Vision Transformer와 CNN의 데이터 효율성과 강인성 비교 연구

CIFAR-10 및 CIFAR-10-C 기반 실험 보고서

Micro-Degree	인공지능 전문가 양성과정
지도교수	백 우 진 교수님
학 과	소프트웨어학과
학 번	202020827
이 름	김경민

목차 (Table of Contents)

1. 서론

이번 Microdegree 과정의 마지막 프로젝트로, 저는 Computer Vision 분야에서 많이 쓰이는 두 계열의 모델인 CNN과 Vision Transformer(ViT)를 직접 비교해 보고 싶었다. 학부 4학년에서 완전히 새로운 알고리즘을 제안하기는 어렵기 때문에, 수업 시간에 배운 내용을 최대한 활용할 수 있는 "비교·분석형" 연구를 목표로 잡았다. 실제 환경에서 들어오는 이미지는 카메라 센서 노이즈, 모션 블러, 조명 변화, 안개·비와 같은 날씨 효과 때문에 깨끗하지 않은 경우가 많다. 이런 상황에서 모델이 깨끗한(Clean) 데이터에서만 잘 동작하는지, 아니면 손상된(Corrupted) 데이터에서도 어느 정도 성능을 유지하는지가 실질적인 관점에서는 더 중요하다. Hendrycks와 Dietterich가 제안한 CIFAR-10-C, ImageNet-C 같은 벤치마크도 이런 관점에서 등장했다. 한편, 최근 Vision Transformer(ViT)는 "이미지를 패치로 쪼개 Transformer에 그대로 넣는다"는 매우 단순한 아이디어로, 충분한 사전학습 데이터가 주어졌을 때 기존 CNN과 비슷하거나 그 이상 성능을 낸다는 것이 알려져 있다. 하지만 ViT와 CNN이 데이터가 부족할 때 어떤 차이를 보이는지, 그리고 노이즈나 블러가 심한 이미지에서 누가 더 강인한지에 대해서, 학부 과정에서 직접 실험해 보고 싶었다. 그렇기에 이번 보고서에서는 다음 두 축을 묶어서 하나의 주제로 잡았다.

1. CNN vs ViT의 성능 차이를 데이터 효율성(data efficiency) 관점에서 비교해 보기

2. Clean vs Corrupted(CIFAR-10-C) 환경에서 두 모델의 강인성(robustness)을 비교해 보기

제가 이번에 실제로 구현하고 실험한 흐름은 "CIFAR-10에서 기본 성능 확인 → 학습 데이터 양을 줄였을 때 성능 변화 관찰 → CIFAR-10-C를 이용해 다양한 corruption과 severity에 대한 강인성 평가" 순서로 진행되며, 이 과정에서 나온 결과를 바탕으로 ViT의 장점과 한계를 정리하고자 한다.

2. 관련 연구 정리

먼저 CNN은 이미 잘 알려진 것처럼, 이미지에서 지역적인 패턴을 추출하는 convolution 연산과 pooling을 이용해 translation equivariance 및 지역 불변성(local invariance)을 가지는 구조이다. ResNet 계열은 skip connection을 도입해서 깊은 네트워크도 안정적으로 학습할 수 있게 만들었고, CIFAR-10이나 ImageNet 분류 문제에서 사실상의 표준 백본으로 쓰이고 있다. Vision Transformer(ViT)는 자연어 처리에서 쓰이던 Transformer 구조를 거의 그대로 이미지에 적용한 모델이다. Dosovitskiy 등은 이미지를 16×16 패치로 나누고, 각 패치를 토큰처럼 다뤄서 Transformer Encoder에 넣는 방식을 제안했다. 이 논문에 따르면 ViT는 대규모 데이터(JFT-300M 등)를 이용해 사전학습을 한 뒤, CIFAR-100, ImageNet 같은 다운스트림 데이터셋에 전이학습을 하면 CNN과 비슷하거나 더 좋은 성능을 낸다. 다만, 이렇게 하려면 수억 장의 이미지와 거대한 컴퓨팅 자원이 필요하다는 문제가 있다.

이 한계를 줄이기 위해 Touvron 등의 DeiT(Data-efficient image Transformers)는 ImageNet 하나만으로도 ViT를 안정적으로 학습시키는 전략과, CNN 교사 모델로부터 지식 증류를 하는 방법을 제안했다. 이를 통해 "데이터 효율적인 ViT 학습"이 가능하다는 것이 보여졌다.

