논문 2024-4-15 http://dx.doi.org/10.29056/jsav.2024.12.15

인공지능 재귀 학습에 따른 모델 붕괴 현상 개선 방안

이승연*, 허석렬**, 이완직**†

Improvement of Model Collapse Phenomenon due to Artificial Intelligence Recursive Learning

Seung-Yeon Lee*, Seok-Yeol Heo**, Wan-Jik Lee**†

요 약

인공지능 분야 중, 사물 인식 기술은 공정 자동화, 자율 주행, 스마트 시스템 구축 등을 위한 기반 기술로 널리 사용되고 있다. 이러한 사물 인식 기술에서는, 인식의 정확도를 높이기 위해 인공지능이 생성한 이미지를 다시 학습 데이터로 활용하는 재귀 학습(Recursive Learning) 기법을 널리 사용한다. 하지만, 이러한 재귀 학습은 모델 붕괴(Model Collapse) 현상이 발생하여 사물 인식의 성능을 저하하는 문제를 발생시킬 수 있다. 이러한 문제를 해결하기 위해, 본 논문에서는 모델 붕괴의 원인이 되는 이미지 품질 저하 과정을 수학적으로 분석하고, 이를 기반으로 특이점 분석과 자연로그 함수를 활용한 생성 이미지 평가 기법을 제안하였다. 또한, 논문의 이미지 평가 기법을 실험적으로 구현하여, 원본 이미지와 서로 다른 손상 비율을 가진 이미지들에 대한 분석을 수행하였다. 이러한 실험과 실험 결과를 통해, 논문에서 제안한 이미지 평가 기법이 재귀 학습의 모델 붕괴에서 발생할 수 있는 이미지 손상을 조기에 감지하여 모델 붕괴 현상을 개선하는 데 충분히 활용될 수 있음을 보였다.

Abstract

Object recognition technology is widely used as a basic technology for process automation, autonomous driving, and smart system construction in the field of artificial intelligence. In such object recognition technology, recursive learning techniques are widely used to improve the accuracy of recognition, in which images generated by artificial intelligence are reused as learning data. However, this recursive learning can cause the problem of model collapse, which reduces the performance of object recognition. To solve this problem, in this paper, we mathematically analyzes the image quality degradation process that causes model collapse, and propose a generated image evaluation technique using singularity analysis and natural logarithm function based on this. In addition, the image evaluation technique in the paper was experimentally implemented, and analysis was performed on images with different damage ratios from the original image. Through these experiments and experimental results, it was shown that the image evaluation technique proposed in the paper can be sufficiently utilized to improve the model collapse phenomenon by early detecting image damage that can occur in model collapse of recursive learning.

한글키워드 : 인공지능, 재귀 학습, 모델 붕괴 개선, 특이점 분석, 이미지 평가 기법

keywords : Artificial Intelligence, Recursive Learning, Model Collapse Improvement, Singularity Analysis, Image Evaluation Techniques

* 부산대학교 IT응용공학과 학생
** 부산대학교 IT응용공학과 교수
† 교신저자: 이완직(email: wjlee@pusan.ac.kr)

접수일자: 2024.11.25. 심사완료: 2024.12.05.
게재확정: 2024.12.20.

1. 서 론

최근, 기계학습을 활용한 인공지능 기술의 눈부신 발전에 따라, 인공지능 관련 기술들이 산업 전 분야에서 많이 사용되고 있다. 이러한 기술 중, 특히 사물 인식 기술은 공정 자동화, 스마트 시스템 구축, 자율 주행 등을 위한 기반 기술로 널리 활용되고 있다. 인공지능 기술에서는 사물 인식의 정확도를 높이기 위해 인공지능이 생성한 이미지를 다시 학습 데이터로 활용되는 재귀 학습(Recursive Learning)이 많이 사용한다.

그렇지만, 이러한 학습 방식은 다양한 부작용을 발생할 수 있으며, 그중에서도 모델 붕괴 현상은 심각한 문제로 인식되고 있다[1]. 모델 붕괴는 인공지능 시스템이 새로운 정보를 학습하면서 이전에 학습한 내용을 손실하거나, 특정 유형의 데이터만 생성하는 문제를 발생하게 하며, 이러한 현상은 인공지능의 실용성과 신뢰성에 부정적인 영향을 미칠 수 있다.

