적대적 도구 추정을 통한 인과 특징 추론과 네트워크의 견고성을 위한 인과 요소 주입 방법

Demystifying Causal Features on Adversarial Examples and Causal Inoculation for Robust Network by Adversarial Instrumental Variable Regression[1)]

김준호*, 이병관*, 노용만
Image and Video Systems Lab, School of Electrical Engineering, KAIST, South Korea
{arkimjh, leebk, ymro}@kaist.ac.kr

요 약

적대적 예시의 인과관계는 아직 명확히 밝혀지지 않았으며 네트워크의 취약성을 보완하기 위한 중요한 논점입니다. 본 논문에서는 적대적으로 훈련된 네트워크의 취약점을 인과적 관점에서 파헤치기 위해 적대적 도구변수 회귀 방법을 제안합니다. 이 방법을 통해 알려지지 않은 교란 요인으로부터 분리된 편향되지 않은 적대적 예측의 인과관계를 추정할 수 있습니다. 우리의 접근 방식은 인과 특징 분류기 (즉, 가설 모델)과 인과적 특징을 찾는데 방해가 되는 극단적 반증들(즉, 테스트 함수) 사이의 제로섬 최적화 게임을 활용하여 내재 된 인과적 특징을 밝히는 것을 목표로 합니다. 우리는 광범위한 분석을 통해 추정된 인과적 특징이 적대적 견고성에 대한 올바른 예측과 높은 연관성을 가지며 반증들은 이를 크게 벗어나게 만드는 극단적 특징을 보인다는 것을 입증하였으며 이러한 특성을 이용해 적대적 견고성을 향상시키기 위한 CAusal FEatures(CAFE) 방법을 제안합니다.

1. 서 론

적대적 예시는 인간이 구별할 수 없는 섭동을 추가하여 딥 네트워크를 교란하여 현실 세계에서 시스템이 오작동하도록 하는 문제를 일으킵니다. 이를 해결하기 위해 적대적 예시가 유발하는 오류와의 인과관계를 파악하기 위한 많은 연구가 진행되었으나, 여전히 근본적인 원인에 관한 연구가 부족합니다. 특히 기존 연구에선 학습된 연관성에서 가짜 상관관계가 쉽게 유도되기 때문에 편향된 관점이 존재할 수 있는 환경에서 진행되어 적대적 예시의 진정한 인과관계를 해석할 수 없습니다.

인과적 관점에서 적대적 취약점의 출처를 설명하고 진정한 인과관계를 추론하기 위해서는 주어진 적대적 예시의 데이터 집단에 대한 단순한 연관성 분석을 넘어 인과관계를 추정하는 개입 지향적 접근 방식 (즉, 인과 추론)을 사용해야 합니다. 이를 위한 효과적인 도구 중 하나는 도구변수로 인과관계 추론의 내생성을 높이는 미지의 교란 요인으로부터 편향되지 않은 환경을 제공합니다.

구체적으로, 그림 1과 같이 인과 추론을 위한 데이터 생성 프로세스(DGP)[1]를 고려할 때, 알려지지 않은 교란변수 U의 존재는 인과 추정자 h(즉, 가설 모델)가 처리 변수 T와 결과 Y 사이의 인과관계를 추정하는 것을 방해하는 백도어 경로를 생성하여 가짜 상관관계를 유도할 수 있습니다 (T←U→Y). 이를 적대적 환경에서 해석하면, 교란변수 U가 적대적 기원에 대한 모호한 해석을 쉽게 유발하여 적대적 사례와 정답 레이블 사이에 가짜 상관관계를 생성할 수 있습니다. 본 논문의 저자들은 네트워크 f에서 파생된 중간 특징 표현에 개입하여 적대적 예시와 자연 예시 사이의 DNN의 특징 공간에서 특징 변이를 Z로 정의하고 도구변수로 채택함으로써, 편향되지 않은 h로부터 진정한 인과관계를 추정합니다 (Z→T→Y).

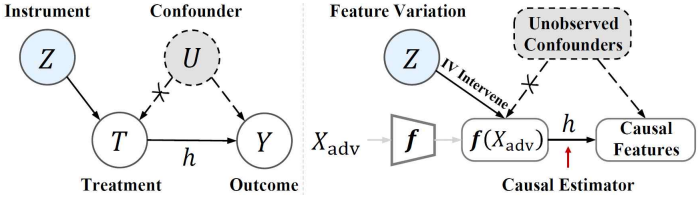

그림 1 도구변수를 사용한 데이터 생성 프로세스(DGP). Z를 통해 알 수 없는 교란변수 U를 배제하고 변수 T와 결과 Y 사이의 인과관계를 추정함.

