딥러닝 체계의 적대적 공격에 대한 데이터 증강 기법 기반의 방어기법 연구 동향

이재현, 박휘랑, 조영호(교신저자)
국방대학교 관리대학원 국방과학학과 컴퓨터공학과/사이버전 협동전공
e-mail: dlwowogus@gmail.com, sharku7@gmail.com, youngho@kndu.ac.kr

A Survey on Data Augmentation Technique-based Defenses against Adversarial Attacks on Deep Learning Systems

Jaehyun Lee, Hweerang Park, Youngho Cho
Dept. of Defense Science (Computer Engineering and Cyberwarfare Major),
Korea National Defense University

요 약

최근 딥러닝(Deep Learning)은 컴퓨터 비전(Computer Vision), 자연어 처리(Natural Language Processing), 음성 인식(Voice Recognition) 등 다양한 분야에서 뛰어난 성능을 보이고 있다. 하지만, 적대적 공격(Adversarial Attacks)은 딥러닝을 대상으로 큰 위협이 되고 있다. 그러므로, 안전한 딥러닝 모델을 생성하기 위해서는 적대적 공격을 방어하는 기법에 대한 이해가 매우 중요하다. 데이터 증강기법(Data Augmentation Techniques)은 딥러닝 모델의 성능 향상을 위해 사용되는데, 이를 방어기법으로 활용한 적대적 공격에 대한 방어기법 연구가 활발히 이루어지고 있다. 따라서, 본 연구에서는 데이터 증강 기법을 활용한 방어 기법들의 최신 국외 연구 동향을 조사하여 정리하였다. 이를 통해, 딥러닝 모델에 데이터 증강기법을 적용하는 연구자들의 이해를 높여 안전한 딥러닝 모델을 생성하는데 도움을 주고자 한다.

1. 서론

최근 딥러닝(Deep Learning)은 컴퓨터 비전(Computer Vision), 자연어 처리(Natural Language Processing), 음성 인식(Voice Recognition) 등 다양한 분야에서 복잡한 문제를 해결하는 데 우수한 성능을 보였지만, 딥러닝 모델을 공격하는 적대적 공격(Adversarial Attacks)으로 인해 딥러닝 시스템의 성능이 크게 저하되고 신뢰성을 훼손하는 등의 보안 위협이 대두되고 있다[1-2]. 따라서, 딥러닝의 성공 만큼이나 적대적 공격에 대한 방어는 중요한 연구 분야가 되었다[3].

방어기법 중에서는 딥러닝 모델의 성능을 높이기 위해 흔히 사용되는 데이터 증강기법(Data Augmentation Techniques)을 사용할 수 있다[4]. 원래 데이터 증강기법은 훈련데이터를 다양한 방법으로 변환하여 데이터수를 늘려 모델의 성능을 높이기 위해 제안된 기술이나, 데이터 증강기법을 사용하면 훈련데이터가 다양해지므로 모델의 성능을 높이면서 공격을 완화시키는 효과가 있다[5]. 데이터 증강기법에 대한 연구들을 이해하여 적절히 딥러닝 시스템에 적용하면 적대적 공격을 효과적으로 방어할 수 있으므로, 데이터 증강기법을 활용한 최신 방어기법 연구들의 정리가 필요하다[6].

따라서, 본 논문에서는 데이터 증강을 활용한 방어 기법의 국외의 최신 연구 동향에 대해 조사하고 분류함으로써, 연구자들로 하여금 데이터 증강기법을 활용한 적대적 공격 방어에 대한 이해를 돕고, 안전하고 신뢰성 있는 딥러닝 모델 구축을 돕는 것이 목표이다.

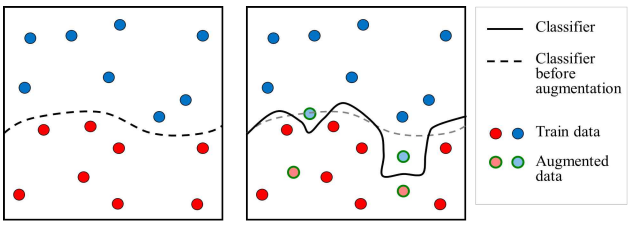
이후 논문 구성은 다음과 같다. 2장에서는 데이터 증강기법의 의미와 분류를 소개한다. 3장에서는 데이터 증강기법을 사용하여 방어에 활용되는 내용을 기술한다. 4장에서는 조사 내용을 분석하여 결론 및 향후연구를 기술한다.

