# 엔트로피 기반 유사 라벨 생성을 통한 멀티 소스 블랙박스 도메인 적응 알고리즘

유채화 O, 강제원
이화여자대학교 전자전기공학과 스마트팩토리융합전공
chyoo@ewhain.net, jewonk@ewha.ac.kr

## 요 약

비지도 도메인 적응 (UDA)는 라벨링 된 소스 도메인의 지식을 전달하여 라벨링 되지 않은 타겟 도메인을 위한 예측 모델을 학습한다. 그러나 일반적인 UDA 는 소스 데이터나 소스 모델의 세부사항에 대한 접근이 가능하다는 가정이 존재하고, 이는 데이터 및 모델 보호와 저장 문제를 야기하기 때문에 실용적이지 못하다. 최근 소스 모델 인터페이스만이 제공되어, 소스 모델의 예측만을 사용하여 타겟 모델을 학습하는 블랙박스 도메인 적응 (BDA)의 필요성이 대두되었다. 기존의 BDA 연구에서는 하나의 소스 모델만을 이용하는 문제로 효과적인 도메인 적응을 이루기 어려웠다. 더욱이 다수의 소스 도메인이 있는 경우 타겟 모델의 도메인 적응 (DA)을 위해서 소스 개수만큼 개별적으로 타겟 모델을 학습하고 그 중에 가장 우수한 성능을 제공하는 소스 도메인을 선택해야 하는 어려움이 따른다. 본 논문에서는 이 문제를 해결하기 위하여, 다수개의 소스 모델의 예측이 존재할 때 기존의 BDA 방식보다 우수한 성능을 제공하는 알고리즘을 제안한다. 제안 기법에서는 소스 모델의 예측에서 개별 소스와 타겟 도메인 간의 상관성 이외에도 서로 다른 소스 도메인 사이의 관계를 추정하여 타겟 모델의 학습을 위한 유사 라벨을 생성한다. 여러 DA 벤치마크에 대해 수행한 실험 결과를 통해 제안하는 유사 라벨 생성만으로도 기존 방법 대비 0.6-2%의 적응 성능 향상을 보인다.

## 1. 서론

비지도 도메인 적응 (unsupervised domain adaptation, UDA)은 도메인 차이가 있는 소스 도메인 (source domain)의 라벨링 된 데이터의 도움을 받아, 타겟 도메인 (target domain)의 라벨링이 없는 데이터에 대한 예측 함수 학습을 목표로 한다. UDA 는 물체 인식 [1], 의미론적 분할 [2], 물체 탐지 [3]등의 태스크에 대해 활발히 연구되어 오고 있다. 기존의 UDA 셋팅에서는 타겟 도메인을 적응시키는 데 소스 데이터에 대한 접근이 필요하다 [4],[5].

최근 들어서는, 데이터 보안의 필요성이 증가함에 따라 소스 도메인의 접근이 없이 소스 모델에 대한 접근만을 통해 도메인 적응이 가능한 소스-프리 도메인 적응 (source-free domain adaptation, SFDA) 방법이 연구되고 있다 [6],[12]. 나아가 블랙박스 도메인 적응 (black-box domain adaptation, BDA)은 그림 1 과 같이 타겟 도메인의 적응을 위해서 소스 모델의 예측 값만 제공되는 환경에서 도메인 적응을 수행하는 기법이다 [7]. Google, Tencent AI 플랫폼 같은 오픈 AI 인터페이스가 제공될 경우, 인터페이스의 예측 값들을 타겟 모델 학습에 바로 사용함으로써 데이터를 보호하는 한편 블랙박스 모델 제공자들이 상업적 이윤을 얻을 수 있다.

