GalLoP: Learning Global and Local Prompts for Vision-Language Models

**1. Introduction**

- multiple prompt를 사용한 prompt ensembling은 class specific, transformer-layer specific한 diverse prompt를 학습하는 것이 목적임

- 하지만 기존 method들은 global visual 표현에만 집중해 specific region을 활용하지 못함

- local representation을 이용해 이를 개선한 method들도 제안되었지만, 모든 local feature를 사용하기 때문에 irrelevant, noisy feature도 포함하게 되고, pre-trained CLIP의 global image-text alignment와도 맞지 않아, 높은 성능을 보이지 못함

- 본 논문의 GalLoP method는 global, local visual representation을 모두 활용하며, prompt dropout, multiscale strategy를 제안해 prompt diversity를 얻음

- 또한 local prompt와 local region 간의 sparse alignment와 linear projection을 제안해, fine-grained semantic을 학습하게 되고, text-to-image matching도 개선했음

**3. Combining global and local prompts with GalLoP**

- GalLoP은 m개의 global prompt와 n개의 local prompt를 학습함

- 각 global prompt는 기존 CLIP과 동일하게 global visual feature zg와의 similarity를 계산해 학습함

- local prompt의 경우, L개의 local visual feature zl 중 similarity가 높은 top-k local feature와의 average similarity를 사용하는 sparse strategy를 적용하므로, irrelevant, background feature의 영향을 제거하고 class semantic feature에 집중하게 됨

- 그러나 CLIP은 global image feature와 text 간 align하도록 pre-train되었기 때문에, local feature와 text의 alignment를 향상시키기 위해 local visual feature에 linear projection을 적용한 후 similarity를 계산함

- multiple prompts의 diversity를 위해, 두 가지 training technique을 제시했음

- prompt dropout은 global prompt마다 batch내의 image에 대한 random dropout을 적용해, prompt별로 서로 다른 image subset을 학습해 diversity를 얻음

- multi scaling은 local prompt에 적용되며, n개의 prompt가 앞서 언급한 top-k similarity의 k로 서로 다른 값을 사용함

- 작은 scale 값 k를 사용한 prompt는 small visual concept, 큰 k 값을 사용한 prompt는 large visual concept을 학습하게 됨

- inference 때에는, global image feature와 global prompt 간 similarity 평균과, local image feature와 local prompt 간 similarity 평균을 더한 값을 최종 class similarity로 계산해 classification을 수행함

**4 Experimental results**

- 다른 prompt learning method들과 비교했을 때, GalLoP은 가장 높은 ID classification 성능을 보이는 동시에, domain generalization과 OOD detection에서도 가장 높은 성능을 나타냈음

- ablation study를 통해, 본 논문에서 제안한 global prompt와 local prompt combination, local feature의 sparsity strategy, prompt dropout, multiscale strategy의 효과를 알 수 있음

**Pros & Cons**

1)

- domain-invariant, specific feature를 추출하지 않으면서도 prompt diversity를 얻어내, classification과 generalization 간 trade-off를 개선시키므로 범용적인 method로 보임

1)

- 각 class, image마다 foreground region 크기가 다른데, multi-scale

Soft Prompt Generation for Domain Generalization

**1. Introduction**

- distribution shift 문제를 다루기 위해 VLM에 prompt learning을 적용한 기존 domain generalization method들은 fixed prompt나 residual prompt를 학습하기 때문에, diversity와 target transferability에 한계가 있음

- 반면, 본 논문은 generative model을 활용해 domain-specific prompt를 생성하는 방식으로 이러한 문제들을 개선함

**4. Method**

- training stage 1에서는 각 domain에 해당하는 domain prompt label vdi를 일반적인 prompt tuning과 같은 방식으로 학습함

- second stage에서는 Generative Model Pre-training 과정을 통해 CGAN을 학습함

- generator에 noise z와 함께 image embedding을 입력해, generated prompt를 생성함

- discriminator는 앞서 학습한 domain prompt label을 1, generated prompt를 0로 예측함

- pre-training이 완료된 CGAN은 domain-invariant, domain-specific feature 정보를 갖게 돼 여러 domain 간 consistency를 유지하면서도, input variation에 따른 prompt diversity 또한 얻게 됨

- inference time에는 generated domain, instance-specific prompt를 사용해 classification을 수행하므로, target domain에 대해 generalization을 가짐

**5 Experiments**

- 다른 Clip-based DG methods, prompt tuning methods와 비교했을 때는 가장 높은 성능을 보였지만, distribution shift가 큰 dataset의 경우 traditional DG method나 zero-shot CLIP에 못 미치는 accuracy를 나타냄

