1. Introduction

- test set에서 볼 수 있는 real-world의 unknown classes를 다루기 위해 OOD detection task이 관심을 받게 됨

- ID-specific features를 학습한 model을 바탕으로, input data가 ID classes에 얼마나 가까운 지를 나타내는 ID-ness를 측정하기 위한 scoring function을 고안해내는 것이 OOD detection의 대표적인 method임

- 이러한 방법은 그림의 갈색 별 모양의 hard-to-distinguish OOD sample처럼 특정 ID class와의 거리가 가깝고, ID class들까지의 거리 차이가 큰 sample은 감지하기 어렵다는 문제가 있음

- 따라서 generalizable representation을 갖는 CLIP model을 활용해, ID dataset을 학습하지 않고 ID class name 정보만으로 OOD data를 감지하는 zero-shot OOD detection을 시도하는 연구들이 제시되었음

- ZOC method는 CLIP image encoder로부터 얻은 image feature를 학습한 image description generator에 넣어 unseen label set을 생성함

- 그 다음, CLIP 학습에 사용된 seen label set과 unseen label set의 합집합을 CLIP text encoder에 넣고, CLIP feature space에서 image feature와의 similarity를 계산해 zero-shot OOD detection을 수행함

- MCM method는 특정 image embedding과 모든 ID class의 textual embedding 간의 유사도를 구해, maximum probability를 계산해 OOD uncertainty를 판단함

- 하지만 ZOC는 추가적인 dataset이 필요하고, ID class 수가 큰 dataset을 사용할 때 성능이 낮아지는 문제가 있으며, MCM은 여전히 hard-to-distinguish OOD sample을 다루기 어려움

- 본 논문에서는 CLIP의 open world knowledge를 활용해, ID-ness가 높은 sample에 대해서도 OOD detection을 수행함

2. Related Work

2.3. Out-of-Distribution Detection

- OOD detection을 위한 score function은 대부분 probability, logits, feature로부터 얻음

- probability의 경우, baseline method인 MSP score는 maximum softmax probability를 ID score로 사용함

- probability와 특정 class의 p(y|x) template 사이의 KL divergence가 가장 작은 class의 값을 score로 사용하는 방법도 있음

- MaxLogit은 maximum logit을 score로 사용하며, logit은 probability와 달리 unnormalized하기 때문에, class 개수가 많거나 fine-grained class를 갖는 경우에 probability가 disperse되는 문제를 개선할 수 있음

- 또는 class centroid나, low-dimensional embedding과 feature 간 차이를 이용하는 방법도 있음

- 위의 방법들은 image만으로 ID score를 계산하며, score를 통해 ID-ness를 판단하지만, CLIPN은 image와 text를 모두 고려하며, 바로 OOD sample을 감지한다는 차이가 있음

3. Methodology

3.3. Prompt Design

- CLIPN은 pre-training 시에 기존 CLIP의 positive prompt 외에도 추가적으로 “no” prompt를 사용함

- input의 class name text t가 주어지면, L개의 hand-crafted “no” prompts를 갖는 “no” prompt tool에서 랜덤하게 prompt를 선택하고, embedding layer를 통해 얻은 token feature vector를 “no” text encoder에 입력함

- 이 때 token feature는 learnable parameter를 통해 학습 가능함

3.4. Training Loss Design

- CLIPN은 image feature와 correct “no” prompt feature를 가까워지도록 학습하는 Image-Text Binary-Opposite Loss (ITBO)와, positive prompt feature와 “no” prompt feature를 멀어지도록 학습하는 Text Semantic-Opposite Loss (TSO)를 loss function으로 사용함

- i, j를 mini-batch data의 index라고 할 때, i번째 이미지와 j번째 “no” text 간의 match-ness를 mij로 표현함

- mij=1은 matched yet unrelated 경우를 나타내며, image와 “no” text가 관련되지 않음을 의미함

- mij=0은 reversed matched 경우를 나타내며, “no” text가 image와 반대되는 의미를 가짐을 의미함

- ITBO loss는 왼쪽 그림과 같으며, pnoij는 i번째 image와 j번째 “no” text 간의 matched probability를 의미함

- 같은 sample pair의 image feature와 “no” text feature 사이 거리는 멀어지도록 하고, 다른 pair의 경우 서로 가까워지도록 함

- TSO loss는 같은 sample pair의 positive text feature와 “no” text feature 간의 거리가 멀어지도록 만들어, opposite semantic을 반영할 수 있음

3.5. Inference algorithm of CLIPN

- pij는 ID class probability를 의미하고, 각 ID class의 positive prompt feature와 image feature 유사도의 softmax를 계산해 구할 수 있음

- pnoij는 ITBO에서 사용한 값과 같은 값을 사용하며, 여기서는 각 ID class의 negative prompt와 image feature 간의 matching probability를 의미하고, class마다 독립적으로 계산됨

- Competing-to-win 알고리즘은 먼저 가장 큰 pij를 갖는 class의 pnoij 값을 찾음

- 이 때 1-pnoij가 i번째 image와 class j의 positive text 간의 matching probability를 의미하기 때문에, 1-pnoij가 0.5 이상이면 input x의 ID-ness가 1이 되고, 반대의 경우 0이 돼 OOD라고 판단함

- 하지만 maximum ID probability 값을 가진 class가 명확하지 않은 경우에는 OOD detection에 실패할 수 있기 때문에, agreeing-to-differ algorithm이 추가적으로 제시됨

- OOD에 해당하는 경우를 unknown class로 생각할 때, C+1개의 class probability를 갖게 되고, 이때 각 class probability를 (1-pnoij)\*pij 로 계산함

- probability 합이 1이 되도록 하는 pc+1을 구할 수 있고, 이 값이 maximum ID probability보다 큰 경우 OOD로 판단하게 됨

- 두 알고리즘 모두 추가적인 training 없이 ID class name만을 inference 때 사용하기 때문에, zero-shot으로 OOD detection task를 수행하게 됨

4. Experiment

4.2. Results on Zero-shot OOD Detection

- large-scale dataset과 small-scale dataset에 대해 OOD detection을 수행한 결과, 두 경우 모두 CLIPN에 ATD 알고리즘을 이용했을 때, sota method인 MCM 대비 평균적으로 더 좋은 성능을 보임

- 또한 하나의 dataset에 대한 in-domain OOD detection의 경우 MSP method 대비 높은 성능을 보임

4.3. Ablation Study

- ITBO loss만을 사용했을 때의 성능이 sota method보다 높고, ITBO loss와 TSO loss를 함께 사용했을 때는 더 높은 성능을 보이기 때문에, 두 loss 모두 효과적임을 알 수 있음

- hand-crafted “no” text만을 사용할 때보다 learnable “no” text까지 함께 적용했을 때 더 뛰어난 성능을 보임

- CLIPN-A의 경우 높은 threshold일 때의 MSP method보다도 mis-classified OOD samples 수가 적음

- image feature의 positive text feature와의 유사도, negative text feature와의 유사도에 대해 커널 밀도 함수 추정을 한 결과, CLIP과 달리 CLIPN은 두 유사도가 뚜렷한 차이를 보임

- 또한 correct class text의 경우 positive text와의 유사도가 더 높고, false class text의 경우 negative text와의 유사도가 더 높음

- 따라서 image와 “no” text 간의 matching을 학습한 것을 알 수 있음