이미지 데이터 기반의 빠른 반사실적 예제 생성 기법 연구

김태형*, 김종국*
*고려대학교 전기전자공학부
magental@korea.ac.kr, jongkook@korea.ac.kr

A Study of Image Data Based Fast Counterfactual Instances Generation Method

Tae-Hyeong Kim*, Jong-Kook Kim*
*School of Electrical Engineering, Korea University

요 약

인공지능 기술이 사회 전반에 적용되면서 인공지능에 대한 인간의 이해도 역시 중요해지고 있다. 이러한 필요성을 기반으로 설명 가능한 인공지능(XAI) 분야 연구가 현재 활발히 진행되고 있다. 이 중 입력의 변화를 통하여 반사실적 대안을 제시하는 반사실적 예제 기반의 설명은 피처 수가 많아지는 이미지 데이터에서 연산량이 크게 증가하는 단점이 있다. 본 연구에서는 이러한 단점을 해결하고자 이미지의 추상화된 피처 영역에서 프로토타입 피처를 이용한 반사실적 예제를 생성하는 기법을 제안한다. 나아가 이러한 이미지 형식의 반사실적 예제를 활용할 분야를 제시하고자 한다.

1. 서론

인공지능에 대한 발전에 힘입어 의료, 국방, 금융, 법률 등 다양한 사회적 영역에서 인공지능을 활용하고자 하는 시도가 증가하고 있다. 하지만 심층 신경망 기반 인공지능(Deep Neural Network Based AI)은 의사결정 과정을 사용자가 이해할 수 없다는 문제점이 존재한다. 따라서 이러한 인공지능을 활용하는 시스템은 추론 과정에 대한 투명성과 공정성을 사용자에게 제공할 필요성이 크다. 유럽 연합은 GDPR 조항에서 이러한 점을 명기하여, 인공지능 기반 상품 혹은 서비스가 개인 데이터를 사용하여 의사 결정을 수행할 때 그 과정에 대해 사용자가 이해할 수 있는 근거를 제공할 것을 규제한 바 있다.[1] 이러한 사회적 수요에 힘입어 설명 가능한 인공지능(eXplainable AI, XAI)에 대한 연구가 활발히 진행되는 중이다.

반사실적 예제를 생성하는 기법은 설명 가능한 인공지능 모델 중 예제를 바탕으로 사용자에게 추론의 근거를 설명하는 기법에 속한다. 반사실적 추론은 원인 X 에 대해 작은 변화를 가했을 때, 결과 Y 와는 다른 결과 Y'를 생성하는 가상의 인과관계를 의미한다.[2] 반사실적 설명은 반사실적 추론을 바탕으로 하는 설명 방법이다. 인공지능 모델이 사용자가 원하는 추론값을 도출하기 위해 사용자가 가진 데이터를 최소한으로 변형하여 반사실적 예제를 생성하는 기법이다.[3] 이 때 데이터 피처의 변화도는 모델이 다른 추론에 도달하기 위한 조건을 의미한다. 사용자는 사용자가 보유한 데이터와 이에 대한 반사실적 예제들을 비교함으로써 모델의 의사결정 과정에 대한 설명을 얻게 된다.

하지만 이미지 데이터에서는 이미지의 픽셀의 수를 데이터 피처로 볼 수 있다. 따라서 피처의 변동을 통한 반사실적 예제 생성 방법은 이미지 데이터에서 연산량이 크게 증가하는 경향이 뚜렷하다. 이는 기본적으로 이미지 데이터 영역에서 현행 연구가 상대적으로 더딘 발전속도를 불러오는 결과를 야기한다. 특히 이미지 기반 설명가능한 모델의 활용[4, 5]에 있어 연산량 증대로 인한 예제 제공 속도의 감소는 치명적이다. 따라서 컴퓨터 비전 분야에서 반사실적 예제를 활용하려면 생성 속도에 대한 특수한 전략이 불가피하다. 본 연구는 기존의 반사실적 예제 생성 연구를 토대로, 이미지 데이터 기반 인공지능 모델에 적용할

수 있는 반사실적 예제 생성 기법에 대해 제안하고자 한다. 나아가 본 연구에서 제안한 방법이 이미지 데이터상에서 기존 방식에 비해 반사실적 예제를 얼마나 빠르게 생성할 수 있는지 정량적으로 보이고자 한다.

