DCGAN의 학습 기준을 분석하기 위한 Grad-CAM 기반의 XAI 접근 방법

옥진주[1]
[1]이화여자대학교 경제학과 학부생

pearlrich@ewhain.ac.kr

An XAI approach based on Grad-CAM to analyze learning criteria for DCGANS

Jin-Ju Ok[1]
[1]Dept. of Economics, Ewha Womans University

요 약

생성형 인공지능은 학습의 기준을 파악하기 어려운 모델이다. 그 중 DCGAN을 분석하여 판별자를 통해 생성자의 학습 기준을 판단할 수 있는 하나의 방법을 제안하고자 한다. 그 과정에서 XAI 기법인 Grad-CAM을 활용하여 학습 시에 모델이 중요시하는 부분을 분석하여 적합한 학습과 학습에 적합하지 않은 데이터를 분석하는 방법을 소개하고자 한다.

1. 서론

생성형 인공지능은 생성한 데이터에 상응하는 정답이 없는 모델이기에, 모델의 정확도와 같은 명확한 기준이 존재하기 어려운 모델이다. 이에 대해, 현재 다양한 연구가 진행되고 있는 XAI 분야의 기법을 사용하여 분석할 수 있는 방향을 제시하고자 한다. 이 과정에서 GAN의 특징인 판별자와 생성자의 관계로 판별자를 통해 생성자의 성능을 파악할 수 있음을 소개하고, DCGAN의 특징인 CNN(Convolution neural network)에 따라 특징을 분석해 낼 수 있는 XAI 기법인 Grad-CAM을 이용하여 분석을 진행하였다.

2. 관련 연구 및 이론

XAI와 시각화 기법

XAI (Explainable Artificial Intelligence)는 설명 가능한 인공 지능으로, 인공 지능 모델의 동작과 결정 과정을 사람들이 이해할 수 있는 방식으로 설명하는 기술과 분야를 의미한다. 모델 신뢰성 향상과 투명성 증가 그리고 윤리적 고려 사항을 해결하고자 하는 목표가 있다. 이중 대표적인 XAI의 방법으로 시각화 기법이 있다. 이 방법론은 시각화 기법을 활용해 모델의 동작을 이해하고 예측에 영향을 미치는 요소들을 시각적으로 표현한다. 설명 방법론에는 대표적으로 LIME, SHAP, Grad-CAM 등이 있으며 각각 다른 방법으로 시각화를 수행한다. DCGAN의 판별자를 설명하기 위해서 OA-GAN은 대표적인 Model-Agnostic 방법론인 LIME을 사용하여 판별자를 시각화해 과적합을 제한하는 모델을 개발했다. 하지만, CNN에 특화되어 있고 계층과 중요도를 나누어 볼 수 있는 Grad-CAM으로 판별자의 추가적 해석이 필요하다. 이에 더해, 단순히 판별자의 해석뿐 아니라, 학습분포에 따른 특징을 판별자의 시각화를 통해 도출해 낼 수 있으며, 이를 통해 생성자 학습의 적절성을 분석해볼 수 있다.

GAN(Generative Adversarial Networks)과 DCGAN

GAN은 이미지를 생성하는 생성자와 생성된 이미지의 사실 여부를 판단하는 판별자로 구성되어 있다. 이러한 두 개의 모델은 서로 대립하는 방식으로 학습이 되는데, 생성자는 판별자를 생성된 데이터를 진짜 데이터로 속이기 위해 학습하고, 판별자는 가짜 데이터를 판별하기 위해 학습된다. 경쟁적인 훈련 과정에 의해 생성자는 판별자를 속일 수 있을 정도의 가짜 데이터를 생성할 수 있는 성능 높은 생성 모델로 학습이 끝나게 된다. 또한 이 과정에서 진짜 데이터의 학습분포를 생성자가 따르게 된다. DCGAN(Deep Convolutional Generative Adversarial Networks)은 GAN의 성능을 높이기 위해, CNN이 GAN의 생성자와 판별자의 층에 추가 되어 이미지 특징을 추출과 공간적 구조를 더 잘 파악할 수 있게 되었다.

