**1. Introduction**

-

**2. Related Works**

**-** 초기의 unsupervised domain adaptation method는 domain간의 feature 분포 차이를 최소화해 domain gap을 줄이는 domain-invariant representation에 집중했음

- moment matching 방식은 kernel embedding, hidden activation등의 분포 통계의 moment 차이를 줄이도록 학습하는 adaptation layer를 추가했음

- adversarial learning 방식은 feature generator가 domain을 예측하는 domain classifier는 속이는 동시에, source sample의 label은 정확하게 예측하도록 학습시켰음

- 하지만 이러한 feature alignment 방식은 feature space distortion 문제와 class discriminability loss 문제가 발생할 수 있음

- 최근에는 source domain sample을 domain-invariant factor와 domain-specific factor로 나눠causal representation을 학습하는 방식, source label과 target pseudo-label을 함께 학습하는 self-training 방식이 자주 사용됨

- 특히 vision transformer의 cross-attention, self-attention을 활용해 transferable feature를 학습하지만, vision modality에만 적용 가능하다는 한계점이 있음

- 반면, vision-language model은 대규모의 image-text pair로 pre-training하기 때문에 transferable multimodal representation을 학습할 수 있음

- VLM을 UDA에 적용하기 위해서는 pre-trained knowledge를 어떻게 활용할 지에 대한 문제와 source knowledge를 target domain에 어떻게 transfer할 지에 대한 문제를 해결해야 함

- VLM을 이용해 UDA를 다룬 대표적인 두 가지 방식이 있는데, 첫 번째 방식은 우선 source domain으로 image encoder를 fine-tuning하고, VLM의 zero-shot prediction으로 생성한 pseudo-label을 바탕으로 target domain을 unsupervised learning으로 학습함

- 하지만 pseudo-label이 manually designed prompt로부터 얻어지고, fine-tuning이 pre-trained knowledge를 잊게 만드는 문제가 있음

- 두 번째 방식은 prompt tuning을 frozen VLM에 적용해 disentangled semantic and domain representation을 학습하며, 학습 파라미터 수가 작고 pre-trained knowledge를 유지한다는 장점이 있음

- DAPrompt는 각 domain information을 담고 있는 domain-specific prompt를 학습하고, pseudo label을 이용해 학습한 target-specific prompt로 target domain sample을 예측함

- 하지만 source knowledge 대부분이 source-specific prompt에 encoding되기 때문에, 그림과 같이 target domain으로의 knowledge transfer에 한계가 있음

- 실제로 source supervision loss를 제외하고 학습했을 때에도 target performance 결과에 큰 차이가 없음

- 또한 single modality prompting 방식은 VLM의 multi-modal knowledge를 활용하지 못함

**3. Proposed Method**

- UDA 문제는 각 domain의 marginal distribution P(X)는 다르지만, conditional distribution P(YlX)는 domain에 따라 변하지 않는 covariate shift 상황을 가정함

- 본 논문에서는 이를 바탕으로 모든 domain이 prompt를 공유하는 domain-agnostic textual prompt를 사용해 P(YlX)를 모델링함

- class k의 textual prompt tk는 왼쪽 식과 같이 표현하고, N개의 learnable context p와 k번째 class name CLSk로 구성됨

- 그러나 domain-agnostic prompt를 바로 학습하는 데 어려움이 있음

- 우선, image encoder는 domain adaptation을 고려하지 않고 pre-train되었기 때문에 visual embedding이 domain-biased된 상태임

- 또한, 같은 class에 속하더라도 object color, size 등의 intra-class variation이 있기 때문에 class-level textual prompt을 모든 sample에 대해 적용하기 어려움

- 이러한 문제들을 해결하기 위해, 본 연구는 visual prompt를 추가해 domain-agnostic visual representation을 나타낼 수 있도록 하며, textual prompt가 각 image의 context 정보 또한 반영할 수 있도록 함

**Language-Guided Visual Prompting**

- text prompt tk는 text encoder에 입력되어 J개의 encoder layer를 지나 (N+Lk,D) 크기의 embedding을 출력하는데, CLIP은 J번째 layer의 마지막 위치에 해당하는 sk만 사용함

