소스-프리 도메인 적응 연구동향

황의원 [1]
[1] 연세대학교 디지털헬스케어학부 교수

uiwon.hwang@yonsei.ac.kr

A Trend of Source-free Domain Adaptation

Uiwon Hwang[1]
[1]Division of Digital Healthcare, Yonsei University

요 약

딥러닝의 발전으로 인공지능의 실세계 응용이 다방면으로 확대되고 있다. 하지만 학습에 사용된 소스 도메인 데이터와 테스트에 사용된 타겟 도메인 데이터 간의 분포 차이로 인해 모델의 성능이 크게 저하될 수 있다. 이를 극복하기 위해 도메인 적응 방법이 제안되었으나, 소스 도메인 데이터에 접근할 수 없는 경우 적용에 한계가 있다. 이에 대응하여 소스 데이터가 필요 없는 소스-프리 도메인 적응 기술과 실시간으로 적응하는 테스트 시간 적응 방법이 연구되고 있다. 본 논문은 최신 소스-프리 도메인 적응 및 테스트 시간 적응 방법의 동향을 파악하고 각 방법론의 기술적 특징을 분석하고자 한다.

1. 서론

인공지능 및 딥러닝이 빠르게 발전하면서 이를 실세계에 적용하는 노력이 지속되고 있다. 예를 들어, 스마트 팩토리를 위한 머신비전 [1], 자율주행 [2], 질병진단 [3], 추천시스템 [4] 등에서 인공지능의 실제 적용 사례가 늘고 있다. 하지만, 실세계 데이터의 특성으로 인해 인공지능 모델의 성능이 크게 저하되는 경우가 빈번하게 발생하고 있다 [5]. 도메인 차이(Domain shift) 문제 [6]는 대표적인 실세계 데이터 문제 중 하나로, 인공지능 모델 학습에 사용된 소스 도메인 데이터와 예측에 사용되는 타겟 도메인 데이터의 분포에 차이가 있는 경우 타겟 도메인 데이터에 대한 인공지능 모델의 성능이 크게 저하되는 문제를 의미한다. 예를 들어, 날씨가 맑은 날에 취득된 데이터로 학습한 자율주행 모델이 비가 오는 날에 적용될 때, 도메인 차이로 인해 자율주행 성능이 크게 저하될 수 있다.

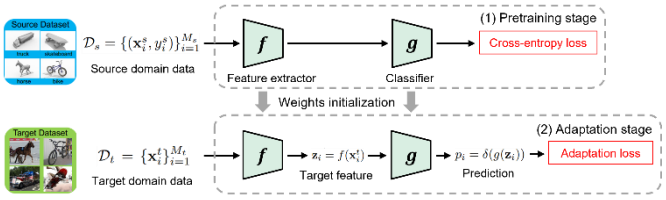
이러한 도메인 차이 문제를 해결하기 위해 도메인 적응 방법이 제안되어 왔다. 이는 소스 도메인 데이터와 학습된 모델을 이용하여 라벨이 없는 타겟 도메인 데이터에 대한 인공지능 모델의 성능 향상을 목표로 한다. 하지만, 실세계 시나리오에서는 비용 및 개인정보 보호를 이유로 소스 도메인 데이터에 접근 불가능한 경우가 많다. 예를 들어, 병원에서 취득된 환자 데이터의 경우 민감한 개인정보가 포함되어 있어 병원 밖으로 반출이 불가능한 경우가 많다. 소스 도메인 데이터에 접근이 불가능한 경우 소스 도메인 데이터를 필요로 하는 도메인 적응 방법은 사용하지 못할 수 있다.

최근에는 그림 1과 같이 소스 도메인 데이터 없이 소스 도메인 데이터로 학습된 모델만으로 도메인 적응이 가능한 소스-프리 도메인 적응 (Source-free domain adaptation, SFDA) 방법이 제안되고 있다. 소스-프리 도메인 적응 기술은 빠르게 발전하여 현재 소스 도메인 데이터를 사용하는 도메인 적응 기술보다 높은 성능을 보이고 있다. 또한, 인공지능 모델에 도메인 차이가 존재하는 테스트 데이터가 입력되었을 때, 최소한의 연산을 이용하여 실시간으로 테스트 데이터에 적응하기 위한 테스트 시간 적응 (Test-time adaptation, TTA) 방법이 활발히 연구되고 있다. 이러한 방법을 통해 더욱 현실적인 시나리오에서 도메인 차이를 효과적으로 극복할 수 있다.

따라서 본 논문에서는 최근 제안된 소스-프리 도메인 적응 및 테스트 시간 적응 방법의 동향을 파악하고, 각 방법론의 기술적 특징을 분석하고자 한다.

2. 본론

본 논문에서는 소스 도메인 데이터 없이 효과적으로 도메인 차이를 극복하는 도메인 적응 방법인 AaD와 SF(DA)^{2}를 소개하고, 실시간으로 소스-프리 도메



(그림 1) 소스-프리 도메인 적응 개요

인 적응을 수행하는 Tent, DeYO 를 소개한다.

