공학석사학위논문

학습분포 외 샘플탐지를 위한 교차 주의집중 트랜스포머 기반 대조표현학습

Cross Attention Transformer-based Contrastive Representation Learning for Out-of-distribution Detection

2023년 2월

서울대학교 대학원

전기정보공학부

정 재 호

공학석사학위논문

학습분포 외 샘플탐지를 위한 교차 주의집중 트랜스포머 기반 대조표현학습

Cross Attention Transformer-based Contrastive Representation Learning for Out-of-distribution Detection

2023년 2월

서울대학교 대학원

전기정보공학부

정 재 호

학습분포 외 샘플탐지를 위한 교차 주의집중 트랜스포머 기반 대조표현학습

Cross Attention Transformer-based Contrastive Representation Learning for Out-of-distribution Detection

지도교수 최 진 영

이 논문을 공학석사 학위논문으로 제출함

2023년 2월

서울대학교 대학원

전기정보공학부

정 재 호

정재호의 공학석사 학위 논문을 인준함

2023년 2월

위 원 장: 고 형 석 (인)

부위원장: 최 진 영 (인)

위 원: 정 교 민 (인)

요 약

이상치 탐지 문제 중 하나인 OOD(Out-of-Distribution) 샘플 탐지는 입력으로 들어오는 샘플에 대해 모델 학습에 사용된 데이터(known data within in-distribution)인지, 학습하지 못한 데이터(unknown data within OOD)인지 판단하는 문제이다. 이는 잘못된 판단이 사고율로 직결되는 의료나 시스템 안전과 같은 분야에서 매우 중요한 요소이다. 이러한 문제를 해결하기 위해 여러 방법들이 제안되었지만, 현실적으로 OOD 데이터의 정의와 범위 등이 모호하다. 따라서 모델 학습을 위해 OOD 샘플을 사용하는 것은 한계가 있다. 이 한계를 극복하기 위해 OOD 샘플들을 학습에 사용할 수 없는 일반적인 상황에서의 OOD 샘플을 탐지하는 연구가 많이 진행되었다.

OOD 샘플을 탐지하기 위해 비전 트랜스포머(Vision Transformer)가 가장 좋은 성능을 보여 왔다. 최근 클래스 대표 특성과 이미지 데이터 특성 간의 교차주의집중 트랜스포머 (cross-attention transformer)를 이용하여 OOD 탐지 성능을 향상시키는 연구가 진행되었다. 본 연구에서는 이 교차주의집중 트랜스포머를 이용하여 표현 학습(representation learning)을 적용한 OOD 탐지 방법을 제안하였다. 구체적으로 설명하면, 클래스와 각 이미지간의 상관 관계를 인코딩 하는 교차주의집중 트랜스포머가 대조표현학습(contrastive representation learning) 방법을 통해 학습 데이터의 특성 표현을 학습함으로서 보다 더 강하게 상관관계를 학습하도록 한다.

따라서, 본 논문에서 제안하는 학습 방법은 2단계 학습(2-stage learning) 방법으로, 첫 번째 단계에서 교차주의집중 트랜스포머를 사전훈련 하기 위해 대조표현학습 방법을 적용한다. 두 번째 단계에서는 사전학습된 트랜스포머에 분류기(classifier)를 붙여서 정밀학습(fine-tuning)하는 방법이다. 제안된 방법은 오직 학습 데이터셋인 ID(in-distribution) 샘플로만 학습을 하였다. OOD 탐지는 OOD 샘플 탐지 이외에도 ID 샘플에 대해서 다중 클래스 분류(mult-class classification)를 동시에 수행하는 문제이기에, 최종적으로 테스트 샘플들이 인코더와 분류기를 통해 나온 신뢰 점수

(confidence score)를 기반으로 OOD 샘플인지 아닌지 판단하면서 동시에 ID 샘플들이 어느 클래스에 속하는지 다중 클래스 분류까지 한다.

주요어: Out-of-Distribution 탐지, 비전 트랜스포머, 교차주의집중, 표현학습
학번: 2021-20714

차 례

표 차례

그림 차례

제 1 장 서론

최근, 딥 뉴럴 네트워크(Deep Neural Networks)가 컴퓨터 비전(Computer Vision), 자연어 처리(Natural Language Processing), 그리고 음성 인식(Speech Recognition)과 같은 분야에서 전례없는 성공과 성능을 보여주면서 많은 딥러닝(Deep Learning) 분야 연구자들에게 주목을 받고 있다. 딥 뉴럴 네트워크가 높은 성능과 결과를 보여주지만, 딥 네트워크의 결과에 대해 지나치게 높은 신뢰(over-confidence)를 보여주는 경향이 있으며 [2], 이미지 분류(image classification) 문제에서 주로 사용되는 ReLU 활성화(activation) 함수 기반의 네트워크들에서 높은 신뢰도의 예측을 하게 되는 결과가 있다 [3]. 이와 같은 현상들은, 이상치 탐지(Anomaly Detection) 분야에서 모델의 성능과 직결되는 중요한 문제로, 모델의 입력으로 학습 데이터 분포가 아닌 다른 분포의 데이터가 들어오게 되는 경우 해당 샘플에 대해 모델이 높은 신뢰도를 가지고 잘못된 예측이나 결과를 출력하게 되고, 이는 시스템 안전(system safety)과 관련된 의료 진단이나 자율 주행과 같은 분야에 치명적인 결과를 불러올 수 있다.

이상치 탐지 중 하나인 학습 분포 외 샘플(OOD; Out-of-Distribution) 탐지는 입력 샘플이 학습 데이터 분포 내의 샘플(ID; In-distribution)인지 아닌지 판단하는 것과 동시에 입력 샘플이 학습 데이터 분포 내의 샘플이라면 학습 데이터셋의 클래스들 중 하나로 분류하는 다중 클래스 분류(multi-class classification)를 해결하는 문제이다. 즉, OOD 탐지는 기존의 이미지 분류에 해당 입력 샘플이 학습 분포 내의 샘플인지 아닌지 판단하는 이진 분류(binary classification)가 추가된 문제라고 볼 수 있다. 따라서, OOD 샘플을 탐지하는 것은 앞서 언급한 시스템 안전과 관련된 분야에서 매우 중요한 문제이다.

OOD 샘플들을 탐지하기 위해, 기존 연구들은 모델을 학습시키는 방법 혹은 스코어 함수(score function)를 제안하였으며, 해당 방법들은 주로 컨벌루션 신경 계층망 (CNN; Convolutional Neural Network) [4] 기반의 모델을 사용하였다. 하지만, [5], [6] 저자들은 주의 집중(attention) 기반 모델이 전역(global) 정보를 CNN

기반의 모델보다 더 잘 포착하는 특징을 이용하여 비전 트랜스포머(ViT; Visiont Transformer) [7] 기반의 OOD 샘플을 탐지하는 방법을 제시하였다. 비전 트랜스포머(ViT)는 자연어 처리 분야에서 널리 사용되는 트랜스포머 구조를 비전 분야에 적용한 모델로, 이미지를 여러 개의 패치로 나눈 후에 패치들 간의 연관성을 보면서 학습을 하도록 설계된 자가주의집중(self attention) 기반의 모델이다. 따라서, 비전 트랜스포머는 이미지의 전체적인 맥락과 주의집중(attention)을 이용하여 이미지 패치들로부터 추출한 특성별 상관관계(feature-wise correlation)를 학습하기 때문에, CNN보다 더 이미지의 전체적인 맥락을 잘 이해한다고 볼 수 있고, 이러한 특징을 이용하여 OODformer [5]는 처음으로 비전 트랜스포머를 활용한 OOD 탐지를 제안하였다. 또한, [6] 저자들은 비전 트랜스포머가 CNN에 비해 배경 변화(background shift), 스타일 변화(style shift), 질감 변화(texture shift), 그리고 손상으로 인한 변화(corruption shift) 같은 분포 변화(distribution shift)에서 일반화가 더 잘 되는 것을 실험적으로 입증하였다. 따라서, [5], [6] 저자들은 여러 실험 결과와 분석을 통해 비전 트랜스포머 기반의 모델이 CNN 보다 OOD 탐지 성능이 더 우수하다는 결론을 내렸다.

본 논문에서의 핵심 아이디어는 비전 트랜스포머로부터 추출한 특성(feature)을 클래스 특성과 같이 활용하여 모델이 학습 데이터의 특성 표현(feature representation)을 학습하도록 하는 것이다. 다시 말해, 모델이 하나의 이미지에서 자가주의집중(self attention) 기반으로 학습하는 것이 아닌, 하나의 이미지 내의 다른 패치들 사이에서의 자가주의집중 이외에 추가로 클래스 대표 특성과 이미지 특성 사이의 상관관계를 학습하는 교차주의집중 트랜스포머(cross attention transformer; CAT) [1] 기반의 대조표현학습(contrastive representation learning) 방법을 제안한다. 본 논문에서는 제안된 방법을 CAT-based CoReL이라 칭한다. ImageNet [8] 데이터셋으로 사전학습된(pre-trained) 비전 트랜스포머를 CIFAR-10/-100 [9] 데이터셋으로 미세조정(fine-tuning)을 한 모델을 기준 모델(baseline)으로 한다. 이 모델을 특성 추출기(feature extractor)로 사용하여 교차주의집중 트랜스포머를 구성하고 대조표현학습을 통해 사전학습시킨 후에 분류기를 추가로 학습한다. 이때, 교차주의집중 트랜스

포머와 분류기는 ID 데이터만 이용해서 학습을 진행한다.

본 논문에서 제안한 CAT-based CoReL의 contribution은 다음과 같다.

- OOD 데이터셋을 학습에 활용할 수 없는 상황에서 OOD 샘플 탐지를 위한 교차주의집중 기반 대조표현학습 방법을 제안한다.
- 다양한 OOD 데이터셋과 어블레이션 연구(ablation study)를 통해 OOD 탐지 성능을 비교하고 분석하여 제안한 방법의 효용성을 보였다.
- ID 데이터셋이 CIFAR-10, OOD 데이터셋이 SVHN, Places365, Texture인 far OOD 탐지 경우, 기존 최고 성능인 [10] 방법에 비해 -0.31% (FPR95 평균값), +0.3% (AUROC 평균값)이 향상된 1.85% (FPR95 평균값), 99.65% (AUROC 평균값)을 달성하였다. 게다가, CIFAR-10(ID) vs CIFAR-100(OOD)인 near OOD 탐지에서 -0.68% (FPR95), +0.04% (AUROC) 만큼 향상된 6.21% (FPR95), 98.56% (AUROC)를 달성하였다.

제 2 장 관련 연구

2.1 컨벌루션 신경 계층망 기반 out-of-distribution(OOD) 탐지

MSP(Maximum Softmax Probability) [11]는 컴퓨터 비전, 자연어 처리, 그리고 음성 인식 분야에서 out-of-distribution(OOD) 샘플 탐지를 위한 방법과 평가 지표(metric)를 처음으로 제시한 베이스라인 논문이다. In-distribution(ID) 데이터로 잘 사전 학습(well pre-trained)된 뉴럴 네트워크(neural network)는 ID 테스트 데이터의 소프트맥스 확률값(softmax probability)이 OOD 데이터의 값보다 높게 나올 것이라는 개념으로, 해당 논문에서는 입력으로 들어온 이미지 샘플에 대해서 각 클래스의 소프트맥스 확률값 중 가장 큰 값을 신뢰 점수(confidence score)로 이용하여 특정 임계치(threshold)를 기준으로 신뢰도 점수가 임계치보다 높게 나오면 ID 샘플, 낮게 나오면 OOD 샘플로 판단하는 방법을 제안하였고, 신뢰 점수 함수인 스코어 함수(score function)는 식(2.1)과 같다.

$$S_{MSP}(x; f) = \max_k \frac{\exp(f_k(x))}{\sum_{j=1}^{C} \exp(f_j(x))}. \tag{2.1}$$

이 때, C는 데이터셋의 클래스 개수이며, $f_i(x)$는 입력 데이터 x에 대한 뉴럴 네트워크의 i번째 로짓(logit)값이다.