2. 관련 연구

반사실적 예제 생성은 [3]에서 제안하는 최적화 문제를 바탕으로 한다. [3]에서는 손실함수에 대한 최적화 문제 (1)를 제안한다. 원본 입력 x 가 모델에 의해 원본 추론 $y = f(x)$를 도출한다면, 반사실적 예제 x'은 원 입력값 x 와 각 피쳐 관점에서 최대한 유사해야 한다. 그 거리값은 수식 (2)로 정의된다. 반사실적 예제에 대한 모델의 추론 $f'(x)$와 목표값 y' 간의 거리 역시 최소화되어야 한다.

$$\arg\min_{x'}\max_{\lambda} L(x,x',y',\lambda) = \lambda(f(x')-y')^2 + d(x,x') \quad (1)$$

$$d(x,x') = \sum_{j=0}^{p}\frac{|x_j - x'_j|}{MAD_j} \quad (2)$$

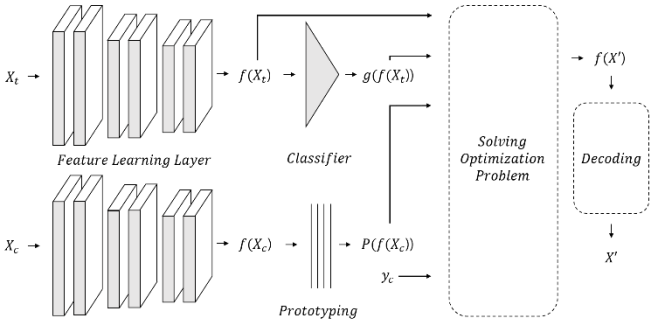
이 최적화 문제를 바탕으로 하여, 반사실적 예제가 정규분포에 따르도록 좀 더 최적화 문제를 세분화하거나[6] 전체 데이터셋의 프로토타입을 바탕으로 더 빠르게 반사실적 예제를 생성하는[7] 후속 연구가 등장하였다.

[4,8,9]에서는 이미지 데이터에 대한 반사실적 예제 생성 방법을 제안하였다. [4]는 이미지 분류 모델의 가장 추상화된 피쳐 영역에서 피쳐 쿼리를 그리디하게 탐색하여 반사실적 예제 생성에 사용한다. [8,9]는 모델의 의사결정을 내릴 때 각 피쳐들 간의 상대적 중요도를 표현하는 방법인 Saliency map 을 이용한다. 위 Saliency map 을 토대로 변동시킬 이미지 픽셀을 정한 후, 이를 휴리스틱한 방법 혹은 생성 모델을 이용한 방법으로 변화를 주어 반사실적 예제를 생성한다.

반사실적 예제를 인간이 식별 가능한 유의미한 데이터로 보는 견해에 따라, 생성된 반사실적 예제를 다른 영역에 활용하는 시도 역시 발견된다. 반사실적 예제를 통한 설명 모델을 사람에게 교육하는 방향으로 활용하거나[4], 반사실적 예제를 인공지능 학습에 직접 적용하여 성능 향상을 꾀하는[5] 연구가 진행되었다.

3. 방법 제안

본 연구는 [4]의 방법 일부를 기조로 하여 이미지의 추상화된 피쳐 영역에서의 반사실적 예제 생성 기법



(그림 1) 제안 기법의 모식도

을 제안한다. 그림 1 과 같이 이미지 분류 모델은 피쳐 학습층과 분류층으로 크게 구분할 수 있다. 이미지 쿼리 X 에 대하여 이미지의 추상화된 피쳐 영역은 피쳐 학습층의 결과물 $Z=f(X)$로 정의한다. 쿼리에 대한 모델의 추론값 집합은 $g(Z)=g(f(X))$로 정의한다.

단순히 추상화된 피쳐에 무작위한 변화를 주는 방식은 비효율적인 연산량 증대를 야기할 수 있으므로 방향성을 잡아 줄 프로토타입 피쳐를 사용한다. 생성하고자 하는 반사실적 예제 쿼리 X'에 대한 목표값이 y_c 라고 할 때 입력 X 에 대한 목표값은 y_c 와 다름을 전제로 한다. 목표값 라벨을 가진 이미지 데이터셋 X_c 는 수식 (3)과 같이 정의한다.

