적대적 생성 신경망의 풀린 특징 공간에서의 잠재벡터 조작을 통한 분류기에 대한 반사실 설명*

나승협 O1 이성환 1

1 고려대학교 인공지능학과

naash@korea.ac.kr, sw.lee@korea.ac.kr

Counterfactual Explanation for Classifier Through Latent Vector Manipulation in Disentangled Feature Space of Generative Adversarial Network

Seung-Hyup Na O1 Seong-Whan Lee 1

1 Department of Artificial Intelligence, Korea University

요 약

반사실 설명이란 기존 입력에 어떤 변화를 주어야 그 결과가 반대 사실인지 설명하는 것이며, 이는 블랙박스 형태의 인공지능 분류기에 대해 변형된 입력인 예시를 통해 간접적으로 설명한다. 반사실은 입력으로부터 최소한으로 변화되어 사용자가 쉽게 현실화할 수 있어야 하고 동시에 실제 사례(데이터)처럼 그럴듯해야 한다. 두 가지 특성을 같이 고려하기 위하여 적대적 생성 신경망의 의미정보가 풀려 있는 특징 공간 $\mathcal{W}$를 활용한다. 클래스 정보의 선형 변형 가능성을 활용하여 입력의 잠재벡터를 조작한다. 결론적으로, 입력으로부터 클래스 정보만을 최소한으로 변화시켜 분류기의 결정경계를 넘는 반사실을 현실적이고 그럴듯한 반사실을 생성한다.

1. 서 론

인공지능 기술이 발전됨에 따라 현업에서 인공지능 모델에 대한 활용이 높아지고 있으며, 특히 중요한 의사결정을 대체하고 업무의 자동화를 이루어 일의 효율성을 높이고 있다. 이에 따라, 인공지능에 대한 해석가능성(interpretability)에 대한 수요가 증가되어 설명 가능 인공지능(XAI: eXplainable Artificial Intelligence) 기술이 활발히 개발되고 있다. 이는 인공지능에 대해 사람이 이해할 수 있는 형태로 설명을 제공하는 기술이다. 하지만, 복잡한 신경망 구조를 띄고 있는 블랙박스 형태의 인공지능 모델에 대해 설명하는 것은 쉽지 않으며 이를 간접적으로 할 수 있는 방법이 반사실 설명이다. 반사실 설명 기술이란 인공지능 분류기의 판단에 반하는 판단, 즉 반대 사실을 얻으려면 기존 입력에서 어떠한 것들을 변화시켜야 하는지 설명하는 것이다.

기존 방법론들에서는 효과적인 반사실을 반사실 클래스로 판단되는 최소한으로 변경된 입력이라고 정의하며 이는 사용자가 쉽게 현실화시킬 수 있는 반사실이다. 수학적으로 식 (1)과 같이 정의된다[1].

$$\hat{x}^* = \underset{\hat{x}}{\operatorname{argmin}}\, d(x, \hat{x}) \;\; \text{s.t. } f(\hat{x}) = \hat{y} \qquad (1)$$

$x, y, \hat{x}, \hat{y}$는 각각 입력, 입력 클래스, 반사실, 반사실 클래스를 나타낸다.

하지만, 최근에는 데이터 매니폴드(manifold)에 대한 근접성(실제 데이터와 비슷하여 그럴듯한)이 중요 특성으로 떠오르고 있다[2]. 위에서 언급한 두가지 특성은 서로 트레이드 오프(trade-off) 관계에 있기 때문에 동시에 고려하여 반사실을 생성하는 것은 쉽지 않은 문제이다. 이에 대한 근거를 들자면, 직관적으로, 입력은 해당 클래스 데이터 매니폴드에 속해있고 이를 반사실 클래스 데이터 매니폴드에 포함되게끔 하려면 입력에 변화를 주어야만 하기 때문에 입력에 대한 변화량이 증가할 수밖에 없다.