ODIN [12]은 MSP의 방법에 온도 스케일링(temperature scaling)과 입력 전처리(input preprocessing)를 적용하여 ID와 OOD 샘플간의 신뢰 점수 차이를 크게 하여 보다 정확하게 OOD 탐지를 하기 위한 방법을 제안하였다. 온도 스케일링과 입력 전처리는 각각 식(2.2)와 식(2.3)에 해당된다.

$$S_i(x; T) = \frac{\exp\left(f_i(x)/T\right)}{\sum_{j=1}^{C} \exp\left(f_j(x)/T\right)}. \tag{2.2}$$

$$\tilde{x} = x - \epsilon \cdot sign(-\nabla_x log(S(x;T))). \tag{2.3}$$

식(2.2)에서 $f(\cdot)$은 신경망으로 $f_i(x)$는 입력 데이터 x에 대한 i번째 로짓값이며, C는 클래스 개수이다. 온도 스케일링은 식(2.2) [12]과 같이 네트워크의 출력인 로짓값에 하이퍼파라미터(hyperparameter)인 T로 나눈 후에, 소프트맥스를 적용한 방법이다. 온도 스케일링 이외에도, 저자들은 입력 전처리를 적용하였으며 식(2.3) [12]에서 입력 데이터 x에 작은 변화(small perturbation)를 가하는 방법으로 ϵ는 작은 변화의 크기를 조정하는 하이퍼파라미터로 OOD 테스트셋 이외의 다른 OOD 데이터셋을 검증 데이터셋(validation dataset)으로 사용하여 해당 하이퍼파라미터 값을 정한다. 따라서, 온도 스케일링과 입력 전처리 방법을 통해 ID와 OOD 샘플 간의 신뢰 점수 차이를 크게 벌어지게 하여 보다 더 정확하게 OOD 샘플을 탐지하는 방법을 제안하여, 기존 베이스라인 방법(MSP) [11] 보다 OOD 탐지 성능을 더 높이게 되었다. ODIN의 스코어 함수는 다음 식(2.4)과 같다.

$$S_{ODIN}(\tilde{x}; f) = \max_i \frac{\exp\left(f_i(\tilde{x})/T\right)}{\sum_{j=1}^{C} \exp\left(f_j(\tilde{x})/T\right)}. \tag{2.4}$$

마할라노비스(Mahalanobis) [13]는 사전 학습된 소프트맥스 기반 신경 계층망의 특성(feature)이 클래스 조건부 가우시안 분포(class-conditional Gaussian distribution)를 따를 것이라는 가정 하에, 분포와 임의의 데이터 한 점 사이의 거리인 마할라노비스 거리(mahalanobis distance)를 이용하여 OOD 탐지를 위한 새로운 스코어 함수를 제시한 논문이다. 해당 저자들은 마할라노비스 거리를 구하기 위해 사전 학습된 컨벌루션 신경 계층망 모델의 모든 층(layer)으로부터 학습 데이터 분포들의 중심(mean)과 공분산(covariance)를 구한 뒤, 마할라노비스 거리를 계산하여 신뢰 점수로 사용하였다. 마할라노비스 거리를 계산하기 위해 학습 데이터의 각 클래스의 평균과 공분산을 다음 식(2.5) [13]과 같이 각각 계산한다.

$$\hat{\mu}_c = \frac{1}{N_c}\sum_{i:y_i=c} f(x_i), \quad \hat{\Sigma} = \frac{1}{N}\sum_{c}\sum_{i:y_i=c}(f(x_i)-\hat{\mu}_c)(f(x_i)-\hat{\mu}_c)^T. \tag{2.5}$$

N_c는 레이블(label) c에 속하는 학습 데이터의 개수이며, $f(\cdot)$는 완전 연결 계층(fully-connected layer) 바로 직전의 특성 출력이고, N은 학습 데이터의 총 개수이다. 저자들은 학습 데이터로부터 가우시안 분포의 중심과 공분산을 구한 후, 마할라노비스 거리 기반의 신뢰 점수 함수를 다음 식(2.6) [13]과 같이 계산한다.

$$S_{Mahalanobis}(x; f) = -\max_{c}(f(x)-\hat{\mu}_c)^T\hat{\Sigma}^{-1}(f(x)-\hat{\mu}_c). \tag{2.6}$$

식(2.6)을 이용하여 테스트 샘플과 가장 가까운 클래스 분포와의 거리가 신뢰 점수로 사용이 되어 해당 신뢰 점수가 낮으면 ID 샘플 반대면 OOD 샘플로 분류하게 된다. 이외에도, 그들은 ODIN [12]에서 제안한 입력 전처리 방법을 추가로 적용하여 OOD 탐지 성능을 올렸다.

이상치 노출(OE; Outlier Exposure) [14]은 기존의 OOD 탐지와는 다른 관점으로 OOD 탐지하는 방법을 제안한 논문이다. 기존의 OOD 탐지에서 모델은 오직 ID 샘플로만 학습이 된 후에 OOD 샘플을 탐지한다. 이는 학습된 모델에게 주어진 정보는 ID 샘플 밖에 없으며 현실적으로도 ID 데이터셋 이외의 OOD 데이터셋을 이용하여 모델을 학습시키는 것은 제한적이고, 만약 학습을 시킨다고 해도 모델이 학습에 사용되는 OOD 데이터셋에 편향될 수 있는 문제가 있다. 하지만, [14] 저자들은 기존의 OOD 탐지 문제 정의에 부합하면서 앞서 언급한 문제점들을 야기시키지 않도록 하는 학습 방법을 제안하였다. 그들은 학습에 사용할 ID 데이터셋과 보조 데이터셋(auxiliary dataset)을 이용하여 모델을 학습시키는데, 보조 데이터셋은 ID 데이터셋과 추론시 사용할 OOD 데이터셋과 겹치지 않는 데이터셋을 사용하였고 테스트에 사용되는 OOD와는 달리 학습에 사용되는 보조 데이터셋은 OOD 샘플 역할로 모델 학습에 있어 일종의 규제(regularization) 역할을 한다. 따라서, 모델은 ID 데이터셋에 대해서는 신뢰 점수가 높게, 보조 데이터셋에 대해서는 신뢰 점수가 낮게 출력되도록 학습된다. 그들이 제안하는 손실 함수는 다음 식(2.7) [14]과 같다.

$$\mathcal{L} = \mathbb{E}_{(x,y)\sim D_{in}}[\mathcal{L}_{CE}(f(x), y) + \lambda \mathbb{E}_{x_{out}\sim D_{out}^{OE}}[\mathcal{L}_{OE}(f(x_{out}), f(x), y)]]. \quad (2.7)$$

학습시 사용할 손실 함수는 사용자가 목적과 해결하고자 하는 문제에 따라 정할 수 있다. 예를 들어, 최대 소프트맥스 확률 기반 탐지기(maximum softmax probability detector)를 사용한다면, ID 데이터셋에 대한 손실 함수는 교차 엔트로피(cross entropy)이고, OOD 데이터셋에 대한 손실 함수는 모델의 출력값이 균등 분포(uniform distribution)가 되도록 쿨백-라이블러 발산(KL-divergence; Kullback-Leibler divergence)을 이용한다. 식(2.7)에서 $f(\cdot)$는 모델이고 $\mathcal{L}_{CE}$, $\mathcal{L}_{OE}$는 위에서 설명하였듯이, 각각 교차 엔트로피 손실 함수와 OOD 데이터셋에 대한 모델의 출력값이 균등 분포가 되도록 하는 손실 함수이다.

에너지(Energy) [15]는 기존의 확률 기반의 신뢰 점수(i.e., 소프트맥스 신뢰 점수)가 지나친 신뢰(over-confidence) 문제에 취약하다는 점을 지적하면서 OOD 탐지를 위한 하이퍼파라미터가 없는 새로운 에너지 스코어 함수를 제안하였다. 저자들은 에너지 기반 모델(EBM; Energy-based Model) [16]에서 입력 데이터 x를 비확률적 스칼라(non-probabilistic scalar)값으로 사상(mapping)시키는 에너지 함수를 딥 뉴럴 네트워크 분류기에 적용하기 위해, 다음과 같은 수식 2.8, 2.9들을 이용하여 해당 논문에서 제시한 에너지 함수와 스코어 함수를 유도한다.

$$p(y = i|x) = \frac{\exp(f_i(x)/T)}{\sum_j^C \exp(f_j(x)/T)}. \quad (2.8)$$

$$p(y|x) = \frac{\exp(-E(x,y)/T)}{\int_{y'} \exp(-E(x,y')/T)} = \frac{\exp(-E(x,y)/T)}{\exp(-E(x)/T)}. \quad (2.9)$$

식(2.8) [15]은 사후 확률(posterior probability)을 소프트맥스를 이용하여 나타낸 수식이며, 식(2.9) [15]은 에너지 함수와 확률 분포간의 관계를 표현하는 수식이다. 따라서, 저자들은 식(2.8)과 식(2.9) 간의 관계를 연결지으며 입력 데이터에 대한 에

너지를 $E(x, y) = -f_i(x)$로 정의하였고, 이는 곧 뉴럴 네트워크 분류기의 출력인 로짓값을 에너지로 정의한 것을 의미하며 다음 식(2.10) [15]과 같이 정의된다.

$$E(x; f) = -T \cdot \log \sum_{j=1}^{C} \exp(f_j(x)/T). \tag{2.10}$$

C는 데이터셋의 클래스 개수이고, $f_j(x)$는 입력 데이터 x에 대한 뉴럴 네트워크의 j 번째 출력 로짓값이며, T는 온도 매개변수(temperature parameter)이다. 논문 저자들은 온도 매개변수 값을 $T = 1 \sim 1000$에 따른 실험을 진행한 결과, 큰 온도 매개변수 값을 사용하게 되면 ID와 OOD 샘플 간의 에너지 점수로 잘 구별이 안 된다는 것을 실험적으로 입증하였고, $T = 1$을 기본값으로 설정하였다. 이는 그들이 제안한 에너지 점수가 매개변수로부터 자유롭다(parameter-free)는 것을 의미한다. 따라서, 에너지 신뢰 점수는 최종적으로 다음 식(2.11)과 같다.

$$S_{Energy}(x; f) = -\log \sum_{j=1}^{C} \exp(f_j(x)). \tag{2.11}$$

ReAct [17]는 Rectified Activations의 줄임말로, 저자들은 ID와 OOD 샘플들이 모델의 활성화(activation) 출력값들의 차이가 크게 나타나는 현상을 사후 관찰 분석(post hoc analysis)을 통해 발견하여 활성화를 특정 임계치(threshold)로 절단(rectification)하는 방법을 제안한 논문이다. 저자들은 OOD 샘플이 ID 샘플과는 다른 활성화 패턴을 보여주는 원인으로 모델을 학습시키는 내부 메커니즘 중 하나인 배치 정규화(batch normalization) [18]로 근거를 들었다. 특정 학습 데이터셋으로 모델을 학습 시킬때, 학습 단계에서 모델은 학습 데이터셋에 의해 평균(running mean)과 분산(running variance)을 추정하고 추론 단계에서는 학습 단계에서 구한 배치 통계(batch statistics)를 학습 테스트 데이터셋에 적용한다. 하지만, 이러한 배치 통계가 학습 데이터셋 즉, ID 데이터셋에 의해 추정되고 맞추어져 있어 학습된 모델에 OOD 샘플이 들어오게 되면 입력 OOD 샘플과 ID 샘플의 배치 통계가 불일치(mismatch)하게 되고, 따라서 OOD 샘플과 ID 샘플 간의 활성화 값의 차이가 생기는 결과가 생긴다고 설명하였다. 게다가, 저자들이 제안한 방법을 기존의 OOD 탐지 스코어

함수에 적용하면 기존 성능보다 더 올릴 수 있음을 보였고, 모델 내부의 메커니즘이 OOD 탐지에 영향을 미친다는 결론을 내렸다.

