유사도 가중치를 적용한 FSPGD 공격 성능 개선

권세은, 박은솔, 박미소, 신용구*
고려대학교
{senikk1118, espark_82, miso419, *ygshin92}@korea.ac.kr

Enhanced FSPGD using Similarity-Weighted Mask

Se Eun Kwon, Eun Sol Park, Mi So Park, and Yong Goo Shin*
Korea University

요 약

본 논문에서는 시맨틱 세그멘테이션 공격 기법인 FSPGD에서 사용되는 이진 마스크 대신 유사도 가중치 마스크를 도입하여, 적대적 특징 맵 내에서 유사한 특성을 지닌 위치에 보다 집중적으로 교란을 가하는 방식을 제안한다. 이 방법은 특징 벡터 간 유사도에 비례하여 공격 세기를 조절함으로써 공격 효과를 향상시킨다. PASCAL VOC 2012 데이터셋을 대상으로 한 실험 결과, 제안한 기법은 기존 FSPGD 대비 최대 1.64% 낮은 mIoU를 기록하며 더 강력한 공격 성능을 입증한다. 이는 본 연구가 마스크 설계의 가중치 조정을 통해 전이 기반 공격의 효율을 향상시킬 수 있음을 보여준다.

Ⅰ. 서론

딥러닝 기술은 급속히 발전하고 있으나, 이에 수반되는 모델의 안정성과 보안성 보장에 대한 연구는 아직 충분하지 않다. 적대적 공격은 입력 데이터를 악의적으로 조작하여 딥러닝 모델이 잘못 예측하도록 유도하는 기법이다. 시맨틱 세그멘테이션 모델을 대상으로 한 적대적 공격 연구가 활발히 진행되고 있으며, 공격 대상 모델의 정보를 활용할 수 없는 블랙박스 환경에서는 전이성을 향상시키기 위한 다양한 기법들이 제안되고 있다. 대표적으로, 모델의 내부 특징 표현의 차이를 확대하거나 여러 공격 기법을 결합하여 사용하는 방식 등이 있다. FSPGD[3]는 인코더의 중간 특징 맵을 공격하는 시맨틱 세그멘테이션 공격으로 기존 공격 대비 우수한 성능을 보이지만, 벡터의 유사 정도를 고려하지 않고 공격한다는 점에서 한계가 존재한다. 본 연구에서는 FSPGD에 유사도 정보를 반영한 가중치 마스크를 도입하여 공격 효과를 강화하고, 시맨틱 세그멘테이션 모델에 대한 블랙박스 공격 성능을 향상시키고자 한다.

Ⅱ. 본론

가. FSPGD

FSPGD[3]는 반복적인 공격 과정을 통해 하나의 적대적 이미지를 생성하는 PGD 기반 기법으로, 시맨틱 세그멘테이션 모델의 중간 계층에서 추출한 특징 맵 (Feature map)을 대상으로 공격을 수행한다. 초기에는 랜덤노이즈를 추가하여 적대적 이미지를 설정하고, 이후 손실값을 증가시키는 방향으로 이미지를 점진적으로 갱신한다. 본 기법은 두 가지 손실 함수를 정의하여 공격 성능을 효과적으로 향상시키며 전체 구조는 그림 1에 제시되어 있다.

외부 손실 함수 L_{ex}은 정규화된 원본 특징 맵 f_x 과 적대적 특징 맵 f_a 간의 벡터 유사도 연산으로 정의된다. 이를 통해 특징 맵 간 유사도가 낮아지고, 모델의 내부 표현이 효과적으로 교란된다.

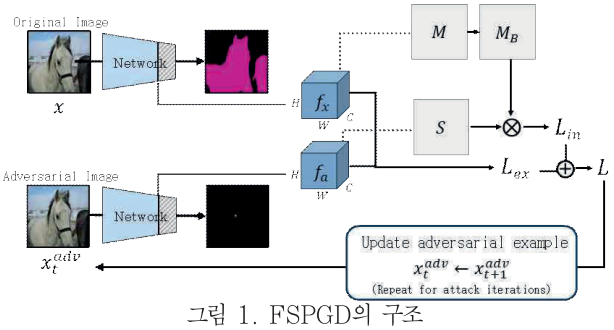

그림 1. FSPGD의 구조

$$L_{ex} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{f_x(i)}{|f_x(i)|}\right)^T \frac{f_a(i)}{|f_a(i)|} \tag{1}$$