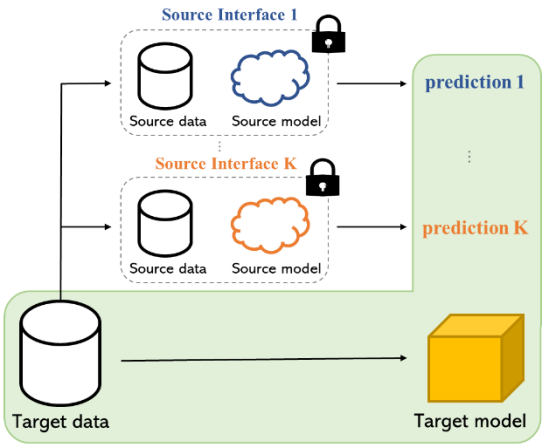


**그림 1 제안 방법의 가정 시나리오. 멀티 소스 블랙박스 도메인 적응의 경우 타겟 모델 학습 시에 연두색 박스 안의 타겟 데이터와 다수 개의 소스 모델 예측만이 접근 가능하다.**

블랙박스 도메인 적응에 대한 최근의 연구들은 [7],[8] 주로 단일 소스 모델 인터페이스를 이용하여

타겟 도메인으로 적응하는 방법을 제안하였다. 하지만 이러한 접근법의 가정은 도메인 적응을 위해서 가장 상관관계가 높은 소스 모델 인터페이스가 무엇인지 알아야 한다는 것이다. 이보다 어려우면서도 더 실용적인 시나리오는 소스 모델의 인터페이스 집합이 존재하는 상황에서의 도메인 적응으로, 각각의 소스 모델은 타겟에 대해 각기 다른 정도로 상관관계가 존재하며 그 상관성에 따라 학습 지식을 전이해 타겟 도메인의 학습에 사용하는 것이다.

본 논문에서는 멀티 소스를 사용한 블랙박스 도메인 적응 방법을 제안한다. 소스 데이터나 모델의 세부 사항에 대해 접근이 불가능하기 때문에, 유일하게 사용 가능한 정보인 소스 모델의 예측의 엔트로피를 사용하여 소스 도메인과 타겟 도메인 간의 상관성을 추정한다. 추정된 상관성을 바탕으로 소스와 소스 간의 유사성을 반영한 거리를 계산하고 이는 유사 라벨을 생성하는 가중치를 정의하는 데 사용된다. 소스 모델의 예측과 가중치 사이의 가중합으로 유사 라벨을 생성해 타겟 모델의 학습에 사용한다.

## 2. 관련 연구

### 2.1 비지도 도메인 적응

UDA 는 다양한 태스크에 대해 사용되고 있지만 [1],[2],[3], 기존의 UDA 는 라벨링 된 소스 데이터에 대한 접근이 필요하다 [4],[5]. 이는 실제 어플리케이션에 적용될 때 개인 정보 보호에 대한 우려가 제기될 수 있다. 이와 다르게 [6]은 소스 데이터에 대한 접근이 없이도 적응 가능한 새로운 소스-프리 도메인 적응 방법을 제안했지만, 여전히 잘 디자인된 소스 모델에 대한 접근이 필요하다는 가정이 존재한다.

최근 들어 보편화되고 있는 기성 인터페이스 [9]와 데이터 및 모델의 보호 문제 [8] 때문에 블랙박스 도메인 적응에 대한 연구가 진행되고 있다. 그러나, 이러한 방법들은 타겟 도메인의 적응에 가장 최상의 소스 모델 인터페이스가 제공된다는 가정 아래 제안되었기 때문에, 블랙 박스 인터페이스가 다수 개 존재할 수 있는 실제 상황과는 맞지 않는다. 따라서, 본 연구에서는 보다 현실적이고 어려운 블랙 박스 도메인 적응을 위한 방법을 제안한다.

### 2.2 멀티 소스 도메인 적응

멀티 소스 도메인 적응 (multi-source domain adaptation, MSDA)은 다수개의 소스 모델로부터 통합된 지식을 전이하여 기존의 UDA 보다 우수한 성능을 제공한다. 이 때 주요 과제는 도메인 간 관계 학습하는 것이다. [9]은 모든 소스 도메인이 타겟 도메인에 동등하게 중요하다고 가정했다. 이와 달리 [11]은 타겟 도메인의 라벨링 되지 않은 데이터를 사용하여, MSDA 를 위한 통합된 소스 도메인을 개별 소스 도메인 간의 가중합으로 표현한다.