GradNorm [19]은 특성 공간(feature space)이 아닌 기울기 공간(gradient space)에서의 정보를 활용하여 OOD 탐지 하는 방법을 제안한 논문이다. ID 데이터셋에 의해 모델 가중치가 학습 되는데, 이는 ID 샘플들을 특정 클래스로 잘 분류하도록 모델 가중치가 학습이 된다는 것을 의미한다. 따라서, 저자들의 핵심 아이디어는 학습된 모델은 ID 샘플의 소프트맥스 출력 분포와 균등 분포 사이의 쿨백-라이블러 발산의 가중치에 대한 기울기의 크기가 크게 나오고, OOD 샘플은 낮게 나온다는 것이다. ODIN [12]에서도 입력 전처리를 위해 기울기를 활용하였으나, GradNorm은 기울기를 OOD 탐지를 위해 직접적으로 활용한 방법이며 이를 기반으로 저자들이 제안한 스코어 함수는 다음 식(2.12) [19]과 같다.

$$S_{GradNorm}(x; f) = \| \frac{\partial D_{KL}(\mathbf{u} \parallel softmax(f(x)))}{\partial \mathbf{W}} \|_p . \tag{2.12}$$

$f(\cdot)$은 모델, $D_{KL}(\cdot)$은 쿨백-라이블러 발산, $\mathbf{u}$는 균등 분포, $\mathbf{W}$는 가중치 행렬(weight matrix), $\|\cdot\|_p$는 L_p-norm이다.

OpenHybrid [20]는 open set recognition(OSR) [21], [22] 문제를 해결하기 위한 학습 방법을 제안한 논문이다. OSR은 OOD 탐지와 매우 유사한 문제로, 차이점은 테스트셋이 주어진 클래스(known class)와 주어지지 않은 클래스(unknown class) 모두 포함한다는 가정하에 모델의 환경이 열린 공간(open set)이라는 것이다. 예를 들어, 모델이 4개의 클래스를 가진 데이터셋으로 학습이 되면 추론시 해당 모델은 입력 샘플을 자신이 학습한 클래스 중 하나로 분류를 하려고 할 것이며, 이는 모델이 학습하지 못한 샘플(unknown sample)이 입력으로 들어왔을 때 OOD로 탐지를 못하게 되는 상황이 생길 수 있다. 따라서, OSR은 현실적으로 OOD 샘플을 탐지하기 위해 열린 공간으로 확장하여 모르는 샘플도 분류할 수 있게 unknwon 클래스를 추가하여, 모델이 열린 공간을 인지하도록 학습시킨 후 ID와 OOD 샘플을 탐지한다. [20] 저자들은 open set에서의 OOD 탐지를 위해 입력 데이터를 특성 공간으로

임베딩하는 인코더 $f(\cdot)$, ID 샘플을 분류하는 분류기, 그리고 OOD 샘플을 탐지하기 위한 확률 밀도 추정기(density estimator)로 네트워크를 구성하여 ID 샘플을 분류하는 분류기와 OOD 샘플을 판단하기 위한 확률 밀도 추정기 사이의 결합적인 표현(joint representation)을 end-to-end 방식으로 학습하는 방법을 제시하였다. 따라서, open set 문제에서 저자들이 제안한 학습 방법으로 OOD 샘플에 높은 가능도(higher likelihood)를 부여하는 문제를 해결하였다.

OECC [23]는 이상치 노출(OE) [14]의 아이디어를 기반으로 OOD 탐지를 위한 새로운 학습 방법을 제안하였다. 저자들은 손실 함수를 최적화 기반으로 설계하였는데 식(2.13) [23]와 같다.

$$\begin{aligned}\min_{\theta}(&\mathbb{E}_{(x,y)\sim D_{in}}[\mathcal{L}_{CE}(f_\theta(x),y)] \\ &+\lambda_1(A_{tr}-\mathbb{E}_{x\sim D_{in}}[\max_{i=1,\cdots,K}\exp(z_i)/\Sigma_{j=1}^{K}\exp(z_j)])^2 \\ &+\lambda_2\Sigma_{x\sim D_{out}^{OE}}\Sigma_{i=1}^{K}\left|1/K-\exp(z_i)/\Sigma_{j=1}^{K}\exp(z_j)\right|). \end{aligned} \tag{2.13}$$

$\mathcal{L}_{CE}(\cdot)$는 교차 엔트로피 손실 함수, A_{tr}은 모델의 학습 정확도(train accuracy), D_{in}는 학습 데이터셋, D_{out}^{OE}는 보조 데이터셋(auxiliary dataset), z_i는 모델의 i번째 출력 로짓값, $f_\theta(\cdot)$는 θ로 매개변수화 된 모델이다. 식(2.13)의 두번째 항은 모델의 소프트맥스 출력이 A_{tr}이 되도록 규제(regularization)하는 역할을 하고, 세번째 항은 모델의 출력이 균등 분포가 되도록 규제 역할을 한다. 위 손실 함수를 기반으로 저자들은 2단계 학습 방법을 다음과 같이 제시한다. 우선, A_{tr}을 추정하기 위해 적정 수준까지 모델을 학습시킨 후에, A_{tr}을 고정시키고 식(2.13) 이용하여 미세 조정을 한다. 따라서, 해당 논문 저자들은 제안한 학습 방법으로 이미지와 텍스트 분야에서 OOD 샘플을 탐지하는 방법을 제안하였다.

2.2 트랜스포머 기반 out-of-distribution(OOD) 탐지

트랜스포머 구조는 자연어 처리 분야에서 널리 사용되는 구조로 [24]에서 주의 집중 메커니즘(attention mechanism)을 이용하여 자연어 시퀀스(sequence)를 학습

하고 예측하기 위해 제안되었다. 주의집중은 한 시퀀스 내의 토큰(token)들이 서로 얼마나 연관되어 있는지를 나타내는 주의집중 지도(attention map)를 통해 시퀀스 내에서 전반적인 연관도를 제공한다.

트랜스포머의 동작 원리는 단어들의 시퀀스가 입력으로 들어오게 되면 해당 단어들은 토큰화 과정을 통해 계산이 가능한 벡터로 임베딩을 하게 되고, 이 과정을 단어 임베딩(word embedding)이라 한다. 단어 임베딩이 된 후, 토큰들은 트랜스포머 입력으로 들어가게 되고 해당 토큰들의 각 위치마다 위치정보가 있어야 하므로 위치 인코딩(position encoding)을 하게 된다. 이후, 주의집중을 하기 위해 각 토큰들마다 유사도를 구하고 싶은 토큰을 선형 사상(linear mapping)시킨 쿼리(query)와 유사도를 구하고 싶은 토큰과 비교할 토큰을 선형 사상시킨 키(key), 그리고 쿼리와 키로부터 구한 유사도를 곱할 밸류(value)를 추출해야 한다. 이들을 추출하기 위해서 선형 사상을 위한 가중치 W^Q, W^K, W^V가 필요하며 해당 가중치들을 이용하여 각 토큰들마다 쿼리, 키, 밸류를 추출한다. 쿼리, 키, 밸류를 추출한 이후에, 식(2.14) [24]처럼 쿼리와 키 사이의 유사도를 스케일 된 내적과 정규화를 시킨 결과에 밸류를 곱하여 주어진 시퀀스에 대한 맥락 특징을 얻는다.

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{\alpha}})V. \tag{2.14}$$

쿼리, 키, 밸류가 각각 Q, K, V로 표기되었고, α는 상수로 키의 차원을 나타낸다. 다중-헤드(multi-head) 주의집중은 주의집중 기반 모델에서 주로 사용되며 주의집중을 병렬적으로 처리하는 구조를 가진다. 만약, N개의 헤드와 피드포워드(feed-forward) 가중치 W_{ff}가 존재할 때, 식(2.15) [24]처럼 계산된다. $Multihead(\cdot)$ 함수는 다중-헤드 형태로 주의집중을 하는 함수이며, $head_i$는 식(2.14)에서 QW_j^Q, KW_j^K, VW_j^V가 입력으로 들어가서 나온 결과이며, $Concat(\cdot)$ 여러 개의 텐서(tensor)를 하나의 텐서로 쌓는 함수로 $Concat(\cdot)$을 이용하여 여러 번의 계산을 한 번에 처리할 수 있다.

$$Multihead(Q, K, V) = Concat(head_1, head_2, ..., head_N)W_{ff}. \tag{2.15}$$

트랜스포머 도입 이후로 컴퓨터 비전 분야에 트랜스포머를 적용하는 논문들이 있었고, 그 중 비전 트랜스포머(ViT) [7]는 이미지를 여러 개의 패치로 나눈 후에 이를 하나의 시퀀스로 간주하여 입력으로 사용한 모델 구조이다. 비전 트랜스포머는 컨벌루션 신경 계층망 기반 네트워크보다 우수한 분류 성능을 보여주었으며, 전역(global) 정보를 보다 더 잘 포착하여 분포 이동(distribution shift)에서 일반화 성능이 우수하다는 것이 입증되었다 [6], [5]. [6]의 저자들은 다양한 분포 이동에서의 비전 트랜스포머의 성능 테스트에 의해 비전 트랜스포머는 배경(background)과 질감(texture)보다 형태(shape)와 구조(structure)에 대해 강한 귀납 편향(inductive bias)을 가지고 학습을 한다는 것을 확인하였다. 이는 비전 트랜스포머가 컨벌루션 신경 계층망 모델보다 분포 이동에서의 일반화 능력이 좋다는 것을 의미하며, 따라서 비전 트랜스포머가 OOD 일반화 성능도 우수하다는 것을 의미한다.

하지만, 이러한 우수한 성능과 일반화 능력에도 불구하고, 비전 트랜스포머는 이미지를 패치로 나눈 후에 토큰화 과정을 거친 후에 모델의 입력으로 들어가게 되므로 엄청난 계산량과 메모리를 필요로 하다는 문제점이 있다 [25], [26]. [25], [26]의 저자들은 계산량은 줄이면서 성능은 유지 혹은 좋아지게 하기 위해서 기존의 자가 주의집중이 아닌 교차 주의집중(cross attention)을 적용한 방법을 제안하였다.

[26] 방법은 2개의 브랜치(branch)를 사용하여 교차 주의집중을 사용하였는데, 하나의 작은 크기의 패치들을 다루는 브랜치이고 나머지 하나는 큰 크기의 패치들을 다루는 브랜치이다. 그들은 다른 크기의 패치들을 다루는 브랜치들로부터 다중-규모(multi-scale)의 특성 표현들(feature representations)을 추출한 뒤 교차 주의집중을 하는 방법을 통해, 계산량은 감소시키면서 성능을 향상시켰다. 해당 논문에서 제안한 방법이 우리의 방법과 유사하지만, 그들은 분류 문제에 적용했다는 것과 교차 주의집중의 입력이 우리의 입력과 다르다.