예를 들어, Generative Adversarial Networks (GANs)는 이미지 생성과 변형에서 매우 효과적인 모델로 알려져 있다. 그러나 GAN의 훈련 과정에서 종종 발생하는 모드 붕괴(mode collapse) 현상은 생성기(generator)가 훈련 데이터의 특정 모드, 예를 들어 특정 유형의 이미지만 생성하는 문제를 초래한다[2]. 이에 따라 생성된 이미지의 다양성이 크게 제한되고, 사용자가 기대하는 결과를 충족하지 못하는 상황이 발생할 수 있다. 이는 데이터 세트의 불균형, 모델의 표현력 부족, 그리고 훈련 방법의 불균형 등이 원인으로 작용한다.

또한, 강화 학습(Reinforcement Learning) 분야에서도 재앙적 망각(catastrophic forgetting) 현상이 큰 문제점으로 지적된다[3]. 강화 학습에서는 에이전트가 새로운 환경에서 행동을 학습하는 과정에서 이전에 습득한 정보가 손실되는 경우가 종종 발생한다. 예를 들어, 로봇이 물체를 잡는 방법을 배운 후, 새로운 작업인 물체 던지기를 학습할 때 이전의 잡기 기술을 잊어버릴 수 있다. 이는 학습 경험의 축적 및 재활용의 제한과 가중치 업데이트 과정에서의 문제로 인해 발생하며, 결과적으로 다양한 작업에서의 성능 저하를 초래할 수 있다.

재귀 학습으로 인한 모델 붕괴 현상은 여러 요인에 기인한다. [4]에서는 생성적 적대 신경망(GANs)을 통해 고품질 이미지를 생성하는 과정에서 발생할 수 있는 데이터 손실과 왜곡 문제를 제기하였다. GANs는 생성기와 판별기(discriminator) 간의 경쟁을 통해 이미지를 생성하지만, 이 과정에서 원본 이미지의 고유한 특징이 점차 희석될 수 있다. 특히, 모델이 반복적으로 변형된 이미지를 생성하는 경우, 재귀적인 처리가 품질 저하를 가속할 수 있다.

이러한 문제를 해결하기 위해, 본 논문에서는 우선 모델 붕괴 현상을 발생시키는 이미지 품질 저하 과정을 수학적으로 분석하고, 이를 기반으로 특이점 분석과 자연로그 함수를 활용한 생성 이미지 평가 기법을 제안한다.

2. 관련 연구

인공지능 재귀 학습에서 발생하는 모델 붕괴 문제를 해결하기 위한 다양한 접근 방식이 모색되고 있으며, 메모리 기반 기술 및 데이터 출처 추적 방법론의 개발이 이루어지고 있다. 예를 들어, DeepFake 기술과 같은 응용 분야에서는 생성된 콘텐츠의 출처를 명확히 하고 신뢰성을 확보하기 위한 연구가 활발히 진행되고 있다. 이러한 연구들은 인공지능 생성 이미지의 품질을 보장하고 사용자에게 신뢰할 수 있는 정보를 제공하기 위한 필수 요소로 자리 잡고 있다.

[5]의 연구는 고해상도의 자연 이미지를 생성하기 위해 대규모의 GAN 훈련을 시도하면서, 모델 붕괴를 완화하는 방법을 제안한 연구이다. 이 논문에서는 하이퍼파라미터 조정과 새로운 아키텍처를 통해 GAN의 성능을 개선하고, 훈련 과정에서 모델 붕괴가 발생하지 않도록 하는 방법을 소개하였다. 이 연구는 대규모 데이터를 다룰 때, GAN 모델이 더 다양한 이미지를 생성할 수 있는 환경을 조성하는 데 초점을 맞추고 있다.

[6]에서는 Deep Convolutional GAN (DCGAN) 기법을 제안하여, 이미지 생성에서 비지도 학습의 가능성을 탐구하였다. 이 연구는 GAN 구조에 컨볼루션 레이어를 도입하여 고해상도의 이미지를 생성하면서도 모델 붕괴를 방지하는 방법을 제시하였다. DCGAN은 비지도 학습을 통해 더 다양한 이미지 패턴을 학습할 수 있으며, 이는 모델 붕괴를 줄이는 데 효과적이다.