한편, ViT와 CNN의 robustness 비교에 대한 연구도 진행되어 왔다. CIFAR-10-C, ImageNet-C 같은 데이터셋은 원래의 테스트 이미지에 노이즈, 블러, 날씨, 디지털 왜곡 등 19가지 정도의 common corruption을 severity 1~5 단계로 적용해, 모델이 얼마나 성능을 잃는지 측정하는 용도로 사용된다. 최근 논문들을 보면, ViT의 아키텍처 설계(예: patch embedding, convolutional FFN)와 학습 시 사용하는 augmentation 전략에 따라 common corruption에 대한 강인성이 크게 달라진다는 결과가 보고되고 있다. 이번 보고서는 이런 거대 규모 연구를 재현하는 수준까지는 아니지만, 학부생 입장에서 실제로 CNN과 ViT를 구현해서 CIFAR-10 / CIFAR-10-C에서 직접 실험해 보고, 논문들에서 이야기하는 "데이터 효율성"과 "robustness"의 감을 체감하는 것을 목표로 했다.

3. 실험 환경 및 설정

3.1 데이터셋

실험에는 기본적으로 CIFAR-10과 CIFAR-10-C 두 가지를 사용했다.

항목	CIFAR-10	CIFAR-10-C

이미지 크기	32×32 RGB 컬러 이미지	CIFAR-10 test 이미지 기반 동일 크기
클래스 수	10 개 (비행기, 자동차, 개, 고양이 등)	CIFAR-10 과 동일
데이터 구성	Train 50,000 장 / Test 10,000 장	Corruption 이 적용된 Test 세트
Corruption 종류	없음	19 종 (noise, blur, weather, digital 등)
Severity 단계	없음	각 corruption 마다 5 단계 severity
목적	기본 이미지 분류 학습/평가	모델의 강건성 평가(Robustness Test)

실제 구현에서는 `data/` 하위에 CIFAR-10을 내려받고, CIFAR-10-C tar 파일을 따로 받아서 지정된 디렉터리에 풀어놓은 뒤, 공통 모듈(`cifar_common.py`)에서 이 경로를 읽어 data loader를 구성하도록 했다. 학습 및 평가 코드는 세 개의 Jupyter 노트북으로 나누었고, `run_all.sh` 스크립트로 한 번에 순차 실행되도록 만들었다.

3.2 모델 구조

모델은 CNN과 ViT를 각각 한 종류씩 선택해 비교했다.

항목	CNN: ResNet-18	ViT: vit_tiny_patch16_224
모델 소스	torchvision.models.resnet18(pretrained=False)	timm.create_model('vit_tiny_patch16_224', pretrained=True, num_classes=10)
Pretrained	사용 안 함	ImageNet-1k pretrained 사용
최종 레이어 수정	fc → 10 클래스	num_classes=10 지정
입력 해상도	32×32	224×224
Train 변환	RandomCrop(32, padding=4)RandomHorizontalFlip()ToTensor()정규화(CIFAR-10 mean/std)	RandomResizedCrop(224)RandomHorizontalFlip()정규화(ImageNet mean/std)
Test 변환	동일(32×32 기반, ToTensor + 정규화)	Resize(224)CenterCrop(224)정규화(ImageNet mean/std)
Optimizer	SGDlr=0.01momentum=0.9weight_decay=5e-4	AdamWlr=3e-4weight_decay=0.05
Batch size	128	64
학습 방식	From-scratch	Fine-tuning

모든 실험은 NVIDIA RTX 4090 GPU를 사용해 진행했으며, PyTorch, timm, torchvision을 포함한 Python 환경을 conda로 고정하여 재실행이 가능하도록 했다.

3.3 실험 시나리오

실험은 크게 세 단계로 구성했다.