[25]의 저자들은 교차 주의집중을 이용하여 패치들 내부와 패치들 간의 정보를 활용하는 방법을 제안하였다. 그들이 제안한 방법이 컨벌루션 신경 계층망과 비슷한 구조를 가지도록 비전 트랜스포머를 구성함으로서 비전 트랜스포머의 전역 정보 포착과 컨벌루션 신경 계층망의 지역 정보 포착하는 특징들을 모두 활용하여 분류,

탐지, 분할(segmentation) 문제들에 적용하였다. 따라서, 위 논문들을 통해 주의집중 메커니즘과 특성들을 결합하는 방법에 따라 비전 트랜스포머의 성능을 향상시킬 수 있음을 의미한다.

앞서 언급한 비전 트랜스포머의 특징과 우수한 성능을 기반으로 OOD 탐지에 적용한 논문들이 있다 [5], [10]. 그 중에서, OODformer [5]는 처음으로 비전 트랜스포머를 이용하여 OOD 샘플을 탐지한 논문이다. 저자들은 이미지의 전체적인 맥락과 이미지 패치들로부터 추출한 특성별 상관관계(feature-wise correlation)를 학습하는 비전 트랜스포머가 특성 추출(feature extraction)에 있어 우수하다는 점을 활용하여, 비전 트랜스포머를 일종의 인코더(encoder)로 활용하고 트랜스포머 뒤에 분류기(classifier)를 붙여서 지도 학습(supervised learning)을 통해 모델을 학습시켰다. 그들은 마할라노비스 거리 기반의 신뢰 점수를 이용하여 OOD 샘플을 탐지하였으며, 특히 near OOD 탐지에서 비전 트랜스포머의 우수한 능력을 입증하였다. 이 밖에도, 저자들은 비전 트랜스포머의 OOD 데이터 탐지에 대한 일반화 능력(generalizability)을 다양한 실험을 통해 입증하였다.

OODformer [5] 이외에도, [10]은 OOD 탐지 중 near OOD 탐지 문제 해결에 초점을 둔 논문이다. near OOD는 OOD 데이터가 ID 데이터와 의미론적으로 매우 유사한(semantically close) 경우에 대한 탐지 문제로 기존의 OOD 탐지인 far OOD 탐지 문제보다 더 어려운 문제이다. near OOD 탐지의 대표적인 문제로 ID 데이터가 CIFAR-10 [9] 일 때, CIFAR-100 [9]가 OOD 샘플로 탐지하는 문제이며 반대의 경우가 해당 문제보다 더 어렵다. 일반적으로 CIFAR-100이 CIFAR-10보다 학습이 어렵고 클래스 개수가 많기에, CIFAR-100 (ID) vs. CIFAR-10 (OOD)가 더 어려운 문제이다. [10] 저자들은 대규모 데이터셋(large-scale dataset)으로 사전 학습된 비전 트랜스포머를 ID 데이터셋(e.g., CIFAR-10, CIFAR-100 [9])으로 미세 조정(fine-tuning)한 모델이 near OOD 탐지 성능을 크게 향상 시킨다는 것을 실험적으로 입증하였고, 추가적으로 일부의 OOD 샘플을 사용할 수 있다는 가정 하에, 퓨-샷 이상치 노출(few-shot outlier exposure) 방법을 이용하여 성능을 더 향상시키는 방법을 제안하였다. 저자들은 제안한 방법으로 near OOD 탐지 문제에서 최고 성능을 달성하였다.

2.3 특성 표현 학습(Feature Representation Learning)

컨벌루션 신경망 계층 모델을 학습시키는 방법으로 지도 학습(supervised learning)이 널리 사용되며, 해당 방법은 이미지넷 데이터셋 [8]에서 높은 정확도의 성능을 보여준다. 하지만, 공개된 데이터셋(public dataset) 이외의 데이터셋을 사용하여 학습시킬 경우, 데이터셋에 레이블(label)을 달아주기에는 시간과 비용이 많이 든다는 단점이 있다. 따라서, 레이블을 사용하여 학습할 수 없는 상황에서 비지도 학습(unsupervised learning)을 이용해야 하는데, 비지도 학습을 통해 학습시킨 모델의 성능이 지도 학습에 비해 성능이 낮고 이를 해결하기 위해 여러 연구와 시도들이 있었다 [27], [28].

비지도 학습 중 자기 지도 학습(self-supervised learning)의 대표적인 방법으로 contrastive learning이 있다. contrastive learning은 샘플 간의 특성(feature)의 차이(contrast)를 학습하는 방법으로 하나의 기준 샘플(anchor sample)로부터 데이터 증강(data augmentation)을 통해 얻어진 유사한 샘플들을 positive 샘플로 정하고 그 이외의 나머지 샘플들은 negative 샘플로 정하여 임베딩 공간(embedding space)에서 positive 샘플끼리는 가깝게 negative 샘플과는 멀어지도록 샘플들의 특성 표현(feature representation)을 학습하는 방법이다. 따라서, contrastive learning을 통해 사전 학습을 하게 되면 모델이 샘플들의 다양한 특성 표현을 학습할 수 있어 지도 학습에 비해 강건(robust)하게 학습이 된다는 장점이 있다. SimCLR [27]에 의해 지도 학습의 성능과 비교될 만큼 향상되었으며, 해당 저자들은 레이블이 주어지지 않은 상황에서 모델이 다양한 특성 표현을 학습할 수 있는 방법을 제안하였다. 하지만, 해당 방법은 레이블이 주어져 있지 않기에 의미론적으로 같은 클래스의 샘플들에 대해서도 네거티브로 인식이 될 수 있다는 단점이 있다.

앞서 언급한 문제를 해결하기 위해 주어진 데이터셋의 레이블을 활용한 supervised contrastive learning 방법이 제안되었다 [29]. 저자들은 기존의 contrastive 손실 함수에 레이블을 추가하여 같은 레이블의 샘플끼리는 가깝게, 다른 레이블의 샘플과는 멀어지게 임베딩 되도록 특성 표현 학습을 한다. 이러한 레이블을 이용하는

방법은 구분하기 어려운 positive와 negative 샘플들(hard positive/negative samples)에 대해서도 모델이 샘플들의 특성 표현을 잘 학습할 수 있게 하도록 도움을 주는 역할을 한다. 따라서, supervised contrastive learning은 지도 학습과 자기 지도 학습 방법을 결합한 학습 방법으로 볼 수 있다.

자기 지도 학습을 이용하여 OOD 탐지를 수행한 선행 연구들 중 하나인 CSI [30]는 자기 지도 학습인 SimCLR [27]에 초점을 맞추어 OOD 탐지를 위한 학습 방법과 스코어 함수를 제안한 논문이다. 데이터 증강은 자기 지도 학습에 있어 필수 요소로 사용되는데 증강도니 샘플들(augmented samples)은 기존의 데이터에서 변형(transformation)에 의해 증강된 샘플들이다. 따라서, 증강된 샘플은 기존 데이터에서 이동된 샘플(shifted sample from an original sample)로 볼 수 있고 이를 일종의 OOD 샘플로 간주하여 학습에 사용한다. 저자들은 이를 con-SI 손실 함수(contrasting shifted instance loss)라고 정의하였다. 이외에도, 자기 지도 학습 방법은 기존의 지도 학습과는 달리 분류기가 없는 구조이다. 저자들은 모델이 주어진 입력 데이터에 대해 어느 변형이 적용되었는지 구분하는 분류기를 추가하여 손실 함수를 구성하였는데 이를 cls-SI 손실 함수(classifying shifted instance loss)로 정의하였다. 해당 손실 함수는 모델이 이동된 샘플에 대해 보다 잘 구별할 수 있도록 학습을 시키는 역할을 한다. 따라서, 최종 손실 함수는 con-SI와 cls-SI를 밸런스 하이퍼파라미터(balancing hyper-parameter)를 사용하여 구성되며, 이를 모델 학습에 사용하였다. 저자들은 OOD 탐지를 위해 제안한 손실 함수들을 기반으로 코사인 유사도(cosine similarity) 스코어를 제시하였다.

SSD [31]는 레이블이 주어지지 않은 상황에서의 OOD 샘플 탐지에 초점을 두고 자기 지도 학습을 이용한 OOD 탐지 방법을 제시하였다. 저자들은 모델이 데이터의 특성 표현을 잘 학습하게 하는 방법과 학습된 특성을 이용하는 것이 레이블이 주어지지 않은 상황에서 OOD 탐지 방법의 핵심이라고 언급하며 이를 위해, 자기 지도 학습과 마할라노비스 거리를 활용한 OOD 탐지 방법을 제안한 것이다. 게다가, 저자들은 레이블과 OOD 데이터를 활용할 수 있는 2가지 경우에 대해서도 확장할 수 있음을 보여주었고, OOD 데이터를 활용할 수 있는 경우를 퓨-샷(few-shot) OOD

탐지 문제에 대해서 정의하였다.

kNN [32]은 OOD 샘플들이 ID 샘플로부터 멀리 떨어져 있을 것이라는 가정 하에, 기존의 OOD 탐지 방법들은 특성 공간(feature space)에서의 분포 가정(distributional assumption) 기반으로 이루어진다는 것을 지적하며, non-parametric 방법 중 하나인 k-최근접 이웃(k-NN; k-Nearest Neighbor) 방법을 이용하여 분포 가정이 없는 OOD 샘플 탐지 방법을 제시한 논문이다. 저자들은 특성 공간에서 k-최근접 이웃 방법으로 OOD 샘플을 탐지하기 위해 모델이 학습 데이터의 특성 표현을 잘 학습하는 것이 핵심이라고 언급하였고, 자기 지도 학습 방법으로 학습된 모델에 유클리디안 거리(Euclidean distance) 기반의 k-최근접 이웃 방법을 적용하였을 때, 분포 가정 없이 OOD 샘플을 효과적으로 탐지할 수 있다는 것을 실험적으로 입증하였다. 따라서, 저자들은 특성 표현 학습과 적절한 거리를 이용하는 것이 OOD 샘플 탐지를 위한 중요한 요소라고 결론지었다.

제 3 장 방법

3.1 OOD 탐지 문제 정의

$f(\cdot)$를 임의의 뉴럴 네트워크, $\{(x_i, y_i)\}_{i=1,\cdots,N} \in X^{ID}$는 학습 데이터셋이라고 할 때, y_i는 i번째 데이터의 레이블로 $y_i \in \{1, \cdots, C\}$이며 C는 학습 데이터셋의 총 레이블 개수이다. 뉴럴 네트워크 $f(\cdot)$는 해당 학습 데이터셋으로만 학습이 되었다고 가정한다.

테스트 입력 샘플 **x**이 탐지기(detector) $G(\cdot)$의 입력으로 들어왔을 때, 해당 입력 샘플이 다음과 같은 정의에 의해서 ID(in-distribution) 샘플인지 OOD(out-of-distribution) 샘플인지 탐지기는 판단하게 된다.

$$G(\mathbf{x}; \lambda, f) = \begin{cases} in, & \text{if } S(\mathbf{x}; f) > \lambda \\ out, & \text{if } S(\mathbf{x}; f) \leq \lambda \end{cases} \tag{3.1}$$

λ는 임계치(threshold)로, 입력 샘플의 스코어 $S(\mathbf{x}; f)$가 임계치보다 크면 ID 샘플로, 낮으면 OOD 샘플로 판단한다.

