant(대각지배 매트릭스)이다. 만약 off-diagonal이 $k_{i,j} \ \forall i, j \in [n], \ i \ne j \ satisfy \ 0 \le k_{i,j} \le \frac{k_*}{n^3+2n^2-2n-3}$이면, (6)에서 주어진 $J_b(S)$는 monotone submodular이다. Diagonal dominance 조건은 결과 표현을 단순화하는 것의 목적으로 주어진 커널 매트릭스 확인이 쉽다. 또한 필요한 프로토타입의 수 m_*를 결정하면 그 조건이 크게 약해진다는 것에 주목했다. 게다가 bounds(8)이 모두 m의 단조롭게 감소하는 함수여서 MMD는 더욱더 단순화되므로 조건은 m_*에 대해서만 확인될 필요가 있다. **정리2**의 일반적인 접근 방법이 커널에서 임의로 인덱싱 된 최대 entries를 허용하므로 위 조건이 필수조건이 아님은 확인해야한다. 아마 실전에 더 중요한 것은 **추론3**에 의해 표현된 diagonal dominance 조건이 적절히 선택된 파라미터를 가진 커널에 의해 만족된다는 관찰 결과이다. 우리는 RBF 커널과 powers of positive 표준 커널의 예제를 제공한다.

예제4(Radial basis function Kernel): $k_{i,j} = k(x_i, x_j) = \exp(-\gamma \parallel x_i - x_j \parallel)$인 값들을 가진 RBF 커널 K은 중복되지 않는 points로 구성된 샘플 X로 평가된다. off-diagonal 커널 entries는 γ이 증가함에 따라 단조롭게 감소한다. 따라서 $\exists\, \gamma^*$는 $\gamma \geq \gamma^*$를 만족시킨다.

예제5(Powers of Positive Standardized Kernels): 양의 값만 가지는 커널 매트릭스 G는 단일 대각선이 $g_{i,i} = 1 \; \forall \, i \in [n]$인 경계 $0 \leq g_{i,j} < 1$로 표준화되었다. 커널 power를 $K \; with \; k_{i,j} = g_{i,j}^p$라고 정의한다. off-diagonal 커널 entries는 p이 증가함에 따라 단조롭게 감소한다. 따라서 $\exists\, p^*$는 $p \geq p^*$를 만족시킨다.

여기 설명된 예제 외에도 다양한 파라미터화 된 커널 함수에 대해 유사한 조건을 열거할 수 있으며 모델 기반 커널들을 쉽게 확인할 수 있다. 위 예제들을 통해 얻은 것은 **추론3** 조건이 지나치게 제한적이지 않다는 것이다. Submodular 함수의 제약된 최대화는 일반적으로 NP-hard이지만 단순 탐욕 알고리즘 기반 선택에서 최적 성능을 보이는 것으로 알려졌으며 이는 강력한 이론적 보장을 바탕으로 나온 것이다.

정리6(Nemhauser et al, 1978): 모든 함수 F가 정규화 되었고, 단조로운 submodular인 경우에 탐욕 알고리즘에서 얻어진 집합 S_*는 적어도 목표치의 상수 분수 $(1 - \frac{1}{e})$를 달성한다. 또한 만약 P=NP가 아니면 다항 시간 알고리즘은 더 나은 근사 보장을 제공할 수 없다. 탐욕 알고리즘의 추가적인 장점은 학습 시간에 의해 만들어지는 프로토타입 m_*의 개수를 결정할 필요가 없기 때문에 커널이 적절한 조건을 충족한다고 가정하면 여전히 의미 있는 결과를 돌려주면서 어떤 m_*에서도 훈련을 중단할 수 있다. 탐욕 알고리즘에 대한 설명은 아래와 같다.

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.

Model Criticism

프로토타입 샘플 선택에 의하면 MMD-critic은 criticism이라 불리는 모델인 프로토타입에 의해 잘 설명되지 않는 데이터 points의 특징을 나타낸다. 이 데이터 points들을 목적 함수(5)의 최대값으로써 선택된다. 즉, 데이터셋과 프로토타입 간의 유사성을 가장 많이 이탈하는 곳이다. 비용 함수를 고려해보자.

$$L(C) = \sum_{l \in C} \left| \frac{1}{n} \sum_{i \in [n]} k(x_i, x_l) - \frac{1}{m} \sum_{j \in S} k(x_j, x_l) \right| \qquad (9)$$

절대값은 샘플의 밀집도가 underfit하면 $f(x) > 0$이고 overfit이면 $f(x) < 0$가 되는데 이 모두를 우리가 측정한다는 것을 보여준다. 그래서 우리는 그 값의 부호보다는 이탈하는 정도에 초점을 맞춘다. 다음 정리는 (9)가 C의 선형 함수임을 보여준다.