본 논문에서는, 이러한 기존 연구를 기반으로 재귀 학습으로 인한 이미지 모델 붕괴 현상의 원인을 분석하고, 실제 환경에서 쉽게 활용할 수 있는 해결 방안을 제시하고자 한다.

3. 모델 붕괴 현상의 원인 분석

3.1. 이미지 압축 및 양자화에 의한 손실

이미지를 디지털화하고, 압축하는 과정에서 발생하는 기본적인 손실은 다음과 같이 정의할 수 있다. 특히 JEPG와 같은 손실 압축 방법은 다음과 같은 단계를 포함한다:

1) 변환

이 과정에서는 이미지를 주파수 영역으로 변환하는데, 예를 들어, 가장 기본적인 변환 방법으로는 이산 코사인 변환(DCT)이 사용된다.

$$Y(u,v)=\sum_{x=0}^{N-1}\sum_{y=0}^{M-1}f(x,y)\cos\left[\frac{\pi(2x+1)u}{2N}\right]\cos\left[\frac{\pi(2y+1)v}{2M}\right] \quad (1)$$

2) 양자화

이 단계에서는 DCT 계수를 양자화하여 데이터 크기를 줄인다. 이 과정에서 정보 일부가 손실되는데, 양자화는 다음과 같이 표현할 수 있다:

$$Y_q(u,v)=\left[\frac{Y(u,v)}{Q(u,v)}+0.5\right] \quad (2)$$

이 식에서 $Y(u,v)$는 DCT 계수, $Q(u,v)$는 양자화 행렬의 해당 값, $Y_q(u,v)$는 양자화된 계수를 표현한다.

3) 부호화

양자화된 값을 인코딩하여 저장하는 단계인데, 이 과정에서도 원본 이미지의 정보를 일부 손실한다.

3.2 반복 처리에 의한 누적 손실

이미지를 반복적으로 처리하면, 각 처리 단계에서 추가적인 손실이 발생하는데, 여기서는 이러한 추가적인 손실에 의한 누적 손실, 즉 전체 손실을 정의한다.

1) 손실함수 정의

먼저 손실함수를 정의하면, 손실함수 L은 모델의 예측값과 실제 값 y간의 차이를 측정하는 함수다. 이 함수는 평균 제곱 오차(MSE)를 계산하는 식으로 정의할 수 있다.

$$L(y,\hat{y})=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \quad (3)$$

(3) 식에서 각각 y는 실제 값, $\hat{y}$는 예측값, n은 데이터 포인터 수를 나타낸다. 이때 평균제곱오차는 예측값과 실제 값의 차이를 제곱하여 평균을 내는 값으로써, 이 값이 클수록 예측이 실제 값과 다르다는 것을 의미한다.

2) 반복 처리 모델

이미지를 x로 표현하고, 반복적인 처리를 T번 수행하는 모델을 가정한다. 각 단계에서 이미지가 변환될 때마다 손실이 발생한다. 이를 수학적으로 표현하면 다음의 식 (4)와 같다.

$$\hat{y}^{(t)} = f(\hat{y}^{(t-1)}) \qquad (4)$$

이 식에서 f는 변환함수이며, $\hat{y}^{(t)}$는 t번째 처리 후의 예측 결과이다. 또 $\hat{y}^{(0)}$은 초기 입력 이미지에 대한 예측값이다.

3) 손실 누적

각 단계의 손실은 다음과 같이 나타낼 수 있다.

$$L^{(t)} = L(y, \hat{y}^{(t)}) = L(y, f(\hat{y}^{(t-1)})) \qquad (5)$$

이 공식은 t번째 처리 단계에서의 손실을 의미한다. 즉, 실제 값 y와 t번째 단계에서의 예측값 $\hat{y}^{(t)}$의 차이를 측정한다. 반복적으로 손실이 발생하므로, 전체 손실 $L_{total}(t)$은 다음과 같이 계산된다.

$$L_{total}(t) = \sum_{t=1}^{T} L^{(t)} \qquad (6)$$

4) 추가 손실 발생

각 단계에서 변환함수 f가 불완전하거나 이미지 정보의 일부를 손실시키는 경우, 각 단계에서 발생하는 손실은 다음과 같이 설명할 수 있다.