3.2 교차주의집중 트랜스포머(Cross Attention Transformer) [1]

그림 3.1은 교차주의집중 트랜스포머(cross attention transformer)의 구조를 나타내는 그림이다. 교차주의집중 트랜스포머 [1]는 교차주의집중 트랜스포머 블록(cross attention transformer block)이 12개의 층(layer)으로 구성된다. 교차주의집중 트랜스포머의 입력으로 학습 데이터셋의 각 클래스를 대표하는 특성 시퀀스 평균(class-wise feature sequence mean)과 이미지 특성 시퀀스(image feature sequence)가 들어오게 되고, 교차주의집중 트랜스포머는 특성 표현 학습을 위한 인코더(encoder)로 사용된다. 교차주의집중을 이용하여 모델은 학습 데이터셋의 클래스 대표와 관련된 특징을 학습 데이터로부터 추출하여 학습한다. 교차주의집중 트랜스포머의

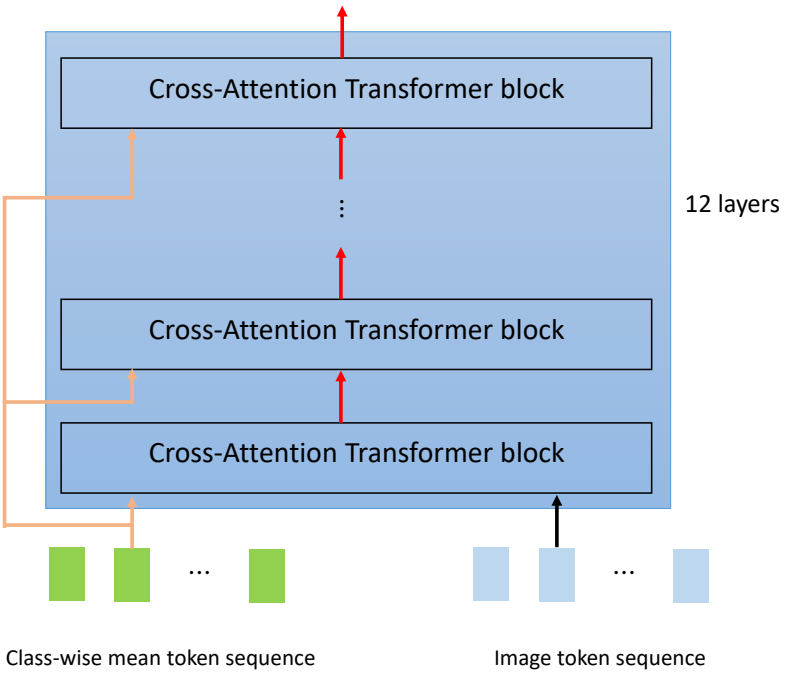

그림 3.1: **교차주의집중 트랜스포머.** 교차주의집중 트랜스포머의 구조를 나타내는 그림으로 교차주의집중 블록이 12개의 층으로 구성되어 있다.

핵심은 학습 데이터셋의 각 클래스마다 대표 특성 시퀀스를 추출(extraction)한 후에 이미지 특성 시퀀스 간에 교차 주의 집중을 하는 것이다. 이때, 각 클래스의 대표 특성 시퀀스는 비전 트랜스포머로 추출한 각 클래스의 모든 이미지 특성에 대해 평균을 취한 시퀀스이다. 학습 데이터셋의 클래스 대표 특성 시퀀스와 이미지 특성 시퀀스를 추출하기 위해 비전 트랜스포머 B_16(ViT_B_16) [7]을 특성 추출기(feature extractor)로 사용하며, 학습 데이터셋에 맞게 미세 조정(fine-tuning)을 하였다.

기존의 트랜스포머 [24]의 인코더-디코더(encoder-decoder) 주의 집중처럼 한쪽 시퀀스로부터 쿼리(query)를 그리고 다른 시퀀스로부터는 키(key)와 밸류(value)를 추출하여 교차주의집중을 하였다. 그림 3.6은 교차 주의집중의 블록을 나타내는 그

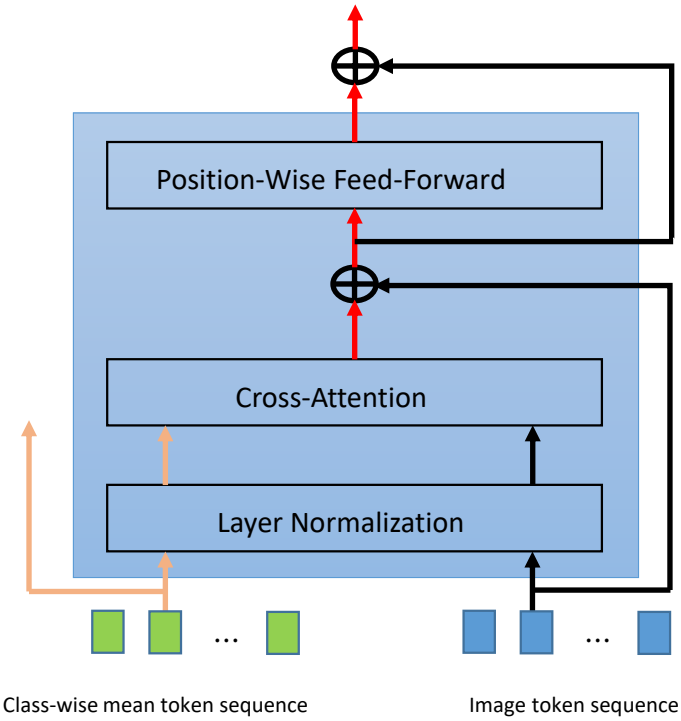

그림 3.2: **교차주의집중 트랜스포머 블록(cross attention transformer block).** 교차주의집중 트랜스포머에서 1개 블록을 나타낸 그림이다. 매 블록마다 클래스별 평균 토큰 시퀀스가 입력으로 들어간다.

림이다. 교차 주의 집중 모델의 첫 번째 블록에 특성 시퀀스 평균과 이미지 특성 시퀀스가 입력으로 들어가게 되고, 그 이후 블록부터는 이전 블록의 출력과 클래스 특성 시퀀스 평균이 입력으로 들어가게 된다.

3.3 교차주의집중 트랜스포머 기반의 대조 표현 학습

본 논문에서 교차주의집중 트랜스포머 기반 대조표현학습(CAT-based CoReL) 방법을 제안한다. 제안하는 방법은 특성 표현 학습의 대표적인 방법인 supervised contrastive learning [29] 방법을 사용하여 교차주의집중 트랜스포머 [1]를 사전학습

시킨 후, 학습된 모델에 분류기(classifier)를 추가적으로 학습하는 2단계 학습(2-stage learning)하는 방법이다. 따라서, 본 논문에서는 제안한 방법을 통해 OOD 샘플을 탐지하는 것을 목표로 한다. 3.3.1절에서 supervised contrastive learning을 적용한 교차주의집중 트랜스포머 학습 방법에 대해 설명하고, 3.3.2절에서는 사전학습된 교차주의집중 트랜스포머에 분류기(classifier)를 추가적으로 학습하는 2단계 학습(2-stage learning) 방법을 설명한 뒤, 3.3.3절에서 OOD 탐지를 위해 사용한 스코어 함수(score function)에 대해 설명한다.

3.3.1 교차주의집중 트랜스포머 기반 대조표현학습

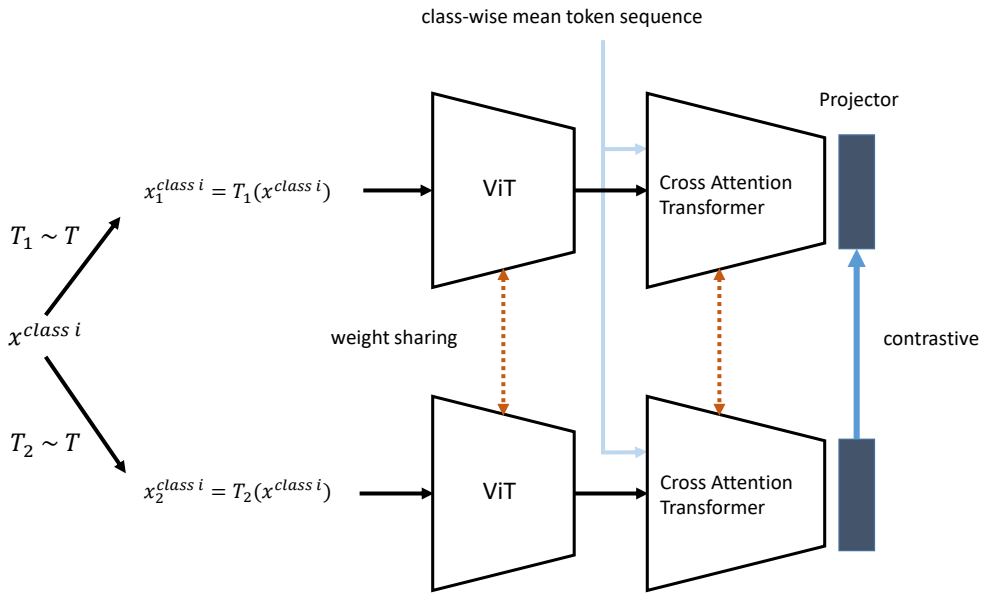

그림 3.3: **CAT-based CoReL의 전체적인 개요.** 본 논문에서 제안하는 CAT-based CoReL 방법을 나타낸 그림이다. T는 데이터 증강(data augmentation) 할 때 데이터에 가하는 변형들(transformations)의 집합이고, T_1과 T_2는 집합 중 하나이다.

그림 3.3은 교차주의집중 트랜스포머에 supervised contrastive learning을 적용

하여 학습시키는 방법에 대한 개요이다. $x^{class\ i}$는 class i에 속하는 학습 샘플이고, $x_1^{class\ i}$와 $x_2^{class\ i}$는 $x^{class\ i}$로부터 각각 T_1과 T_2에 의해 변형되어 증강된 샘플들(augmented samples)이다. Supervised contrastive learning을 위해 데이터 증강은 필수적인데, 그 이유는 모델에 데이터의 다중 관점(multi view)을 제공하기 때문이다. 다시 말해, 특정 데이터로부터 변형되어 증강된 샘플들은 해당 데이터의 다양한 특성 정보를 가지고 있다고 볼 수 있고, 모델이 데이터 증강된 샘플로 데이터의 풍부한 특성 표현을 학습할 수 있게 된다.

$x^{class\ i}$로부터 증강된 샘플들이 비전 트랜스포머를 거쳐 이미지 특성 시퀀스가 추출되고, 클래스별 대표 특성 시퀀스(class-wise mean token sequence)와 함께 교차주의집중 트랜스포머의 입력으로 들어간다. 이때, 앞서 교차주의집중 트랜스포머에서 설명한 것처럼, 비전 트랜스포머는 특성 추출기의 역할을 하기에 학습 데이터셋으로 미세 조정을 한 후 특성 추출을 위해 가중치를 고정시켜 학습시키지 않고, 교차주의집중 트랜스포머만 학습을 진행한다. 최종적으로 교차 주의 집중 모델의 출력이 프로젝터(projector)를 거쳐 임베딩 공간(embedding space)로 임베딩 된 후에, 같은 레이블(label)의 샘플끼리는 가까워지게 다른 레이블의 샘플과는 멀어지도록 학습이 된다. 이때, 프로젝터의 출력은 정규화(normalize)되어서 나온다.