먼저 AaD (Attracting and Dispersing) [7]는 소스-프리 도메인 적응 문제를 비지도 클러스터링 문제로 정의하고, 예측 일관성을 고려한 손실함수를 제안하였다. 구체적으로, 소스 도메인 데이터로 사전학습된 모델에 특정 타겟 도메인 데이터 $\mathbf{x}_i$를 입력하였을 때, 출력 레이어의 입력으로 사용되는 특징 벡터를 $\mathbf{z}_i$라 하면 전체 타겟 도메인 중 가장 가까운 K개 특징 벡터들을 close neighbor set C_i으로 정의하고, 현재 배치 내 다른 특징 벡터들을 예측값을 background set B_i 으로 정의하였다. 그리고 두 집합을 위한 로그 가능도를 정의하고 상한을 유도하면 아래 식과 같다:

$$L_i(C_i, B_i) = -\log\frac{P(C_i)}{P(B_i)} \leq -\sum_{j\in C_i}\boldsymbol{p}_i^T\boldsymbol{p}_j + \lambda\sum_{m\in B_i}\boldsymbol{p}_i^T\boldsymbol{p}_m \quad (1)$$

여기서, $\boldsymbol{p}_i$는 특정 타겟 도메인 데이터의 예측값, $\boldsymbol{p}_j$는 close neighbor set 에 속하는 타겟 도메인 데이터의 예측값, 그리고 $\boldsymbol{p}_m$는 background set 에 속하는 타겟 도메인 데이터의 예측값이다. 식 (1)을 해석해 보면, 첫번째 항은 특정 타겟 도메인 데이터와 C_i에 속하는 데이터 간의 예측값이 일관성을 갖도록 서로 당기는 (Attracting) 역할을 수행하고, 두번째 항은 특정 타겟 도메인 데이터와 B_i에 속하는 데이터 간의 예측값이 달라지도록 밀어내는 (Dispersing) 역할을 수행한다. 이렇게 특징 벡터 간의 거리를 활용하여 높은 소스-프리 도메인 적응 성능을 보였다.

SF(DA)^2 (Source-free Domain Adaptation Through the Lens of Data Augmentation) [8]은 데이터 증강의 관점에서 소스-프리 도메인 적응을 해석하며 방법을 제안하였다. 먼저 사전학습된 모델의 특징 공간에서 인접한 특징 벡터는 유사한 의미 정보를 갖는다는 가정과 같은 클래스를 갖는 데이터는 서로 비선형적인 변형을 통해 변환이 가능하다는 가정을 가지고, 특징 벡터 간의 거리를 기반으로 증강 그래프를 정의하였다. 그리고 spectral clustering [9]과 유사하게 증강 그래프에서의 분할을 찾는 방식으로 spectral neighborhood clustering (SNC) 손실함수를 아래와 같이 유도하였다:

$$L_{\text{SNC}}(\boldsymbol{p}_i) = -\frac{2}{K}\sum_{j\in C_i}\boldsymbol{p}_i^T\boldsymbol{p}_j + \sum_{m\in B_i}(\boldsymbol{p}_i^T\boldsymbol{p}_m)^2 \quad (2)$$

SNC 손실함수와 AaD 의 손실함수와 유사하지만, AaD 손실함수가 로그 가능도로부터 유도된 것과 달리 SNC 손실함수는 spectral clustering 으로부터 손실함수를 유도하였고, 그 결과 두번째 항에 제곱이 추가되었다. 도메인 적응 성능을 더욱 향상시키기 위해 최소한의 계산과 메모리 오버헤드를 사용하면서 사전지식 없이 데이터를 증강하는 효과를 내기 위한 implicit feature augmentation (IFA) 손실 함수를 제안하였고, 특징 공간의 서로 다른 방향이 구별되는 클래스 의미 정보를 학습하도록 하는 feature disentanglement (FD) 손실 함수를 제안하였다. 최종적으로 제안하는 세가지 손실함수를 모두 합하여 소스-프리 도메인 적응을 수행하였으며, 여러 데이터셋에서 가장 높은 성능을 보였다. 또한 IFA 손실함수를 기존 소스-프리 도메인 적응 방법에 적용하였을 때, 클래스 불균형이 심한 데이터에서 성능을 크게 향상시킴을 확인하였다.

Tent (Test Entropy Minimization) [10]는 최소한의 연산으로 라벨이 없는 타겟 도메인 데이터에 대한 테스트 시간 적응을 수행하기 위해 제안된 방법이다. 사전학습된 모델에 특정 타겟 도메인 데이터가 입력되면 예측값의 엔트로피 $H(\boldsymbol{p}_i)$를 손실함수로 모델을 업데이트한다. 다만 모델의 모든 파라미터를 업데이트하는 것은 많은 연산량을 필요로 하므로, 배치 정규화 층의 파라미터만을 업데이트하는 방식을 채택하였다. 이러한 방법을 통해 실시간에 가깝게 소스-프리 도메인 적응이 가능함을 보였으며, 테스트 시간 적응 시나리오에서 좋은 성능을 보였다.