$$X_c = \{\, x \mid g(f(x)) = y_c \,\} \quad (3)$$

이때 프로토타입 피쳐는 이미지 데이터셋에서 랜덤하게 뽑은 n 개의 데이터를 토대로, n 개 데이터의 피쳐값의 중앙값과 가장 가까운 L2 거리값을 지닌 단일 피쳐로 정의한다. 이는 수식 (4)로 나타내어진다.

$$Proto(f(X_c)) = Proto(Z_c) = \arg\min_{z_c} |z_c - z_{mean}|^2 \quad (4)$$

$$z_{mean} = \sum_{j=1}^{n}\frac{f(x_c^j)}{n}$$

$$z_c = f(x_c)$$

$$x_c \in \{x_c^1, x_c^2, \dots, x_c^n\} \subset X_c$$

이렇게 구한 프로토타입 피쳐 $Proto(Z_c)$는 추상화된 피쳐 영역에서 클래스 c 에 속하는 대표성을 띄게 된다. 따라서 반사실적 피쳐 Z'의 전체 영역에 변동을 줄 수 있는 가이드라인으로 삼을 수 있다. 이 가정을 토대로 다음 최적화 문제 (5)를 풀이한다.

$$\arg\min_{a} \mathcal{L} = \lambda \mathcal{L}_{ce} + \mathcal{L}_{dis} \quad (5)$$
$$\mathcal{L}_{ce} = cross\ entropy\ loss\ between\ g(Z')\ and\ y_c$$
$$\mathcal{L}_{dis} = \sum |Z - Z'|^2$$
$$Z' = aZ + (1-a)Proto(Z_c),\ a \in [0,1]$$

최적화 문제 (5)를 만족하는 반사실적 피쳐 Z'는 분류층에 입력되었을 때 그 결과 $g(Z')$ 가 목표 추론 값 y_c 와 최대한 유사하다. 동시에 Z'는 원본 피쳐 Z 와 최대한 유사한 특성을 지닌다. 따라서 반사실적 피쳐 Z'는 반사실적 예제의 정의에 부합한다.

마지막으로 최적화 문제 (5)를 만족하는 반사실적 피쳐 Z'를 이미 학습된 디코딩 계층을 통해 복원한다. 이 때 디코딩 계층은 반사실적 피쳐 Z'가 속하는 클래스에 대한 통계적 정보를 이미 학습하고 있으므로 피쳐를 추가로 정제하는 효과가 있다. 이 디코딩 계층을 통하여 반사실적 예제 X'를 생성한다.

4. 실험

본 연구는 반사실적 예제의 생성 속도에 초점을 두어 그 생성 시간을 기존 반사실적 예제 생성 기법과 비교하고자 한다. 실험 GPU 모델은 TITAN Xp 를 사용하였다. 데이터셋은 CIFAR-10 을 사용하였다. CIFAR-10 은 50000 개의 트레이닝 이미지, 10000 개의 테스트 이미지로 구성된 데이터셋이며 총 10 개의 클래스로 분류된다. 데이터셋에 속하는 모든 단일 이미지들은 3x32x32 픽셀 크기를 가진다.

실험에 사용한 분류 모델은 사전에 CIFAR-10 이미지를 학습한 CNN 기반의 ResNet-34 를 사용하였다. ResNet-34 의 피쳐 학습층에서 가장 추상화된 피쳐는 512 개의 피쳐 수를 가진다. 반사실적 예제의 정확한 생성을 위해 분류 모델 학습에 사용한 트레이닝 데이터셋을 반사실적 예제의 원본 이미지로 설정하였다. 최적화 문제 (5)에 사용한 상수 λ 는 임의의 최적의 값으로 설정하였다.

본 실험은 의사결정 모델을 사전학습한 후, 데이터셋에서 랜덤하게 추출한 이미지들에 본 연구와 기존의 반사실적 생성 예제 방법들을 적용하였다. 각 반사실적 예제의 목표 클래스는 원본 이미지가 속하는 클래스의 바로 다음 클래스 값으로 정의하였다. 비교하고자 하는 기존 반사실적 예제 생성 모델[3, 7]은 [10]에서 제공되는 라이브러리를 활용하였다.

