Segmentation 기반 적대적 공격 동향 조사

홍윤영[1], 신영재[2], 최창우[1], 김호원[1]
[1]부산대학교 정보융합공학과
[2]부산대학교 사물인터넷 연구센터
hyy0238@pusan.ac.kr, yeongjae@islab.re.kr, changwoo@islab.re.kr,
howonkim@pusan.ac.kr

Research Trends of Adversarial Attacks in Image Segmentation

Yoon-Young Hong[1], Yeong-Jae Shin[2], Chang-Woo Choi[1],Ho-Won Kim[1]
[1]Dept. of Information and Convergence Engineering, Pusan National University
[2],Internet of Things Research Center, Pusan National University

요 약

컴퓨터 비전에서 딥러닝을 활용한 이미지 분할 기법은 핵심 분야 중 하나이다. 이미지 분할 기법이 다양한 도메인에 사용되면서 딥러닝 네트워크의 오작동을 일으키는 적대적 공격에 대한 방어와 강건함이 요구되고 있으며 자율주행 자동차, 질병 분석과 같이 모델의 보안 취약성이 심각한 사고를 불러올 수 있는 영역에서 적대적 공격은 많은 관심을 받고 있다. 본 논문에서는 이미지 분할 기법에 따른 구별방법과 최근 연구되고 있는 적대적 공격의 방향성을 설명하며 향후 컴퓨터 비전 분야 연구의 효율성을 위해 중점적으로 검토되고 있는 연구주제를 설명한다

1. 서론

최근 딥러닝 및 컴퓨팅 성능의 발달로 딥러닝 네트워크가 다양한 도메인에서 적용되고 있다. 특히, 컴퓨터 비전 분야에서 객체 탐지 및 이미지 분할(Image Segmentation) 등에 관한 연구가 활발히 이루어지고 있으며, 뛰어난 성능을 보여주는 모델이 많이 발표되고 있다. 하지만 C.Szegedy 등[1]이 실험을 통해 증명한 딥러닝 네트워크의 보안 취약성은 자율 주행에서 객체의 잘못된 인식이나 질병의 오분류와 같이 큰 사고로 이어질 수 있는 분야에 대해 제약사항이 되었다. 이러한 문제를 해소하기 위해 적대적 공격/방어 기법의 연구가 활발히 진행되고 있다. 본 논문은 컴퓨터 비전 분야의 핵심기술 중 하나인 이미지 분할에 대한 적대적 공격 기법의 연구 동향에 대해 알아본다.

2. 이미지 분할

이미지 분할에서는 3가지 기법이 존재한다.

이미지 의미론적 분할 (Image Semantic Segmentation) [2, 3]은 컴퓨터 비전 분야에서 가장 핵심적인 분야로, 이미지의 각 픽셀의 클래스를 개체 수와 무관하게 동일한 색상으로 칠한다.

이미지 인스턴스 분할(Image Instance Segmentation)[4, 5]은 장면에서 관심 있는 각각의 개별 객체를 감지하고 분할하는 것으로, 의미론적 분할과는 다르게 동일한 클래스 개체라도 서로 다른 인스턴스로 분할한다.

이미지 판옵틱 분할(Image Panoptic Segmentation) [6]의 목표는 픽셀 단위의 클래스 레이블과 인스턴스 ID를 모두 인코딩하여 고유한 값을 이미지의 모든 픽셀에 할당하는 것으로, 의미론적 분할과 인스턴스 분할을 조합한 작업이다.

의미론적 분할 및 인스턴스 분할은 주로 CNN 기반의 네트워크를 사용하여, 픽셀 단위로 분류를 하며[7, 8], 최근엔 Transformer[9]를 기반으로 하는 ViT 기반의 연구[10, 11]가 진행되어오고 있다. 판옵틱 분할의 경우, 최근 ViT 기반의 연구가 활발히 진행 중이다. 특히 Bowen Cheng[12] 등이 제안한 모델은 판옵틱 분할 뿐만 아니라, 의미론적 분할, 인스턴스 분할 데이터 세트에서도 SOTA(State-of-Arts)를 달성하였다.