- 하지만 다른 위치도 의미있는 context 정보를 encoding하고 있으므로, 본 논문은 domain, class agnostic한 semantic 정보를 얻기 위해 앞에 N개의 embedding stilt도 이용함

- visual embedding의 경우, CLIP은 visual encoder에서 얻은 feature map z와 z를 global average pooling해 얻은 zbar를 multi-head self-attention layer에 입력해 v와 vtilt embedding을 얻음

- CLIP은 그 중 class token position에 해당하는 (1,D) 크기의 v만을 사용하는데, 본 연구에서는 나머지 spatial position에 해당하는 (HW,D) 크기의 vtilt도 이용해 spatial, semantic 정보를 얻도록 함

- visual prompting 방식 중 prompt를 encoder input 단계에 넣는 pre-model prompting은 모든 vision backbone에 적용하기 어렵기 때문에, visual embedding을 바탕으로 prompt를 생성하는 post-model prompting 방식을 이용함

- visual embedding v를 query로 사용하고, domain, class-agnostic text embedding stilt를 key, value로 사용해 decoder의 cross-attention 연산을 거쳐 v\*를 얻음

- 그 후, residual connection으로 v를 더해 최종적으로 domain-agnostic visual embedding v’을 생성함

- text prompt의 경우에도 같은 방식으로 post-model prompting을 적용해 줌

- class-level text embedding sk를 query로 사용하고, vtilt를 key, value로 사용해 같은 파라미터를 공유하는 decoder에서 cross-attention을 수행함

- 최종 text embedding sk’을 얻게 되고, 이 값은 각 sample의 image feature의 영향을 받아 생성되었기 때문에 instance-dependent한 특성을 갖게 됨

**3.3. Auxiliary Regularizations**

- textual embedding sk’이 image의 domain-specific semantic을 encoding해 domain-biased될 수 있으므로, instance-discrimination contrastive loss를 추가로 제안했음

- IDC loss는 같은 domain 내의 서로 다른 image 간의 textual embedding sk’ 차이를 maximize해 domain-specific한 정보를 제거하도록 유도함

- batch 내에서 같은 sample일 때는 positive pair, 다른 모든 sample의 경우 negative pair가 되고, pair 간의 similarity를 contrastive하게 학습함

- 이 때 similarity는 text embedding sk’와 visual embedding v’ 간 cosine similarity를 모든 class에 대해 평균해 계산하기 때문에, unsupervised하게 target domain sample에도 적용 가능함

- 또한 visual embedding v’의 domain-invariance를 강화하기 위해 Semantic-Consistency Regularization도 제안했음

- SC loss는 FixMatch 연구와 유사하게 strong augmentation을 적용한 input x에 대한 예측 확률을 높이도록 유도함

- target data의 경우 pseudo-label을 바탕으로 confident sample에 대해서만 학습함

- unconfident target sample도 활용하기 위해, information maximization 기반의 IM loss를 추가해 줌

- IM loss의 첫 번째 항은 예측확률이 one-hot label에 가까워지도록 하며, 두 번째 항은 예측 확률이 class에 대해 고르게 나타나도록 만들기 때문에, 예측이 globally diverse하고 locally confident해지도록 도와줌

- 전체 training objective는 source supervised loss, confident pseudo-label에 대한 target loss, 앞서 언급한 auxiliary regularization을 모두 더해 구함

- encoder는 freeze한 상태로, input text prompt p와 prompting module G를 학습하게 됨

**4. Experiments**

- UDA 실험 결과, 다른 method와 비교해 높은 average accuracy를 나타냄

- 특히 single-modality prompt-based model인 DAPrompt와 AD-CLIP보다 뛰어난 성능을 보여 multi-modal prompting의 효과를 알 수 있음

- 또한 vision backbone으로 ResNet과 ViT를 사용한 두 경우 모두 baseline과 뚜렷한 차이를 나타냄

- 그림의 옅은 색의 왼쪽, 진한 색의 오른쪽 부분은 각각 visual prompting 이전과 이후의 visual embedding을 시각화한 모습이고, visual prompting 이후 domain gap이 줄어들고, 명확한 cluster를 형성하는 것을 볼 수 있음