손실 함수는 [29]에 따라, 식(3.2)과 같다.

$$\mathcal{L}_{supcon} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a / \tau)}. \tag{3.2}$$

식(3.2)에서 τ는 온도(temperature) 하이퍼파라미터이고, $\cdot$ 은 내적을 의미하며, z_i는 증강된 입력 샘플이 비전 트랜스포머와 교차 주의 집중 모델 그리고 프로젝터를 거쳐서 나온 정규화된 출력 벡터(normalized output vector)이다. i는 $I \equiv 1, \cdots, 2N$ 중 하나의 인덱스이며, I는 기존의 배치 크기 N개에서 데이터 증강에 의해 $2N$개로 증강된 배치(augmented batch)이다. $A(i) \equiv I \setminus \{i\}$는 증강된 배치 I에서 i번째 인덱스를 제외한 집합이다. $P(i) \equiv \{p \in A(i) : \tilde{y_p} = \tilde{y_i}\}$는 증강된 배치 안에서 i번째 인덱스와 같은 레이블을 가지는 모든 positive 샘플들의 인덱스들의 집합이다.

식(3.2)을 통해 z_i와 z_p의 내적 연산은 두 벡터간의 코사인 유사도(cosine similarity)를 의미하고 같은 레이블의 샘플끼리의 유사도(similarity)를 구하는 것이다. 따라서, 손실 함수를 최소화하는 방향으로 학습이 진행이 되는 것은 같은 레이블의 샘플간의 유사도를 최대화하는 방향으로 학습한다는 의미이다.

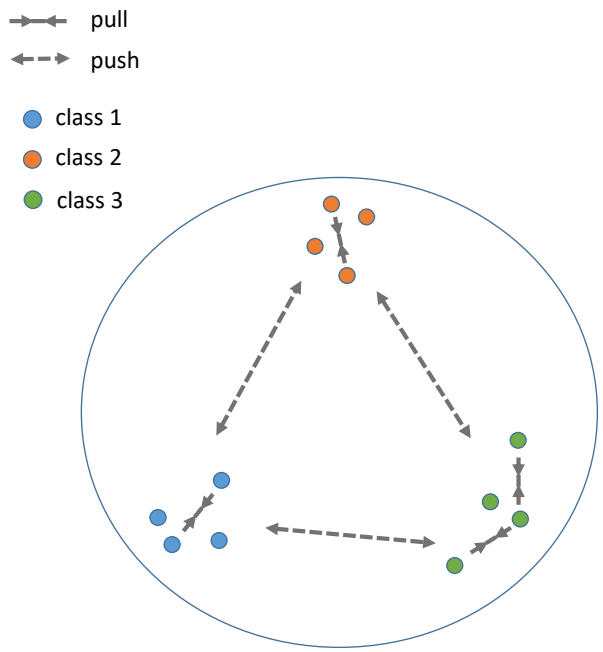

그림 3.4: **교차주의집중 트랜스포머의 임베딩 공간.**

그림 3.4는 임베딩 공간에서 샘플들이 CAT-based CoReL 과정을 통해 임베딩되는 과정을 나타낸 그림이다. 식(3.2)에서 손실 함수는 학습 데이터의 레이블을 반영하기 때문에, 임베딩 공간에서 같은 클래스의 샘플끼리는 가깝게, 다른 클래스의 샘플과는 멀게 임베딩되도록 학습이 진행된다. 따라서, CAT-based CoReL 방법을 통해 모델이 학습 데이터 특성 시퀀스와 클래스별 대표 특성 시퀀스간의 교차 주의집중을 한다는 것은 곧, 모델이 증강된 학습 데이터 특성 시퀀스와 클래스별 대표 특성 시퀀스간의 교차 주의 집중을 통해 학습 데이터셋의 클래스 대표와 관련된 특성 표현들을 학습한다는 것이다.

3.3.2 2단계 학습 방법

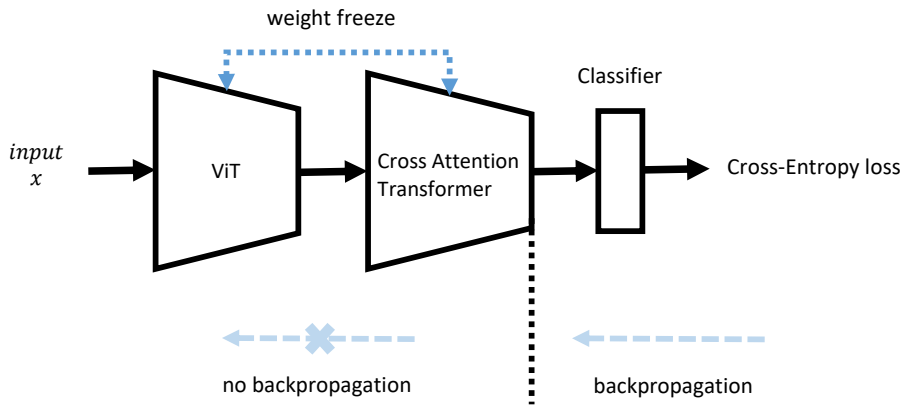

그림 3.5: **CAT-based CoReL 방법의 2단계 학습 과정 개요.**

그림 3.5은 인코더를 사전학습시킨 후 분류기를 학습시키는 2단계 학습 과정을 나타낸 개요이다. [29] 저자들이 제안한 학습 방법에 따라, 분류기를 학습하기 위해 학습된 비전 트랜스포머와 교차 주의 집중 모델의 모든 가중치들을 고정한(weight freeze) 후, 이미지 분류를 위해 분류기를 학습시킨다. 이때, [29] 저자들은 인코더를 학습시킬 때 분류기도 동시에 학습할 수 있다고 언급하였으나, 본 연구의 실험 환경에서는 인코더와 분류기를 동시에 학습을 진행하였을 때, 각각 학습을 진행한 것보다 성능이 낮으며 분류기 학습이 불안정한 것을 확인하였다. 따라서, 본 논문에서는 인코더와 분류기를 각각 학습을 진행하였다.

분류기 학습을 위한 손실 함수로 교차 엔트로피(cross entropy) 손실 함수를 사용하여 학습을 진행하였다.

3.3.3 OOD 탐지를 위한 스코어 함수

본 논문에서 OOD 탐지를 위한 스코어 함수로 에너지 스코어(energy score) [15]를 사용하였고, 식(2.11)과 같다. 교차주의집중 트랜스포머 구조상 이미지 특성 시퀀스와 클래스 대표 특성 시퀀스가 입력으로 들어간다. 따라서, OOD 샘플에 대한

스코어를 구할 때 하나의 이미지 특성 시퀀스와 ID 데이터셋의 모든 클래스 대표 특성 시퀀스를 입력으로 들어가고, 그 중 가장 높은 스코어를 OOD 샘플 스코어로 사용하는데, 이는 OOD 샘플 탐지 문제에서 특정 임계치(threshold)보다 크면 ID 샘플로, 낮으면 OOD 샘플로 탐지하는 기존의 정의에 따라 정한 것이다. 예를 들어, CIFAR-10 [9]이 ID 데이터셋이고 OOD 데이터셋이 SVHN [33]이라고 하면, 임의의 OOD 샘플 하나의 이미지 특성 시퀀스와 10개의 ID 데이터셋 클래스 대표 특성 시퀀스가 입력으로 들어간다. 해당 OOD 샘플에 대해 각 클래스 대표 특성 시퀀스와 대응하는 10개의 스코어가 나오게 되고, 그 중 가장 큰 스코어를 해당 OOD 샘플의 스코어로 결정한다.

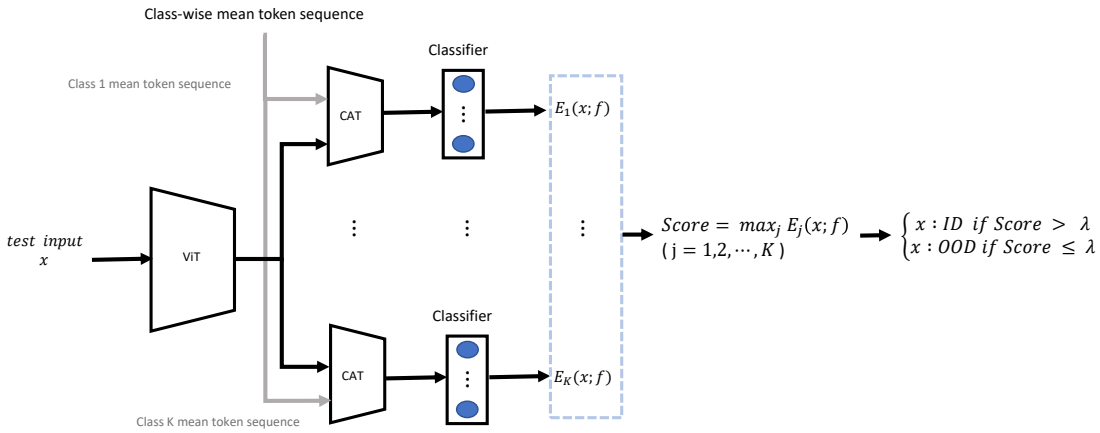

그림 3.6: **CAT-based CoReL의 OOD 샘플 탐지 방법.** 제안하는 방법을 통해 OOD 샘플을 탐지하는 과정을 나타낸 그림이다.

제 4 장 실험

4.1 실험 세부 사항

본 절에서는 OOD 탐지를 위해 사용한 데이터셋과 평가 지표(evaluation metric) 그리고 학습 세부 사항에 대해 설명한다. 본 논문에서는 기존의 OOD(out-of-distribution) 탐지를 위해 사용한 모든 데이터셋들로 제안한 방법의 성능에 대해 평가하였다.

4.1.1 실험 데이터셋

본 논문에서 ID 데이터셋으로 CIFAR-10 데이터셋 [9]을 사용하여 모델을 학습시킨 후 OOD 샘플들을 탐지하는 기존의 CIFAR [9] 데이터셋 기반 OOD 탐지 성능 평가를 위한 OOD 데이터셋과 ImageNet [8] 데이터셋 기반 OOD 탐지 성능 평가를 위한 OOD 데이터셋 모두에 대해 제안한 방법의 성능 평가 실험을 진행하였다.

따라서, 본 논문에서 far OOD 탐지를 위한 테스트 데이터셋으로 SVHN [33], LSUN(resize/crop) [12], TinyImageNet(resize/crop) [12], Texture [34], 그리고 Places365 [35] 총 7개의 OOD 테스트 데이터셋을 사용하였으며, near OOD 탐지를 위한 데이터셋으로 CIFAR-100 데이터셋을 사용하였다. 따라서, 총 8개의 테스트 데이터셋으로 제안한 방법의 성능을 평가하였다. LSUN(resize/crop)과 TinyImageNet(resize/crop)은 ODIN [12] 저자들이 LSUN [36] 그리고 TinyImageNet [37] 데이터셋에 대해 각각 랜덤 크롭(random crop)과 크기 조정(resize)을 통해 이미지의 크기를 32×32로 구성하여 만든 데이터셋이다. 실험 결과 표에 crop과 resize 데이터셋을 표기할 때 (C), (R)로 표기하도록 한다. 예를 들어, TinyImageNet(resize) 경우, ImageNet(R)로 표기한다.

4.1.2 평가 지표(Evaluation Metric)

OOD 탐지 성능을 확인하기 위해 FPR 95, AUROC, AUPR 3가지 평가 지표를 사용한다. FPR 95는 TPR(True Positive Rate)이 95% 일 때의 FPR(False Positive Rate)을 나타내는 지표이다. FPR은 $FPR = \frac{FP}{FP+TN}$ 같이 구해지며, FP는 false positive, TN은 true negative이다. 해당 지표는 모델이 전체 negative 샘플들 중 positive 샘플로 잘못 분류한(i.e., OOD 샘플을 ID 샘플로 잘못 분류한 경우) FP의 비율을 나타내는 지표로서 모델이 OOD 샘플을 OOD가 아닌 ID로 잘못 판단하는 척도를 나타내는 것을 의미하며, 해당 지표값이 작을수록 모델이 OOD 샘플을 잘 탐지하는 것을 의미한다.