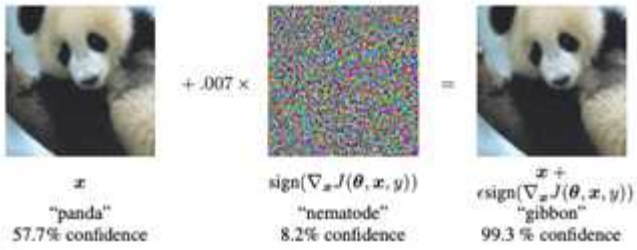
3. 적대적 공격 기법

3.1 초기 적대적 공격 기법

딥러닝 모델은 비선형적인 특성을 띤다고 알려졌지만, Goodfellow 등[13]은 LSTM, ReLU, Maxout Network 등 주로 사용되는 기술로 인해 선형성을 띠고 있다고 밝혔다. 이러한 선형적인 특성으로 인

해 딥러닝 모델은 취약성을 가지고 있음을 밝혔고 눈에 보이지 않는 섭동(perturbation)을 삽입하여, 원래 클래스를 다른 클래스로 분류하는 적대적 공격 FGSM(Fast Gradient Sign Method)을 제안하였다. FGSM 수식은 (1)과 같다.

$$\eta = \epsilon^* sign(\nabla_x(J(\theta, x, y))) \quad (1)$$



(그림 1) FGSM 실험 결과

η는 추가할 섭동, θ는 모델 파라미터, x, y는 각각 입력 이미지와 입력 이미지의 클래스, J은 목적 함수로, 손실함수의 기울기를 통해 입력 이미지의 클래스에서 멀어지는 방향으로 섭동을 생성하여 모델의 오분류를 일으킨다.

Carlini와 Wagner[14]는 적대적 예제를 방어하기 위해 제안된 Defensive distillation[15]을 공격하는 C&W Attack 기법을 제안하였다. C&W Attack은 원본 이미지를 공격자가 지정한 클래스로 오분류하도록 하는 표적 공격으로, 새로운 목적 함수를 정의하였으며 (2)와 같다.

$$\begin{aligned} \min \quad & ||\eta||_p + c \cdot g(x+\eta) \\ s.t. \quad & x+\eta \in [0,1]^n \end{aligned} \quad (2)$$

CNN 기반 모델과 이미지를 파편화시켜 각 파편 간의 관계를 학습하는 ViT(Vision Transformer)[25]의 Self-Attention 기법은 구조적인 특성이 다르므로 기존의 방법들로 적대적 공격이 어렵다고 알려져 있다. Mahmood[16] 등은 이러한 특성을 가진 ViT 기반의 분류모델을 CNN 모델과 앙상블 하여 적대적 예제를 생성하는 Self-Attention Gradient Attack을 정의하였으며 (3)과 같다.

$$\begin{aligned} x_{adv}^{(i+1)} &= x_{adv}^{(i)} + \epsilon * \text{sign}(G_{\text{blend}}(x_{\text{adv}}^{(i)})) \\ G_{blend}(x_{adv}^{(i)}) &= \sum_{k \in K} \alpha_k \frac{\partial L_k}{\partial x_{adv}^{(i)}} + \sum_{v \in V} \alpha_v \phi_v \odot \frac{\partial L_v}{\partial x_{adv}^{(i)}} \\ \phi_v &= (\Pi_{l=1}^{n_l}[\sum_{i=1}^{n_h}(0.5\, W^{(\text{att})_{l,i}} + 0.5\text{I})]) \odot \text{x} \end{aligned} \quad (3)$$

기존 FGSM의 공격 기법을 변형하여 CNN과 ViT의 기울기를 앙상블 하는 G_{blend} 함수를 사용하였다. G_{blend}는 적대적 예제에 대한 손실함수의 기울기를 앙상블 하는데, 이때 ViT의 어텐션(Attention) 정보를 얻기 위해 ϕ_v를 사용한다. ϕ_v는 각각 어텐션 헤드 i, 어텐션 레이어 l 마다, 어텐션 가중치 행렬 $W_{l,i}^{(\text{att})}$를 원본 이미지 x마다 곱하여 적대적 예제를 생성한다. 위 공격 기법을 통해 생성한 적대적 예제는 ImageNet을 학습한 VIT 기반 이미지 분류 모델에서 91%의 공격 성공률을 보여주었다.

