MPLA: Multimodal Foundation Model-based Pseudo-Label Allocation for SFDA

2024-2 AJOU SOFTCON

이 름 이휘수 **지도교수** 황원준

연구 배경 및 관련 연구

❖ *본 연구는 Workshop on Image Processing and Image Understanding (IPIU 2025)에 게재 예정이다.*

◆ **SFDA 연구의 필요성:** Domain Adaptation은 소스 도메인에서 학습한 모델을 타겟 도메인에 적응시키는 것을 말한다. 기존에는 소스 데이터와 타겟 데이터를 모두 사용하는 Supervised Domain Adaptation (SDA)와 소스 데이터와 타겟 이미지만을 사용하는 Unsupervised Domain Adaptation (UDA)에 대한 연구가 활발하게 진행되었다. 하지만 현실적인 상황에서 소스 데이터는 정보보호 및 저작권 문제가 있을 수 있고, 리소스가 제한된 장치에서는 활용하기 어렵다는 문제를 갖고 있다. 이에 따라 최근에는 소스 데이터와 타겟 라벨 없이 오직 Pre-trained 소스 모델과 타겟 이미지만을 이용하여 DA를 수행하는 Source-free Domain Adaptation (SFDA)에 대한 연구의 필요성이 제기되었다.

◆ **Multimodal Foundation Model (MFM):** Foundation 모델은 대규모 데이터에서 훈련하여 범용성을 갖춘 모델을 말한다. 특별히 MFM은 시각 및 언어 등 다중 모달리티를 통합한 Foundation Model이다. 이러한 MFM은 소스 데이터가 부재한 상황에서 타겟에 대한 어느 정도 정확한 정보를 제공해줄 수 있다는 점에서 SFDA에 활용될 수 있고, 2024년도에 들어서는 SFDA의 방향성이 이러한 MFM을 활용하는 방향으로 발전하고 있다. 이후 언급될 *CLIP(2021)*과 *BLIP(2023)*은 MFM 중 하나이다.

◆ **Source Hypothesis Transfer for UDA (SHOT, 2020, PMLR):** SHOT은 SFDA의 대표적인 모델로 SFDA의 시초가 된다. SHOT의 핵심은 이상적인 타겟 모델의 출력은 individually certain, globally diverse 해야 한다고 가정하고, 수식 (1)과 같이 Information Maximization (IM) Loss를 사용하여 모델을 학습하는 것에 있다. 하지만 타겟에 대한 정보가 부족한 상황에서 단순히 IM Loss를 줄이는 것은 타겟 feature를 부정확하게 정렬할 우려가 있다. 이에 따라 본 연구에서는 Foundation 모델과 함께 SHOT의 문제를 개선하고, Foundation 모델의 지식을 극대화하는 MPLA framework를 제안한다.

제안 기법

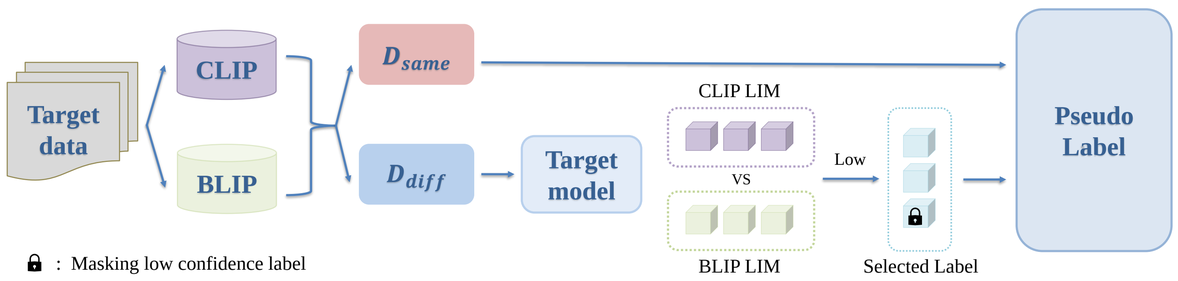


Fig 1. MPLA framework

◆ **MPLA framework:** MPLA는 Foundation 모델의 지식을 기반으로 SFDA를 달성하는 framework이다. MPLA은 다음의 3가지 단계로 구성된다. Fig 1에서 전체 MPLA의 과정을 확인할 수 있다.