Model	Wachter et al.[3]	Looveren et al.[7]	Ours
ResNet-34	100.56	35.34	**1.47**

<표 1> 반사실적 예제 생성 속도 비교 (time/instance)

실험 결과는 <표 1>과 같이 나타났다. 본 연구에서 제안한 방법이 [3, 9]에 비해 예제 생성 속도가 월등히 빠른 경향을 보인다. 본 실험에 사용된 단일 이미지 데이터는 3x32x32 개의 픽셀 크기를 가지므로 총 3072 개의 피쳐를 가진 데이터로 볼 수 있다. 따라서 최적화 문제를 단순히 풀이하여 피쳐를 변동하는 기존의 반사실적 문제는 이미지 기반 데이터에서 이미지의 픽셀 크기에 비례해 연산량이 증가한다. 하지만 본 실험에 사용된 모델의 추상화된 피쳐는 512 개의 피쳐 수를 가지게 된다. 본 연구에서 제안하는 방법은 추상화된 피쳐 영역에서 변동을 시행하므로 기존 방법보다 더 빠르게 예제를 생성할 수 있다. 또한 모델의 분류층의 추상화된 피쳐 크기에 영향을 받으므로 이미지의 픽셀 크기에 상대적으로 견고하다. 마지막으로, 본 제안에서는 정의된 프로토타입을 통한 변동이 이루어지므로 최적화 문제를 무작위적 초기 방향으로 풀이하는 기존 방법들에 비해 속도적 이점을 보유한다.

5. 결론 및 향후 연구 방향

본 논문은 앞서 밝힌 연구를 통해 이미지 데이터셋 기반의 반사실적 예시를 빠르게 생성하는 방식을 제안하였다. 이러한 반사실적 예시 생성 속도 향상은 반사실적 예시 기반 모델 설명 기법 연구에 있어 필요불가결하므로 이에 대한 개선이 필수적이다. 또한 산업 각 계층에 컴퓨터 비전 분야의 수요가 높아지는 바, 사용자에게 인공지능 모델에 대한 빠른 설명을 제공할 수 있다. 동시에 교육, 인공지능 학습 등[4,5] 다양한 영역에서 부가적 이점을 안겨줄 것으로 기대된다.

본 논문에서는 이에 더불어 다음과 같은 연구 방향을 제안한다. 반사실적 예제는 의도적으로 입력에 변동을 줌으로써 인공지능 모델의 추론을 변화시킨다는 성질을 지니고 있다. 따라서 반사실적 예제를 제한된 형태의 적대적 공격으로 분류할 수 있다는 전제하에, 반사실적 예제를 현재 학습이 진행중인 인공지능에 실시간으로 적용하여 적대적 학습의 효과를 수행할 것으로 예상된다. [3,11]

Acknowledgement

이 논문은 2016 년도 정부(교육부)의 재원으로 한국연구재단 기초연구사업의 지원을 받아 수행된 연구임(NRF-2016R1D1A1B04933156)

참고문헌

[1] A. Adadi and M. Berrada, Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI), in *IEEE Access*, vol. 6, pp. 52138-52160, 2018.

[2] D. Lewis, Counterfactuals. Basil Blackwell, Oxford, 1973.

[3] S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr, Harv. JL & Tech., 31:841, 2017.

[4] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh and S Lee, Counterfactual visual explanations, In Proceedings of the ICML, 2019, PMLR 97:2376-2384.

[5] CH. Chang, GA. Adam and A. Goldenberg, Towards Robust Classification Model by Counterfactual and Invariant Data generation, In Proceedings of the CVPR, 2021.

[6] S. Dandl, C. Molnar, M. Binder and B. Bischl. Multi-Objective Counterfactual Explanations. Parallel Problem Solving from Nature – PPSN XVI. PPSN 2020. Lecture Notes in Computer Science, vol 12269. Springer, Cham. 2020.

[7] A. Van Looveren and J. Klaise, Interpretable Counterfactual Explanations Guided by Prototypes. In: Machine Learning and Knowledge Discovery in Databases. Research Track. ECML PKDD 2021. Lecture Notes in Computer Science, vol 12976. Springer, Cham. 2021.

[8] R. Fong and A. Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In Proceedings of the ICCV, 2017. pp. 3449-3457

[9] CH. Chang, E. Creager, A. Goldenberg and D. Duvenaud, Explaining image classifiers by counterfactual generation. arXiv preprint, arXiv:1807.08024, 2018.

[10] A. Van Looveren, J. Klaise and V. Giovanni and C. Alexandru. Alibi Explain: Algorithms for Explaining Machine Learning Models, Journal of Machine Learning Research, v22, 181, 1-7, 2021

[11] IJ. Goodfellow, J. Shlens and C. Szegedy, Explaining and Harnessing Adversarial Examples, In Proceedings of the ICML, 2015