최근에는 데이터 보안 이슈가 증가함에 따라 소스 데이터의 접근 없이 소스 모델의 최적 조합을 찾는 멀티 소스 소스-프리 도메인 적응 방법이 제안되었다 [12]. 그러나 이러한 모든 방법에는 타겟 도메인으로의 적응 과정 동안 소스 데이터 혹은 소스 모델에 대한 접근이 가능하다는 가정이 존재한다. 이와 달리 본 연구에서는 이러한 접근이 모두 불가능한 시나리오에 적합한 멀티 소스 도메인 적응 방법에 대해 제안한다.

## 3. 제안 방법

제안하는 멀티 소스 블랙박스 도메인 적응은 다수 개의 소스 도메인에 대해 학습된 모델들의 예측만을 사용하여 라벨링이 없는 타겟 도메인을 위한 새로운 타겟 모델을 적응시키는 문제이다. 본 연구에서는 입력 샘플 $x$로부터 라벨 $y$를 예측하는 분류 모델을 적응시키는 상황에 대해 설명한다.

일반적인 MSUDA 셋팅에서, 각각의 소스 도메인이 $\mathcal{S}_k \triangleq \{(x_i^{\mathcal{S}_k}, y_i^{\mathcal{S}_k})\}_{i=1}^{|\mathcal{S}_k|}$, $x_i^{\mathcal{S}_k} \in \mathcal{X}^{\mathcal{S}_k}$, $y_i^{\mathcal{S}_k} \in \mathcal{Y}^{\mathcal{S}_k}$ 로 정의되는 $K$개의 소스 도메인 $\mathcal{S} = \{\mathcal{S}_k\}_{k=1}^{K}$ 과 타겟 도메인으로부터 라벨링 되지 않은 데이터 $\mathcal{T} \triangleq \{x_i^{\mathcal{T}}\}_{i=1}^{|\mathcal{T}|}, x_i^{\mathcal{T}} \in \mathcal{X}^{\mathcal{T}}$ 가 존재한다. 이와 동시에, 각각의 소스 도메인 데이터에 대해 학습된 $K$개의 소스 모델 집합 $\mathcal{F} = \{f_{\mathcal{S}_k}\}_{k=1}^{K}$ 이 존재한다고 가정한다. 이 때 각각의 소스 모델은 $f_{\mathcal{S}_k} : \mathcal{X}^{\mathcal{S}_k} \rightarrow \mathcal{Y}^{\mathcal{S}_k}$로 정의된다. 제안 연구에서는 먼저 소스 도메인 간 라벨 공간이 동일하다고 가정한다 ($\mathcal{Y}^{\mathcal{S}_1} = \cdots = \mathcal{Y}^{\mathcal{S}_K} = \mathcal{Y}$). 이와 함께, 제안 연구에서 제안하는 블랙박스 도메인 적응 셋팅에서는 어떤 소스 데이터 $\mathcal{S}_k$도 접근이 불가능하다. 또한 SFUDA [6],[12]와 다르게 소스 모델의 모델 구조나 파라미터 등의 어떠한 세부사항도 알 수 없다고 가정한다. 대신, 제안 연구에서는 $K$개의 소스 모델 $\mathcal{F}$ 의 타겟 샘플 $\mathcal{X}^{\mathcal{T}}$에 대한 예측만을 사용하여 타겟 도메인으로 적응된 타겟 모델 $f_{\mathcal{T}} : \mathcal{X}^{\mathcal{T}} \rightarrow \mathcal{Y}$를 구하는 것이 목표이다.

우선 제안 연구의 시나리오에서 타겟 모델의 학습을 위해 사용할 수 있는 유일한 정보는 $K$개의 소스 모델 $\mathcal{F}$ 로부터 얻은 타겟 샘플에 대한 예측이다. 본 절에서는 소스 모델의 예측에서 소스 도메인과 타겟 샘플 간의 상관도를 측정하고, 상관도를 반영한 가중치를 계산하여 이를 사용한 예측의 가중합을 생성하는 방법을 소개한다. 생성된 예측의 가중합은 타겟 모델 학습을 위한 유사 라벨로 사용된다.

