ViT 기반 모델의 강건성 연구동향

신영재[1,†], 홍윤영[1], 김호원[1,‡]
[1]부산대학교 정보융합공학과
luther11949@gmail.com, hyy0238@pusan.ac.kr, howonkim@pusan.ac.kr

A Research Trends on Robustness in ViT-based Models

Yeong-Jae Shin[1,†], Yoon-Young Hong[1], Ho-Won Kim[1,‡]
[1]Dept. of Information and Convergence Engineering, Pusan National University

요 약

컴퓨터 비전 분야에서 오랫동안 사용되었던 CNN(Convolution Neural Network)은 오분류를 일으키기 위해 악의적으로 추가된 섭동에 매우 취약하다. ViT(Vision Transformer)는 입력 이미지의 전체적인 특징을 탐색하는 어텐션 구조를 적용함으로 CNN의 국소적 특징 탐색보다 특성 픽셀에 섭동을 추가하는 적대적 공격에 강건한 특성을 보이지만 최근 어텐션 구조에 대한 강건성 분석과 다양한 공격 기법의 발달로 보안 취약성 문제가 제기되고 있다. 본 논문은 ViT가 CNN 대비 강건성을 가지는 구조적인 특징을 분석하는 연구와 어텐션 구조에 대한 최신 공격기법을 소개함으로 향후 등장할 ViT 파생 모델의 강건성을 유지하기 위해 중점적으로 다루어야 할 부분이 무엇인지 소개한다.

1. 서론

컴퓨터 비전 분야에서 CNN(Convolution Neural Network) 기반의 모델들은 예측 결과를 교란하기 위해 악의적으로 추가된 매우 작은 크기의 노이즈, 즉 섭동(perturbation)에 취약한 것으로 알려져 있다[1].

CNN 기반의 모델은 커널을 이용해 이미지의 부분적인 특성을 추출하는 것과 달리 ViT (Vision Transformer)[2] 모델은 트랜스포머(Transformer)의 어텐션(Attention) 구조를 적용하여 이미지의 전체적인 특성을 파악하기 때문에 특정 픽셀에 섭동이 추가되는 공격 기법에 비교적 강건한 특성을 가진다고 밝혀졌다.[3]

하지만 ViT의 섭동에 의한 보안 취약성 연구가 활발히 진행됨에 따라 ViT 대상으로 한 새로운 공격기법이 공개되면서, 어텐션 구조도 적대적 공격에 결코 안전하지 못하다는 것이 밝혀지고 있다.

본 논문은 ViT이 적대적 공격으로부터 강건함을 유지하기 위한 방법과 어텐션 구조에 대한 최신 공격기법에 대한 연구를 소개함으로써 모델의 강건성과 ViT 기반 모델을 대상으로 적대적 공격의 연구 방향성을 소개한다.

2. 본론

2.1 ViT 강건성 분석

Xiafeng 등[4]은 ViT 모델의 구조를 요소별로 분석하여 강건성 유지의 긍정적 요소와 부정적 요소를 밝히고 있다. ViT가 적대적 공격으로부터 강건함을 향상시키기 위한 5가지 방법은 다음과 같다.

- Convolutional stem과 같은 입력 이미지 패치의 Low-level feature 사용[5]
- 트랜스포머 블록의 공간 해상도별 단계적 구성
- 적절한 어텐션 헤드 갯수 선택
- Convolution FFN(Feed-Forward Network)
- CLS(Classification Token)를 전역 평균 풀링으로 대체[6]

CeiT[5] 모델에서 사용된 I2T(Image to Tokens)와 같은 방법은 이미지의 Low-level의 패치 임베딩을 통해 이미지의 상세한 시각 정보를 활용할 수 있고 표준 정확도에도 이점이 있음을 밝히고 있다. Low-level의 특성을 활용하는 것은 섭동에 의한 교란에 강건함을 주는 요소임을 Xiafeng 등[4]은 실험적 결과로 증명하고 있다.

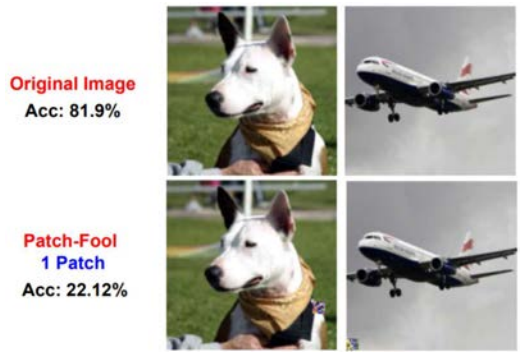
트랜스포머 블록의 공간 분해능을 점진적으로 감

소시키는 것은 다양한 크기의 특성맵이 활용될 수 있도록 하는 방법이 되고, 기존의 FFN처럼 하나의 토큰만 인코딩하는 것이 아닌 이웃된 토큰까지 인코딩하는 컨볼루션 FFN을 사용하는 방법 역시 한가지 정보만을 이용하는 것이 아닌 여러가지 특성을 동시에 활용하여 강건성과 표준 정확도에 도움을 주는 기법이다.