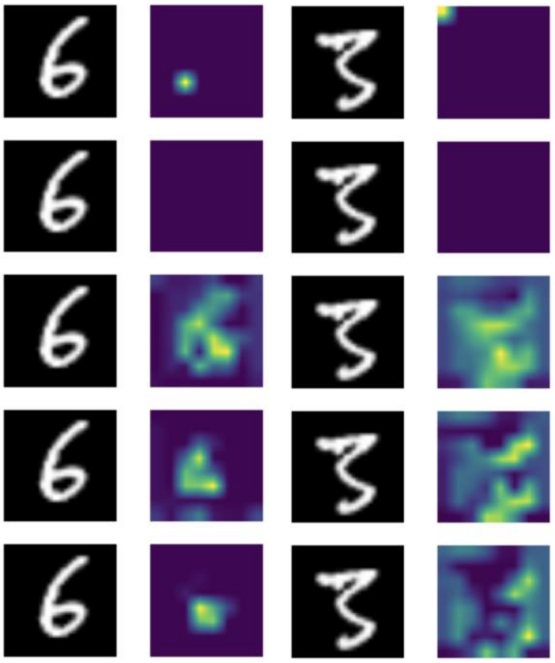
각각의 학습 데이터를 모방하여 확률 분포를 만들어낼 때, 모방된 생성자의 확률 분포는 판별자가 식별하는 특징과 높은 관련성이 있다. 가짜 데이터와 진짜 데이터를 구분해낼 수 있는 확률이 0.5가 되는 지점에 판별자가 근접하게 되기 때문에 D(x)는 사실상 특징을 구분하는 기준으로 작용하고 생성자의 분포가 원본 데이터의 분포를 모사하는 기준이 된다. 경쟁이 끝난 판별자는 생성자가 만들어 낼 이미지에 대한 특징에 대한 예측을 가능하게 한다. 이를 통해, 판별자의 분석이 생성자의 성능과 연결될 수 있다는 것을 알 수 있다.

Grad-CAM(Gradient-weighted Class Activation Mapping)

Grad-CAM은 딥러닝 모델의 예측 결과를 해석하고 시각화하는 기술이다. 특히 합성곱 신경망 (CNN)을 사용하는 이미지 분류 모델의 동작을 이해하는 데 도움을 준다. 기본적으로, Grad-CAM은 모델이 특정 클래스에 대한 예측을 내리는 데 어떤 부분이 중요한 역할을 하는지 시각적으로 파악할 수 있게 해주는 기법이다. 이를 위해 Grad-CAM은 모델의 예측을 생성하는데 얼마나 중요한 역할을 하는지를 판단하기 위해 그래디언트 정보를 사용한다. 모델의 컨볼루션 레이어의 출력에 대한 클래스 점수에 대한 그래디언트를 계산하여 컨볼루션 레이어의 출력에 대한 중요도 가중치를 나타낸다. 이 가중치를 원본 입력 이미지의 각 위치에 적용하여 "Activation Map"을 생성한다. 이 맵은 모델의 예측과 관련하여 입력 이미지의 각 부분의 중요도를 시각화한 것이다. Grad-CAM은 모델 내부의 동작을 더 잘 이해하고 모델의 신뢰성을 높이는 데 도움이 되는 유용한 도구 중 하나이다.

3. 실험 및 결과

실험은 MNIST 데이터에 64*64 이미지로 DCGAN 모델을 훈련하였다. 판별자의 경우 Conv2d, LeakyReLU, Conv2d, BatchNorm2d, LeakyReLU, Conv2d, BatchNorm2d, LeakyReLU, Conv2d, Sigmoid 의 순서로 모델을 구성하였고, 생성자의 경우 ConvTranspose2d, BatchNorm2d, ReLU, ConvTranspose2d, BatchNorm2d, ReLU, ConvTranspose2d, BatchNorm2d, ReLU, ConvTranspose2d, BatchNorm2d, ReLU, ConvTranspose2d, Tanh 순으로 모델을 구성하였다.



(그림 1) 원본 이미지와 Grad-CAM 분석 이미지

위 그림은 판별자의 학습을 5회, 10회, 15회, 20회, 25회의 기준으로 나누어 원본이미지와 Grad-CAM으로 분석한 이미지이다. 좌측 2개의 열을 보면, 판별자가 약 15회 학습되었을 때 가장 특징을 뚜렷하게 인식하고 있음을 알 수 있다. 위와 같이 특징을 선명하게 인식하는 모델과 유사한 학습 횟수의 모델들을 비교하면 가장 잘 인식된 모델을 찾을 수 있게 되며, 이 모델은 18회를 학습하였을 때 가장 판별자가 특징을 잘 해석할 수 있음을 분석을 통해 알 수 있었다. 이 과정에서, 같은 특징을 반복적으로 잡아내는 샘플이미지도 있을 수 있어 적절한 샘플이미지를 선정하여 학습도에 따라 다른 activation map을 그릴 수 있도록 하여야 한다. 이와 같이, 시각화를 통해 분석을 하게 되면 잘 학습된 판별자를 통해 잘 학습된 생성자를 찾을 확률을 높일 수 있다.

반면, 우측 2개의 열을 보면 데이터가 잘못 학습되고 있음을 알 수 있다. 이때의 특징은, 중점이 되는 숫자 이미지에 집중하는 것이 아닌 그 주변에 집중하여 특징을 구별해 내는데 있다. 이는, activation map의 tensor에서 70% 이상의 이미지 데이터를 중요하게 받아들인다면, 그 데이터를 학습과정에서 삭제하여 더욱 정교한 학습을 하도록 발전시킬 수 있다.

참고문헌

[1] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, & Yoshua Bengio, “Generative Adversarial Networks”, Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.

[2] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 618-626

[3] J. Kim and H. Park, "OA-GAN: Overfitting Avoidance Method of GAN oversampling based on xAI," 2021 Twelfth International Conference on Ubiquitous and Future Networks (ICUFN), Jeju Island, Korea, Republic of, 2021, pp. 394-398