먼저 타겟 샘플 $x^{\mathcal{T}} \in \mathcal{X}^{\mathcal{T}}$에 대해 $k$번째 소스 모델 예측 $p_{\mathcal{S}_k} = f_{\mathcal{S}_k}(x^{\mathcal{T}})$ 로 정의한다. 소스 모델 $f_{\mathcal{S}_k}$이 타겟 샘플 $x_i^{\mathcal{T}}$에 대해 높은 판별성을 가질수록 $p_{\mathcal{S}_k}$는 낮은 엔트로피를 가진다 [13]. 따라서, 본 연구에서는 블랙 박스 소스 모델과 타겟 샘플 간의 상관도를 측정하기 위해 식 (1)으로 정의된 엔트로피 값을 구한다.

$$H_{S_k} = \sum_{c=1}^{|y|} p_{s_k}(c)\, log\, p_{s_k}(c). \quad (1)$$

이후 엔트로피 값이 가장 작은 소스 $\hat{S}$를 식 (2)와 같이 정의한다. 이 때의 소스 $\hat{S}$가 입력 타겟 샘플 $x^T$를 판별하는데 가장 가까운 분포를 가진 도메인이라고 생각할 수 있다.

$$\hat{S} = \underset{S_k \in S}{\operatorname{argmin}} H_{S_k}. \quad (2)$$

얻어진 최상의 소스 $\hat{S}$를 기준으로 각 소스 간의 상관도를 식 (3)의 거리 값으로 나타낸다.

$$d_{S_k} = KL(p_{S_k} || p_{\hat{S}}), \quad (3)$$

이 때 $KL(.)$은 KL divergence 이다.

계산된 거리 값을 바탕으로 유사 라벨을 생성하기 위해 사용될 각각의 가중치를 식 (4)와 같이 정의한다. 따라서 사용하는 가중치는 $0 \le \alpha_k \le 1$, $\sum_{k=1}^{K} \alpha_k = 1$ 조건을 만족한다.

$$\alpha_k = 1 - \frac{e^{d_{S_k}}}{\sum_{j=1}^{n} e^{d_{S_k}}}. \quad (4)$$

최종적으로 아래 식과 같이 타겟 샘플 $x^T$에 대해 타겟 모델을 학습하는 데 사용할 유사 라벨 $p_S$은 식 (4)의 가중치를 사용한 소스 모델 예측들의 가중합으로 정의된다.

$$p_S = \sum_{k=1}^{K} \alpha_k p_{S_k}. \quad (5)$$

타겟 모델은 식 (6)의 타겟 모델의 예측과 유사 라벨 간의 크로스 엔트로피 손실을 사용하여 학습된다.

$$L = -\mathbb{E}_{x^T \in X^T} \sum_{c=1}^{|y|} \mathbb{I}\{p_S = c\} \log \delta_c\big(f_t(x^T)\big), \quad (6)$$

이 때 $\delta(.)$ 는 softmax 연산이고 $u \in \mathbb{R}^C$ 에 대해 $\delta_c(u) = \frac{\exp(u_j)}{\sum_{i=1}^{C} \exp(u_i)}$ 이다.

## 4. 실험

### 4.1 실험 환경

제안하는 DA 방법의 효과를 검증하기 위해 물체 인식을 위한 2 가지 벤치마크 데이터셋을 사용했다. Office [14]는 Amazon (A), DSLR (D)와 Webcam (W)의 세 가지 도메인으로 구성되며 각각의 도메인은 총 31 개의 물체 카테고리로 분류된다. Office-Home [15]은 Art (Ar), Clipart (Cl), Product (Pr)과 Real-world (Re)의 네 가지 도메인으로 구성되면 각각의 도메인은 65 가지의 객체 클래스를 포함한다. 모든 실험에서 한 가지 도메인을 타겟 도메인, 나머지 도메인을 다수개의 소스 도메인으로 설정했다.