- ✓ **MPG:** Multimodal Foundation Model-based Prediction Generation
- ✓ **SPLA:** Consensus-based Static Pseudo-Label Allocation
- ✓ **DPLA:** LIM-based Dynamic Pseudo-Label Allocation

◆ **MPG:** MPG는 CLIP과 BLIP으로부터 타겟 샘플에 대한 클래스 예측을 얻고, 신뢰성 있는 예측과 신뢰성 없는 예측을 구분하는 단계이다. CLIP의 경우 모델에 타겟 샘플과 각 클래스에 대한 텍스트를 입력하여 가장 높은 유사도를 갖는 클래스를 예측 클래스로 추출한다. BLIP의 경우 먼저 모델에 타겟 샘플과 클래스 정보를 담은 텍스트를 입력하여 이미지에 대한 설명인 캡션을 생성하고, S-BERT를 이용하여 캡션과 클래스에 대한 임베딩을 생성한다. 생성된 임베딩을 통해 코사인 유사도를 측정하여 가장 높은 유사도를 갖는 클래스를 예측 클래스로 추출한다. 모든 타겟 샘플 $x_t \in X_t$에 대하여 CLIP의 예측인 $\widehat{y_c}$와 BLIP의 예측인 $\widehat{y_b}$를 생성하고, 이를 바탕으로 수식 (2)와 같이 2개의 데이터셋을 구축한다. CLIP과 BLIP의 예측이 동일한 경우 충분히 신뢰성 있는 예측으로 평가하여 D_{same} 데이터셋을 구축하고, CLIP과 BLIP의 예측이 다를 경우 신뢰성이 떨어지는 예측으로 D_{diff} 데이터셋을 구축한다.

◆ **SPLA:** SPLA는 충분히 신뢰할 수 있는 예측인 D_{same} 데이터셋의 예측을 Pseudo-Label로 할당하여 타겟 모델을 훈련하는 단계이다. 생성된 Pseudo-Label을 수식 (3-4)와 같이 Cross-Entropy Loss로 학습하고, SHOT의 IM Loss와 결합하여 학습한다. 이때, Foundation Model의 신뢰성 있는 예측은 SHOT의 IM Loss가 더 효과적으로 동작하도록 유도하고, IM Loss는 Foundation Model의 신뢰성 있는 예측을 강화하게 된다.

◆ **DPLA:** DPLA는 신뢰성이 다소 떨어지는 예측인 D_{diff} 데이터셋의 예측에 대해 배치 단위로 서로 다른 Label set인 CLIP과 BLIP의 Label set 중 더 정확한 Label set을 찾아서 동적으로 Pseudo-Label을 할당하는 단계이다. 이때, 더 정확한 Label set을 선택하기 위해 Label Information Maximization Loss (LIM) 척도를 정의한다.

➢ **가정:** LIM을 정의하기에 앞서 2가지 가정이 필요하다. 첫째는 소스 모델과 Foundation 모델로부터 지식을 전이 받은 타겟 모델은 이상적인 Label set에 대해 정답확률의 Entropy가 작게 측정되어야 한다는 것이다. 둘째는 셔플된 데이터셋은 클래스 분포의 무질서도가 증가하여 배치 단위에서 측정하는 Diversity가 커야 한다는 것이다. 즉, 정답에 근접하는 Label set은 Entropy는 작게, Diversity는 크게 측정되어야 할 것이라고 가정한다.