3.3.1. Clean Baseline (CIFAR-10)

항목	내용
목적	epoch 수 변화에 따른 기본 성능 비교
epoch 설정	5, 10, 20
데이터	CIFAR-10 clean train/test
학습 방식	CNN / ViT 각각 독립 학습
기록 방식	각 epoch 설정에서 최고 test accuracy 저장

3.3.2. 데이터 효율성 실험

항목	내용
목적	학습 데이터 양 감소 시 성능 저하 정도 비교
train data ratio	1.0, 0.5, 0.2, 0.1
데이터 샘플링	train set 에서 비율만큼 랜덤 샘플링
epoch 수	20 epoch 고정
비교 방식	각 비율에서 CNN/ViT 의 best test accuracy 비교

3.3.3. Robustness 실험 (CIFAR-10-C)

항목	내용
목적	다양한 corruption·severity 상황에서 모델 강인성 평가
학습 데이터	CIFAR-10 clean train (40 epoch, ratio=1.0 모델 사용)
평가 데이터	CIFAR-10-C (19 종 corruption × severity 1~5)
측정 지표	각 corruption 의 test accuracy
분석 방식	• corruption-wise 비교 (gaussian, shot, brightness, fog, motion_blur 등) • severity-wise 평균 정확도(1~5)

Clean 데이터로 학습한 모델이 각종 corruption에 얼마나 버티는지 평가한다.

학습에는 CIFAR-10 clean train set만 사용 (40 epoch, ratio=1.0 기준 모델 사용)

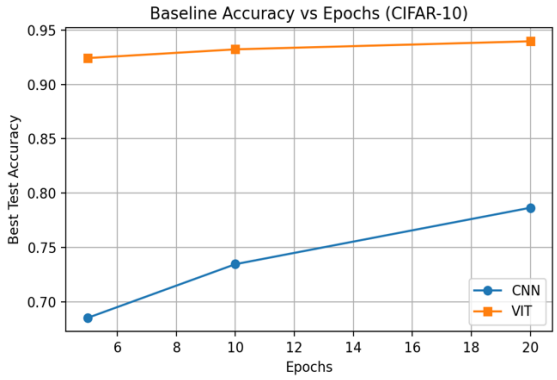
CIFAR-10-C의 여러 corruption과 severity=1~5에 대해 test accuracy 측정

corruption-wise(예: gaussian_noise, shot_noise, brightness, fog, motion_blur)와 severity-wise(1~5 평균) 결과를 정리해 그래프와 표로 분석

4. 실험 결과

4.1 Clean Baseline 결과 -> 40 epoch 비교 수정 필요

먼저 CIFAR-10 clean 데이터에서 epoch 수를 달리하며 CNN과 ViT를 학습했을 때의 결과는 다음과 같다. (정확도는 test set 기준 best 값이다.)



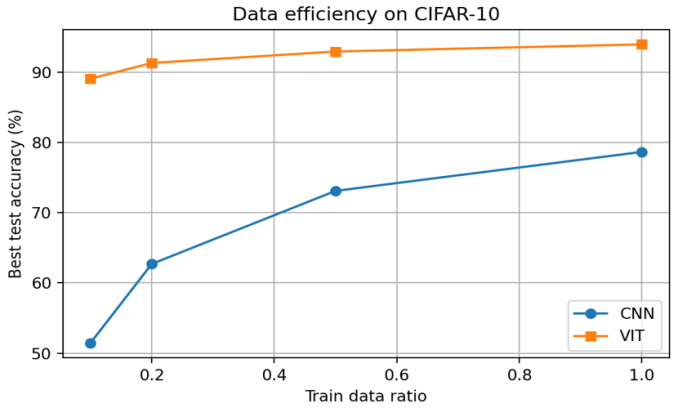
Epochs	CNN Accuracy	ViT Accuracy
5	0.6853 (약 68.5%)	0.9242 (약 92.4%)
10	0.7346 (약 73.5%)	0.9322 (약 93.2%)
20	0.7865 (약 78.7%)	0.9396 (약 94.0%)

epoch이 늘어날수록 두 모델 모두 성능이 증가하긴 하지만, ViT는 5 epoch 수준에서 이미 92%대 정확도를 찍고 이후에는 소폭만 올라가는 반면, CNN은 5→20 epoch까지 비교적 꾸준히 올라가도 78%대에 머무른다. 즉, 같은 학습을 줬을 때 ViT가 CNN보다 약 15~20%p 정도 높은 정확도를 보였고, 수렴 속도도 훨씬 빠르게 나타났다. 이는 "대규모 ImageNet 사전학습 → 소규모 데이터셋 전이학습" 세팅에서 ViT가 강력하다는 기존 보고와도 일관된 결과이다.

4.2 데이터 효율성 결과

다음으로, 학습에 사용하는 데이터 양을 줄였을 때 두 모델의 성능이 어떻게 변하는지 살펴보았다. 여기서는 epoch 수를 20으로 고정하고, train data ratio만 조절하였다.