3.2 이미지 분할 모델에 대한 적대적 공격 기법

초기의 이미지 분할 모델에서의 적대적 공격 기법 [17]은 FGSM을 변형하여 정의하였으며, 공격 공식은 (4)와 같다.

$$\begin{aligned} \xi^{(0)} &= 0 \\ \xi^{(n+1)} &= \text{Clip}_\xi\{\xi^{(n)} - \alpha \text{sign}(\nabla_x J_{cls}(f_\theta(x+\xi^{(n)}), y^{\text{target}})\} \end{aligned} \quad (4)$$

ξ는 원본 이미지에 삽입되는 섭동, $J_{cls}(\cdot)$은 목적 함수로 픽셀별 분류 결과가 나오며, f_θ은 이미지 분할 모델, α는 섭동의 양 조절하는 하이퍼 파라미터, y^{target}은 오분류 일으키도록 대상 클래스로, 기존 FGSM에서 l_∞ norm을 추가하여 y^{target}으로 오분류를 일으키는 적대적 예제를 생성한다.

공격 대상으로 오분류 일으키는 초기의 공격 기법과는 다르게, Xie[18] 등은 의미론적 분할 모델 또는 객체 탐지 모델들이 각각의 클래스를 별도로 분류한다는 점에 기인하여 모든 출력 클래스에 대하여 오분류 일으키는 DAG 알고리즘을 제안하였다.

Treu[19] 등은 옷의 질감 정보를 이용하여 이미지에서 사람이 구분하지 못하도록 하는 공격 기법인 FashionAdv를 정의하였다. FashionAdv는 사람이 있는 영역 P만을 사용하며, $w \times h$ 크기의 이미지 X 내의 사람 수에 따라 P_k로 정의한다. $M_k, \overline{M_k} \in [0,1]^{w,h}$는 공격 가능한 영역(e.g. 옷)과 공격 불가 영역(e.g. 얼굴, 머리, 손 등)을 나타내는 이진 마스크를 의미하며, 전체 이미지 X의 픽셀 단위 곱셈을 통해 attack region을 설정한다. 적대적 이미지는 그림 2와 같이 쉽게 보이지 않는 최소한의 변화와 자연스러움이 중요한데 본 논문에서는 패션 이미지 S와 적대적 이미지 $\widetilde{X}$를 합성하여 두 이미지 간의 거리를 줄이는 방법을 사용하였으며 (5)와 같이 정의하였다.

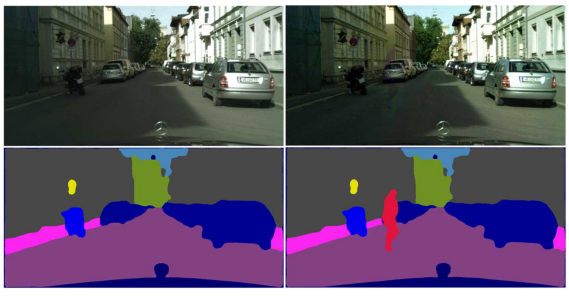
$$\text{argmin}_{\widetilde{X}} \sum_{k=1}^{K} (D_1(\widetilde{P}_k, P_k) + D_2(\widetilde{P}_k, S)) \quad (5)$$

(그림 2) FashionAdv 실험 결과

또한 강건성를 보장하기 위해 EOT 기법[22]를 사용하였으며, 출력되는 적대적 이미지가 모델의 분할 성능이 최소화 하도록 최적화하여 생성한다. EOT는 외부에서 적용될 수 있는 노이즈, 촬영 거리, 각도 등의 변수에도 강건할 수 있는 적대적 공격을 수행하는 기법이다.