이 관계를 고려하여 현실적이면서 그럴듯한 반사실을 생성하기 위하여, 적대적 생성 신경망(GAN: Generative Adversarial Network)의 의미정보가 풀린(disentangled) 특징 공간에서는 그림 1처럼 의미 정보를 선형적으로 변화시킬 수 있다는 특성[3]을 활용한다. 입력으로부터 반사실 클래스 데이터처럼 보이게끔 하는 클래스 정보만을 최소한으로 변경시켜 분류기의 결정경

* 이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2022-0-00984, 플러그앤플레이 방식으로 설명가능성을 제공하는 인공지능 기술 개발 및 인공지능 시스템에 대한 설명 제공 검증)

계를 넘겨 반사실을 생성한다.

관련 연구로는, [4]와 [2]가 있다. [4]에서는 본 연구와 같이 GAN의 특징 공간에서 잠재벡터로부터 이미지를 생성하고 분류기에 통과시켰을 때의 반사실 클래스에 대한 신뢰점수를 높이는 방향으로 잠재벡터를 변화시키는 목적함수를 최적화하는 방법을 채용한다. 하지만, 극솟값에 빠져 유효한 반사실을 찾지 못하거나 여러 손실항들의 작용을 종합적으로 고려하지 않는다는 단점이 있다.

[2]에서도 GAN의 특징 공간을 활용하긴 하지만, 이 공간은 의미정보가 풀린 공간이 아니며, 또한 입력에서 변경시킬 때 특징들의 관계를 종합적으로 고려하지 않으며 개별적으로 클래스 신뢰도 점수에 대한 영향을 보고 변경시킨다.

2. 방법론

방법론을 요약하면 다음과 같다. 특징 공간에서 입력으로부터 의미 정보를 조작시킬 방향의 기준이 되는 참조 반사실을 생성한다. 입력과 참조 반사실에 대한 잠재벡터를 유도하고 두 벡터의 내삽(interpolation)을 진행하며 분류기의 결정경계를 넘은 벡터를 찾는다.

2.1 참조 반사실 생성

조건부 적대적 생성 신경망인 [5]을 활용하였으며, 클래스 정보가 조건으로서 입력된다. 판별기의 영향으로 생성기의 출력 샘플은 타겟 클래스 매니폴드에 속하게 되며 일관성 손실 $\mathcal{L}_{cyc}$에 의해 입력으로부터 변화를 제약 받게 된다. 따라서, 출력된 샘플은 입력 기준에서 반사실 데이터 매니폴드로의 방향을 나타내는 참조 반사실로 활용하기에 적합하다. 설명하고자 하는 분류기의 클래스 판단 기준을 생성기에 반영시켜야 하기 때문에, 실제 데이터에 대해서 생성기와 판별기를 학습시킬 때 라벨은 분류기에 의해 예측된 클래스를 사용한다. 이 신경망 구조에서 인코더 $\mathcal{E}$와 디코더 $\mathcal{D}$사이의 공간 $\mathcal{W}$는 공간 $\mathcal{Z}$처럼 구체 형태의 분포인 가우시안 분포를 띄어야 한다는 제약 조건이 존재하지 않고 입력공간으로부터 $\mathcal{E}$를 통해 학습된 조각 연속 맵핑(piecewise continuous mapping) 되며, 생성기가 $\mathcal{D}$를 통해 이미지를 생성할 때 더 실제와 같은 데이터를 생성하기 위해서 의미 정보들이 얽히지 않고 풀린 형태로 맵핑이 되게 된다.[3] 우리는 처음으로 이 특징 공간을 통해 반사실 설명을 생성한다.

2.2 잠재벡터 도출

$\mathcal{W}$공간에서 입력에 대한 잠재벡터를 구하기 위해서 최적화 기반 GAN 반전(optimization-based GAN inversion) 프로세스를 식 (2)와 같이 진행한다.

$$w^* = \arg\min_w \|\mathcal{D}(w) - x\|_2^2 + \|\mathcal{E}(\mathcal{D}(w), \hat{y}) - \mathcal{E}(x, \hat{y})\|_2^2 \quad (2)$$

각 손실항은, 잠재벡터 w가 입력 공간 및 특징 공간에서 디코딩 및 인코딩 후 입력 값에 대한 결과와 같게끔 강제하는 항이다. 참조 반사실에 대한 잠재벡터는 간단히 $\hat{w} = \mathcal{E}(x, \hat{y})$를 통해 구할 수 있다.

