비전 트랜스포머 모델의 적대적 공격 취약성 연구: 완전 미세 조정과 프롬프트 튜닝 비교 고찰

강민준, 박태진, 이재구*
국민대학교

*jaekoo@kookmin.ac.kr

The Vulnerability to Adversarial Attack on Vision Transformers Trained with Full Finetuning vs Visual Prompt Tuning

Minjun Kang, Taejin Park, Jaekoo Lee*
College of Computer Science, Kookmin University

요 약

대규모 데이터로 사전 학습된 심층학습 모델을 하위 과업(Downstream Task)에 미세 조정(Finetuning)하는 방식은 뛰어난 성능을 보이며 다양한 방법들이 연구되고 있다. 컴퓨터 비전 분야에서 비전 트랜스포머(Vision Transformer)를 효과적으로 미세 조정하는 방법 중 하나인 시각적 프롬프트 조정(Visual Prompt Tuning)은 적은 매개변수만을 학습하여 완전 미세 조정(Full Finetuning)과 동등하거나 높은 과업 성능을 달성하였다. 그런데, 뛰어난 성능을 보이는 심층학습 모델들은 사람의 눈으로 감지하기 어려운 미세한 잡음을 입력에 추가함으로써 모델을 오작동시키는 적대적 예제 공격에 취약하다. 본 논문에서는 미션 크리티컬(Mission Critical) 분야 중 하나인 질병 분류 과업을 수행하기 위해 비전 트랜스포머를 완전 미세 조정 또는 시각적 프롬프트 조정으로 미세 조정한 후, FGSM(Fast Gradient Sign Method)을 이용한 적대적 공격에 대한 취약성을 살펴보았다.

Ⅰ. 서 론

대규모 데이터로 사전 학습된 모델을 하위 과업(Downstream Task)에 미세 조정(Finetuning)하는 방식은 자연어 처리 분야에서 뛰어난 성능을 보이며 표준으로 자리잡았다[1]. 특히, BERT[2], GPT-3[3] 등 자연어 처리 분야의 최첨단 모델들은 트랜스포머(Transformer)[4] 구조를 기반으로 하여 뛰어난 성능을 달성하였다. 이러한 성공에 힘입어 컴퓨터 비전 분야에서도 트랜스포머 구조를 적용한 비전 트랜스포머(Vision Transformer)[5] 모델이 다양한 하위 과업에 뛰어난 성능을 보이며 사전 학습된 백본(Backbone) 모델로 사용되고 있다.

그러나 실제로 사전 학습된 대형 모델을 하위 과업에 미세 조정하는 것은 다양한 도전 과제를 내포하고 있다. 이러한 대형 모델 전체를 완전 미세 조정(Full Finetuning)하는 것은 일반적인 방식이지만, 비용이 많이 들고 컴퓨팅 자원에 따라 제한적일 수 있다. 특히, 비전 트랜스포머는 컴퓨터 비전 분야에서 널리 사용되는 합성곱 신경망(Convolution Neural Network)에 비해 큰 구조를 가지고 있어 미세 조정에 더 많은 데이터를 필요로 한다. 이에 따라, 사전 학습된 모델을 하위 과업에 효과적으로 미세 조정하는 방법인 프롬프트(Prompt) 튜닝 방법이 대두되고 있다. 프롬프트 튜닝 방법 중 사전 학습된 비전 트랜스포머의 매개변수를 고정시키고, 적은 양의 학습 가능한 매개변수인 프롬프트와 모델의 선형 출력층(Linear Head) 만을 미세 조정하는 시각적 프롬프트 조정(Visual-Prompt Tuning)[6] 방법이 최근 제안되었다.

한편, 심층학습 모델들은 뛰어난 성능을 보이지만 미세한 잡음에도 성능이 크게 하락할 수 있다[7]. 이렇게 사람이 알아채지 못할 정도의 미세한 잡음을 입력에 추가하여 모델의 출력을 변화시키는 것을 적대적 예제 공격이라 한다. 이러한 적대적 예제 공격은 모델이 출력에 대한 높은 신뢰도(Confidence)를 유지한 채로 오작동하도록 하기 때문에 자율주행, 질병 진단과 같은 미션 크리티컬(Mission Critical) 분야에서 특히 치명적이다.