Chan 등[20]은 CNN 모델 기반의 의미론적 분할 모델을 대상으로, 공격 클래스만 오분류 일으키는 적대적 공격 기법을 제안하였다. Semantically Stealthy Adversarial Attack은 먼저, i번째 픽셀이 $\hat{y_k}$ 클래스에 속하는 확률인 $p_{i,\hat{y_k},k}$는 공격 형태에 따른 하이퍼 파라미터 T_k를 곱함으로써 크로스-엔트로피 기반의 적대적 손실함수 L를 정의하였으며 마지막으로, 원본 이미지에 해당하는 손실함수 L_{reg}를 앙상블 하여 아래와 같은 공격 기법을 (6)과 같이 정의하였다.

$$
\begin{aligned}
&L(\hat{l_x},\hat{y}) = \min \sum_{i=0}^{n}\sum_{k=0}^{n} T_k(-\log(p_{i,\hat{y_k},k}) + (1-T_k)(-\log(p_{i,\hat{y_k},k})) \\
&L_{reg}(l_{reg},y) = \min \sum_{i=0}^{n}\sum_{k=0}^{m} -\log(p_{i,y_k,k}) \\
&L_{\text{total}} = L + \lambda_0 * L_{reg}
\end{aligned}
\tag{6}
$$

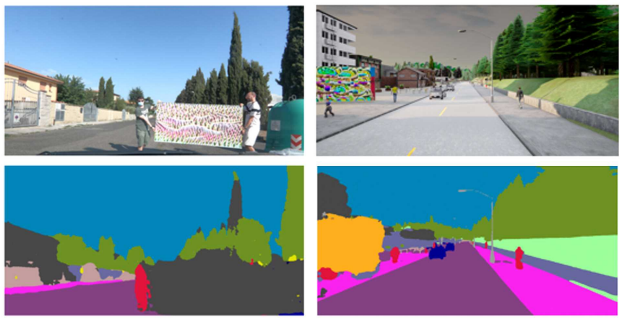


(그림 3) Stealthy 적대적 공격 결과

Nesti 등[21]은 CNN 기반의 의미론적 분할 모델에 대한 적대적 공격 기법을 현실에 적용하기 위해 가상 데이터를 기반으로 한 적대적 패치 공격 기법을 제안하였다. 공격에 사용되는 가상 데이터는 자율주행 시뮬레이터인 Carla 내에서 생성하였으며, EOT 기법을 활용하여 현실에 적용 가능한 패치를 생성하였다. Scene-specific attack은 강력한 공격을 위해, 적대적 패치를 삽입한 이미지 $\tilde{x}$에 대하여, 올바르게 분류되는 픽셀 $L_M^{\tilde{x}}$와 오분류된 픽셀 $L_{\bar{M}}^{\tilde{x}}$ 에 대하여 크로스 엔트로피 L_{CE} 손실함수를 앙상블하여 사용하였으며 (7)과 같이 정의하였다.

$$
\begin{aligned}
&L_M^{\tilde{x}} = \sum_{i \in \tau} L_{CE}(f_i(\tilde{x}), y_i), \qquad L_{\bar{M}}^{\tilde{x}} = \sum_{i \notin \tau} L_{CE}(f_i(\tilde{x}), y_i). \\
&\nabla_\delta L(f(\tilde{x}), y) = \gamma \cdot \frac{\nabla_\delta L_M^{\tilde{x}}}{\|\nabla_\delta L_M^{\tilde{x}}\|^2} + (1-\gamma) \cdot \frac{\nabla_\delta L_{\bar{M}}^{\tilde{x}}}{\|\nabla_\delta L_{\bar{M}}^{\tilde{x}}\|^2}
\end{aligned}
\tag{7}
$$

적대적 공격의 기울기 $\nabla_\delta L(f(\tilde{x}), y)$는 하이퍼 파라미터 γ에 의해 공격 방향이 설정된다. Scene specific attack은 위의 수식과 EOT 기법 병행하여 현실에서도 적용 가능한 적대적 패치를 생성한다.