2. 관련연구

2.1 딥러닝 분야에서 데이터 증강기법

딥러닝 분야에서 데이터 증강기법은 훈련 데이터셋의 크기와 품질을 향상시키는 방법을 말하며, 데이터 증강기법을 적용하면 더 성능이 좋은 딥러닝 모델을 구축할 수 있다[4]. [그림 1]은 데이터 증강기법을 적용했을 때 딥러닝 모델의 변화를 간단하게 표현한 그림이다; (a)는 증강데이터(Augmented data)가 없을 경우의 데이터 분류를 나타낸 것이며, (b)는 증강데이터(Augmented data)를 추가한 후 더욱 향상된 데이터

분류를 표현한 것이다. 데이터 증강기법을 사용하면, 분류기(Classifier)의 모호한 결정 경계(Decision boundary)가 구체화되어 증강하지 않았을 때보다 더 높은 분류성능을 갖게 된다.



[그림 1] 데이터 증강기법의 적용 전(a)과 후(b)의 비교

2.2 데이터 증강기법의 발전

Krizhevsky 등[7]은 딥러닝 모델의 성능을 높이기 위해 이미지 분류 분야에서 최초로 데이터 증강기법을 사용했다. 최초의 데이터 증강기법은 [그림 2]와 같이 이미지를 회전하거나 늘리는 등의 기본적인 기법이다[8].

[그림 2] 이미지 분야에서 기본적인 데이터 증강기법 예시 (맨 왼쪽: 기본 데이터, 오른쪽 10개: 증강된 데이터)[8]

이후, MixUp등과 같은 이미지 결합 기법이 제안되었다[9]. 이 기법들은 여러 개의 이미지를 결합하여 새로운 이미지를 생성한다. 생성된 이미지는 원본 이미지들의 특징을 결합하여 새로운 정보를 갖게 되고, 결합된 이미지들을 추가로 사용하여 모델을 훈련하면 모델의 일반화 성능이 개선된다.

2.3 적대적 공격 소개

적대적 공격은 크게 회피 공격(Evaision attacks)과 포이즈닝 공격(Poisoning attacks)으로 구분한다. 회피 공격은 추론 단계(Inference phase)에서 악의적인 공격자가 학습된 딥러닝 모델에 대해 조작된 적대적 예제(Adversarial examples, Purtubated test sample)를 입력하여 모델이 잘못된 예측을 하도록 만드는 공격이며, 포이즈닝 공격은 훈련 단계(Training phase)에서 딥러닝 모델의 훈련데이터를 악의적으로 조작하여 모델의 성능을 저하시키는 공격이다[1-3, 10, 11].

2.4 기존 연구

데이터 증강기법에 대한 동향 연구들이 활발히 진행되어왔다. Bayer 등[4]은 이미지 분야에서 증강기법의 분류를 제시하였고, 훈련 과정에서 과적합(Overfitting)을 완화하는 관점에서 증강기법을 분석하였다. Shorten과 Khoshgoftaar[5]은 텍스트 분류 분야의 증강기법을 조사하여 100개 이상의 데이터 증강기법을 12개로 분류하였다. Chen 등[6]은 자연어 처리 분야에서 11개의 데이터 증강기법을 조사하여 분석하였다.

기존 연구들은 다양한 분야에서 데이터 증강기법을 정리하였지만 보안 관점에서 데이터 증강기법을 정리한 동향 연구는 없었다. 그러므로 본 논문은 데이터 증강기법을 활용하여 적대적 공격을 방어하는 관점에서 연구 동향을 조사하고 분석하고자 한다.

3. 데이터 증강기법을 활용한 방어기법

3장에서는 적대적 공격에 대한 데이터 증강기법 기반의 방어기법들을 적대적 공격의 유형에 따라 1) 회피 공격 방어를 위한 증강기법과 2) 포이즈닝 공격 방어를 위한 증강기법 두 가지로 나누어 소개한다[1-3].