본 논문에서는 생물 의학 데이터 집합인 HAM10000[8]에 사전 학습된 비전 트랜스포머 모델을 미세 조정하는 방법에 따른 적대적 예제 공격에 대한 취약성을 탐구하였다. 제안 실험 설정에서 하위 과업 미세 조정 방법은 완전 미세 조정과 시각적 프롬프트 조정을 사용하였으며, 적대적 예제 공격은 FGSM(Fast Gradient Sign Method)[7]을 이용하였다.

Ⅱ. 본 론

본 논문에서는 사전 학습된 비전 트랜스포머를 의료 사진 데이터 집합에 완전 미세 조정과 시각적 프롬프트 조정 두 가지 방법으로 미세 조정 방법별 질병 분류 과업 성능을 비교하였다. 시각적 프롬프트 조정은 학습 가능한 매개변수인 프롬프트 p를 다양한 위치와 개수로 삽입하여 수행할 수 있다. 프롬프트 p는 비전 트랜스포머의 임베딩(Embedding) 차원과 동일한 차원을 가지며, 모델에 추가적으로 삽입하여 학습한다. 프롬프트 p를 첫 번째 층에만 삽입하는 것을 얕은(Shallow) 시각적 프롬프트 조정, 모든 층에 삽입하는 것을 깊은(Deep) 시각적 프롬프트 조정이라고 한다.

질병 분류 과업은 미션 크리티컬 분야 중 하나이며, 적대적 예제 공격이 모델의 출력을 변화시킬 경우 치명적일 수 있다. 이러한 맥락에서, 우리는 비전 트랜스포머 모델의 미세 조정 방법에 따른 취약성을 평가하기 위해 대표적 적대적 예제 공격인 FGSM을 사용하여 실험을

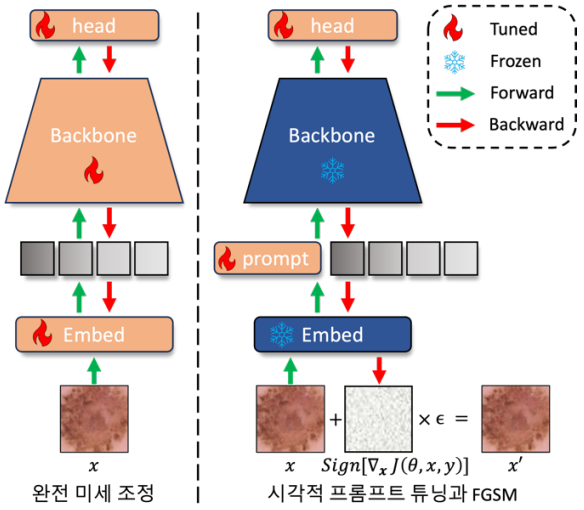

그림 1. 시각적 프롬프트 조정과 적대적 예제 생성

수행하였다. FGSM은 아래의 수식 (1)과 같이, 입력 사진 x를 모델에 순전파(Forward)하고, 그 결과와 모델 매개변수 θ 및 정답값 y를 함께 사용하여 손실 함수 $J(\theta,x,y)$로부터 x에 대한 기울기(Gradient)를 구한다. 이후, 기울기에 사인 함수 $Sign[\cdot]$을 적용하여 부호화한다. 부호화된 기울기에 잡음 강도 ϵ을 곱해 x에 더함으로써 적대적 예제 x'을 생성한다.

$$x' = x + \epsilon \cdot Sign[\nabla_x J(\theta,x,y)] \quad (1)$$

우리는 [그림 1]과 같이, 두 가지 미세 조정 방법을 각각 수행하고, FGSM을 이용하여 미세 조정 방법별 적대적 예제 공격에 대한 취약성을 분석하였다.