(그림 4) (좌) 현실에서의 공격 실험 결과
(우) Carla 상에서의 공격 실험 결과

4. 결론

딥러닝 기술이 지속적으로 발전하고 있음에 따라 자율 주행 시스템 및 산업 현장 등 다양한 분야에 활용되고 있지만, 적대적 공격 등 AI 역기능 공격에 취약한 것으로 밝혀졌다. 이러한 AI 역기능 공격을 방어하기 위해서는 다양한 적대적 예제를 추가로 학습하는 것이 효과적인 방법이라고 알려졌다. 하지만, 본 논문에 따르면 이미지 분할의 적대적 공격은 의미론적 분할 공격과 CNN 기반의 모델에 대해서만 연구가 이뤄지고 있으며, 현재 활발히 연구되고 있는 판옵틱 분할과 ViT 기반의 모델에 대해서는 적대적 예제에 대한 연구 결과가 나오지 않고 있다. 본 논문은 향후 인공지능과 밀접한 실생활에 큰 위협이 될 수 있는 적대적 공격에 대한 방어 연구를 수행하기 위한 연구 지표가 될 것이다.

사사(Acknowledgement)

이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2019-0-01343,융합보안핵심인재양성사업)

참고문헌

[1] C. Szegedy et al., “Intriguing properties of neural networks,” arXiv preprint arXiv:1312.6199, 2013.

[2] X. He, R. S. Zemel, and M. A. Carreira-Perpinán, “Multiscale conditional random fields for image labeling,” in Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004., 2004, vol. 2, p. II - II.

[3] M. Cordts et al., “The cityscapes dataset for semantic urban scene understanding,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 3213 - 3223.

[4] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, “Simultaneous detection and segmentation,” in European conference on computer vision, 2014, pp. 297 - 312.

[5] T.-Y. Lin et al., “Microsoft coco: Common objects in context,” in European conference on computer vision, 2014, pp. 740 - 755.

[6]Kirillov, Alexander, et al. "Panoptic segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[7] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431 - 3440.

[8] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961 - 2969.

[9] A. Vaswani et al., “Attention is all you need,” Advances in neural information processing systems, vol. 30, 2017./

[10] J. Jain et al., “SeMask: Semantically Masked Transformers for Semantic Segmentation,” arXiv preprint arXiv:2112.12782, 2021.

[11] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in European conference on computer vision, 2020, pp. 213 - 229.

[12]Cheng, Bowen, et al. "Masked-attention mask transformer for universal image segmentation." arXiv preprint arXiv:2112.01527 (2021).

[13] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” arXiv preprint arXiv:1412.6572, 2014.

[14] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in 2017 ieee symposium on security and privacy (sp), 2017, pp. 39 - 57.

[15] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, “Distillation as a defense to adversarial perturbations against deep neural networks,” in 2016 IEEE symposium on security and privacy (SP), 2016, pp. 582 - 597.

[16]Mahmood, Kaleel, Rigel Mahmood, and Marten Van Dijk. "On the robustness of vision transformers to adversarial examples." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[19] Fischer, Volker, et al. "Adversarial examples for semantic image segmentation." arXiv preprint arXiv:1703.01101 (2017).

[20]Xie, Cihang, et al. "Adversarial examples for semantic segmentation and object detection." Proceedings of the IEEE international conference on computer vision. 2017.

[21]Treu, Marc, et al. "Fashion-guided adversarial attack on person segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

[22]Athalye, Anish, et al. "Synthesizing robust adversarial examples." International conference on machine learning. PMLR, 2018.

[23]Chen, Zhenhua, Chuhua Wang, and David Crandall. "Semantically Stealthy Adversarial Attacks Against Segmentation Models." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022.

[24]Nesti, Federico, et al. "Evaluating the robustness of semantic segmentation for autonomous driving against real-world adversarial patch attacks." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022.

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