AUROC는 Area Under a Receiver Operating Characteristic curve의 줄임말로 TPR(세로축)과 FPR(가로축) 대한 ROC 곡선의 면적을 의미한다. 좋은 OOD 샘플 탐지기일수록 TPR이 크고 FPR이 작게되어 곡선의 면적이 1에 가까워지게 되고, 안 좋은 탐지기일 경우 면적이 0.5가 된다. 모델이 테스트 샘플에 대해 얼마나 잘 예측하는지를 설명하는 지표이며 해당 값이 클수록 좋은 분류기임을 의미한다.

AUPR은 Area Under the Precision-Recall curve의 줄임말로 precision(세로축)과 recall(가로축)에 대한 곡선의 면적이다. $precision = \frac{TP}{TP+FP}$은 모델이 positive라고 예측한 샘플들 중에서 실제 positive인 샘플들에 대한 비율이며, $recall = \frac{TP}{TP+FN}$은 실제 positive 샘플들 중에서 모델이 예측한 positive 샘플의 비율로 AUPR은 모델의 positive 샘플에 대해 얼마나 잘 예측하는지에 대한 척도이다. 본 논문에서 ID 샘플을 positive 샘플로 간주한 AUPR(in)을 평가 지표로 사용하였고, 해당 지표값이 클수록 모델의 성능이 좋다는 것을 의미한다.

4.1.3 학습 세부 사항(Training Details)

본 논문에서 제안한 CAT-based CoReL 방법으로 교차주의집중 트랜스포머를 학습시키기 위해 비전 트랜스포머 ViT_B/16 [7]을 특성 추출기로 활용하여 교차주

의집중 트랜스포머의 입력으로 들어갈 학습 데이터셋의 클래스별 대표 특성 시퀀스(class-wise mean token sequence)와 이미지 특성 시퀀스를 추출한다. 비전 트랜스포머는 ImageNet [8] 데이터셋으로 사전 학습되어 있어, ID 데이터셋인 CIFAR-10 [9]으로 미세 조정(fine-tuning)을 한 이후에 특성 추출기로서 사용하였다. 이때, 학습률(learning rate)을 0.001, 배치 크기는 10, SGD(Stochastic Gradient Descent) 옵티마이저와 모멘텀(momentum)을 0.9로 설정하였고, 교차 엔트로피(cross entropy) 손실 함수를 이용하여 비전 트랜스포머 ViT_B/16을 3 에폭(epoch)으로 미세 조정하였다.

[7]에 따라, 입력 이미지 크기를 224×224로 조정한 뒤에 비전 트랜스포머 입력으로 들어가고, 이미지 패치 크기를 16×16으로 설정하여 입력 시퀀스의 길이는 클래스 토큰 1개를 포함한 197($\frac{224}{16} \times \frac{224}{16} = 196$)이 되며, 특성 차원(feature dimension)은 768이 된다. 따라서, 비전 트랜스포머의 출력 차원은 (배치 크기, 197, 768)이 되고, 해당 출력 토큰 시퀀스는 클래스별 대표 특성 시퀀스와 함께 교차주의집중 트랜스포머의 입력으로 들어가게 된다. 이때, 클래스별 대표 특성 시퀀스의 차원은 (학습 데이터셋의 클래스 개수, 197, 768)이 된다.

교차주의집중 트랜스포머를 학습시키기 위한 하이퍼파라미터들(hyper-parameters)은 SupCon [29]에 기반하여, 초기 학습률을 0.1로 코사인 어닐링 스케줄러(cosine annealing scheduler)를 사용하여 학습률을 조정하고, 가중치 감쇠(weight decay)는 0.0001, 그리고 SGD 옵티마이저와 모멘텀 0.9를 기본으로 설정하였다. 학습 파라미터 설정 이외에 추가로, 프로젝터의 임베딩 차원(embedding dimension)은 128차원, 배치 크기는 64, 온도(temperature) 파라미터는 0.1, 그리고 학습 에폭은 500으로 기본값으로 설정하여 학습을 진행하였다.

교차주의집중 트랜스포머를 사전훈련 시킨 뒤, 분류기를 학습시키기 위한 학습 파라미터는 다음과 같이 설정하였다. 학습 에폭은 10, 학습률 0.005, 가중치 감쇠는 0.005로 학습을 진행하였으며, 초기 학습률에서 5번째 에폭에서 0.5만큼 학습률 감소(learning rate decay)를 적용하였다. 옵티마이저는 교차주의집중 트랜스포머 학습에 사용된 값을 똑같이 사용하여 학습을 진행하였다.

	Energy [15]			OE [14]			OECC(MD) [23]			ReAct [17]			GradNorm [19]			CAT-based CoReL(ours)		
	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑
SVHN	35.59	90.96	97.64	4.8	98.4	89.4	-	99.2	98.4	49.77	92.18	93.67	17.76	96.66	-	2.74	99.49	98.33
LSUN(R)	27.58	94.24	98.67	5.59	98.94	99.79	-	99.8	99.5	17.94	96.98	97.56	-	-	-	6.82	98.6	98.46
LSUN(C)	8.26	98.35	99.66	2.89	99.49	99.9	-	-	-	16.99	97.11	97.48	0.23	99.87	-	2.18	99.55	99.53
ImageNet (R)	-	-	-	-	-	-	-	99.6	99.4	-	-	-	-	-	-	13.34	96.77	96.01
ImageNet (C)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	8.41	98.22	98.01
Places365	40.14	89.89	97.3	17.3	96.2	87.3	-	-	-	43.97	91.33	91.66	57.85	85.2	-	2.61	99.52	99.47
Texture	52.79	85.22	95.41	12.2	97.7	91	-	-	-	47.96	91.55	95.4	37.71	90.76	-	0.2	99.95	99.97

표 4.1: CIFAR-10(ID) 컨벌루션신경망 기반 OOD 탐지 방법들과의 성능비교 결과표.

	SSD+ [31]			SSD_k+ (k=5) [31]			CSI(ensemble) [30]			knn+ [32]			CAT-based CoReL(ours)		
	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑
SVHN	0.2	99.9	99.9	1.9	99.6	99.8	-	97.9	-	0.34	99.87	-	2.74	99.49	98.33
LSUN(R)	-	-	-	-	-	-	-	97.7	-	-	-	-	6.82	98.6	98.46
LSUN(C)	-	-	-	-	-	-	-	-	-	1.36	99.51	-	2.18	99.55	99.53
ImageNet (R)	-	-	-	-	-	-	-	97.6	-	-	-	-	13.34	96.77	96.01
ImageNet (C)	-	-	-	-	-	-	-	-	-	-	-	-	8.41	98.22	98.01
Places365	-	-	-	-	-	-	-	-	-	26.9	94.45	-	2.61	99.52	99.47
Texture	7.7	98.5	97.3	3.6	99.2	98.9	-	-	-	9.56	98.21	-	0.2	99.95	99.97

표 4.2: CIFAR-10(ID) 대조표현학습기반 OOD 탐지 방법들과의 성능비교 결과표.

4.2 실험 결과(Main Results)

본 논문에서 제안한 CAT-based CoReL을 이용한 OOD 탐지 성능 결과를 각각 ID 데이터셋이 CIFAR-10 일 때 여러 컨벌루션 신경 계층망 기반 OOD 탐지 방법, 대조표현학습 기반 OOD 탐지 방법, 그리고 비전 트랜스포머 기반 OOD 탐지 방법들과 비교하였다. 표 4.1는 CIFAR-10 [9]이 ID 데이터셋인 경우에 대해 제안한 방법과 컨벌루션 신경 계층망 기반 OOD 탐지 방법들의 성능을 비교한 결과이다. 표 4.2(ID 데이터셋이 CIFAR-10)는 대조표현학습 기반의 OOD 탐지 방법들과 비교한 결과이며 표 4.3(ID 데이터셋이 CIFAR-10)는 비전 트랜스포머 기반의 OOD 탐지 방법들과 비

	OODformer(MD) [5]			OODformer(MSP) [5]			Exploring(MD) [10]			Exploring(MSP) [10]			CAT-based CoReL(ours)		
	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑
SVHN	-	99.5	-	-	-	-	1.9	99.58	99.82	4.03	98.77	99.82	2.74	99.49	98.33
LSUN(R)	-	99.2	-	-	97.6	-	-	-	-	-	-	-	6.82	98.6	98.46
ImageNet (R)	-	98.8	-	-	96	-	-	-	-	-	-	-	13.34	96.77	96.01
Places365	-	-	-	-	-	-	4.54	98.51	99.95	10.13	97.14	50.92	2.61	99.52	99.47
Texture	-	-	-	-	-	-	0.05	99.97	99.83	1.73	99.59	99.92	0.2	99.95	99.97

표 4.3: CIFAR-10(ID) 비전트랜스포머기반 OOD 탐지 방법들과의 성능비교 결과표.

교한 결과이다. 각 표의 결과들은 해당 논문들에서 보고한 결과들을 기입하였으며, '-'는 논문에서 해당 데이터셋에 대해 OOD 탐지 성능 평가를 보고하지 않아 결과가 없는 것을 의미한다. ↓는 낮을수록, ↑는 높을수록 성능이 우수하다는 것을 의미한다. 데이터셋에서 (C)와 (R)은 각각 crop과 resize를 의미한다.(e.g., ImageNet(R)은 TinyImageNet(resize) [12]를 LSUN(C)는 LSUN(crop) [12]을 의미한다.)

CIFAR-10이 ID 데이터셋인 경우, 7개의 데이터셋에 대해 제안한 방법이 5.18% (FPR95 평균값), 98.87% (AUROC 평균값)의 결과를 얻었다. 또한, 최고 성능 방법 [10]과 비교하였을 때, 3개의 OOD 데이터셋(SVHN [33], Places365 [35], Texture [34])에 대해 CAT-based CoReL은 2.16%(FPR95 평균값), 99.35%(AUROC 평균값)을 달성하였고 [10] 방법에 비해 -0.31%(FPR95 평균값), +0.3%(AUROC 평균값) 만큼 결과가 향상되었다.

본 논문에서 Far OOD 탐지 외에, near OOD 탐지에 대한 성능 평가도 진행하였다(표 4.4). Near OOD 탐지 문제는 ID 데이터셋과 의미론적으로 매우 유사한 클래스들을 가지고 있는 데이터셋을 탐지하는 문제이다. 대표적인 예로 CIFAR-10(ID) vs CIFAR-100(OOD) 또는 CIFAR-100(ID) vs CIFAR-10(OOD)가 있으며 기존의 OOD 탐지보다 더 어려운 문제이다. 제안한 방법을 통해 CIFAR-10(ID) vs CIFAR-100(OOD) 경우의 6.21% (FPR95), 98.56% (AUROC)를 달성하여 기존 SOTA 대비 -0.68% (FPR95), +0.04% (AUROC)의 향상을 가져와 최고 성능(state of the art)을 달성하였다. 이외에도 [38]와 [10]에 따르면, near OOD 탐지 문제 다음으로 CI-

	CIFAR-100(OOD)		
	FPR95↓	AUROC↑	AUPR↑
OE [14]	28	93.3	76.2
OECC(MSP) [23]	23.8	94.9	82
OpenHybrid [20]	-	95.1	-
SSD+ [31]	38.5	93.4	92.3
SSD_k+(k=5) [31]	34.6	94	92.9
CSI-ensemble [30]	-	92.2	-
knn+ [32]	37.98	-	-
Exploring(MD) [10]	6.89	98.52	98.7
Exploring(MSP) [10]	9.76	97.79	97.73
CAT-based CoReL(ours)	6.21	98.56	98.33

표 4.4: CIFAR-10(ID) vs CIFAR-100(OOD) near OOD 탐지 성능 비교 결과 표.