[12]와 동일하게, 소스 모델은 모두 미리 학습된 ResNet-50 [16]으로 초기화한 뒤 각각의 소스 데이터에 대해 Office 는 100 epoch, Office-Home 50 epoch 로 학습시켰다. 타겟 모델도 소스 모델과 동일한 구조의 ResNet-50 를 사용하되 모든 벤치마크에 대해 30 epoch 동안 학습시켰다.

동일 실험 환경에서 기존 방법과 제안 방법을 비교했다. 제안 연구와 비슷하게 Dis-tune [7]은 블랙박스 셋팅에서 타겟 도메인으로의 비지도 적응을 수행한다. 그러나 이 방법은 한 번에 하나의 소스로부터의 적응을 시도한다. 공정한 비교를 위하여 [7]을 앙상블을 통해 멀티 소스 환경으로 확장하고 제안 기법과 성능을 비교하였다. 구체적으로, 개별 소스 모델의 예측으로부터 적응된 개별 타겟 모델의 출력 값의 평균을 최종 출력으로 하여 적응 성능을 비교했다. 이 방법을 Dis-tune-Ens 로 명명했다.

### 4.2 실험 결과

표 1 은 Office 데이터셋에 대해 3 가지 적응 태스크의 결과를 보여준다. A, D 와 W 는 각각 Amazon, DSLR 과 Webcam 의 약자이다. 적응 태스크의 소스 도메인과 타겟 도메인은 각각 화살표의 왼쪽과 오른쪽에 표시되어 있다. 표시된 다수개의 소스 도메인으로부터 화살표 먼저 단일 소스 방법에 대해, Source-best 와 -worst 는 타겟 도메인에 대한 적응 없이 소스 모델로 타겟 데이터를 테스트한 결과이다. 블랙 박스 셋팅의 Dis-tune [7] 방법으로 적응시킨 경우 최적과 최악의 소스를 선택했을 경우 객체 인식률이 각각 4.3%와 10% 증가한다. 그러나 이 방법은 소스가 하나일 때를 가정하기 때문에 여러 개의 소스 인터페이스가 존재할 때 소스 별로 개별적으로 적응을 수행해야 하는 비효율이 존재한다.

반면에 여러 개의 소스가 존재하는 환경으로 확장시킨 Dis-tune [7]-Ens 의 경우 소스 별 타겟 도메인과의 상관성을 고려하지 못하기 때문에 Dis-tune [7]-worst 대비 평균 3% 만의 성능 회복이 이루어졌다.

표 1. Office 에 대한 도메인 적응 결과표

| Source | Method | A,D →W | A,W →D | D,W →A | Avg. |
|---|---|---|---|---|---|
| Single | Source-best | 96.3 | 98.4 | 62.5 | 85.7 |
| | Source-worst | 75.6 | 80.9 | 62.0 | 72.8 |
| | Dis-tune [7]-best | 98.1 | 98.7 | 73.1 | 90.0 |
| | Dis-tune [7]-worst | 85.1 | 91.0 | 72.4 | 82.8 |
| Multiple | Dis-tune [7]-Ens | 95.9 | 98.4 | 63.3 | 85.8 |
| | Ours | **96.1** | **99.6** | **67.3** | **87.7** |

표 2. Office-Home 에 대한 도메인 적응 결과표

| Source | Method | Ar, Cl, Pr →Rw | Ar, Cl, Rw →Pr | Ar, Pr Rw →Cl | Cl, Pr, Rw →Ar | Avg. |
|---|---|---|---|---|---|---|
| Single | Source-best | 74.1 | 78.9 | 46.2 | 65.8 | 66.3 |
| | Source-worst | 64.8 | 62.8 | 40.9 | 55.3 | 55.5 |
| | Dis-tune [7]-best | 80.5 | 83.1 | 54.1 | 69.1 | 71.7 |
| | Distune [7]-worst | 76.7 | 75.1 | 47.6 | 60.7 | 65.0 |
| Multiple | Dis-tune [7]-Ens | 80.7 | 76.9 | **52.9** | **70.0** | 70.1 |
| | Ours | **81.3** | **79.6** | 51.9 | 69.9 | **70.7** |

반면, 제안 방법은 소스 모델의 예측으로부터 소스 도메인과 타겟 도메인과의 상관성을 모델링함으로써 Dis-tune [7]-Ens 보다 평균 약 2% 높은 성능 회복을 보였다.