- 그림과 같이 domain 간의 명확한 prompt clustering이 나타나므로, domain knowledge를 효과적으로 학습한 것으로 볼 수 있음

**Pros & Cons**

1)

- prompt generation을 통해 unseen target domain에 대해서도 transferability를 갖는다는 점에서 기존 method들과 차별됨

1)

- stage 1에서는 prompt에 class token을 추가한 후 classification task로 domain prompt label을 학습한 반면, stage 2에서는 class label을 모르는 single image로부터 class token을 제외한 domain prompt만을 생성함

- 따라서 discriminator에 입력되는 real prompt와 generated prompt를 동등한 기준으로 볼 수 없음

- generated prompt는 class 정보까지 포함하고 있음

2)

- inference stage에서 domain, instance-specific prompt를 생성하기 위해, 매 image마다 generator를 거쳐야 함

Split to Merge: Unifying Separated Modalities for Unsupervised Domain Adaptation (CVPR 2024)

**1. Introduction**

- VLM의 image, text feature 간에 존재하는 modality gap을 통해, 각 feature가 modality-specific cue을 담고 있음을 알 수 있음

- vision branch는 class-specific visual pattern, text branch는 semantic detail에 집중함

- 따라서 양 쪽의 장점을 모두 활용하는 multimodal adaptation framework가 UDA에 더 적합함

- vision branch fine-tuning과 text prompt tuning을 함께 진행하는 기존 방식은 pre-trained CLIP의 alignment를 방해할 수 있으므로, 본 논문은 visual feature를 language associated component LAC, vision-associated component VAC 두 가지로 disentangle하는 방식을 제안함

**3. Method**

- 먼저 linear layer로 구성된 text separator, vision separator에서 image feature를 각각 와 로 분리하며, orthogonal loss를 통해 disentanglement를 유도함

- LAC의 경우, 와 text feature 간 similarity로 class logit 를 계산한 후, target sample은 zero-shot으로 구한 pseudo-label 와의 KL divergence loss로 학습하고, source sample은 label과의 cross-entropy loss로 학습함

- 따라서 knowledge distillation을 이용해 pre-trained CLIP의 semantic content를 target domain에 전달하게 됨

- VAC의 경우, 가 두 개의 linear classifier를 지나 bottleneck feature 와 class logit 를 출력함

- source sample은 을 바로 사용해 LAC처럼 label과의 cross-entropy loss로 학습함

- target sample일 때는 우선 와 를 합친 target ensemble logit 를 왼쪽 식과 같이 계산함

- 이 때 weight generator가 를 입력 받아 출력한 w를 weight로 사용하기 때문에, training dataset variation과 epoch 변화에 맞춰 유연하게 VAC와 LAC의 영향력을 조절할 수 있음

- 그 다음, 와 VAC의 bottleneck feature 를 바탕으로 K-mean based cluster centroid 를 계산하고, 와 가장 가까운 의 k가 pseudo-label 가 됨

- 따라서 visual feature의 locality structure를 활용할 수 있음

- 최종적으로 target sample의 VAC는 와 간 cross-entropy loss로 학습됨

- 또한 VAC와 LAC를 구별하는 modality discriminator를 source data로 학습해, target data에 대해서도 domain-invariant separation을 유지하도록 bce loss를 추가해 줌

- Inference time에는 fixed weight w를 사용해 계산한 logit 이 최종 예측값이 됨

**4. Experiments**

- 다른 CLIP backbone non-tuning method들에 비해 높은 성능을 나타냈지만, domain gap이 큰 경우에는 full-tuning method인 PADCLIP에 비해 낮은 성능을 보였음

- 왼쪽 그림과 같이 modality separation으로 인해 LAC, VAC가 분리되었고, text feature가 LAC에 가깝게 분포함

- 또한 오른쪽 그림에서 VAC의 domain alignment와 locality structure를 확인할 수 있음

**Pros & Cons**

1) 대부분의 CLIP-based UDA method들은 pre-training때처럼 image, text feature alignment를 목표로 하는 것과 달리, modality-specific 정보를 고려해 image feature를 disentangle한다는 점에서 다름

1)

- LAC를 학습할 때 CLIP의 기존 text feature와 manual prompt를 그대로 사용하기 때문에, pre-trained knowledge에 의존하게 됨

2)

- visual feature에는 vision modal과 language modal 둘 다 필요로 하는 정보가 존재할 텐데,

- LAC와 VAC를 orthogonal loss, discriminator loss로 완전히 분리시키는 것은 strict한 제약으로 보임