내부 손실 함수 L_{in}은 적대적 특징 맵 내에서 유사한 벡터 간 유사도로 정의된다. 이때 유사한 벡터를 얻기 위해 원본 특징 맵을 2차원 (H×W, C)으로 재구성한 뒤, 전치 행렬과의 곱을 통해 각 위치 간 유사도를 포함하는 행렬 M을 계산한다.

$$M(p,q) = \left(\frac{f_x(p)}{|f_x(p)|}\right)^T \frac{f_x(q)}{|f_x(q)|} \tag{2}$$

이후 일정 임계값 τ 이상의 유사도를 갖는 요소를 기준으로 이진 마스크 M_B을 생성한다.

$$M_B(p,q) = \begin{cases} 1, \text{ if } M(p,q) > \tau \\ 0, \textit{ otherwise} \end{cases} \tag{3}$$

적대적 특징 맵에서도 동일한 방식으로 유사도 행렬 S를 계산하고, 해당 마스크를 적용하여 선택된 벡터 쌍 간의 유사도를 내부 손실로 산출한다. 이 손실 함수는 이미지 내의 문맥 정보를 효과적으로 교란하여 시맨틱 세그멘테이션 성능을 저하시킨다.

$$S(p,q) = \left(\frac{f_a(p)}{|f_a(p)|}\right)^T \frac{f_a(q)}{|f_a(q)|} \tag{4}$$

		Target Models			
Source Model	Attack Method	Source Model	PSP Res101	DV3 Res101	FCN VGG16
PSP Res50	PGD	7.72	54.73	59.41	45.70
	SegPGD	5.41	54.10	58.95	45.43
	CosPGD	**1.84**	56.63	64.37	45.99
	DAG	65.82	62.67	66.22	38.91
	NI	7.71	33.49	38.52	32.94
	DI	6.41	32.00	35.25	37.34
	TI	18.28	64.50	69.60	36.80
	FSPGD	3.39	22.24	16.84	19.75
	Proposed	3.38	**21.24**	**15.61**	**19.20**

		Target Models			
Source Model	Attack Method	Source Model	PSP Res101	DV3 Res101	FCN VGG16
DV3 Res101	PGD	9.75	59.36	55.54	47.48
	SegPGD	7.18	58.96	46.53	46.53
	CosPGD	**2.73**	58.83	58.64	47.25
	DAG	67.55	67.07	67.58	39.48
	NI	9.49	36.41	35.62	35.62
	DI	7.64	34.11	31.66	40.99
	TI	27.16	65.79	65.13	37.98
	FSPGD	3.28	11.42	13.45	21.49
	Proposed	3.29	**11.39**	**13.35**	**19.85**

표 1. 기존 시맨틱 세그멘테이션 모델 공격과 제안하는 기법의 성능 비교

$$L_{in} = \frac{1}{2}\frac{1}{K}\sum_{p=1}^{N}\sum_{q=1}^{N} M_B(p,q) \otimes S(p,q) \tag{5}$$

최종 손실 L은 공격이 진행됨에 따라 외부 손실의 비중은 점차 증가시키고, 내부 손실의 비중은 감소시킴으로써 효과적인 교란을 유도한다. 여기서 t는 현재 공격 횟수이고 T는 총 공격 횟수이다.

$$L = \lambda_t L_{ex} + (1-\lambda_t) L_{\text{in}}, where\, \lambda_t = \frac{t}{T} \tag{6}$$

나. 제안하는 방법

본 논문에서는 벡터 간 유사도가 높을수록 더 강한 공격을 수행할 수 있도록, FSPGD의 내부 손실 계산에 사용되는 이진 마스크에 벡터 유사도를 가중치로 부여한다. 기존 이진 마스크 방식은 임계값 τ을 기준으로 유사성 여부만을 구분하여 유사도의 세밀한 차이를 반영하지 못하는 한계가 있다. 본 연구에서는 이 한계를 극복하고자, 유사도가 높은 벡터 쌍일수록 내부 손실에 더 크게 반영되도록 설계함으로써, 적대적 이미지 생성 시 해당 영역에 집중적인 공격이 이루어지도록 한다. 이를 통해 공간적 문맥 정보를 효과적으로 교란함으로써, 기존 대비 더 우수한 전이성과 공격 성능을 달성할 수 있음을 실험을 통해 입증한다.