광범위한 분석을 통해 적대적 사례에서 특징 시각화를 활용하여 적대적 예제에 대한 인과적 특징을 해석하고 인과적 특징을 사람이 인식할 수 있는 방식으로 보여줍니다. 또한, 추정된 인과적 특징을 도입하여 네트워크에 강인성을 효율적으로 주입할 수 있는 CAusal FEatures (CAFE) 방법을 제시하고 효과적으로 공격을 방어함을 실험적으로 보였습니다.

2. 방 법

적대적 예시에서 밀접한 관련이 있는 내재된 인과적 특징을 추정하는 적대적 도구변수 회귀 추정 방법에 대해 설명합니다. 이를 위해 먼저 GMM을 사용한 비모수적 도구변수 회귀 추정을 살펴볼 것입니다. 그리고 이를 적대적 환경에 접목시킨 후 네트워크의 적대적 인과 특징을 파악할 것입니다.

2.1 비모수적 도구변수 회귀 추정

1) 본 논문은 The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023에서 통과된 동 제목 논문의 요약본임을 밝힙니다.

우리는 먼저 조건부 모멘트 제한 (CMR)[2]에서 문제를 시작합니다. CMR은 수식적으로 가설 공간 H에 대한 인과 추정기라고 불리는 가설 모델 h를 사용하여 다음과 같이 정의할 수 있습니다:

$$E_T[\psi_T(h)|Z] = \mathbf{0}, \quad (1)$$

여기서 $\psi_T: H \to R^d$는 변수 T에 대한 일반화된 잔여 함수를 나타내며 회귀문제에 관한 결과 오류로 간주 되는 $\psi_T(h) = Y - h(T)$를 나타냅니다. $\mathbf{0} \in R^d$는 제로 벡터를 나타내며 d는 결과 Y의 차원을 나타냅니다.
여기서 도구변수 Z를 사용하여 CMR을 만족하는 가설 모델 h를 찾으면, 다음 공식에 따라 h를 사용하여 인과 추론을 시도하는 도구변수 회귀 추정을 수행할 수 있습니다: $E_T[h(T)|Z] = \int_{t \in T} h(t)dP(T=t|Z)$, 여기서 P는 조건부 밀도 측정값을 나타냅니다. 우리는 딥 네트워크와 같은 복잡한 모델에서의 가설 모델 h를 찾기 위해 GMM[3]을 선택하였습니다. 가설 모델과 그 반증에 대한 오류를 나타내는 모멘트를 선택하면, GMM은 이를 사용하여 CMR의 단순한 제약을 넘어 가설 모델에 무한한 모멘트 제한을 제공합니다. 수식 (1)을 확장하면 GMM은 다음과 같이 m:H×G→R로 표시되는 모멘트로 작성할 수 있습니다:

$$m(h,g) = E_{Z,T}[\psi_T(h) \cdot g(Z)] = E_Z[E_T[\psi_T(h)|Z] \cdot g(Z)] = 0, \quad (2)$$

여기서 연산자 $\cdot$는 내적 곱을 말하고, g ∈ G는 테스트 함수 공간 G에서 무한 모멘트 제한을 생성하는 역할을 나타내며 R^d의 차원을 갖습니다. 테스트 함수 g를 통해 가설 모델 h에 대한 편향된 추정치를 쉽게 자극하는 도구변수의 극단 부분을 쉽게 포착하여 일반화를 위해 가능한 모든 반사실적 사례를 고려함으로써 가설 모델 h가 보다 진정한 인과 관계를 추론할 수 있도록 도와줍니다. 더 나아가, 저자는 앞선 한계에 기인하여 최대 모멘트 제한을 구성하고, 폐형 식에서 $\sup_{g \in G} m(h,g)$로 표시되는 도구변수의 극한 부분에만 집중하여 무한 모멘트를 효율적으로 처리합니다. 따라서 가설 모델 h와 테스트 함수 g 사이의 제로섬 게임으로 생각되는 최소-최대 최적화를 사용하여 GMM을 다시 작성할 수 있습니다:

$$\min_{h \in H} \sup_{g \in G} m(h,g) \approx \min_{h \in H} \max_{g \in G} E_{Z,T}[\psi_T(h) \cdot g(Z)]. \quad (3)$$