➢ **LIM:** LIM은 필터링 벡터 **F**를 통해 정의된다. 수식 (5-8)와 같이 필터링 벡터 **F**를 통해 Label set 내에 속한 클래스로 확률벡터 p를 필터링하고, 이를 이용하여 SHOT의 IM과 유사하게 Label Entropy Loss와 Label Diversity Loss를 정의한다. 이후 수식 (9)와 같이 두 Loss를 더하여 LIM을 정의하게 된다. SHOT의 경우는 전체 클래스에 대한 IM을 측정하여 두 Label set에서 동일하게 측정된다. 즉, SHOT의 IM은 Label과 무관하지만, DPLA의 LIM은 필터링을 통해 Label set의 IM을 측정할 수 있게 된다. 이렇게 정의된 LIM은 Label set이 얼마나 더 정답에 근접하는 Label set인지를 측정할 수 있는 척도가 된다. 이후 수식 (10)과 같이 LIM Loss가 더 작은 Label set을 각각의 배치에서 동적으로 Pseudo-Label로 할당한다. 이렇게 하여 DPLA는 Foundation Model의 신뢰성 없는 예측을 해소하고, SHOT의 문제를 해결하는 역할을 한다.

➢ **Masking Threshold:** LIM을 통해 선택된 Label set이라 할지라도 Label set 내의 모든 Label이 정답이 되는 것은 어렵다. 이에 따라 수식 (11)과 같이 Masking Threshold τ를 두어 예측 확률값이 τ보다 낮을 경우 이 Label을 Masking하여 Pseudo-Label이 잘못 할당되는 문제를 해결한다. 이후에는 위의 과정에서 얻어진 최종적인 Pseudo-Label을 수식 (12-13)과 같이 Cross-Entropy Loss로 학습하고, SHOT의 IM Loss와 결합하여 학습한다.

◆ **최종 MPLA 학습 알고리즘:** MPLA의 학습 알고리즘은 우측의 Algorithm 표에서 확인할 수 있다.

$$\mathcal{L}_{\text{ent}}(f_t;\mathcal{X}_t) = -\mathbb{E}_{x_t\in\mathcal{X}_t}\sum_{k=1}^{K}\delta_k(f_t(x_t))\log\delta_k(f_t(x_t))$$
$$\mathcal{L}_{\text{div}}(f_t;\mathcal{X}_t) = \sum_{k=1}^{K}\hat{p}_k\log\hat{p}_k \tag{1}$$

$$\mathcal{D}_{same} = \left\{\left(x_t^i,\hat{y}_{\text{c}}^i\right) \mid \hat{y}_{\text{c}}^i = \hat{y}_{\text{b}}^i, \forall_{x_t^i}\in\mathcal{D}\right\}$$
$$\mathcal{D}_{diff} = \left\{\left(x_t^i,\hat{y}_{\text{c}}^i,\hat{y}_{\text{b}}^i\right) \mid \hat{y}_{\text{c}}^i \neq \hat{y}_{\text{b}}^i, \forall_{x_t^i}\in\mathcal{D}\right\} \tag{2}$$

$$\mathcal{L}_{spla}(f_t;\mathcal{X}_s,\mathcal{Y}_s) = -\mathbb{E}_{(x_s,y_s)\in\mathcal{X}_s\times\mathcal{Y}_s}\sum_{k=1}^{K}q_k\log\delta_k(f_t(x_s)) \tag{3}$$

$$\mathcal{L}(f_t;\mathcal{X}_s,\mathcal{Y}_s) = \mathcal{L}_{\text{ent}} + \mathcal{L}_{\text{div}} + \mathcal{L}_{spla} \tag{4}$$

$$\mathbf{F} = \begin{cases} 1 & \text{if } k\in\mathbf{L} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$p^f = p\odot\mathbf{F} \tag{6}$$