$$M_W(p,q) = \begin{cases} M(p,q), & \text{if } M(p,q) > \tau \\ 0, & otherwise \end{cases} \tag{7}$$

다. 실험 및 결과

본 연구에서는 시맨틱 세그멘테이션에서 사용되는 데이터셋 PASCAL VOC 2012의 테스트 이미지는 1,449장을 사용하였다. 모델 구조와 깊이에 따른 전이성을 평가하기 위해 소스 모델 PSPNet-ResNet50[2, 4], DeepLabv3-ResNet101[1, 2]을 사용하였고, 타겟 모델로 PSPNet-ResNet50, DeepLabv3-ResNet50, PSPNet-ResNet101, DeepLabv3-ResNet101, FCN-VGG16을 사용하였다. 파라미터는 FSPGD와 동일하게 최대 교란값 ϵ는 8/255, 스텝 사이즈 α는 2/255, 공격반복횟수 T는 20, 마스크 생성을 위한

(a) 원본 이미지 (b) 정답 이미지 (c) 적대적 예제 (d) 공격 후 예측 결과

그림 2. 공격 후 결과

임계값 τ는 cos(π/3)로 설정하였다. 특징 맵 추출 계층은 FSPGD 논문의 특징 맵 추출 계층별 성능 비교 결과에 따라 PSPNet-ResNet50은 layer3_2, DeepLabv3-ResNet101은 layer3_10로 하였다.

공격 성능은 시맨틱 세그멘테이션 평가 지표인 평균 교집합-합집합 비율 (mean Intersection over Union, mIoU)로 평가하였고, mIoU (%)가 낮을수록 더 강한 공격 성능을 의미한다. 제안하는 기법과 기존 공격 기법들의 성능 비교 결과는 표 1에 제시되어 있으며, FSPGD 외의 기법들에 대한 정보는 FSPGD[3] 논문에서 확인할 수 있다. PSPNet-ResNet50을 소스 모델로 사용한 경우, 기존 FSPGD 대비 평균 0.93%p, DeepLabv3-ResNet101을 소스 모델로 사용한 경우, 평균 0.59%p 낮은 mIoU를 보여 공격 성능의 향상을 입증하였다. 그림 2 는 원본 이미지, 해당 이미지의 예측 결과, 제안한 기법으로 생성된 적대적 이미지와 그 예측 결과를 제시한다.

Ⅲ. 결론

본 연구에서는 세그멘테이션 공격 기법인 FSPGD 구조를 기반으로, 기존의 이진 마스크를 유사도 기반 마스크로 대체함으로써, 공격 성능과 전이성을 향상시키는 개선된 방법을 제안하였다. 본 방식은 유사도가 높은 영역에 정교한 공격을 수행함으로써 같은 객체 내의 내부 유사도를 효과적으로 교란한다. 이는 향후 특징 맵 기반 전이 가능한 공격에 응용 가능성을 제시할 뿐만 아니라 세그멘테이션 공격의 정밀도와 효율성을 높이는데 기여한다.

ACKNOWLEDGMENT

This work was supported by IITP grant funded by MSIT (Grant#. RS-2025-02263277, 50%). This work was supported by ITRC support program supervised by the IITP and funded by MSIT (Grant#: IITP-2025-RS-2023-00258971, 50%)

참 고 문 헌

[1] L. C., Chen, Y., Zhu, G., Papandreou, F., Schroff, and H., Adam. "Encoder-decoder with atrous separable convolution for semantic image segmentation." Proceedings of the European conference on computer vision. 2018.

[2] K., He, X., Zhang, S., Ren, and J., Sun. “Deep Residual Learning for Image Recognition.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[3] E. S., Park, M. S., Park, and Y. G., Shin. “FSPGD: Rethinking Black-box Attacks on Semantic Segmentation.” arXiv preprint arXiv:2502.01262. 2025.

[4] H., Zhao, J., Shi, X., Qi, X., Wang, and J., Jia. "Pyramid scene parsing network." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.