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 사용해 prompt diversity를 가짐

- 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 적용함

- 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

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

**Limitations**

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

**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, vision-associated component 두 개로 disentangle하는 방식을 제안함

**3. Method**