정리7(The criticism function L(C) is a linear function of C): 우리는 다양한 비평 포인트를 선택하도록 돕는 regularizer의 추가로 성능을 향상시킬 수 있다는 것을 발견했다. $2^{[n]} \mapsto \mathbb{R}$이 정규화 함수를 표현한다고 하고, 우리는 이 비용 함수를 최대화하는 비평 포인트들을 선택한다.

$$\max_{C \subseteq [n] \backslash S, |C| \leq c_*} L(C) + r(K, C) \qquad (10)$$

여기서 $[n]\backslash S$가 프로토타입을 포함하지 않는 모든 인덱스들을 나타내고 c_*는 원하는 비평 포인트들의 수를 의미한다. 다행히 (5)의 선형성으로 인해 최적화 함수(10)는 정규화 함수가 submodular일 때 submodular이다. 이 때 우리는 비평 선택에 다양성을 포함시킬 수 있는 regularizer 사용을 권장하며 log-determinant regularizer(Krauses, 2008)를 이용했을 때 성능이 가장 좋은 것을 발견했다. $C \times C$인 인덱스 쌍에 해당하는 K의 부분 매트릭스를 $K_{c,c}$라고 정의하고, 따라서 log-determinant regularizer는 아래와 같이 나온다.

$$r(K, C) = \log\det K_{c,c} \qquad (11)$$

이 방법은 여러 연구자들이 이론적 측면, 적용적 측면에서 모두 greedy 최적화가 최적화를 위해 가장 효과적이라는 것을 발견했다. 따라서 우리는 탐욕 알고리즘을 적용했다.

Related Work

현재 데이터를 요약하는 프로토타입을 선택하는 기술에 관한 연구는 굉장히 많지만 여기서 우리는 연구와 가장 관련된 레퍼런스들의 개요를 소개한다. K-medoid clustering은 데이터 포인트들의 대표적인 부분집합을 선택하기 위한 K-means clustering과 비슷한 고전적인 기술이고, 다양한 반복 알고리즘을 사용하여 해결할 수 있다. 빅데이터 시대가 되면서 데이터 요약 문제에 더 많은 관심이 집중되면서 이미지 요약(Simon, 2007) · 문서 요약(Lin & Bilmes, 2011)를 포함한 여러 영역에 대한 새로운 비용 함수와 알고리즘에 대한 연구가 진행되었다. 또한 cover digraph 접근법(Priebe, 2003)과 해석가능한 분류를 위한 프로토타입 선택(Bien & Tibshirani, 2011)과 같은 분류를 위해 조정된 set cover 문제의 변형에 관해 연구하는 분야도 생겨났다.

Submodular / Supermodular 함수는 조합 최적화 문제에서 연구되고 있으며, submodular는 베이지안 모델링에서 근사 추론을 위해 적용되었다(Koyejo, 2014). (6)에 필요한 기술적인 조건은 커널 유사성 점수의 평균에 기인한다. 특히 모든 평균을 합계로 대체하는 (6)의 analogue는 장면이나 문서 요약에 사용되었다고 잘 알려진 submodular 함수와 동일하다. 이 함수는 커널의 값이 양의 값을 가질 때, 즉, 추가적인 diagonal dominance 조건을 필요로 하지 않을 때 submodular로 알려져 있다. 반면에 평균화는 바람직한 균형 효과를 가지므로 합계를 사용할 때, 실무자는 유사한 밸런스를 맞추기 위해 추가 정규화 파라미터 λ 를 조정해야 한다.

Results

본 논문에서 손글씨인 USPS와 ImageNet 데이터셋을 사용하여 MMD-critic 기법에 대한 결과를 제시한다. 여기서 우리는 USPS 데이터셋의 기준과 비교하여 예측 성능 측면에서 프로토타입을 정량적으로 평가했고, 또한 사람 피험자 파일럿 연구로부터 얻은 최초 결과를 제시한다. 이 결과는 MMD-critic이 특히 사람들의 이해하는 데 유용한 모델 비평임을 보여주었다. 모든 데이터들은 **추론3**의 조건을 만족시키는 RBF 커널에 적용되었다.