FAR vs Places365 탐지 문제가 어렵다고 언급하였으며 [10] 저자들은 CIFAR-10 vs Places365에서 FPR95 4.54%, AUROC 98.51%를 달성하였으나, CAT-based CoReL을 통해 2.61% (FPR95), 99.52% (AUROC)을 달성하여 -1.93% (FPR95), +1.01% (AUROC) 만큼 결과가 향상되었다.

따라서, 학습 데이터와 클래스간의 상관관계를 인코딩하는 교차주의집중 트랜스포머 기반 대조표현학습 방법이 OOD 샘플 탐지에 효과적임을 실험을 통해 입증하였다.

4.3 어블레이션 연구(Ablation Study)

본 절에서는 제안한 CAT-based CoReL의 인코더(encoder) 교차주의집중 트랜스포머 학습 에폭, 분류기 학습 에폭에 대한 OOD 탐지 성능 비교 실험을 진행하였다. 모든 어블레이션 실험들은 CIFAR-10을 ID 데이터셋에서 진행하였고 임베딩 차원 128, 온도(temperature) 0.1, 배치 크기 64, 학습 에폭 500을 기본값으로 설정한 상

황에서 각 하이퍼파라미터 값들을 변경하면서 실험을 진행하였다. 예를 들어, 학습 에폭에 대한 OOD 탐지 성능을 평가한다고 하면 학습 에폭을 제외한 모든 파라미터 값들을 기본값으로 고정한 뒤 학습 에폭을 변경하면서 실험을 진행하였다.

4.3.1 분류기 학습 에폭(training epoch)에 따른 OOD 탐지 성능

표 4.5은 분류기의 학습 에폭에 따른 OOD 탐지 성능을 비교한 결과이다. 분류기를 초기 학습률 0.005, 가중치 감쇠(weight decay) 0.005로 50 에폭까지 학습을 시킨 뒤 각 10 에폭마다 에너지 스코어 [15]를 이용하여 OOD 탐지 성능을 평가하였다. 분류기 학습 에폭이 증가함에 따라 분류 정확도(classification accuracy)가 증가하지만, OOD 탐지 성능이 하락되는 경향이 있다. 분류기의 정확도와 OOD 탐지 성능간에 trade-off가 있다는 것을 실험적으로 발견하였으며, 학습 시간 및 OOD 탐지 성능 등을 함께 고려하였을 때 분류기 학습을 위한 에폭을 10으로 설정하여 실험을 진행한 후 제안한 방법의 성능을 평가하였다.

분류기 학습 에폭	10 epoch		20 epoch		30 epoch		40 epoch		50 epoch	
Test acc.	98.86		98.93		98.94		98.94		98.96	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
CIFAR-100	6.21	98.56	6.29	98.51	6.3	98.49	6.31	98.48	6.31	98.48
SVHN	2.74	99.49	2.73	99.49	2.73	99.49	2.75	99.48	2.74	99.48
LSUN(C)	2.18	99.55	2.18	99.55	2.19	99.54	2.18	99.55	2.18	99.55
LSUN(R)	6.82	98.6	6.83	98.6	6.82	98.6	6.83	98.59	6.82	98.6
ImageNet(C)	8.41	98.22	8.41	98.22	8.43	98.2	8.42	98.21	8.42	98.21
ImageNet(R)	13.34	96.77	14	96.52	14.01	96.51	14.01	96.51	14.03	96.5
Places365	2.61	99.52	2.62	99.52	2.63	99.51	2.62	99.51	2.61	99.52
Texture	0.2	99.95	0.28	99.86	0.28	99.86	0.28	99.86	0.28	99.87
average	5.31	98.83	5.42	98.78	5.42	98.78	5.43	98.77	5.42	98.77

표 4.5: 분류기 학습 에폭에 따른 OOD 탐지 성능 비교 결과.

4.3.2 트랜스포머 학습에폭(training epoch)에 따른 OOD 탐지 성능

표 4.6는 교차주의집중 트랜스포머의 학습 에폭에 따른 OOD 탐지 성능을 비교한 결과이다. 해당 표의 결과들은 모두 에너지 스코어 [15]를 이용하여 OOD 탐지를 수행한 결과이다. 결과에서 보는 것처럼 학습 에폭이 증가할수록 OOD 탐지 성능이 향상된다. 특히, 300 에폭부터 OOD 탐지 성능이 크게 향상이 되는 것을 확인할 수 있는데, 이를 통해 인코더가 입력 샘플들의 특성표현을 잘 학습하기 위해서 충분히 긴 학습 에폭을 필요하다는 것을 확인할 수 있다. SupCon [29]에 따르면 ImageNet [8] 데이터셋을 학습시키기 위해 1000 에폭으로 학습을 진행하였고, knn [32]과 SSD [31]에서도 CIFAR [9] 데이터셋을 500 에폭으로 학습을 시켰다. 따라서, OOD 탐지를 위한 대조표현학습 방법 [32], [31]에 따라 CAT-based CoReL 방법도 500 에폭으로 학습을 진행하였고, 제안된 방법이 비전트랜스포머 기반 OOD 탐지의 최고 성능인 [10]을 넘어서는 결과를 보여준다.

Training epoch	100 epoch		300 epoch		400 epoch		500 epoch	
Test Acc.	99.74		99.36		99.06		98.86	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
CIFAR-100	25.09	94.38	6.65	98.3	6.45	98.45	6.21	98.56
SVHN	11.07	96.91	3.33	99.31	3.04	99.41	2.74	99.49
LSUN(C)	11.96	96.51	2.48	99.42	2.47	99.43	2.18	99.55
LSUN(R)	19.23	94.85	7.96	98.24	7.36	98.3	6.82	98.6
ImageNet(C)	22.78	93.34	9.16	97.82	8.92	97.95	8.41	98.22
ImageNet(R)	26.73	91.05	14.39	96.18	14.04	96.3	13.34	96.77
Places365	31.39	94.36	2.34	99.51	2.67	99.3	2.61	99.52
Texture	47.15	90.38	0.67	99.79	0.23	99.82	0.2	99.95
average	24.43	93.97	5.87	98.57	5.65	98.62	5.31	98.83

표 4.6: 교차주의집중 트랜스포머 학습 에폭에 따른 OOD 탐지 성능 비교 결과.

제 5 장 결론

본 논문에서 OOD 샘플들을 학습에 사용할 수 없는 상황에서의 OOD 샘플을 탐지하기 위해 학습 데이터셋의 클래스와 각 이미지간의 상관관계를 인코딩하는 교차 주의집중 트랜스포머(cross attention transformer)에 대조표현학습(contrastive representation learning) 방법을 적용하여 보다 더 강하게 상관관계를 학습하는 CAT-based CoReL 방법을 제안하였다. CIFAR-10(ID) vs SVHN, Places365, Texture(OOD)인 far OOD 탐지 문제에서 최고 성능(SOTA)인 [10] 방법과 비교했을 때 -0.31%(FPR95 평균값), +0.3%(AUROC 평균값) 만큼 향상된 1.85%(FPR95 평균값)와 99.65%(AUROC 평균값)을 달성하였다. Far OOD 탐지 외에, CIFAR-10(ID) vs CIFAR-100(OOD)인 near OOD 탐지에서도 -0.68%(FPR95 값), +0.04% (AUROC 값) 많큰 향상된 6.21% (FPR95 값), 98.56% (AUROC 값)를 달성하였다.

참고 문헌

[1] Seokho Cho and Jin Young Choi. Within-class cross-attention transformer for out of distribution pattern detection. 2022.

[2] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[3] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 41–50, 2019.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[5] Rajat Koner, Poulami Sinhamahapatra, Karsten Roscher, Stephan Günnemann, and Volker Tresp. Oodformer: Out-of-distribution detection transformer. *arXiv preprint arXiv:2107.08976*, 2021.

[6] Chongzhi Zhang, Mingyuan Zhang, Shanghang Zhang, Daisheng Jin, Qiang Zhou, Zhongang Cai, Haiyu Zhao, Shuai Yi, Xianglong Liu, and Ziwei Liu. Delving deep into the generalization of vision transformers under distribution shifts. *arXiv preprint arXiv:2106.07617*, 2021.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg

Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[9] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[10] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34, 2021.

[11] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[12] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

[13] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

[14] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *Proceedings of the International Conference on Learning Representations*, 2019.

[15] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33:21464–21475, 2020.

[16] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and Fujie Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

[17] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34:144–157, 2021.

[18] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[19] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689, 2021.

[20] Hongjie Zhang, Ang Li, Jie Guo, and Yanwen Guo. Hybrid models for open set recognition. In *European Conference on Computer Vision*, pages 102–117. Springer, 2020.

[21] Abhijit Bendale and Terrance Boult. Towards open world recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1893–1902, 2015.

[22] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1563–1572, 2016.

[23] Aristotelis-Angelos Papadopoulos, Mohammad Reza Rajati, Nazim Shaikh, and Jiamian Wang. Outlier exposure with confidence control for out-of-distribution detection. *Neurocomputing*, 441:138–150, 2021.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[25] Hezheng Lin, Xing Cheng, Xiangyu Wu, Fan Yang, Dong Shen, Zhongyuan Wang, Qing Song, and Wei Yuan. Cat: Cross attention in vision transformer. *arXiv preprint arXiv:2106.05786*, 2021.

[26] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 357–366, 2021.

[27] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[28] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

[29] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

[30] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in neural information processing systems*, 33:11839–11852, 2020.

[31] Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. *arXiv preprint arXiv:2103.12051*, 2021.

[32] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. *arXiv preprint arXiv:2204.06507*, 2022.

[33] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[34] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.

[35] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

[36] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[37] Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*, 2017.

[38] Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam,

Simon Kohl, et al. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020.

ABSTRACT

We focus on a method for detecting OOD samples in case where OOD samples are not available and detection performance improvement. Using the capability that a vision transformer captures global information better than CNN-based models, we propose a transformer-based learning method, called CAT-based CoReL, to learn the feature representation of a training dataset. In CAT-based CoReL, we apply contrastive representation learning using the cross-attention transformer. Specifically, the cross attention transformer that encodes the correlation between the class and each image learns the correlation more strongly by learning the characteristic representation of the training data through the contrastive representation learning.

To evaluate the OOD detection performance of the proposed method, we use 8 OOD datasets adopted in the existing OOD detection. We achieve competitive performance for OOD detection. In the case of far OOD detection where the ID dataset is CIFAR-10, we achieve 1.85% (FPR95 average value) and 99.65% (AUROC average value) which are improved by 0.31% (average FPR95) and 0.3% (average AUROC) comparing to the previous state-of-the-art method. In addition, for near OOD detection, we achieve FPR95 of 6.21% and AUROC of 98.56% impreved by 0.68% (FPR95) and 0.04% (AUROC) in the case of CIFAR-10(ID) vs CIFAR-100(OOD), which is a difficult task than far OOD detection. These results are obtained by using only the ID dataset without using the auxiliary dataset.

keywords: Out-of-Distribution detection, Vision Transformer, Cross-attention

student number: 2020-21108