GAN Data augmentation을 위한 검증 및 설명 가능한 VX-GAN (Validation and eXplainable GAN) 모델

김지하, 박현희*
명지대학교 정보통신공학과

yaki5896@mju.ac.kr, hhpark@mju.ac.kr

Validation and eXplainable GAN(VX-GAN) model for GAN Data augmentation

Jiha Kim, Hyunhee Park*
Myongji Univ, Dept. Information and Communication Engineering

요 약

딥러닝 분야에서 데이터셋의 부족으로 인해 GAN(Generative Adversarial Networks) 모델이 data augmentation에 많이 사용되지만 실제로 그 데이터가 얼마나 유의미할지는 알 수 없다. 따라서 본 논문에서는 만들어진 fake 이미지 샘플이 의미가 있는지를 설명할 수 있는 VX-GAN(Validation and eXplainable - GAN) 모델을 제안한다. 생성자인 generator 모델이 만들어낸 fake 이미지 샘플이 올바른 이미지 샘플인지 검증하여 다시 피드백한다. 판별자인 discriminator의 경우 fake/real을 판단하는 기준이 어떤 영역을 보고 결과를 판단하였는지 사용자에게 설명을 해주는 부분으로 이루어져 있다. 결과적으로 generator 모델은 더 빠르게 학습을 진행하며 이미지를 생성할 수 있게 되며 discriminator가 올바르게 판단할 수 있을 정도의 학습을 임의로 정할 수 있게 된다.

Ⅰ. 서 론

본 논문에서는 딥 러닝의 가장 치명적인 문제점인 black box test의 결과에 대한 신뢰도를 높이면서 data augmentation 관점에서 얼마나 의미 있는 이미지 샘플을 만들어 내는가에 관한 연구를 진행한다. 대표적인 data augmentation 기법은 Generative Adversarial Networks (GAN) [1]이 있다. 해당 기법을 이용하면 어느 정도 실제로 있을 법한 이미지를 만들어낼 수 있다. 하지만 data augmentation의 목적이 부족한 학습 데이터를 보충하기 위한 것이며, 만들어진 이미지 샘플이 정말 의미가 있는지는 알 수 없다. 그림 1 에서 특정 이미지에 노이즈만 추가한 것으로 전혀 다른 이미지로 예측하게 된다. 하지만 설명이 가능한 GAN 기법(xAI-GAN) [2]이 등장하며 기존 GAN과 비교하면 MNIST및 FMNIST 데이터셋을 이용하여 생성한 이미지의 품질이 최대 23.18% 향상됨을 보여주었다. 본 논문에서는 해당 기법에 설명하는 방법을 discriminator 모델에만 적용하고, generator 모델에는 검증 알고리즘을 통해 더 효과적이고 빠른 학습 방법을 제안한다.

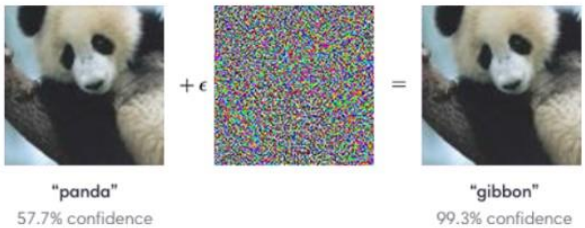

그림 1 노이즈로 인해 잘못된 예측을 하는 예[3]

Ⅱ. 본론

기존 GAN은 discriminator 모델이 real 이미지 샘플과 랜덤한 노이즈의 샘플을 이용하여 학습을 진행한다. 그렇게 discriminator 모델이 어느정도 학습을 진행하면 더 이상의 학습을 중단시킨다. 그 뒤에 generator 모델을 이용하여 랜덤 한 노이즈를 입력으로 이용하여 fake 이미지 샘플을 만든다. 만들어진 fake 이미지 샘플은 discriminator 모델에서 fake/real을 판단하여 결과를 반환한다. Generator 모델은 discriminator 모델이 fake라고 판단한 경우 오차가 증가하며 역전파를 이용하여 generator 모델을 학습한다.

Generator 모델을 검증하는 데에는 유사도 측정, 객체 탐지, 얼굴 탐지 등의 알고리즘을 이용하여 손실 함수를 계산한다. 계산된 손실 함수는 generator 모델이 학습할 때 검증 결과의 오차 값이 가중치에 더해진다. 결과적으로 generator 모델의 오차 역전파의 값이 더 의미 있게 전달될 수 있다.

GAN에서 discriminator 모델에 대한 설명이 필요한 이유는 다음과 같다. GAN 학습에서 discriminator 모델은 어느 정도 학습이 진행되면 학습을 멈추게 된다. 하지만 그 '어느 정도'를 알아내는 방법은 GAN 수식에 의존하여 black box 테스트를 할 수밖에 없다. 따라서 사용자에게 discriminator 모델이 이미지의 어떤 부분을 보고 real 이미지라고 인식하는지 알 필요가 있다.

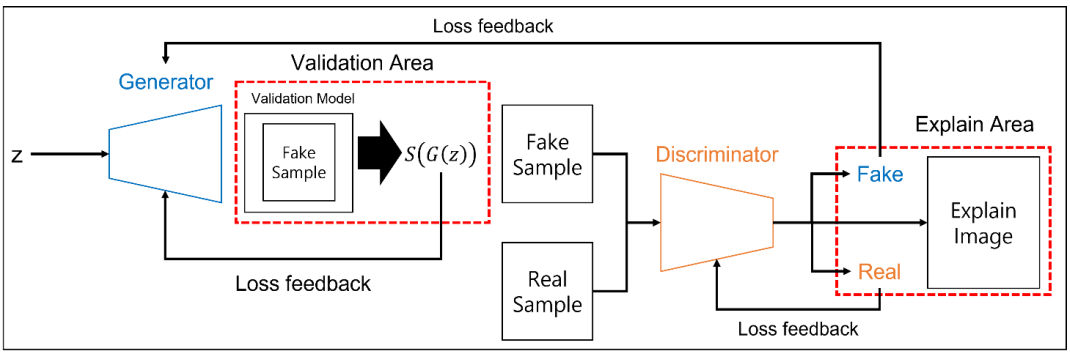

그림 2 본 논문에서 제안하는 VX-GAN 의 구조도

Discriminator 모델을 설명하기 위해서 LIME(Local Interpretable Model-agnostic Explanations)[4] 기법과 SHAP (SHapley Additive exPlanations)[5] 기법들을 적용해 볼 수 있다.