The Nearest Prototype Classifiers: 우리의 주요 관심사는 해석가능한 프로토타입 선택과 비평이지만, 프로토타입은 학습 속도를 높이기 위해서 이웃 검색을 가장 가까운 프로토타입 분류 모델로 제한함으로써 nearest neighbor classifier와 같은 기억 기반 머신 러닝 방법이 유용할 수도 있다. 이 모델은 간접적이긴 하지만 프로토타입의 성능에 대한 객관적인 평가를 제공하고 하이퍼파라미터 설정에 유용하다. 우리는 커널에 의해 유도된 Hilbert 공간 거리를 사용하여 최근접 이웃 분류 모델을 사용한다.

$$\hat{y} = y_{i*}, where\ i^* = \underset{i \in S}{argmin} \parallel \hat{x} - x_i \parallel^2_{H_K} = \underset{i \in S}{argmax}\ k(\hat{x}, x_i)$$

< MMD-critic evaluated on USPS Digit Dataset>

수기로 작성된 숫자 USPS 데이터셋은 7291개의 학습 데이터, 2007개의 테스트 데이터로 0부터 9까지 그레이 스케일된 이미지들로 구성되어 있다. 우리는 2가지 종류의 RBF 커널 (1) global : 모든 데이터 포인트들 간에 계산되는 pairwise 커널 (2) local : 다른 클래스에는 유사성 점수가 0으로 주어지는 커널 들로 고려했다. 커널의 하이퍼파라미터 γ는 평균 교차 검증된 분류 성능을 극대화하는 것을 목적으로 선택되어 다른 모든 실험에 고정된 값으로 사용되었다.

분류: 우리는 MMD-critic을 사용하여 근접 프로토타입 분류 모델로 평가했고, Bien과 Tibshirani (2011)에 쓰였던 기준들과 비교했다. Figure1을 보면 왼쪽은 MMD-critic을 각각 global, local 방식으로 달리 적용하여 다른 기준 모델들과 비교하였다. 그 결과 다른 모델들과 비교할 수 있을 만함 성능을 갖췄음을 보였고, 우리는 MMD가 데이터 정보를 빠르게 요약하기 위해 효과적인 프로토타입 모델을 선정하는데 효과적이라는 점에 주목했다.

선택된 프로토타입과 비평: Figure1의 오른쪽 그림은 무작위로 선택된 프로토타입과 local 방식의 MMD-critic을 사용하여 Criticisms들이다. 여기서 우리는 무작위 방식이 일반적인 숫자들을 포착하는 반면, Criticisms은 명확하게 이상치 값들을 포착했음을 관찰할 수 있다.

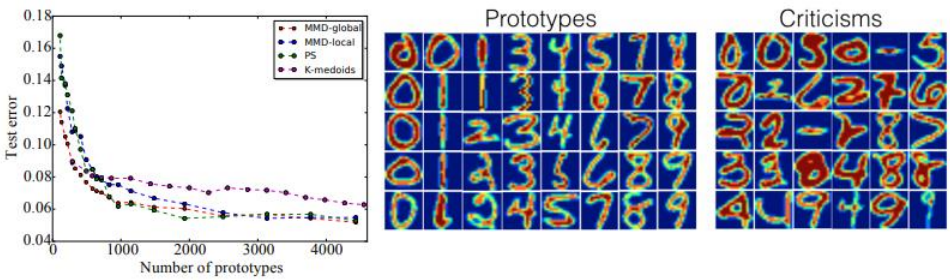

Figure 1: Classification error vs. number of prototypes $m = |S|$. `MMD-critic` shows comparable (or improved) performance as compared to other models (left). Random subset of prototypes and criticism from the USPS dataset (right).

< Qualitative Measure: Prototypes and Criticisms of Images>

이번 절은 ImageNet의 데이터를 사용했으며, 해당 데이터의 각 이미지는 2048 차원 벡터로 임베딩 되었고 1000개의 클래스 중의 하나에 속한다. Figure2에서 보이듯이 MMD-critic은 두 종류의 개 품종에 대해 합리적인 프로토타입과 Criticism을 학습하게 되는데, 왼쪽의 Criticisms은 흑백 사진, 개들이 움직이는 사진 등을 선별했다. 유사하게 오른쪽의 Criticisms을 보면 흔하지 않지만 잠재적으로 빈번한 개들이 의상을 착용한 사진들이 포착되었다.