$$\hat{p}_k = \frac{1}{N}\sum_{i=1}^{N}p_{(i,\,k)}^f \tag{7}$$

$$\mathcal{L}_{lent} = -\mathbb{E}_{x_d\in\mathcal{X}_d}\sum_{k\in\mathbf{L}}p_k^f\log p_k^f$$
$$\mathcal{L}_{ldiv} = \sum_{k\in\mathbf{L}}\hat{p}_k\log\hat{p}_k \tag{8}$$

$$\mathcal{L}_{lim} = \mathcal{L}_{lent} + \mathcal{L}_{ldiv} \tag{9}$$

$$\mathbf{y}_{\text{selected}} = \begin{cases} \mathbf{y}_c & \text{if } \mathcal{L}_{lim}(\mathbf{y}_c) \le \mathcal{L}_{lim}(\mathbf{y}_b) \\ \mathbf{y}_b & \text{otherwise} \end{cases} \tag{10}$$

$$\mathbf{y}_{\text{masked}} = \begin{cases} \mathbf{y}_{\text{selected}} & \text{if } p[\mathbf{y}_{\text{selected}}] \ge \tau \\ \text{ignore} & \text{otherwise} \end{cases} \tag{11}$$

$$\mathcal{L}_{dpla}(f_t;\mathcal{X}_d,\mathcal{Y}_{\text{masked}}) = -\mathbb{E}_{(x_d,y_{\text{masked}})\in\mathcal{X}_d\times\mathcal{Y}_{\text{masked}}}\sum_{k=1}^{K}q_k\log\delta_k(f_t(x_d)) \tag{12}$$

$$\mathcal{L}(f_t;\mathcal{X}_d,\mathcal{Y}_{\text{masked}}) = \mathcal{L}_{\text{ent}} + \mathcal{L}_{\text{div}} + \mathcal{L}_{dpla} \tag{13}$$

Algorithm 1 Training of MPLA

Input: Pre-trained source model $f_s = h_s \circ g_s$,
$D_{same} = \{(x_s, y_s)^i\}_{i=1}^{n_s}$, $D_{diff} = \{(x_d, y_c, y_b)^i\}_{i=1}^{n_d}$ Eq. (2).
number of epochs T, masking threshold τ.
Initialization: Freeze the classifier $h_t = h_s$,
Freeze the backbone $b_t = b_s$,
copy the parameters from g_s to g_t as Initialization.
for $t = 1$ to T **do**
  ========== **Step1** ==========
  for $m = 1$ to n_b **do**
    Sample a batch from D_{same}.
    Update $\mathcal{L}(f_t;\mathcal{X}_s,\mathcal{Y}_s)$ in Eq. (3-4).
  end for
  ========== **Step2** ==========
  for $m = 1$ to n_b **do**
    Sample a batch from D_{diff}.
    Obtain $\mathcal{L}_{lim}(\mathbf{y}_c)$ and $\mathcal{L}_{lim}(\mathbf{y}_b)$ via Eq. (5-9).
    Obtain $\mathbf{y}_{\text{masked}}$ via Eq. (10-11).
    Update $\mathcal{L}(f_t;\mathcal{X}_d,\mathcal{Y}_{\text{masked}})$ in Eq. (12-13).
  end for
end for

실험 결과 및 결론

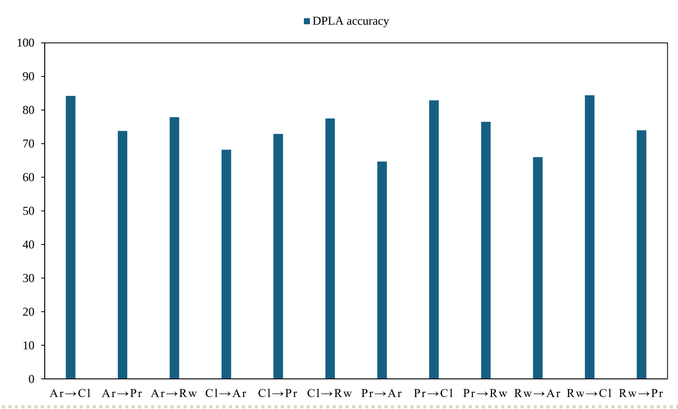
❖ *본 연구와 관련한 자세한 내용과 자료는 우측 하단의 링크에서 확인할 수 있다.*