$$L^{(t)} = L(y, f(\hat{y}^{(t-1)})) \geq L(y, \hat{y}^{(t-1)})) - \epsilon \qquad (7)$$

여기서 ϵ은 각 단계에서 발생하는 추가 손실을 나타낸다. 이는 정보 손실이나 노이즈의 영향을 반영한다. 이에 의해, 전체 손실은 반복 처리 횟수에 비례하여 증가할 수 있다. 이 식은 반복 처리 과정에서 단계마다 추가적인 손실이 발생한다는 점을 표현하므로, 전체 손실은 다음과 같이 증가하게 된다.

$$L_{total}(t) = \sum_{t=1}^{T} L(y, f(\hat{y}^{(t-1)})) + \epsilon \qquad (8)$$

3.3 노이즈 추가

인공지능 모델이 이미지를 처리할 때, 변환이나 필터링 과정에서 노이즈가 추가될 수 있다. 노이즈를 $N(x,y)$라고 하면, 변환된 이미지 $I'(x,y)$는 다음과 같이 표현될 수 있다.

$$I'(x,y) = I(x,y) + N(x,y) \qquad (9)$$

여기서 $I(x,y)$는 원본 이미지이며, 실제 이미지는 노이즈가 포함되므로 이미지 품질이 저하된다. 이와 같은 수학적 과정들이 결합하여 반복적인 이미지 처리 과정에서 품질 저하가 발생하게 된다.

4. 모델 붕괴 방지를 위한 방안

본 장에서는 인공지능 생성물에 대한 평가 및 비교를 위해, 이미지의 특이점 기록과 자연지수 함수를 기초로 한 분포 함수의 표준편차를 관리하는 방법을 제안한다. 이러한 방법을 통해 생성된 이미지의 품질과 정체성을 유지할 수 있다.

4.1. 특이점 기록

인공지능이 생성한 이미지의 고유한 특성과 패턴을 기록하여, 추후 생성된 이미지와의 비교 및 분석에 활용한다. 이는 특정 이미지의 출처와 질을 파악하는 데 활용할 수 있다.

1) 이미지 특이점 기록

이미지의 특이점을 기록하기 위해, 피처를 추출하는 다양한 방법을 사용할 수 있다.

3장에서 언급한 이산 코사인 변환(DCT)을 통해 특이점을 추출할 수 있다. 이미지의 특이점은 이미지의 주파수 성분에서 정보가 집중되는 부분으로 DCT는 이미지의 공간 정보를 주파수 성분으로 변환하여, 저주파수 성분에 주로 이미지의 구조적 정보를, 고주파수 성분에 텍스처 및 세부 정보가 담기도록 한다. 따라서 DCT를 통해 주파수 도메인으로 변환된 계수 $Y(u,v)$에서 특이점을 도출할 수 있다. 예를 들어, $Y(u,v)$가 특정 주파수에서 현저히 높은 값을 가지면 해당 성분에서 이미지의 특이점이 집중됨을 의미한다.

딥러닝 기반 피처 추출의 경우에는 CNN(합성곱 신경망)을 사용하여 이미지에서 중요한 특성(예: 엣지, 색상 분포 등)을 추출한다. 또한 SIFT(Scale-Invariant Feature Transform)/SURF(Speeded-Up Robust Features) 특징을 감지하여 이미지를 표현하는 고유한 특이점을 기록한다[7, 8]. 이 두 방법을 결합하여 특이점을 추출한다면 DCT를 통해 특이점이 있을 가능성이 큰 고주파 영역을 미리 선택함으로써, SIFT/SURF의 계산 범위를 줄이고 처리 속도를 높일 수 있다[9].

2) 특이점 분포 분석

특이점의 분포를 자연지수함수(지수 분포)로 모델링할 수 있다. 이를 통해, 특정 특성이 얼마나 잘 보존되고 있는지 평가할 수 있다. 특이점의 데이터 포인트는 각 특이점에 대해 값과 위치를 기록하는데, 이를 기반으로 분포를 생성할 수 있다. 지수 분포 함수는 다음과 같이 정의되며, λ는 분포의 속도를 나타내는 매개변수이다.

$$f(x|\lambda) = \lambda e^{-\lambda x} \quad (x \geq 0) \qquad (10)$$

4.2 표준편차 관리

자연지수함수를 기초로 하여, 생성된 이미지의 품질 분포를 분석한다. 이를 통해 인공지능이 생성한 이미지의 품질 변동성을 수치로 평가할 수 있으며, 최종적으로 재귀적 합성에 따른 품질 저하를 방지할 수 있다.