실험 로그 상에서 정리된 best test accuracy는 다음과 같다.



ratio	CNN	ViT
1.0	0.7865	0.9396
0.5	0.7310	0.9294
0.2	0.6270	0.9131
0.1	0.5144	0.8907

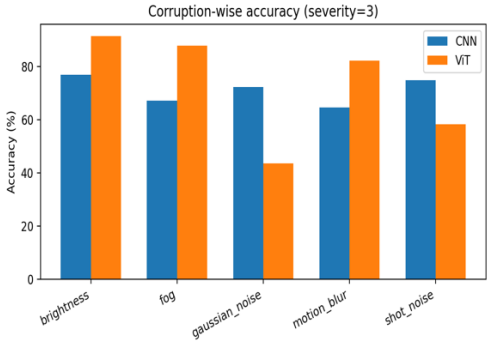
데이터 양을 줄여갈수록 CNN의 성능은 꽤 가파르게 떨어지는 반면, ViT는 10% 데이터만 사용해도 약 89%의 정확도를 유지했다. 1.0→0.1로 줄였을 때를 비교하면, CNN은 78.6%→51.4%로 27%p 정도 떨어지는데 비해, ViT는 93.9%→89.1%로 약 5%p밖에 줄지 않는다.

이 결과만 놓고 보면, 데이터 효율성 측면에서 ViT가 CNN보다 훨씬 유리해 보인다. 다만 여기에는 중요한 전제가 하나 있는데, ViT는 ImageNet으로 이미 사전학습된 가중치를 사용한 데 반해, CNN은 CIFAR-10에서 처음부터 학습했다는 점이다. 결국 이 실험에서의 차이는 "아키텍처 자체의 차이"뿐 아니라 "사전학습 여부"가 크게 섞여 있다고 보는 것이 타당하다. 그래도 실제 프로젝트 관점에서는, 라벨링할 수 있는 데이터가 제한적일 때, 사전학습된 ViT를 가져와 전이학습하는 전략이 매우 강력하다는 메시지를 체감하게 해 준다. DeiT와 같은 연구도 이런 "데이터 효율적인 ViT 학습"을 목표로 하고 있다는 점에서 방향성이 일치한다.

4.3 Robustness(CIFAR-10-C) 결과

4.3.1 Corruption-wise 분석 (severity=3 기준)

CIFAR-10-C에서 대표적인 다섯 가지 corruption(gaussian_noise, shot_noise, motion_blur, brightness, fog)에 대해 severity=3인 경우의 정확도를 뽑아 비교했다. 로그와 그래프를 기준으로 요약하면 다음과 같은 경향이 보였다.



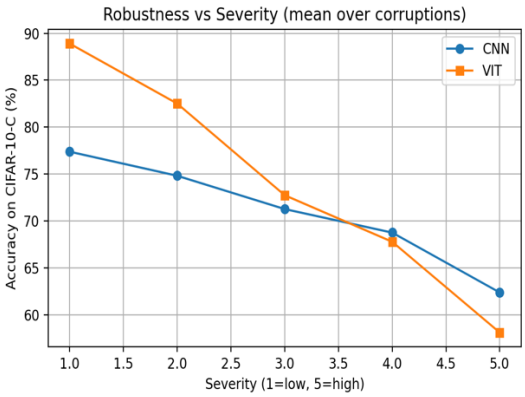
항목	CNN 정확도	ViT 정확도	비고
Gaussian noise	약 72% (0.7229)	약 40%대 중반	CNN이 ViT보다 강인
Shot noise	약 75% (0.7499)	약 50%대 후반	CNN 우세
Motion blur (severity=3)	약 64.7%	약 82%	ViT 우세 (이 경우만 반대)

brightness, fog에서도 비슷하게 ViT가 CNN보다 10~20%p 이상 높은 정확도를 보였다.

정리하면, 조명 변화나 안개, 모션 블러 같은 "부드러운 변형"에 대해서는 ViT가 CNN보다 훨씬 강하지만, Gaussian/shot noise처럼 고주파 노이즈가 섞인 경우에는 오히려 CNN이 훨씬 더 안정적이라는 패턴이 나타났다. 이는 일부 논문에서 ViT가 patch 단위 입력과 self-attention 구조 때문에 특정 노이즈에 민감할 수 있다는 분석과도 어느 정도 맞아떨어진다.