Ⅲ. 실 험

실험에는 피부암 진단에 활용할 수 있는 대표적 생물의학 사진 데이터 집합인 HAM10000을 사용하였다. HAM10000은 다양한 유형의 피부 병변을 현미경으로 촬영한 사진을 포함하고 7개의 클래스로 구성된다. 성능은 질병 분류 과업에 대한 모델의 정확도(Accuracy, ACC, %)를 측정하였다. 백본으로는 Imagenet[9]으로 사전 학습된 비전 트랜스포머 기본 모델을 사용하였다.

[표 1]에 나타낸 실험 결과는 3개의 임의 초깃값을 사용한 각 실험 결과의 평균값이며, 가장 높은 성능을 굵은 글씨로 표기하였다. Baseline은 백본 모델에 각 미세 조정 방법을 수행한 분류 과업 정확도를 나타내며, 적대적 예제 공격의 세기에 따른 취약성을 정확도 변화로 확인하기 위해 FGSM을 ϵ = 0.01, 0.02으로 실험하였다. FGSM은 Baseline 모델이 성공적으로 분류한 데이터에 대해 수행하였다.

실험 결과, Baseline에서 완전 미세 조정 방법보다 시각적 프롬프트 조정 방법이 더 높은 정확도를 보였다. 이는, 적은 양의 데이터로 미세 조정할 때, 시각적 프롬프트 조정이 완전 미세 조정보다 더 효과적임을 시사한다. 또한, 프롬프트 p의 개수가 증가함에 따라 모델 정확도가 향상되는 경향을 보인다. 그러나 FGSM 공격에 대해서는 완전 미세 조정이 시각적 프롬프트 조정보다 강건한 성능을 보였다. 완전 미세 조정 방법은 FGSM의 공격 잡음 ϵ이 증가함에 따라 성능이 다소 떨어지지만, 시각적 프롬프트 조정 대비 적은 하락을 보이며 높은 성능을 유지한다. 반면, 시각적 프롬프트 조정은 Baseline 성능이 완전 미세 조정 방법보다 높았음에도 불구하고 동일한 FGSM 공격에서 완전 미세 조정 방법보다 더 낮은 성능을 보인다. 이는, 시각적 프롬프트 조정이 완전 미세 조정보다 적은 데이터로도 높은 과업 성능을 달성할 수 있지만, 적대적 예제 공격에는 취약함을 나타낸다.

표 1. 미세 조정 정확도와 FGSM 공격 실험 결과

Method	완전 미세 조정	시각적 프롬프트 조정		
		p=5	p=10	p=50
Metric	ACC(%, ↑)	ACC	ACC	ACC
Baseline	67.96	78.84	**80.45**	80.01
ε = 0.01	**57.67**	19.07	21.99	29.81
ε = 0.02	**34.49**	16.03	21.88	33.22

Ⅳ. 결 론

본 논문에서는 비전 트랜스포머의 미세 조정 방법별 성능을 측정하고, 적대적 예제 공격에 대한 취약성을 분석하였다. 실험 결과, 시각적 프롬프트 조정이 완전 미세 조정 방법보다 뛰어난 성능을 달성하지만, 적대적 예제 공격에는 더 취약한 것으로 나타났다. 이는 학습 과정에 적대적 예제 공격을 고려한 방어 기법을 도입할 필요가 있음을 시사한다.

ACKNOWLEDGMENT

본 연구는 과학기술이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.RS-2022-00167194,미션 크리티컬 시스템을 위한 신뢰 가능한 인공지능)

본 연구는 2022년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음"(2022-0-00964)

참 고 문 헌

[1] Bommasani, Rishi, et al. "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258* (2021).

[2] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

[3] Floridi, Luciano, and Massimo Chiriatti. "GPT-3: Its nature, scope, limits, and consequences." *Minds and Machines* 30 (2020): 681-694.

[4] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

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

[6] Jia, Menglin, et al. "Visual prompt tuning." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022.

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

[8] Tschandl, Philipp, Cliff Rosendahl, and Harald Kittler. "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions." *Scientific data* 5.1 (2018): 1-9.

[9] Russakovsky, Olga, et al. "Imagenet large scale visual recognition challenge." *International journal of computer vision* 115 (2015): 211-252.