표준편차는 데이터의 변동성을 나타내므로, 이를 통해 특이점의 일관성을 평가할 수 있다. 표준편차는 다음의 식에 의해 계산되며, 식에서 μ는 평균이고, N은 데이터 포인트 수를 나타낸다.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \qquad (11)$$

4.3 품질 유지 관리

지속적인 이미지 품질을 유지하기 위해서 다음과 같은 세부 기능 수행이 필요하다.

1) 피처(feature) 보존: 특이점의 표준편차가 특정 임곗값 이하로 유지되도록 관리한다. 임곗값을 설정하여 그 이상일 경우, 추가적인 처리(예: 재생성, 변형 조정)를 수행한다.

2) 훈련 데이터 조정: 인공지능 모델이 훈련되는 데이터 세트의 품질을 지속해서 관찰하여, 이상 징후가 발견되면 데이터 세트를 업데이트하거나 추가 학습을 시행한다.

3) 하이퍼파라미터 조정: 이미지 생성 모델의 하이퍼파라미터를 조정하여 품질을 최적화한다. 이때, 표준편차를 고려하여 매개변수를 설정한다.

4) 피드백 루프: 인공지능 모델이 생성한 이미지의 품질에 대한 사용자 피드백을 지속해서 수집하고 이를 모델 개선에 활용한다.

5. 실험 및 결과

4장에서 제안한 모델 붕괴 방지 방법을 실험적으로 검증하고 성능을 확인하기 위하여 이미지 특이점 분포 분석과 두 이미지의 특이점 분포도와 평균 및 표준편차를 비교 분석하는 실험을 수행하였다. 두 실험을 진행하기 위해 먼저 원본

이미지(image_noiseX.jpg)를 준비한 후, 이미지에 노이즈를 추가해주는 Photo Mosh 사이트[10]를 이용해 픽셀 속성으로 10% 노이즈와 50% 노이즈를 갖는 이미지를 생성하였다. 두 실험은 파이썬 3.6과 pycharm 2022 환경에서 진행하였으며 opencv-python, numpy, matplotlib, scipy 라이브러리를 사용하였다.

5.1 실험 절차 A(이미지 특이점 분포 분석)

① 컬러로 되어있는 각 이미지 파일을 모노톤(흑백화면)으로 바꾸어 이미지를 로드한다.
② 이미지를 DCT로 변환하여 저주파와 고주파 성분을 분리한다.
③ DCT에서 고주파 성분이 높은 영역을 강조한 후, 역 DCT를 적용하여 공간 도메인으로 되돌린다.
④ 관심 영역이 강조된 이미지에서 SIFT 알고리즘을 사용하여 특이점과 설명자(이미지의 특이점을 수치로 표현한 벡터)를 검출한다.
⑤ 주어진 특이점 리스트(keypoints)에서 각 특이점의 크기(size)를 추출하여 NumPy 배열로 반환하고 밀도 기준으로 정규화하여 표현한다.
⑥ 특이점 크기의 평균을 계산하고, 0부터 최대 크기까지의 100개의 점을 생성한 뒤, expon.pdf를 이용해 지수 분포의 확률 밀도 함수를 계산한다.

5.2 실험 절차 B(두 이미지의 특이점 분포 분석 및 평균과 표준편차 관리)

① 비교할 두 이미지를 지정된 경로에서 불러온 후 함수를 사용해 각각의 이미지에서 특이점(keypoints)과 설명자를 추출한다.
② 각 특이점의 크기를 추출하여 sizes 리스트에 저장한다. 주어진 데이터인 특이점 크기(sizes)의 히스토그램을 그린 다음, 해당 데이터에 대한 지수 분포를 계산한다.
③ NumPy의 함수를 사용하여 특이점 크기의 평균(mean_size)과 표준편차(std_dev)를 계산한다.