4.3.2 Severity-wise 평균 정확도

모든 corruption에 대한 정확도를 severity별로 평균 내어 비교한 그래프(robustness_severity.png)를 보면, 다음과 같은 추세를 확인할 수 있었다.



ViT의 성능이 꽤 빠르게 떨어져, 가장 심한 severity=5에서는 CNN이 오히려 ViT보다 몇 퍼센트포인트 정도 더 좋은 정확도를 보였다.

그래프를 관찰해 보면, ViT는 초기에는 CNN보다 높은 지점에서 출발하지만 기울기가 더 가파르게 내려가는 반면, CNN은 시작점은 낮지만 내려가는 속도가 상대적으로 완만한 모습을 보인다. 결국 약한 corruption에서는 ViT가 더 강하지만, 손상 강도가 높아질수록 CNN과의 격차가 줄어들고, 심한 구간에서는 CNN이 상대적으로 더 버티는 구조라고 정리할 수 있다.

5. 논의

5.1 데이터 효율성에 대한 해석

데이터 효율성 실험에서 가장 눈에 들어왔던 점은, train data ratio가 0.1(10%)밖에 안 되는 상황에서도 ViT가 89%가 넘는 정확도를 유지했다는 점이다. 반면 CNN은 같은 상황에서 50% 초반까지 떨어졌다. 이 차이는 ViT가 갖고 있는 "표현력" 자체의 차이도 있겠지만, 무엇보다도 대규모 ImageNet 사전학습이 주는 효과가 지배적이라고 생각한다. 실제로 ViT 본 논문에서도 "충분히 큰 데이터셋에서 사전학습을 하면 중간·소형 벤치마크에서 좋은 성능을 낸다"는 결과를 보여주고 있고, DeiT에서는 ImageNet 수준의 데이터와 지식 증류를 이용해 데이터 효율적인 ViT 학습이 가능하다고 보고한다.

이번 실험은 CNN에는 사전학습을 사용하지 않았기 때문에, "공정한 아키텍처 비교"라고 부르기에는 무리가 있다. 그럼에도 실제 프로젝트 상황을 생각해 보면, 오픈소스로 배포되어 있는 사전학습 ViT 모델들을 잘 활용할 수만 있다면, 라벨이 적은 데이터셋에서도 충분히 높은 성능을 빠르게 달성할 수 있다는 점을 몸소 확인했다는 데 의미가 있다고 생각한다.

5.2 Robustness 관점에서 본 CNN vs ViT

CIFAR-10-C 결과를 종합하면, CNN과 ViT는 corruption 종류에 따라 서로 보완적인 강·약점을 가진다고 볼 수 있다. ViT는 brightness, fog, motion_blur와 같이 이미지 전체에 비교적 부드러운 변형이 들어간 경우에 매우 강한 모습을 보였다. 이는 Transformer 구조가 전역적인 문맥 정보를 잘 통합하고, 패치 단위로 넓은 범위를 한 번에 보는 특성과 관련이 있어 보인다. 반대로 Gaussian/shot noise처럼 고주파 성분이 강하고 랜덤한 픽셀 변동이 많은 corruption에는 CNN이 훨씬 안정적이었다. CNN의 convolution 필터가 국소적인 패턴을 보는 구조이고, 일부 필터가 노이즈를 평균 내거나 무시하는 방향으로 학습되는 것과 연관해 볼 수 있다. severity별 곡선에서도, ViT는 초기에 높게 시작하지만 깅힌 손상 구간에서는 급격히 떨어지고, CNN은 낮은 지점에서 시작해서 조금 더 완만하게 감소했다. 실제 환경을 생각해 보면, 일반적인 카메라 환경(조명이나 날씨 변화 정도)에서는 ViT가 더 유리할 수 있지만, 센서 자체가 많이 망가져 있거나 노이즈가 심한 상황에서는 CNN 기반 모델이 여전히 의미가 있을 수 있다는 메시지를 주는 결과였다.

6. 본 연구의 한계

이번 실험에는 다음과 같은 한계가 분명히 존재한다.

첫째, 데이터셋 규모의 한계이다. CIFAR-10과 CIFAR-10-C는 32×32 해상도의 비교적 단순한 데이터셋이기 때문에, 고해상도 실세계 이미지를 다루는 ImageNet 수준의 상황을 그대로 반영한다고 보기는 어렵다.