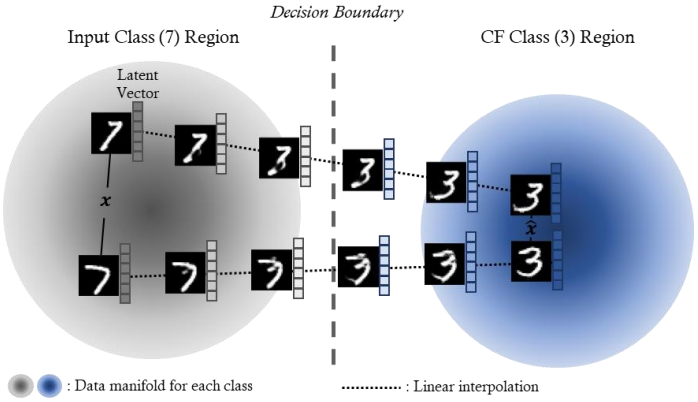

그림 1. 풀려있는 공간에서의 입력과 참조 반사실 사이의 내삽을 통한 잠재벡터 조작

2.3 클래스 정보 조작

$\mathcal{W}$ 공간에서 입력과 참조 반사실에 대한 잠재벡터를 조절하여 현실적이고 그럴듯한 반사실을 얻는 것이 목적이다. 입력을 기준으로 참조 반사실 방향으로, 식 (3)과 같은 선형 내삽을 통해 클래스 정보를 변경시키며 분류기의 결정 경계를 넘었는지 확인한다.

$$\hat{w}^* = (1-\lambda)w^* + \lambda\hat{w},\ \lambda \in (0,1) \quad (3)$$

그림 1과 같이 내삽 변수인 λ를 서서히 증가시키며, 잠재벡터 $\hat{w}^*$를 디코딩 후 분류기를 통과해 경계를 넘었는지 여부를 확인하면서 넘었을 때 λ를 찾는다.

3. 특징 공간 및 조작된 반사실 평가

분류기를 통과해 경계를 넘었는지 여부를 확인하면서 넘었을 때 λ를 찾는다. 본 논문에서는, 숫자에 대한 손글씨 데이터로 이루어진 MNIST 데이터셋과 주택 담보 대출 위험도 평가 데이터로 이루어진 HELOC 데이터셋을 사용하였다. 데이터셋 전체의 80%는 학습에 사용하고 20%인 테스트 샘플에 대해서 실험을 진행했다. 각 데이터셋의 테스트셋에 대한 분류기의 정확도는 각각 98.9%, 74.8%이다.

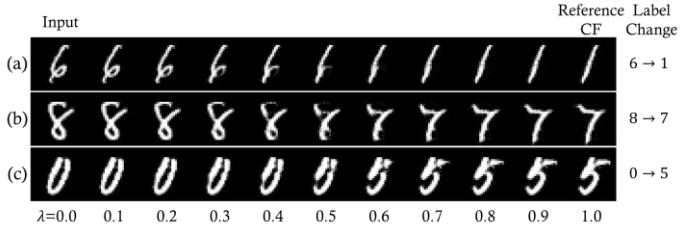

그림 2. MNIST 데이터셋 실험에서, 잠재벡터 조작에 따른 디코딩된 샘플의 부드러운 클래스 정보 변화

그림 2,3과 같이, 우리는 각 데이터셋에 대해서 특징공간에서 내삽 변수 λ를 0에서 1로 0.1씩 증가시키면서 디코딩된 샘플의 변화를 관찰한 결과, 입력에서 참조 반사실로 의미 정보, 즉 클래스 정보가 천천히 변하는 것을 확인하였다. 그림 2에서는 숫자의 형태를 바꾸는 클래스 정보들이 바뀌었으며, 그림 3에서는 해당 데이터셋의 특성대로 각 특성 값들이 단조적으로 증가 또는 감소하면서 클래스 정보가 바뀌었음을 확인하였다. 특징공간 $\mathcal{W}$는 의미 정보가 서로 풀려 있는 공간이기 때문에 이와 같이 선형적으로 잠재벡터가 변함에 따라 클래스 정보가 선형적으로 변한 것이다.