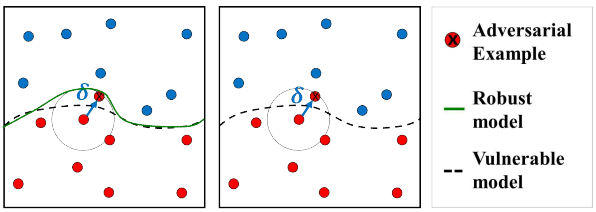
3.1. 회피 공격 방어를 위한 증강 기법

회피공격을 방어하기 가장 좋은 방법은 회피공격 기법으로 생성된 적대적 예제를 미리 훈련하는 것이다. 적대적 예제를 활용하여 훈련데이터를 생성하는 방법은 적대적 훈련(Adversarial Training)과 GAN(Generative Adversarial Networks)이 있다[10, 12-14].

3.1.1 적대적 훈련(Adversarial Training)

Madry 등[10]은 최초로 딥러닝 모델의 강건성(Robustness)을 높이는 적대적 훈련 기법을 제안하였다. 적대적 훈련은 라벨이 올바로 지정된 적대적 예제(Correctly labeled adversarial examples, 이하 증강 예제)를 훈련 데이터에 추가하여 딥러닝 모델의 강건성을 높이는 기법이다[10]. 즉, [그림 4] (a)와 같이 증강 예제(라벨이 올바로 지정된 적대적 예제)를 훈련데이터에 포함하여 학습시키면 이 과정에서 적대적 공격에 강건한 모델(Robust model)을 생성할 수 있다[15].

증강 예제(라벨이 올바로 지정된 적대적 예제)는 데이터에 변조 값(perturbation, δ)이 추가되므로, 적대적 훈련을 통해서 δ만큼 결정 경계가 명확해지는 효과가 있다. 증강 예제(라벨이 올바로 지정된 적대적 예제)를 추가하여 생성한 모델은 더 다양한 데이터를 인식하고, 더 일반화된 예측을 수행하므로, 정확도 등의 성능이 더 높아진다[10, 15].



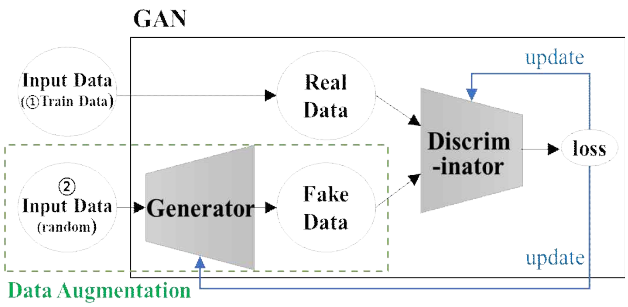
(a) (b)

[그림 4] (a) 올바로 레이블이 지정된 적대적 예제(증강 예제)를 사용한 적대적 훈련(빨강을 빨강으로 학습)과 (b) 적대적 예제를 사용한 회피공격(빨강 데이터를 파랑으로 오분류)

3.1.2 GAN(Generative Adversarial Networks)

GAN은 새로운 데이터를 생성해 내도록 훈련된 생성 모델(Generative Models) 중 하나로, [그림 5]와 같이 생성자(Generator)와 판별자(Discriminator) 두 개의 신경망의 조합으로 구성된다[12]. 생성자 신경망은 ② 랜덤 노이즈를 입력으로 받아 ① 훈련데이터와 유사한 새로운 데이터를 생성하며, 판별자 신경망은 생성된 데이터와 실제 데이터를 구분하는 것이 목적이다[12].

GAN의 훈련 과정은 생성자와 판별자 신경망 간의 적대적 게임을 반복하면서 진행되며, 생성자는 더 실제적인 데이터를 생성하도록 하고, 판별자는 실제 데이터와 생성된 데이터를 더 정확하게 구분하도록 훈련한다. 이때 훈련된 생성자는 데이터를 증강하는 역할을 한다[12-14].



[그림 5] GAN을 사용한 데이터 증강

Samangouei 등[13]은 적대적 훈련을 위해 GAN을 기반으로 Defense-GAN 구조를 제안하였다. GAN을 사용하여 일반적인 증강 데이터를 생성할 수 있다. 즉, 훈련 데이터(Real Data)를 적대적 예제로 하여 생성자 신경망을 훈련하면, 적대적 예제를 추가로 생성한다[11-13]. 추가로 생성된 적대적 예제는 라벨을 올바로 붙여서 증강 데이터로 활용되면 모델의 성능을 높일 수 있다. Li 등[16]은 GAN을 활용하여 정맥인식 데이터를 증강하여 적대적 공격을 방어하였다.