LIME 기법은 어떤 모델에서도 사용할 수 있다는 장점이 있다. LIME은 설명하기 위한 모델이 샘플을 입력했을 때 손실 함수가 최저가 되게 하는 슈퍼 픽셀 조합을 찾는다. 해당 조합을 이용하여 모델이 어떤 영역을 이용하여 결과를 도출했는지 사용자 인터페이스로 보여주게 된다.

SHAP은 특정 feature가 결과에 기여하는 정도를 측정하여 반환한다. 즉, discriminator 모델이 fake/real을 판단하는데 이미지의 어떤 픽셀이 어느 정도의 가치를 가지고 결과에 기여했는지를 인터페이스상에 보여줄 수 있다.

알고리즘 1. VX-GAN의 학습 과정

input: Generator 모델 G
input: 유사도 측정 모델 S
input: Discriminator 모델 D
output: 학습된 generator G
output: 설명된 $eX(D(z))$
1 foreach 노이즈 샘플 z **do**
2 Loss L1 = $\log D(x)$
3 Loss L2 = $\log\left(1-D(G(z))\right)$
Discriminator 모델: L1 과 L2 로부터 Δ_D연산
4 $eX(D(z))$를 이용하여 discriminator의 학습 여부 결정
$G(z)$에서 랜덤 한 노이즈 샘플 생성
5 $S(G(z))$에서의 유사도 측정 오차 값 계산
6 $\Delta'_{G(z)} = \Delta_{G(z)} + learning\_rate * \Delta_{G(z)} + \log\left(S(G(z))\right)$
7 G의 새로운 기울기 $\Delta'_{G(z)}$를 $\Delta_{G(z)}$로 사용
8 end
9 G의 파라미터 θ_G를 $\Delta_{G(z)}$로 업데이트 수행

그림 2 는 본 논문에서 제안하는 VX-GAN 모델의 구조도를 보여준다. 초기 학습 과정에서 discriminator 모델이 fake/real을 판단할 수 있을 정도의 Explain Image와 함께 진행한다. 사용자가 Explain Image를 직접 확인하고 discriminator 모델의 성능을 측정한다. 다음으로 generator 모델이 학습을 진행한다. Generator 모델은 학습을 진행하면서 만들어낸 fake sample을 검증 모델 S를 이용하여 오차를 측정한다. Fake sample은 그대로 다시 discriminator 모델을 통하여 fake/real을 판단 받는다. Fake라고 판단될 경우 모델 S에서 측정된 오차와 함께 가중치를 더하여 모델의 파라미터를 수정한다.

알고리즘 1 은 위의 그림 2 를 의사코드로 설명한 내용이다.

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data(x)}}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z))] \quad (1)$$

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data(x)}}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}\left[\log\left(1-D(G(z))\right) + \log\left(S(G(z))\right)\right] \quad (2)$$

식 (1)은 기존 GAN의 수식이다. Discriminator 모델과 generator 모델은 서로 경쟁을 하며 학습을 진행한다. 하지만 discriminator 모델이 얼마나 학습을 진행하는지, generator 모델은 얼마나 잘 만들어지는지는 알 수 없다. 식 (2)는 본 논문에서 제안하는 VX-GAN 모델의 수식이다. Discriminator 모델은 $eX(D(z))$를 이용하여 사용자가 fake/real을 어떤 기준으로 판단하였는지 알 수 있다(알고리즘 1. Line 4). Generator 모델은 $\log\left(S(G(z))\right)$에서 검증을 통한 오차를 사용하여 학습에 적용한다 (알고리즘 1. Line 6).

Ⅲ. 결론

본 논문에서는 기존의 GAN을 이용하여 data augmentation을 했을 때 만들어진 데이터가 학습 모델에 입력으로 적합한가에 대한 신뢰성을 높이며 generator 모델의 학습이 검증을 통해 빠르게 진행할 수 있는 VX-GAN 모델을 제안한다. 향후 해당 모델을 통하여 부족한 데이터셋을 효과적인 data augmentation 방법을 개선해 나갈 계획이다.

ACKNOWLEDGMENT

이 논문은 2021 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2021-0-00990, 설명가능한 인공지능 기반 무선랜 네트워크 시스템 고도화 핵심 기술 연구)

참 고 문 헌

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., and Bengio, Y. "Generative adversarial nets." In Advances in Neural Information Processing Systems, pp. 2672-2680, 2014.

[2] Nagisetty, V., Graves, L., Scott, J., Ganesh V. "xAI-GAN: Enhancing Generative Adversarial Networks via Explainable AI Systems." arXiv preprint arXiv:2002.10438, 2020

[3] Adversarial Attacks and Data Augmentation[Medium]. (2021.05.021). URL: https://medium.com/analytics-vidhya/adversarial-attacks-and-data-augmentation-1d97296b2d0c

[4] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2016

[5] Lundberg, Scott M., Gabriel G. Erion, and Su-In Lee. "Consistent individualized feature attribution for tree ensembles." arXiv preprint arXiv:1802.03888, 2018