둘째, 사전학습 여부가 서로 다르다는 점이다. ViT는 ImageNet 사전학습 가중치를 사용했고, CNN(ResNet-18)은 CIFAR-10에서 처음부터 학습했다. 따라서 이번 결과는 "사전학습된 ViT + scratch 학습 CNN"이라는 특정 조합에 대한 비교일 뿐, 같은 조건에서의 공정한 구조 비교는 아니다.

셋째, 하이퍼파라미터 탐색을 거의 하지 않았다는 점이다. 학습 epoch, learning rate, weight decay, 증강 정책 등을 grid search나 Bayesian optimization으로 튜닝하지 않고, 비교적 직관적인 값으로만 설정했다. 다른 값 조합에서는 결과가 달라질 여지가 충분히 있다.

넷째, 각 설정별로 하나의 seed만 사용해 학습했기 때문에, 랜덤 초기화나 데이터 셔플에 따른 분산을 고려하지 못했다. 좀 더 엄밀하게 하려면 여러 seed에 대해 평균과 표준편차를 보고 신뢰 구간까지 함께 제시하는 것이 좋다.

7. ViT의 한계와 개선 방안

7.1 이번 실험에서 드러난 ViT의 한계

제가 직접 실험한 범위 내에서 드러난 ViT의 한계는 크게 세 가지로 정리할 수 있었다.

1. 고주파 노이즈에 대한 취약성

Gaussian/shot noise에서 CNN에 비해 정확도가 크게 떨어졌다. 특히 severity가 올라갈수록 ViT의 성능이 급격하게 무너지는 경향이 있었고, 이는 패치 하나에 노이즈가 심하게 걸렸을 때 self-attention이 그 패치에 과도하게 주의를 줄 가능성과 연관되어 있을 것 같다.

2. 사전학습 데이터에 대한 의존성

이번 실험에서는 사전학습된 ViT를 가져다 썼기 때문에 데이터 효율성이 매우 좋게 나왔지만, 사전학습 없이 scratch로 학습했을 경우에는 오히려 CNN보다 성능이 떨어진다는 보고가 많다. 즉, "사전학습 + 전이학습"이라는 조건이 깨지면 ViT의 장점이 크게 약화될 수 있다.

3. 연산량과 메모리 사용량

32×32 CIFAR-10에서는 크게 부담이 되지 않지만, 해상도가 올라가면 self-attention의 복잡도가 $O(N^2)$이기 때문에 CNN보다 연산·메모리 비용이 빠르게 증가한다. 실제 서비스 환경에서 고해상도 이미지를 대량으로 처리해야 한다면, 여전히 CNN 또는 하이브리드 구조가 필요할 수 있다.

7.2 문헌을 참고한 개선 방향

이 한계를 줄이기 위해 기존 논문과 코드들을 살펴보면서, ViT를 더 robust하게 만들기 위한 몇 가지 아이디어를 확인할 수 있었다.

1. Patch embedding / FFN 구조 개선

일부 연구에서는 패치 간에 겹침(overlapping patch embedding)을 넣거나, FFN에 convolution 연산을 섞어서 ViT의 robusteness를 높이고자 한다. 이런 구조를 사용하면 Gaussian noise 같은 corruption에 대한 성능이 개선된다는 결과가 보고되어 있다.

2. 노이즈 중심 데이터 증강

학습 단계에서부터 CIFAR-10-C와 유사한 Gaussian/shot noise, blur 등의 augmentation을 섞어서 학습시키면, 특히 노이즈 타입 corruption에 대한 강인성이 향상된다는 결과가 있다.

3. CNN→ViT 지식 증류(DeiT 방식)

CNN 교사 모델의 출력을 ViT 학생 모델이 따라 하도록 distillation loss를 추가하면, CNN이 가진 inductive bias를 어느 정도 전수받으면서 ViT의 표현력을 유지할 수 있다. DeiT 논문에서도 이런 방식으로 “데이터 효율적인 ViT”를 구현하고 있다는 점을 확인했다.

4. Test-time adaptation / calibration

별도의 재학습 없이, 테스트 시점에 corruption-aware한 adaptation을 수행하거나, 간단한 test-time augmentation을 이용해 corruptions에 대한 성능을 올리는 방법들도 있다.

7.3 내가 해볼 수 있는 현실적인 확장 아이디어

제가 직접 추가로 시도해 볼 수 있겠다고 생각한 확장 방향은 다음과 같다.