DeYO (Destroy Your Object) [11]는 사전학습된 모델에 대해 이미지 내 물체를 파괴하는 변형을 적용하기 전 후의 예측값 간 차이를 측정하는 새로운 신뢰도 수치인 Pseudo-Label Probability Difference (PLPD)를 아래 식과 같이 제안하였다.

$$\text{PLPD}(\mathbf{x}, \mathbf{x}') = \left(\boldsymbol{p}(\mathbf{x}) - \boldsymbol{p}(\mathbf{x}')\right)_{\hat{y}} \quad (3)$$

여기서, $\mathbf{x}$는 특정 타겟 도메인 데이터, $\mathbf{x}'$은 물체를 파괴하는 변형을 적용한 타겟 도메인 데이터, $\hat{y}$은 예측값이 가장 높은 클래스이다. 기존 테스트 시간 적응 방법은 모델 학습에 사용할 데이터를 필터링하기 위해 엔트로피를 신뢰도 수치로 활용하였다. 그러나 데이터셋 내 허위 상관성 (spurious correlation)이 존재하는 경우 엔트로피가 낮은 경우에도 낮은 정확도를 보이는 문제가 존재한다. PLPD 는 엔트로피와 함께 적용되었을 때, 엔트로피만으로 식별할 수 없는 해로운 샘플을 필터링할 수 있다는 것을 보여주었다. 따라서 DeYO 는 샘플 선택 과정에 엔트로피 뿐만 아니라 PLDP 에 대한 임계값을 추가하였다. 추가로, 각 샘플이 모델 업데이트에 미치는 영향을 결정하기 위해 샘플 가중치를 계산할 수 있는데, 엔트로피 뿐만 아니라 PLDP 점수를 추가하여 테스트 시간 적응 성능을 더욱 높였다. 특히, 타겟 도메인 데이터의 분포가 실시간으로 바뀌는 등의 실세계에 가까운 시나리오에

서 기존 방법론 대비 더욱 좋은 성능을 보였다.

3. 결론

본 논문에서는 도메인 적응 분야에서 최근 활발히 연구되고 있는 소스-프리 도메인 적응과 테스트 시간 적응에 대해 탐구하였다. 소스-프리 도메인 적응은 소스 도메인 데이터 없이 학습된 인공지능 모델만으로도 도메인 차이에 강인하게 성능을 향상시킬 수 있음을 알 수 있었고, 타겟 도메인 데이터의 특징 벡터 간 관계를 활용하는 연구가 높은 성능을 보이고 있음을 확인하였다. 또한 테스트 시간 적응은 시간 및 메모리 소모를 최소화하며 소스-프리 도메인 적응을 가능케 하였으며, 최근 엔트로피 기반의 방법론이 높은 성능을 보이고 있음을 확인하였다. 향후에도 소스-프리 도메인 적응 연구가 지속되어 다양한 실세계 시나리오에서 도메인 차이를 더욱 강인하게 극복할 수 있을 것으로 예상된다.

사사문구

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학지원사업 (2019-0-01219), 2024 년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신 사업 (2022RIS-005)의 결과임.

참고문헌

[1] Wang, Jinjiang, Peilun Fu, and Robert X. Gao. "Machine vision intelligence for product defect inspection based on deep learning and Hough transform." Journal of Manufacturing Systems 51 (2019): 52-60.

[2] Grigorescu, Sorin, et al. "A survey of deep learning techniques for autonomous driving." Journal of field robotics 37.3 (2020): 362-386.

[3] Hwang, Uiwon, et al. "Real-world prediction of preclinical Alzheimer's disease with a deep generative model." Artificial Intelligence in Medicine 144 (2023): 102654.

[4] Choi, Sungwoon, et al. "Reinforcement learning based recommender system using biclustering technique." arXiv preprint arXiv:1801.05532 (2018).

[5] Hwang, Uiwon, Dahuin Jung, and Sungroh Yoon. "Hexagan: Generative adversarial nets for real world classification." International conference on machine learning (2019).

[6] Shin, Juhyeon, et al. "Gradient Alignment with Prototype Feature for Fully Test-time Adaptation." arXiv preprint arXiv:2402.09004 (2024).

[7] Yang, Shiqi, Shangling Jui, and Joost van de Weijer. "Attracting and dispersing: A simple approach for source-free domain adaptation." Advances in Neural Information Processing Systems 35 (2022): 5802-5815.

[8] Hwang, Uiwon, et al. "SF (DA) $^ 2$: Source-free Domain Adaptation Through the Lens of Data Augmentation." International conference on learning representations (2024).

[9] HaoChen, Jeff Z., et al. "Provable guarantees for self-supervised deep learning with spectral contrastive loss." Advances in Neural Information Processing Systems 34 (2021): 5000-5011.

[10] Wang, Dequan, et al. "Tent: Fully test-time adaptation by entropy minimization." arXiv preprint arXiv:2006.10726 (2020).

[11] Lee, Jonghyun, et al. "Entropy is not Enough for Test-Time Adaptation: From the Perspective of Disentangled Factors." International conference on learning representations (2024).