5.3 실험 결과 A

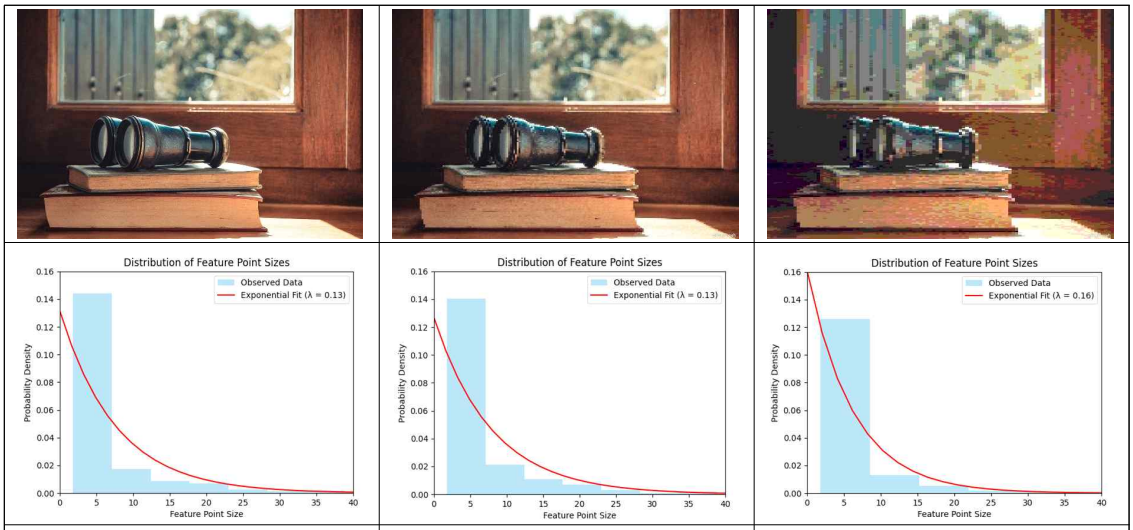
실험 A를 수행한 결과를 표 1과 그림 1에 나타내었다. 표 1과 그림 1에서 보듯이 전반적으로 노이즈가 10%, 50%로 증가할수록 특이점 분포 비율(λ)이 증가하고 변동성이 커짐을 확인할 수 있다. 세부적으로 보면 낮은 노이즈(10%)가 추가되더라도 지수 피팅은 안정적으로 유지되는데, 이는 특이점 크기 분포가 중간 수준의 노이즈에 견고함을 보여준다. 반면에 높은 노이즈(50%) 수준에서는 원본 분포 패턴과 상당한 차이를 보이며, 비율 매개변수도 커진다. 이는 높은 노이즈가 작은 특이점을 덜 보이게 하거나 중간 크기의 잡음을 유발하여 분포를 왜곡함을 의미한다.

표 1. 이미지 특이점 분포도 비교
Table 1. Comparison of the distribution of image singularities

노이즈 수준(%)	비율 매개변수 λ	최대 확률 밀도 (P_max)	모델 안정성 지표 (0~1)
0%	0.13	0.14	1
10%	0.13	0.14	0.9
50%	0.16	0.16	0.5

5.4 실험 결과 B

실험 B를 수행한 결과를 표 2와 그림 2에 나타내었다. 앞 실험 A의 결과와 유사하게 전반적으로 이미지 변환에서 노이즈 수준이 10%, 50%로 증가할수록 특이점 분포와 표준편차에 명확한 변화가 나타나는 것을 확인할 수 있다. 이는 노이즈 강도가 높아짐에 따라 이미지 모델이 견고성을 잃고, 모델 붕괴가 발생할 가능성이 커진다는 것을 의미한다. 이러한 분포의 이동은 안정적인 특징 표현의 상실을 반영하며, 이미지 인식이나 특징 매칭과 같은 후속 작업에 영향을 미친다.



(a) original image (b) noise 10% image (c) noise 50% image

그림 1. 이미지 특이점 분포도 비교
Fig. 1. Comparison of the distribution of image singularities

실험 A에서 확인한 노이즈에 따른 이미지 품질 저하 상태는 인공지능의 재귀 학습 과정에서 발생할 수 있는 모델 붕괴 현상에 대한 중요한 신호가 된다. 이를 활용하여 모델의 적응력을 높이거나(인공지능 모델이 다양한 데이터 상황이나 변화하는 환경에 효과적으로 대응할 수 있도록 하는 것을 의미) 데이터 전처리를 개선함으로써, 이미지 품질을 보존하고 학습의 정확성을 향상할 수 있다.