Office-Home 데이터셋에 대해서도 표 2 에 보이는 것처럼 제안 방법이 다른 방법에 비해 우수한 성능을 보여준다. 표 2 에서 Ar, Cl, Pr 과 Rw 는 각각 Art, Clipart, Product 와 Real-world 의 약자이다. 제안 방법은 Dis-tune [7]-Ens 대비 대다수의 개별 적응 태스크에서 높은 성능을 보이고 평균적으로 0.6%의 성능 향상을 보였다. 이는 Office 보다 가장 좋거나 가장 나쁜 적응 결과 사이의 차이가 적기 때문이다.

## 5. 결론

본 논문에서는 소스 데이터나 소스 모델의 세부 사항에 대한 접근이 제한된 블랙 박스 환경에서 다수 개의 소스 모델 예측 정보만을 사용하는 새로운 UDA 방법을 제안한다. 소스 모델 예측의 엔트로피로부터 개별 소스 도메인과 타겟 도메인 간의 상관도를 추정하여 타겟 모델의 학습을 위한 유사 라벨을 생성한다. 여러 DA 벤치마크에 대해 실험한 결과 기존 방법 대비 평균 0.6-2%의 적응 성능 향상을 보였다. 향후 연구로서 타겟 모델의 예측을 점진적으로 반영한 추가적인 유사 라벨의 개선과 비지도 학습 기반 클러스터링을 추가하면 학습 적응 성능이 더 개선될 것으로 기대된다.

## 감사의 글

이 논문은 2021 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2021-0-02068, 인공지능 혁신 허브 연구개발).

## 참고문헌

[1] Lei Zhang and Xinbo Gao. Transfer adaptation learning: A decade survey, 2020.

[2] M. Chen, H. Xue, and D. Cai. Domain adaptation for semantic segmentation with maximum squares loss. In ICCV, 2019.

[3] S. Kim, et al. Self-training and ad- versarial background regularization for unsupervised domain adaptive one-stage object detection. In ICCV, 2019.

[4] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In ICML, volume 37, pages 97–105, 2015.

[5] Mingsheng Long, et al. Unsupervised domain adaptation with residual transfer networks. In NeurIPS, pages 136–144, 2016.

[6] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In ICML, pages 6028–6039, July 13–18 2020.

[7] Jian Liang, Dapeng Hu, and Jiashi Feng. Distill and Fine-tune: Effective Adaptation from a Black-box Source Model. arXiv preprint arXiv:2104.01539, 2021.

[8] Haojian Zhang, Yabin Zhang, Kui Jia, and Lei Zhang. Unsupervised domain adaptation of black-box source models, 2021.

[9] Arun Reddy Nelakurthi, Ross Maciejewski, and Jingrui He. Source free domain adaptation using an off-the-shelf classifier. In Big Data, pages 140–145, 2018.

[10] Shoushan Li and Chengqing Zong. 2008. Multi-domain sentiment classification. In Proceedings of the 46th ACL. pages 257–260.

[11] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. 2009. Domain adaptation with multiple sources. In Advances in neural information processing systems. pages 1041–1048.

[12] Ahmed, Sk Miraj, et al. Unsupervised Multi-source Domain Adaptation Without Access to Source Data. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[13] Grandvalet, Yves, and Yoshua Bengio, Semi-supervised learning by entropy minimization, CAP 367 (2005): 281-296.

[14] Judy Hoffman, et al. Cycada: Cycle-consistent adversarial domain adaptation. In International conference on machine learning, pages 1989– 1998. PMLR, 2018.

[15] Hemanth Venkateswara, et al. Deep hashing network for unsupervised domain adaptation. In Proceedings ofthe IEEE Conference on Computer Vision and Pattern Recognition, pages 5018–5027, 2017.

[16] Kaiming He, et al. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016.