2.2 적대적 인과 특징 구성

이번 장에서는 식 (3)의 GMM을 적대적 환경에 적용하고 적대적 도구변수 회귀 추정을 통해 적대적 예시의 내재된 인과관계를 밝힙니다. 먼저 특징 변이 Z를 정의하는데 f로 표시되는 적대적으로 훈련된 DNN으로부터 다음과 같이 표현될 수 있습니다:

$$Z = f_l(X_\epsilon) - f_l(X) = F_{adv} - F_{natural}, \quad (4)$$

여기서 f_l은 l번째 중간 계층의 특징 표현을 출력하고, X는 자연 이미지를 나타내며, X_ϵ는 적대적 섭동 ϵ을 가진 적대적 예시를 나타냅니다. 우리는 적대적 특징 F_{adv}가 인과관계의 결과인 Y를 실제로 어떻게 추정하는지 밝혀내고 싶다는 의미에서, 변수 $T = F_{adv}$로 설정하고 반대되는 반사실적 변수 $T_{CF} = F_{natural} + g(Z)$로 설정했습니다. 이러한 구성은 자연 특징을 유지하면서 인과관계를 추론하여 사실적인 결과를 얻을 수 있습니다. 반사실적 변수에서 자연 특징을 빼서 만든 $T' = T_{CF} - F_{natural} = g(Z)$가 가설 모델 h의 출력 Y'를 자연 특징에 추가하여 $Y = Y' + F_{natural} = h(T') + F_{natural}$가 되도록 하여 인과 특징을 복구하기 때문입니다.
이제 적대적 모멘트 제한 (AMR)을 테스트 함수 g에 의해 계산된 반사실적 변수를 포함하여 다음과 같이 새롭게 정의합니다: $E_{T'}[\psi_{T'}(h)|Z] = \mathbf{0}$. 여기서 적대적 환경에서 일반화된 잔여 오류 함수 $\psi_{T'}(h) = Y' - h(T')$는 변환된 인과적 특징 Y'를 보여줍니다. 이를 종합하여 적대적 도구변수 회귀 추정에 적합하도록 GMM과 반사실적 변수을 적용한다면 다음과 같이 작성할 수 있습니다:

$$\min_{h \in H} \max_{g \in G} E_Z[E_{T'}[\psi_{T'}(h)|Z]g(Z)] = E_Z[\psi_{T'|Z}(h)g(Z)]. \quad (5)$$

이를 딥 네트워크에서 활용하기 위해 계산하는 영역을 특징 공간에서 모델 예측의 로그 확률 공간으로 변경합니다: $\Omega(\omega) = \log f_{l+}(F_{natural+}\omega)$, 여기서 f_{l+}는 l번째 중간 계층 이후 네트워크가 분류 확률을 반환하는 후속 네트워크를 나타냅니다. 이에 따라 식 (5)는 로그 확률 공간에 투영된 모멘트로 다음과 같이 수정됩니다:

$$\begin{aligned} &\min_{h \in H} \max_{g \in G} E_Z[\psi^{\Omega}_{T'|Z}(h) \cdot (\Omega \circ g)(Z)] \\ &\quad = E_Z[G_{\log} - (\Omega \circ h)(T') \cdot (\Omega \circ g)(Z)], \end{aligned} \quad (6)$$

여기서 연산자 ∘는 함수 구성을 나타내고, $G_{\log}$는 로그 대상 레이블로 $G_{\log}$ = logG를 만족합니다.
지금까지 적대적 도구변수 회귀 추정을 수행하기 위해 식 (6)에서 AMR에 기반한 GMM, 즉 AMR-GMM을 구성했습니다. 추가적으로 테스트 함수를 명시적으로 정규화하여 인과추론의 불균형한 예측을 피하기 위한 일반화 목적함수로 라데마허 복잡성을 사용했습니다. 따라서 다음과 같이 풍부한 정보량을 갖춘 테스트 함수를 포함한 AMR-GMM의 최종 목표함수를 구축합니다:

$$\min_{h \in H} \max_{g \in G} E_Z[\psi^{\Omega}_{T'|Z}(h) \cdot (\Omega \circ g)(Z)] - |E_Z[Z - g(Z)]|^2. \quad (7)$$