1. 노이즈 기반 증강 실험

ViT 학습 시 Gaussian/shot noise를 섞은 augmentation을 추가하고, CIFAR-10-C 평가에서 특히 노이즈 타입에 대한 성능이 얼마나 개선되는지 확인한다.

2. 사전학습 설정 맞춘 공정 비교

ResNet-18도 ImageNet 사전학습 가중치를 불러와서 CIFAR-10에 fine-tuning 한 뒤, 현재 ViT와 같은 형태의 실험을 다시 돌려 “사전학습 + 전이학습 vs 사전학습 + 전이학습” 비교를 해 본다.

8. 향후 연구 방향

이번 프로젝트는 CIFAR-10 / CIFAR-10-C에 ResNet-18과 ViT tiny를 적용한 비교적 작은 규모의 실험에 머물렀다. 앞으로 이 연구를 조금 더 확장한다면, 다음과 같은 방향을 생각해 볼 수 있다.

1. 데이터셋을 CIFAR-100, Tiny ImageNet 등으로 넓혀, 클래스 수와 난이도가 달라졌을 때 CNN vs ViT의 데이터 효율성과 robustness 차이가 어떻게 변하는지 분석

2. CNN은 WideResNet, ConvNeXt, ViT는 DeiT, Swin Transformer 등 다른 백본들을 추가해, 아키텍처 규모와 설계 차이가 robustness에 미치는 영향을 정리

3. classification 외에 detection, segmentation 같은 다운스트림 태스크에도 corruption을 적용해, “고레벨 비전 태스크에서의 CNN vs ViT 강인성”을 비교

- 주파수 관점(frequency domain)에서 CNN은 고주파, ViT는 저주파 성분에 각각 어떤 특성을 보이는지 분석하는 연구와 연결해, 이번 결과를 이론적으로 해석해 보기

9. 결론

정리하자면, 이번 ‘인공지능 전문가 양성과정’ Microdegree Track을 수강하며 저는 CIFAR-10과 CIFAR-10-C를 이용해 ResNet-18(CNN)과 Vision Transformer(ViT)를 비교하면서, 데이터 효율성과 강인성이라는 두 가지 관점을 중심으로 실험을 진행했다.

실험 결과, 사전학습된 ViT는 적은 데이터만으로도 높은 정확도를 유지하며, 특히 brightness, fog, motion_blur 같은 부드러운 corruption에 대해서는 CNN보다 뚜렷하게 강한 모습을 보였다. 반면, Gaussian/shot noise처럼 고주파 노이즈가 많은 상황에서는 CNN이 훨씬 안정적이었고, corruption severity가 매우 높아지면 평균적으로도 CNN이 ViT를 앞서는 구간이 존재했다.

이러한 결과는 실제 응용에서 다음과 같은 시사점을 준다고 생각한다. 데이터가 상대적으로 깨끗하고, 라벨 수가 많지 않은 상황에서는 사전학습 ViT를 가져와 전이학습하는 것이 매우 좋은 선택이 될 수 있다. 하지만 센서 노이즈가 심하거나 환경이 극단적으로 나쁜 경우에는, 여전히 CNN 또는 노이즈에 특화된 구조, 혹은 CNN-ViT 하이브리드 모델이 필요할 수 있다.

이번 보고서는 거대한 SOTA 연구는 아니지만, 제가 직접 모델을 짜고, 학습 스크립트를 돌리고, 그래프를 그려 보면서 “CNN vs ViT, Clean vs Corrupted”라는 주제를 손으로 느껴봤다는 점에서 큰 의미가 있었다. 이후에는 여기서 얻은 경험을 바탕으로, ViT의 한계를 보완하는 간단한 실험들을 추가로 진행해 보고, 보다 다양한 데이터셋과 태스크로 확장해 보는 것을 개인적인 다음 목표로 삼고자 한다.

10. 코드 및 재현

이번 실험에서 사용한 전체 코드, 학습 스크립트, 모델 설정, 데이터셋 다운로드 방식, 로그 파일, 그리고 결과 그래프들은 모두 Github 저장소에 정리해 두었다. 해당 저장소를 통해 실험을 그대로 재현하거나 추가 실험을 수행할 수 있도록 구조화하였다.

Github Repository

https://github.com/Microdegree-Track/Cnn-ViT-CIFAR-10-Efficiency/tree/master