1. Introduction

-CLIP을 포함한 vision-language model은 web-scale의 image-language pair를 학습해 높은 zero-shot 성능을 보임

-또한 task-agnostic한 방식으로 OOD generalization 성능을 높였음

-하지만 downstream task에 적용하기 위해 fine-tuning을 했을 때, transfer learning 성능은 높아지지만, OOD generalization 성능은 낮아지는 문제가 있음

-이를 해결하려면, pretrained model이 downstream data 분포에 fine-tuning할 수 있는 유연성을 갖추면서, 동시에 분포에 제약이 없는 OOD data에 대해서도 일반화할 수 있어야 함

2. Related Work

-vision-language model은 cross-modal learning을 위해 image와 text를 동일한 embedding space로 학습함

-대표적으로 Masked multi-modal learning을 사용하거나, contrastive learning을 사용해 pre-training을 진행함

-OOD generalization 중 domain shift를 위해 다양한 source-domain training algorithm이 제시되었음

-또한 label space가 변하는 heterogeneous domain generalization이나, unknown domain에 적용하기 위한 open domain generalization 문제로도 확장되었음

-하지만 추가적인 training이 필요하며, pretrained model이 closed set을 가정하기 때문에 한계가 있음

-한편, CLIP과 같은 vision-language pretrained model이 높은 zero-shot generalization 성능을 보였고, downstream task에서의 OOD generalization 능력을 향상시키기 위한 prompt tuning, adapter learning 등의 방식들이 제안됨

-이 방법들 대신, 본 논문에서는 fine-tuning objective와 model optimization 관점에서 OOD 문제를 다룸

3. Methods

3.1. Generalizing CLIP to Out-of-Distributions

-CLIP 모델은 correct image-text pairing을 예측하는 task로 image encoder와 text encoder를 함께 학습하며, test time에는 target class를 text encoder로 embedding해 zero-shot classification을 수행함

- pre-trained CLIP 모델을 train data S = {(x, y)}로 fine-tuning했을 때, OOD test data T = {(x ′, y′)}에 대해 generalization 성능을 보이도록 하는 것이 목적임

- 본 연구는 S와 T 분포가 다른 domain shift 상황과, Y와 Y’이 다른 open class 상황이 동시에 나타나는 challenging한 문제를 다룸

3.2. CLIP Fine-tuning

- 일반적인 fine-tuning은 class c에 해당하는 linear classifier의 parameter Wc와 image feature 간의 유사도에 softmax를 적용해 예측 확률을 구하며, linear classifier와 image encoder를 같이 학습함

- 하지만 이 방법을 CLIP에 적용하면 text 정보는 사용하지 않기 때문에, image-text alignment를 통한 generalization 효과를 얻지 못하고, 또한 train data class에만 맞춰지게 되므로 unseen class에 일반화하기 어려움

-따라서 논문에서는 CLIP의 zero-shot prediction 과정처럼 image embedding Ix와 text prompt embedding Tc 간의 cosine 유사도에 softmax를 적용해 예측 확률을 계산함

- 이 때 text corpus는 image와 달리 downstream data에 의해 제한되기 때문에, text encoder의 overfitting을 막기 위해 image encoder만 fine-tuning 시킴

3.3. Margin Metric Softmax

- 하지만 앞의 식을 그대로 적용하게 되면, label y의 text embedding Ty가 다른 모든 class c의 text embedding Tc를 동등하게 다룸

- 그러므로, pretrained text encoder가 갖고 있는 class 간 semantic relationship을 활용하지 못함

- 따라서 본 논문은 기존 식에 Ty와 Tc간의 distance D(Ty,Tc)를 adaptive margin으로 추가한 Margin Metric Softmax를 제안함

- D(Ty,Tc) 는 1 - S(Ty,Tc) 로 계산하며, text label 간 semantic distance가 멀수록 큰 값을 가짐

- model이 loss를 줄이려 할 때, 거리가 먼 class일수록 분모의 S(Ix,Tc) 값에 높은 가중치가 곱해지기 때문에, S(Ix,Tc)가 더 작아지는 방향으로 image encoder를 학습하게 됨