3.2. 포이즈닝 공격 방어를 위한 증강기법

포이즈닝 공격을 방어하기 위해서 훈련데이터의 정보를 압축하거나 이미지의 노이즈를 제거하는 등의 데이터 변환기법이 사용된다[5]. 이미지 변환기법 중에서 Cutmix 등과 같은 이미지 결합 기법과 방어 기법을 먼저 소개하고, 이미지 변환 기법을 사용한 방어 기법들을 소개한다[4, 17]. 마지막으로 GAN을 활용하여 데이터를 증강하여 포이즈닝 공격을 방어하는 연구에 대해 소개한다[5, 13].

3.2.1 이미지 결합 기법을 활용한 방어기법

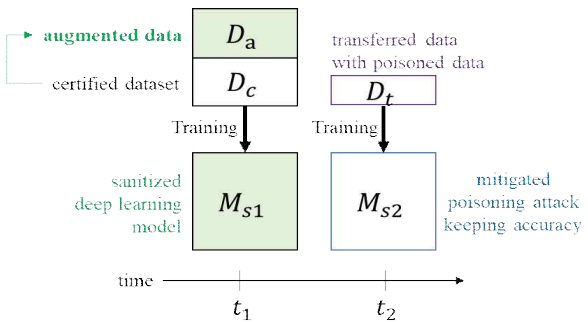
이미지 결합 기법은 대표적으로 Mixup, Cutout, Cutmix가 있다. [그림 6]의 (a)는 기본 이미지이며, (b)는 Mixup, (c)는 Cutout, (d)는 Cutmix 이다[9, 17, 18].

(a) (b) (c) (d)

[그림 6] 이미지 결합 기법을 활용한 데이터 증강[18]

Cutmix는 두 개의 이미지를 랜덤하게 선택하여 임의로 패치를 추출하여 다른 이미지에 삽입해서 생성된 이미지에 새로운 라벨도 함께 생성하여 하나의 새로운 이미지를 만든다[18].

Borgnia 등[19]은 Cutmix를 사용하여 포이즈닝 공격을 방어하는 방법을 제안하였다[18]. [그림 6]과 같이 Cutmix 기법을 사용하여 훈련데이터를 증강(D_a+D_c)하여 생성된 모델(M_{s1})은, 포이즈닝 공격(D_t)이 들어와도 결정경계가 크게 변하지 않아 공격을 완화시키는 효과(M_{s2})가 있다[18, 19].



[그림 7] 데이터 증강기법을 활용한 포이즈닝 공격 방어

3.2.2 이미지 변환 기법을 사용한 방어

Veldanda 등[20]은 포이즈닝 공격 데이터에 특정 노이즈를 추가하는 증강기법을 사용하여 공격 정보를 제거하였다.

Qui 등[21]은 깨끗한 데이터를 증강하여, 강건한 모델을 생성하기 위해 71가지의 이미지 변환 기법을 사용했다. 이미지를 뒤집는 등의 기본 증강기법을 지원하는 Albumentation 라이브러리(65개)[22]와 적대적 예제를 생성하는 FaceBook 라이브러리(6개)[23]을 사용했다.

3.2.2 GAN을 활용한 포이즈닝 공격 방어

GAN 기법은 회피 공격 뿐만 아니라 포이즈닝 공격에 대한 방어에도 활용된다. Chen 등[24]은 두 개의 GAN을 사용하여 포이즈닝 공격을 방어했다. 첫 번째 GAN을 활용하여 오염되지 않은 데이터셋을 기반으로 데이터를 증강하고, 증강된 데이터로 두 번째 GAN의 판별자 신경망을 훈련한다. 훈련된 판별자 신경망은 깨끗한 데이터를 판별하는 훈련을 했으므로, 데이터의 오염 여부를 구분한다[24].

4. 결론

본 논문은 데이터 증강 기법을 활용한 방어기법의 연구 동향을 두 적대적 공격 방법을 기준으로 분류하여 조사하였다. 또한, 데이터 증강 기법과 관련된 공격 기법을 이해하고 이를 방어하기 위한 최신 연구 동향을 제시하였다. 이를 토대로 견고한 딥러닝 모델을 개발하기 위해 데이터 증강 기법을 효과적으로 적용하는 데 도움이 될 것으로 기대된다.