◆ **Setup:** Office-Home 표준 벤치마크 데이터셋에서 모델의 성능을 확인하였다. CLIP 의 경우 VIT-B/32, BLIP의 경우 Inst-BLIP-XL 모델을 사용하였다.

◆ **Result:** 실험결과 기존의 최고 성능을 보이는 2024 CVPR에서 공개된 DIFO-C-B32와 비교하여 SPLA만 수행 시 약 +1.0%, DPLA까지 모두 수행 시 약 +5.0%의 성능 개선을 확인할 수 있었다. Office-Home의 모든 DA에서 최고 성능을 기록했고, State-of-the-art (SOTA)의 성능을 달성함을 확인하였다. DPLA의 Pseudo-Label 할당 정확도는 모든 도메인의 평균으로 약 75.3%를 기록했고, 학습이 진행될수록 할당 정확도가 상승함을 확인하였다. 이는 DPLA의 이론적 가정을 충분히 뒷받침 하는 결과로 해석될 수 있다.

◆ **결론 및 추후 연구:** 본 연구에서는 SFDA 효과적으로 달성하는 MPLA framework와 LIM의 개념을 제안했다. 실험적 결과 MPLA와 LIM의 타당성을 검증할 수 있었고, Office-Home에서 SOTA의 성능을 달성하였다. 추후 연구에서는 LIM에 기반하여 DPLA가 더 정확하게 Pseudo-Label을 할당할 수 있도록 Augmentation 등의 기법으로 D_{diff} 데이터셋의 크기를 늘리고, 클래스 분포를 균일하게 처리하는 방안에 대한 연구가 필요할 것이다.

❖ 자료파일 다운로드
https://github.com/huisu0818/MPLA.git



Method	Venue	FM	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg.
Source	–	–	49.7	69.7	75.0	53.8	64.5	67.2	53.9	44.3	74.4	66.5	48.2	79.2	62.1
SHOT [4]	ICML20	✗	56.7	77.9	80.6	68.0	78.0	79.4	67.9	54.5	82.3	74.2	58.6	84.5	71.9
AaD [9]	NIPS22	✗	59.3	79.3	82.1	68.9	79.8	79.5	67.2	57.4	83.1	72.1	58.5	85.4	72.7
C-SFDA [10]	CVPR23	✗	60.3	80.2	82.9	69.3	80.1	78.8	67.3	58.1	83.4	73.6	61.3	86.3	73.5
CLIP-B32 [2]	ICML21	✓	65.9	88.7	87.9	78.3	88.7	87.9	78.3	65.9	87.9	78.3	65.9	88.7	80.2
Inst-BLIP-XL [3]	NIPS24	✓	77.5	86.8	84.1	80.0	86.8	84.1	80.0	77.5	84.1	80.0	77.5	86.8	82.1
DIFO-C-RN [8]	CVPR24	✓	62.6	87.5	87.1	79.5	87.9	87.4	78.3	63.4	88.1	80.0	63.3	87.7	79.4
DIFO-C-B32 [8]	CVPR24	✓	70.6	90.6	88.8	82.5	90.6	88.8	80.9	70.1	88.9	83.4	70.5	91.2	83.1
SPLA (ours)	–	✓	75.2	90.2	89.8	80.3	90.0	89.7	80.8	75.1	89.9	81.5	75.8	90.6	84.1
MPLA (ours)	–	✓	81.8	93.0	91.3	85.9	92.9	91.3	84.9	82.7	91.4	85.3	83.1	93.3	88.1