Feature	λ										
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
ExternalRiskEstimate	65	66	66	67	67	68	68	69	70	70	71
MSinceOldestTradeOpen	103	115	126	138	151	163	175	187	199	213	224
MSinceMostRecentTradeOpen	0	1	1	1	1	1	2	2	2	2	2
AverageMInFile	46	50	53	57	61	64	68	71	75	78	81
NumSatisfactoryTrades	30	31	33	34	34	35	36	37	38	38	38
NumTrades60Ever2DerogPubRec	0	0	0	0	0	0	0	0	0	0	0
NumTrades90Ever2DerogPubRec	0	0	0	0	0	0	0	0	0	0	0
PercentTradesNeverDelq	92	93	94	95	95	96	97	98	99	99	99
MSinceMostRecentDelq	38	40	42	44	46	47	49	51	52	55	59
MaxDelq2PublicRecLast12M	6	6	6	6	6	6	6	6	6	6	6
MaxDelqEver	6	6	6	6	6	6	6	6	6	6	6
NumTotalTrades	31	32	33	33	34	35	35	36	36	36	36
NumTradesOpeninLast12M	4	4	4	4	4	3	3	3	3	3	3
PercentInstallTrades	47	46	44	43	41	39	38	36	35	34	33
MSinceMostRecentInqexcl7days	3	3	3	3	3	3	3	3	3	3	3
NumInqLast6M	3	3	2	2	2	2	2	2	1	1	1
NumInqLast6Mexcl7days	2	2	2	2	2	2	2	2	2	2	2
NetFractionRevolvingBurden	54	52	50	48	46	44	42	40	38	36	34
NetFractionInstallBurden	98	96	94	91	89	87	85	82	80	77	74
NumRevolvingTradesWBalance	5	5	5	5	5	5	5	5	5	5	5
NumInstallTradesWBalance	3	3	3	3	4	4	4	4	4	4	4
NumBank2NatlTradesWHighUtilization	2	2	2	2	2	2	2	2	2	2	1
PercentTradesWBalance	62	63	65	66	68	69	71	72	74	75	75

그림 3. HELOC 데이터셋 실험에서, 잠재벡터 조작에 따른 디코딩된 샘플의 각 특성들이 부드럽게 단조 증가 및 감소하며 보이는 클래스 정보 변화

그림 4는 조작을 통해 결정경계를 넘은 직후의 반사실을 입력 및 참조 반사실과 비교한 결과이다. 조작된 반사실은 입력으로부터 변화량이 참조 반사실 보다는 더 적고 동시에 참조 반사실과 마찬가지로 반사실 클래스 데이터처럼 보인다. 이는 조작을 통해 클래스 정보만 변화시켜 결정경계 근처의 반사실을 얻었기 때문이다.

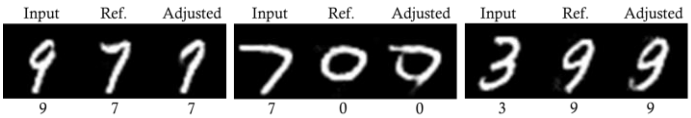

그림 4. 입력 및 참조 반사실 및 조작된 결정 경계 근처의 반사실 샘플의 비교 결과

4. 결론 및 향후 연구

본 연구에서는 적대적 생성 신경망의 의미정보가 풀린 특징 공간에서는 의미 정보를 선형적으로 변화시킬 수 있다는 특성을 활용하여. 현실적이고 그럴듯한 반사실을 생성하였다. 이러한 특징 공간은 활용성이 더 확대될 것으로 기대하며, 풀린 공간에 대한 추가 분석을 통해 더 개선된 반사실을 생성하는 것이 앞으로 연구할 과제로 판단된다.

참 고 문 헌

[1] Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31, 841-887.

[2] Kenny, E. M., & Keane, M. T. (2021). On generating plausible counterfactual and semi-factual explanations for deep learning. *In Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 11575-11585).

[3] Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4401-4410).

[4] Liu, S., Kailkhura, B., Loveland, D., & Han, Y. (2019). Generative counterfactual introspection for explainable deep learning. *In IEEE Global Conference on Signal and Information Processing* (pp. 735-739).

[5] Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., & Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8789-8797).