3. 실험 결과

3.1 인과적 특징의 속성 분석

우리는 적대적 도구변수 회귀 추정의 결과를 보기에 앞서 몇 가지 특징 표현의 결합에 주목합니다. (i) 적대적 특징 (Adv): $F_{natural}$ + Z, (ii) 반사실적 특징 (CF): $F_{natural}$ + g(Z), (iii) 역인과적 특징(CC): $F_{natural}$ + (h ∘ g)(Z), (iv) 적대적인 인과적 특징 (AC): $F_{natural}$ + h(Z). 위의 특징 결합에 따른 분류 정확도로 계산된 적대적 견고성을 통해 적대적 예시의 인과관계를 파악하고자 합니다.
그림 2와 같이 모든 데이터셋 예시에 대한 특징 접합 (즉, Adv, CF, CC, AC)의 분류 정확도를 측정하여 적대적 강건성을 살펴봅니다. 여기서 CF의 적대적 강건성이 CC, AC, 심지어 Adv보다 열등한 것을 관찰할 수 있습니다.

직관적으로 식 (7)을 위반하는 테스트 함수는 특징 표현을 올바른 예측에서 극도로 벗어나는 조건으로 강제하기

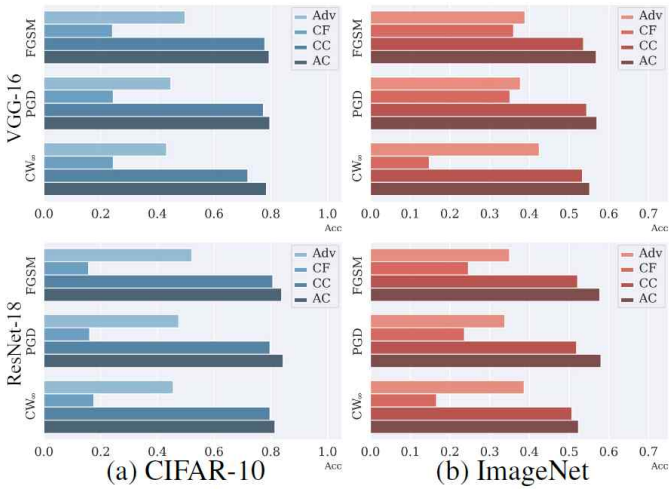

그림 2 세 공격(FGSM, PFD, CW_∞)에 대한 VGG-16 및 ResNet-18의 특징 표현의 적대적 견고성 분석 결과.

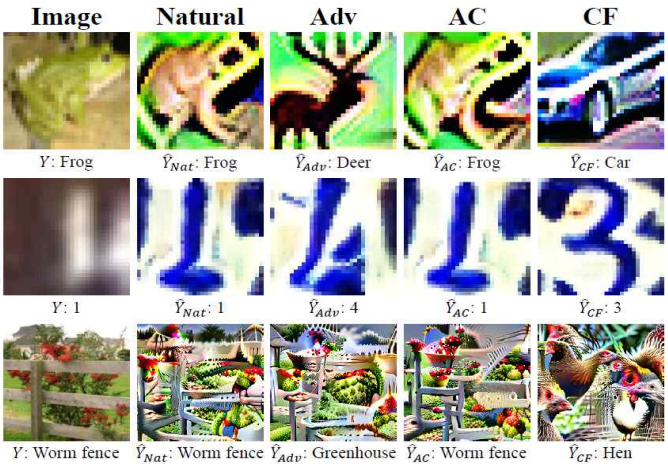

그림 3 특징 표현의 시각화 결과. 위로부터 순차적으로 CIFAR-10, SVHN, ImageNet이 사용됨.

때문에 이는 당연한 결과입니다. CC와 AC에 대한 예측 결과의 경우, 큰 차이로 Adv보다 인상적인 강건성 성능을 보여줍니다. AC는 적대적 섭동으로부터 획득한 특징 변이를 직접 활용하기 때문에 특징 변이에 대한 극도의 경우에 반대 사실을 출력하는 테스트 함수에서 얻은 CC보다 더 나은 적대적 강건성을 나타냅니다. 이러한 견고성은 추정된 인과적 특징이 다양한 유형의 적대적 섭동을 극복할 수 있는 능력을 가지고 있음을 보여줍니다.