반대로 ViT의 강건성에 부정적인 영향을 주는 요소들도 있다. Swin[7]과 같이 지역 정보가 서로 겹치지 않게 제한하는 방법은 높은 계산 효율성과 정확도를 얻을 수 있으나 특정 지역에 집중되게 하는 방식은 강건성에 부정적이라는 것을 확인하였다. 또한, 트랜스포머 블록의 헤드의 수가 증가할수록 강건성과 표준 정확도가 크게 상승하지만 필요 이상의 헤드는 오히려 강건성과 정확도를 감소시키므로 적절한 개수를 선택하는 것이 중요하다.

2.2 ViT 모델 공격기법

Yonggan 등[8]은 ViT는 입력 이미지의 글로벌 상호 작용을 포착하는 것에 중점이기 때문에 CNN 기반 모델의 국소적 탐색보다 특정 픽셀의 섭동에 대해 강건하다는 이전의 연구 사례를 근거로 ViT의 적대적 공격의 핵심은 섭동의 밀도와 강도라고 주장하며 Patch-Fool 공격기법을 제시한다.



(그림 1) Patch-Fool

Patch-Fool 공격기법은 각 픽셀에 대한 섭동의 크기를 제한하지 않고 한 패치 내에서 섭동을 추가한다. (1)은 Patch-Fool 알고리즘을 공식화 한 것으로 E는 섭동, $1_p = \begin{cases} 0, i \neq p \\ 1, i = p \end{cases}$는 원-핫 벡터, p는 적대적 패치를 의미한다.

$$\underset{1 \le p \le n,\, E \in R^{n \times d}}{\arg\max} J(X + 1_p \odot E, y) \quad (1)$$

Patch-Fool은 각 레이어의 토큰과의 중요성을 측정하여 (1)과 같이 가장 영향력 있는 패치 p를 선택하고. (2), (3)와 같이 교차 엔트로피 손실과 계층별 어텐션 손실을 기반으로 최적화하여 섭동이 추가된다. J_{CE}는 크로스 엔트로피 손실, J_{ATTN}은 계층별 어텐션 손실을 의미한다.

$$s_j^{(l)} = \sum_{h,i} a_j^{(l,h,i)} \quad (2)$$

$$J(\widetilde{X}, y, p) = J_{CE}(\widetilde{X}, y) + \alpha \sum_l J_{ATTN}^{(l)}(\widetilde{X}, p) \quad (3)$$

Patch-Fool 기법은 ResNet[9], DeiT[10] 모델로 실험한 결과 어텐션 기반의 모델이 CNN 기반의 모델보다 적대적 공격에 대해 더 강건하지 못할 수 있음을 실험적으로 밝히고 있다.

3. 결론

적대적 공격은 다양한 도메인에 딥러닝이 사용될수록 강력한 위협이 될 수 있다. 특히 컴퓨터 비전 분야에서 ViT 파생모델이 많이 활용되는만큼 앞으로의 연구는 모델의 강건성이 고려된 구조로 설계되어야 할 것으로 보인다. 본 논문은 ViT의 강건성이 유지되는 요소와 ViT의 최신 공격기법이 어떤 방향으로 연구되고 있는지를 알아봄으로써 향후 적대적 공격에 대한 연구 방향성을 제시한다.

사사

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터육성지원사업의 연구결과로 수행되었음(IITP-2022-2020-0-01797)

참고문헌

[1] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).

[2] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

[3] Bhojanapalli, Srinadh, et al. "Understanding robustness of transformers for image classification." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[4] Mao, Xiaofeng, et al. "Towards robust vision transformer." Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition. 2022.
[5] Yuan, Kun, et al. "Incorporating convolution designs into visual transformers." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
[6] Chu, Xiangxiang, et al. "Conditional Positional Encodings for Vision Transformers", arXiv preprint arXiv:2102.10882 (2021).
[7] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
[8] Fu, Yonggan, et al. "Patch-Fool: Are Vision Transformers Always Robust Against Adversarial Perturbations?." arXiv preprint arXiv:2203.08392 (2022).
[9] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[10] Touvron, Hugo, et al. "Training data-efficient image transformers & distillation through attention." International Conference on Machine Learning. PMLR, 2021.