- 따라서 MMS를 통해 cross-modal alignment와 generalization 효과를 얻을 수 있음

3.4. Beta Moving Average

- fine-tuning 때의 OOD generalization 성능도 유지하며, task-agnostic하고 hyperparameter tuning이 필요 없는 Beta Moving Average 또한 제안됨

- fine-tuning의 각 training step에서 얻은 T개 model의 weighted average를 이용한 temporal ensembling을 고려해 볼 수 있음

- 이 때, 초반 step의 model은 generalization 성능은 높지만 task-specific knowledge는 부족하고, 후반 step의 model은 이와 반대임

- 양쪽 model의 장점을 모두 활용하기 위해, 양 끝의 step이 높은 가중치를 가지는 Beta(β, β) (β<1) 분포로 weight를 설정함

- 또한 temporal ensemble은 모든 model을 저장해야 하므로, beta moving average를 이용해 매 step마다 BMA model을 따로 업데이트함

4. Experiments

-domain shift situation, open class situation, 두 가지 함께 발생하는 situation 총 세 가지 scenario로 나눠 experiment 진행함

-domain shift situation의 경우, CLIPood method가 CLIP pretrained model을 사용한 SOTA methods와 비교했을 때 높은 target domain accuracy를 나타냄

-open class situation 실험의 base class와 new class에서의 높은 accuracy 결과를 통해, downstream task performance와 OOD generalization ability 모두 갖고 있음을 알 수 있음

- domain shift와 open class가 함께 발생하는 scenario에서도 모든 test domain에 대해 높은 base class, new class accuracy를 보임

4.4. Analysis of CLIPood

- image, text embedding을 이용하는 metric softmax만을 사용했을 때, linear classifier로 학습하는standard fine-tuning보다 뛰어난 성능을 보임

- Margin Metric Softmax의 경우, fixed margin을 사용한 경우와 margin을 사용하지 않은 경우에 비해 높은 성능을 나타냄

- 또한 그림과 같이 adaptive margin을 사용했을 경우 related class를 예측하는 것으로 보아, semantic class relationship을 유지한다는 것을 알 수 있음

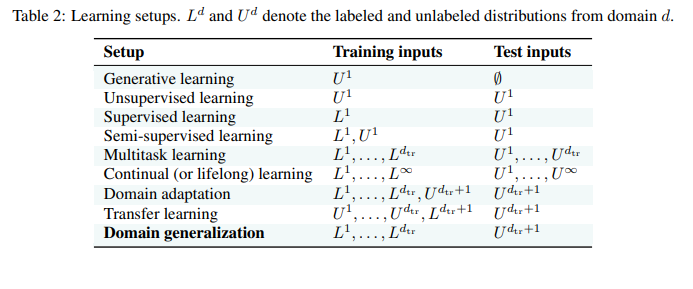
- Beta Moving Average, Exponential Moving Average, uniform average 세 가지의 model optimization 방식을 비교한 결과, EMA는 training 후반부에 집중하기 때문에 new class 정확도는 상대적으로 높지만 base class 정확도는 낮고, AVG는 반대의 경향을 보임

- 반면 BMA는 base class와 new class 둘 다 높은 정확도를 보이기 때문에, pretrained model과 fine-tuning model의 균형을 유지하고 있다는 것을 알 수 있음

5. Discussion

- CLIPood는 image encoder만을 fine-tuning하지만, text modality도 fine-tuning한다면 OOD generalization을 개선시킬 것으로 기대됨

- 또한, 이 방식은 zero-shot performance가 높은 pretrained model의 경우만 적용 가능하므로, 그 외의 경우에도 적용하기 위한 whole pre-training fine-tuning pipeline method가 필요할 것으로 보임



(In search of lost domain generalization, 2021)