3.2 인과적 특징의 시각화 분석

이번 실험에서는 입력 도메인에서 특징 시각화 방법을 활용하여 사람이 인식할 수 있는 방식으로 특징 결합을 해석합니다. 그림 3에서 볼 수 있듯이, 일반적으로 자연 특징의 결과는 대상 객체의 의미적 그림을 나타내는 반면, 적대적 특징(Adv)은 적대적 공격을 받은 표적 객체의 모습을 따라 특징 표현을 강제합니다. CF의 시각화는 대상 객체에 대한 위반된 특징 표현으로 급격하게 변화시켜 가장 크게 특징 공간을 망가뜨립니다. 대조적으로, AC의 시각화는 대상 객체에 대해 의미적 일관성을 나타

표 1 CAFE를 통한 적대적 견고성 및 개선 사항 비교

	Method	CIFAR-10						
		Natural	FGSM	PGD	CW_∞	AP	DLR	AA
VGG	ADV	**78.5**	49.8	44.8	42.6	43.2	42.9	40.7
	ADV_{CAFE}	78.4	**52.2**	**47.9**	**44.1**	**46.4**	**44.5**	**42.7**
	TRADES	**79.5**	50.4	45.7	43.2	44.4	42.9	41.8
	$TRADES_{CAFE}$	77.0	**51.6**	**47.9**	**44.0**	**47.0**	**43.9**	**42.7**
	MART	**79.7**	52.4	47.2	43.4	45.5	43.8	42.0
	$MART_{CAFE}$	78.3	**54.2**	**49.7**	**43.9**	**48.1**	**44.5**	**42.7**
	AWP	**78.0**	51.7	48.2	43.5	47.2	43.4	42.6
	AWP_{CAFE}	77.4	**54.8**	**51.4**	**44.2**	**50.2**	**44.9**	**43.5**
	HELP	**77.4**	51.8	48.3	43.9	47.3	43.9	42.9
	$HELP_{CAFE}$	75.6	**54.4**	**51.4**	**44.6**	**50.4**	**44.8**	**43.7**

내며, 이를 통해 우리는 그 자체로 의미적 정보를 인식하고 인간 관찰자에게 설명할 수 있습니다.

3.3 강인성을 위한 CAusal FEatures

다음으로 획득한 인과적 특징을 통해 딥 네트워크를 학습할 때 효율적으로 강인성을 획득하는 방법을 제안합니다. 다음과 같이 섭동 ϵ를 사용하여 경험적 위험 최소화(ERM)의 형태로 방어 네트워크에 CAusal FEatures (CAFE)를 주입하는 방법을 개발할 수 있습니다:

$$\min_{f\in F} E_S[\max_{\|\epsilon\|_\infty \le \gamma} L_{defense} + D_{KL}(f_{l+}(\hat{F}_{AC}) \| f_{l+}(F_{adv}))], \quad (8)$$

여기서 $L_{Defense}$는 방어 네트워크 f를 달성하기 위한 사전 정의된 손실을 지정하고, S는 데이터 샘플을 나타냅니다. 표 1에서 볼 수 있듯이 CAFE가 5가지 방어 알고리즘의 적대적 견고성을 더욱 향상시킴을 보여줌으로써 저자가 제안한 인과적 특징이 모든 네트워크에서 동작하고, 효과적으로 견고성을 높일 수 있음을 확인합니다.

4. 결 론

본 논문에서는 적대적 예시의 인과관계을 밝히기 위해 인과적 특징을 효과적으로 파악하는 적대적 도구변수 회귀 추정을 제안하고 이를 위해 AMR-GMM을 구축합니다. 가설 모델과 테스트 함수를 사용하여 적대적 예측의 인과 관계를 탐구하고, 이를 통해 사람이 인식할 수 있는 방식으로 의미 정보를 식별합니다. 또한, 방어모델의 강건성을 향상시키기 위한 인과적 특징(CAusal FEatures, CAFE)을 방어에 이식하는 방법을 제안합니다.

5. 참고문헌

[1] Peter CB Phillips and Bruce E Hansen. Statistical inference in instrumental variables regression with i (1) processes. The Review of Economic Studies, 57(1):99–125, 1990. 2

[2] Chunrong Ai and Xiaohong Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. Econometrica, 71(6):1795–1843, 2003. 3

[3] Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. NIPS, 32, 2019. 3, 5