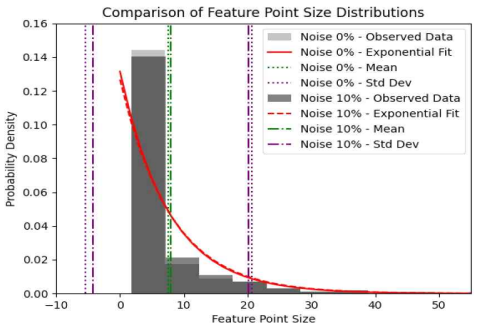
표 2. 이미지 특이점 분포도 및 표준편차 비교
Table 2. Comparison of the distribution of image singularities and standard deviations

노이즈 수준(%)	평균 특이점 크기(μ)	표준 편차 (σ)	비율 매개변수 (λ)
0%	7.60	13.05	0.13
10%	7.89	12.19	0.13
50%	6.22	9.60	0.16

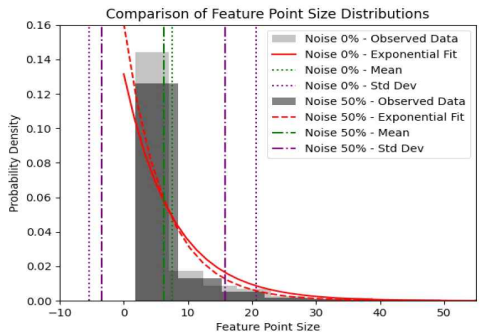
또한 실험 B에서는 서로 다른 노이즈 수준을 주입함으로써, 분포의 평균 및 표준편차에 대응하는 변화를 관찰하여 특징 표현의 불안정성을 확인하였다.

실험 결과에서 확인된 비율 매개변수 λ, 평균 크기 μ, 표준편차 σ, 최대 확률 밀도 P_max와 같은 지표들은 각각 노이즈가 증가할 때, 나타나는 모델의 특징 표현 변화를 수치로 나타내며 해당 값이 변화함에 따라 이미지 모델 붕괴를 감지할 수 있다. λ 값은 특정 크기에 특이점 분포가 집중되었음을 나타내며, 값이 증가하면 왜곡이 심화함을 의미한다. 평균 크기 μ는 특이점 크기 분포의 대푯값으로, 감소 시 큰 특이점이 감지되지 않거나 작은 특이점이 포함되었음을 시사한다. 또한, 표준편차 σ는 분포의 다양성을 나타내며, 값 감소는 크기 분포가 좁아지고 특정 크기에 집중되는 모델 붕괴 징후로 해석된다. 최대 확률 밀도 P_max는 분포의 밀집 정도를 나타내며, 값 증가와 분포 축소는 특정 크기에 과도한 집중을 의미한다.

따라서 노이즈가 10%인 경우 λ와 P_max는 안정적으로 유지되며 μ와 σ의 경미한 변화로 모델의 안정성이 확인되었다. 반면, 노이즈가 50%로 증가하면 λ와 P_max는 증가하고 μ와 σ는 감소하여, 모델의 특이점 분포가 왜곡되고 붕괴 징후가 명확히 드러났다.



(a) noise 10% image



(b) noise 50% image

그림 2. 이미지 특이점 분포도 및 표준편차 비교
Fig. 2. Comparison of the distribution of image singularities and standard deviations

이러한 정량적 지표의 변화를 바탕으로 노이즈 수준과 모델 붕괴 간의 상관관계를 조기에 분석하면, 알고리즘의 성능 저하를 예측할 수 있으며, 이에 따라 데이터 전처리 강화, 특이점 검출 개선, 모델 복원력 향상 등 다양한 대응이 가능하며 결과적으로, 노이즈에 대한 민감도를 낮추고 모델의 안정성을 높일 수 있다.

6. 결 론

인공지능 분야에서는 사물 인식의 정확도를 높이기 위해 인공지능이 생성한 이미지를 다시 학습 데이터로 활용하는 재귀 학습(Recursive Learning) 기술을 널리 사용하고 있다. 그러나 이러한 발전은 모델 붕괴(Mode Collapse)와 같은 심각한 문제를 가진다.