- 또한 instance-conditioned textual prompting으로 인해 textual embedding의 각 class 내에서 diversity가 증가한 것을 알 수 있음

- 다양한 mutual prompting strategy를 적용한 결과, self-attention만을 사용한 independent prompting, linear projection layer를 사용한 uni-directional projection보다 modality 간 bi-directional interaction이 가능한 cross-attention module이 높은 성능을 보였음

- prompting과 loss에 대한 ablation study 결과, textual prompting과 visual prompting을 함께 사용한 mutual prompting이 높은 성능을 보였고, 제안된 모든 regularization 또한 성능 향상에 기여했음

**Limitations**

1)

- visual embedding v’을 domain-invariant하게 만들기 위해 stilt가 domain class-agnostic한 context를 encoding하는 것을 목표로 하는데, DAPrompt 연구처럼 domain-agnostic, domain-specific prompt로 분리하지 않기 때문에 domain-specific한 정보를 학습할 위험이 있을 것 같음

- Semantic-Consistency Regularization에서 augmentation을 이용해 domain-invariance를 학습하는 것만으로는 부족할 것 같음

- textual embedding s’도 domain-biased한 vtilt로 활용하고, Instance-Discrimination Contrastive Loss로 domain-information을 제거하는 것만으로는 부족해 보임

2)

**Code**

- #333 / domain-agnostic context에 해당하는 stilt를 얻기 위해, 우선 모든 K개 class에 대해 (N, D) 크기의 동일한 context vector p를 설정함

- #390 / 그 다음 K개만큼 broadcast한 후, class token을 담고 있는 token suffix와 concat해 (K,L,D) 크기의 prompt를 만들고, text encoder에 input으로 넣어줌

- #441 / text encoder output에서 얻은 K개 embedding 중 첫 번째 embedding의 class token 이전 부분만 사용해 (N,C) 크기의 stilt를 최종적으로 얻음

- #203 / text encoder output에서 CLIP은 token argmax에 해당하는 class token 위치의 text\_embedding\_at\_eos (s)만 사용하지만, DAMP는 모든 위치의 text\_embedding\_all (stilt)도 return 받음

- #250, #304 (CLIP: #73, #240) / image encoder output의 경우에도 유사하게 class token 위치의 v 외에도 모든 위치의 vtilt도 같이 return 받음

- Resnet의 경우, 마지막 attention pool layer에서 CLIP은 query로 첫 번째 위치만 사용하지만, DAMP는 query로 x 전체를 사용함

- ViT에서는 class embedding에 해당하는 (1,H,W) 크기의 global embedding 외에도 spatial region에 해당하는 (C,H,W) 크기의 visual embedding(vtilt)도 return 받음

- #122 / mutual prompting을 수행하는 context decoder는 visual, text embedding에 독립적으로 적용되는 input projection layer, 여러 개의 transformer decoder layer, 마지막 output projection layer로 구성되어 있음

- #443 / visual, text prompt 둘 다 같은 context decoder를 이용해 update됨

- #705 / pseudo-label을 생성할 때, 초반에는 manual한 naïve prompt로부터 얻은 logit을 사용하지만, 학습이 진행될수록 learned prompt로 얻은 logit의 weight를 점점 늘리는 ensemble 방식을 활용함

- 학습 초반에는 source domain knowledge가 부족해 pseudo-label quality가 낮기 때문에, naïve prompt에서 생성된 pseudo-label을 사용해 pre-trained knowledge를 활용할 수 있음

- #708 / 최종적으로, pseudo-label의 maximum probability threshold를 기준으로 한 mask를 적용해 target domain loss를 계산할 수 있음

- # 467 / 각 sample 간 text embedding 거리를 멀어지게 하는 IDCL loss를 계산하기 위해, (B,B) 크기의 matrix를 얻게 됨

- matrix의 element i,j는 i번째 sample의 visual embedding과 j번째 sample의 text embedding 사이의 similarity를 의미함

- #718 / 이를 바탕으로 source, target domain 각각의 contrastive loss를 계산할 수 있음