본 연구는 데이터 증강기법을 활용한 방어기법의 동향과 적대적 공격에 따른 분류를 나누는데 중점을 두어, 각 방어기법들을 심도있게 다루지 못했다. 따라서, 향후 연구로는 본 연구를 확장하여 더 많은 사례를 심도있게 다루는 연구가 필요하다.

참고문헌

[1] A. Kurakin, I. J. Goodfellow and S. Bengio “Adversarial examples in the physical world,” International Conference on Learning Representations, vol. 5, 2017.

[2] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, “Adversarial attacks and defences: A survey,” arXiv preprint arXiv:1810.00069. 2018.

[3] K. Ren, T. Zheng, Z. Qin, and X. Liu, “Adversarial Attacks and Defenses in Deep Learning”, *Engineering*, vol. 6, 2020.

[4] M. Bayer, M. A. Kaufhold and C. Reuter, “A survey on data augmentation for text classification,” ACM Computing Surveys, vol. 55, no. 7, pp. 1-39, 2022.

[5] C. Shorten and T. M. Khoshgoftaar, “A survey on Image Data Augmentation for Deep Learning,” Journal of Big Data, vol. 6, 2019.

[6] J. Chen, D. Tam, C. Raffel, M. Bansal and D. Yang, “An empirical survey of data augmentation for limited data learning in NLP,” Transactions of the Association for Computational Linguistics, vol. 11, pp. 191-211, 2023.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” Communications of the ACM, vol. 60, 2017.

[8] X. Wang, K. Wang and S. Lian, “A survey on face data augmentation for the training of deep neural networks,” Neural computing and applications, vol. 32, no. 19, pp. 15503-15531, 2020.

[9] H. Zhang, M. Cisse, Y. N. Dauphin and D. Lopez-Paz, "mixup: Beyond empirical risk minimization,” International Conference on Learning Representations, vol. 6, 2018.

[10] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” International Conference on Learning Representations, vol. 6, 2018.

[11] I. J. Goodfellow, J. Shlens and C. Szegedy, "Explaining and harnessing adversarial examples," International Conference on Learning Representations, vol. 3, 2015.

[12] I. Goodfellow et al., “Generative adversarial networks,” Communications of the ACM, vol. 63, 2020.

[13] P. Samangouei, M. Kabkab, and R. Chellappa, “Defense-GAN: Protecting classifiers against adversarial attacks using generative models,” arXiv preprint arXiv:1805.06605. 2018.

[14] Y. Deldjoo, T. D. Noia and F. A. Merra, “A survey on adversarial recommender systems: from attack/defense strategies to generative adversarial networks,” ACM Computing Surveys (CSUR), vol. 54, no. 2, pp. 1-38, 2021.

[15] A. Shafahi et al., “Adversarial training for free!,” Advances in Neural Information Processing Systems, vol. 32, 2019.

[16] Y. Li, S. Ruan, H. Qin, S. Deng and M. A. El-Yacoubi, “Transformer Based Defense GAN Against Palm-Vein Adversarial Attacks,” IEEE Transactions on Information Forensics and Security, vol. 18, pp. 1509-1523. 2023.

[17] S. Yun et al., “CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features,” Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6023-6032, 2019.

[18] T. DeVries and G. W. Taylor, “Improved Regularization of Convolutional Neural Networks with Cutout,” arXiv preprint arXiv:1708.04552, 2017.

[19] E. Borgnia et al., “Strong data augmentation sanitizes poisoning and backdoor attacks without an accuracy tradeoff,” 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 2021-June, pp. 3855-3859, 2021.

[20] A. K. Veldanda et al., “NNoculation: Broad spectrum and targeted treatment of backdoored DNNs,” arXiv preprint arXiv:2002.08313, 2020.

[21] H. Qiu et al., “Deepsweep: An evaluation framework for mitigating DNN backdoor attacks using data augmentation,” in Proc. ACM Asia Conf. Comput. Commun. Secur., pp. 363-377, 2021.

[22] A. Buslaev et al., “Albumentations: fast and flexible image augmentations,” Information, vol. 11, no. 2. 2020.

[23] H. Qiu et al., “FenceBox: A Platform for Defeating Adversarial Examples with Data Augmentation Techniques,” arXiv preprint arXiv:2012.01701, 2020.

[24] J. Chen, X. Zhang, R. Zhang, C. Wang and L. Liu, “De-pois: An attack-agnostic defense against data poisoning attacks,” IEEE Transactions on Information Forensics and Security, vol. 16, pp. 3412-3425, 2021.