본 논문에서는, 이러한 모델 붕괴 문제를 해결하기 위해 특이점 기록과 분포 함수 표준편차 관리라는 새로운 방법론을 제안하였으며, 이를 활용하여 생성 이미지에 대한 평가 기법을 구현하고 테스트를 진행하였다. 테스트 결과에서, 특이점 분포와 자연로그 함수를 활용한 이미지 분석 기능이 인공지능 모델의 학습 안정성과 성능 향상에 기여할 수 있음을 확인하였다.

이 논문은 부산대학교 기본연구지원사업(2년)에 의하여 연구되었음

참 고 문 헌

[1] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson and Yarin Gal1, “AI models collapse when trained on recursively generated data,” Nature, Vol. 631, pp. 755-760, July 2024. DOI: https://doi.org/10.1038/s41586-024-07566-y

[2] Srivastava, A., Valkov, L., Russell, C., Gutmann, M. U., and Sutton, C., “VEGAN: Reducing mode collapse in GANs using implicit variational learning,” Advances in neural information processing systems, Vol 30, May 2017. DOI: https://doi.org/10.48550/arXiv.1705.07761

[3] Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O., “Understanding deep learning requires rethinking generalization,” International Conference on Learning Representations (ICLR), April 2017. DOI: https://doi.org/10.48550/arXiv.1611.03530

[4] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y., “Generative adversarial nets,” Proceedings of the 27th International Conference on Neural Information Processing Systems(NeurIPS), Vol. 27, pp. 2672-2680, December 2014. DOI: https://doi.org/10.1145/2969239.2969256

[5] Brock, A., Donahue, J., and Simonyan, K., “Large scale GAN training for high fidelity natural image synthesis,” International Conference on Learning Representations (ICLR), Feburary 2019. DOI:https://doi.org/10.48550/arXiv.1809.11096

[6] Radford, A., Metz, L., and Chintala, S., “Unsupervised representation learning with deep convolutional generative adversarial networks,” International Conference on Learning Representations (ICLR), November 2015. DOI: https://doi.org/10.48550/arXiv.1511.06434

[7] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S., “GANs trained by a two time-scale update rule converge to a local Nash equilibrium,” Advances in Neural Information Processing Systems (NeurIPS), Vol. 30, pp. 6626-6637, December 2017. DOI: https://doi.org/10.48550/arXiv.1706.08500

[8] Odena, A., Olah, C., and Shlens, J., “Conditional image synthesis with auxiliary classifier GANs,” Proceedings of the 34th International Conference on Machine Learning (ICML), Vol. 70, pp. 2642-2651, August 2017. DOI: https://doi.org/10.48550/arXiv.1610.09585

[9] Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y., “Spectral normalization for generative adversarial networks,” International Conference on Learning Representations (ICLR), April 2018. DOI: https://doi.org/10.48550/arXiv.1802.05957

[10] PHOTO MOSH, “Unlimited creative realtime effects for image and video,” Available online: http://photomosh.com, Nov. 2024.

저 자 소 개

이승연(Seung-Yeon Lee)

2020.3-현재 부산대학교 IT응용공학과
2020.09-2021.03 COURR 개발팀 대리
2021.04-2022.04 (주)애정타임 개발팀 대리
2023.05-2023.07 잠실여자고등학교 지구과학 강사
2024.2 부산대학교 IT응용공학과 졸업예정
<주관심분야> 딥러닝, 자율주행, 빅데이터

허석렬(Seok-Yeol Heo)

1986.2 경북대학교 전자공학과 졸업
1991.2 경북대학교 컴퓨터공학과 석사
2008.2 경북대학교 컴퓨터공학과 박사
1992.3-2006.2 밀양대학교 컴퓨터공학부 교수
2012.9-2013.8 Univ. of Texas at Dallas 방문교수
2006.3-현재 부산대학교 IT응용공학과 교수
<주관심분야> IoT, 딥러닝, 빅데이터

이완직(Wan-Jik Lee)

1992.2 경북대학교 통계학과 졸업
1994.2 경북대학교 컴퓨터공학과 석사
2007.2 경북대학교 컴퓨터공학과 박사
1997.3-2006.2 밀양대학교 정보통신학과 교수
2006.3-현재 부산대학교 IT응용공학과 교수
<주관심분야> IoT, 통신프로토콜, AI