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

< Quantitative measure: Prototypes and Criticisms improve interpretability>

또한 우리는 해석가능성의 객관적이고 주관적인 측정을 위해 사람 피험자 파일럿 연구를 수행했다. 위 실험은 ImageNet 데이터를 사용했고, '해석가능성'의 정의는 사용자가 결과를 정확하고 효율적으로 예측할 수 있으면 그 방법을 해석할 수 있다

고 간주했다. 그리고 이를 정량적으로 평가하기 위해 무작위로 추출된 샘플 데이터가 주어졌을 때 피험자가 그 데이터가 속한 클래스를 얼마나 잘 예측할 수 있는지, 얼마나 빠른 시간 내에 수행하는지를 측정했다.

위 실험에서 우리는 4가지 조건을 제시했다. (1) raw 이미지 사용 (2) 프로토타입만 사용한 해석 (3) 프로토타입 모델과 비평을 이용한 해석 (4) 그룹마다 균일한 비율로 데이터 샘플을 추출할 것 이들을 각각 2개의 종류로 100장의 이미지를 포함한다. 이 4가지 조건은 4명의 남성 피험자에게 밸런스를 맞춰 할당했고, 각 문항은 총 21문항으로 처음 3문항은 시험 문항으로 둬 분석 결과에는 배재했다. 피험자들은 정확성과 효율성에 대해 주관적으로 10개의 최대 5점을 가지는 리커트 척도 설문 질문에 응답하도록 요청받았고, 각 설문 조사는 한 쌍의 조건 중에 어느 것이 더 나았는지 비교하는 형식으로 진행되었다.

그 결과 피험자는 (3) 조건에서 M=87.5%, SD=20%라는 가장 좋은 성적을 거두었고, (2) 에서는 유사하게 M=75%, SD=41% 이라는 점수가 나왔다. 반면 나머지 (1), (4) 조건에서는 각각 37%, 36% 감소된 점수를 획득했다. 속도 측면에서 피험자는 (2) 조건에서 질문 당 1.04분으로 가장 적은 시간이 걸렸고 (4) 조건 1.31분, (3) 조건 1.37분, (1) 조건 1.86분을 기록했다. 또한 리커트 척도에서 피험자들은 (3) 조건을 가장 선호했으며, 해당 조건이 수많은 이미지 속에 숨겨진 패턴을 발견하고 시도하여 혼란을 덜 야기했고, 어떤 특징이 중요한지 나타내는 단서가 더 많아 수월했다는 의견을 제시했다. 특히 Prototype과 Criticism을 비교해달라는 질문에서 Criticism을 추가함으로써 원 데이터 내에서 공통된 특징을 더 쉽게 찾을 수 있었다고 전했다. 이를 통해 우리는 프로토타입과 함께 비평을 제공하는 것이 해석가능성을 향상시키는데 탁월한 방향임을 보여주었다.

Conclusion

우리는 ML의 해석가능성을 향상시키기 위해 프로토타입과 비평 선택 목적의 확장 가능한 프레임워크인 MMD-critic을 소개한다. 우리가 아는 선에서 위 방법은 ML에 대한 설명을 생성하기 위해 BMC 프레임워크를 활용하는 첫번째 연구이다. 또한 MMD-critic은 기존 방법에 비교될 만한 경쟁적인 성능을 보였다. 프르토타입과 함께 비평이 주어졌을 때 사람 피험자 연구에서 사람은 데이터 분포가 잘 설명되어야 하는 예측 작업을 더 잘 수행할 수 있음을 입증했다. 이는 비평과 프로토타입이 복잡한 데이터 분포의 해석 가능성을 향상시키기 위한 조치임을 시사한다. 향후 연구에서 커널 선택에 대한 영향과 submodularity를 위한 커널 매트릭스의 weaker 조건과 같은 MMD-critic 특징들을 더 탐구하기를 기대한다. Submodular 최적화 목적의 분산 알고리즘에 대한 최근 연구를 통해 더 큰 데이터 적용을 연구하고, 비평과 프로토타입이 인간의 이해에 어떻게 영향을 미치는지에 대한 대규모 사람 피험자 연구를 진행하고자 한다.