합성 이미지 데이터의 품질평가 지표에 대한 고찰

맹소정, 이상민, 이상복*

한국정보통신기술협회

msj911@tta.or.kr, minuri33@tta.or.kr, *jangpo@tta.or.kr

A Study on Evaluation Metrics for Synthetic Image Data

Maeng So Jeong, Lee Sang Min, Lee Sang Bok*

Telecommunications Technology Association

요 약

인공지능 성능 향상 및 지능화 서비스 개발을 위해 대규모의 고품질 학습 데이터 확보가 중요해지고 있다. 그러나 민감 정보 포함 이슈, 시·공간 제약 때문에 데이터 확보가 제한적인 도메인에서는 데이터 취득이 어려운 문제로 대두되고 있다. 이에 인공적으로 데이터를 생성하는 합성 이미지 데이터 구축이 주목받고 있으며 연구가 활발히 진행되고 있다. 이러한 합성 이미지 데이터는 실제 데이터와 유사한 수준의 학습 성능 구현을 위하여 충실도와 다양성 측면에서 체계적인 품질관리가 요구되고 있는데, 본 논문에서는 품질관리 시 고려되어야 할 평가 지표를 살펴보고, 지표별 특징 및 한계점에 대해 고찰하고자 한다.

Ⅰ. 서 론

최근 딥러닝 기술이 발전함에 따라 고품질의 학습 데이터 확보의 중요성이 더욱 강조되고 있다. 하지만 민감 정보를 포함하고 있거나 시·공간 제약으로 인한 재현의 어려움으로 데이터 확보가 제한적인 자율주행, 의료, 국방 등의 도메인에서는 데이터 수집에서부터 병목이 발생하는 문제를 겪고 있다. 이에 생성적 적대 신경망(GAN)이나 변이형 오토 인코더(VAE)와 같은 딥러닝 기술을 사용하여 기존 데이터로부터 인공적으로 데이터를 생성하는 합성 이미지 데이터를 구축함으로써 데이터의 부족으로 인한 한계를 극복하는 방안이 주목받고 있다. 이러한 합성 이미지 데이터는 실제 데이터의 모든 주요한 속성, 상관관계, 특성을 유사하게 유지하고 편향되지 않도록 해야 학습에 효과적이며 그렇지 못한 경우에는 학습 효과가 낮아 체계적인 품질관리가 필요하다. 본 논문에서는 합성 이미지 데이터의 품질을 측정하고 관리하기 위해 고려되어야 할 평가 지표를 살펴보고, 지표별 특징 및 한계점에 대해 고찰하고자 한다.

Ⅱ. 본론

합성 이미지 데이터의 품질은 실제 데이터 셋과 얼마나 유사한지를 나타내는 충실도(Fidelity)와 데이터가 얼마나 다양하게 생성되었는지를 나타내는 다양성(Diversity) 측면에서 관리되어야 한다. 이러한 두 가지 성질을 고려한 평가 지표는 다음과 같다.

1. IS(Inception Score)

IS는 사전학습 모델인 Inception 모델을 이용하여 합성 이미지의 충실도와 다양성을 동시에 평가하는 지표로, 실제 이미지 없이 합성 이미지만으로 평가한다. 먼저, 아래 산식의 확률 분포 $p(y|x)$를 계산하여 충실도를 판단한다. 이때, $p(y|x)$의 엔트로피가 IS 값에 영향을 미치는데, 엔트로피는 확률 분포의 불확실성을 의미하고 랜덤 변수 x의 클래스 예측이 어려울수록 높게 측정되는 특징이 있다.[1] 다음으로 주변 확률 $p(y)$를 통해 다양성을 판단하며, 합성 이미지가 다양할수록 데이터의 분포가 균등하다고 해석할 수 있다. 마지막으로 앞서 계산한 $p(y|x)$와 $p(y)$의 확률 분포 차이값인 KL-divergence를 계산하는데, 이 값의 기대값이 클수록 IS 값이 높아져 합성 이미지 데이터가 잘 만들어졌다고 판단한다.

$$IS = \exp(E_x KL(p(y|x) \| p(y)))$$

x = 합성 이미지의 집합, y = 클래스, E_x = 기댓값
$p(y|x)$ = 합성 이미지가 Inception 모델에 의해 예측된 클래스 y의 확률,
$p(y)$ = 클래스 y의 확률 분포

하지만 IS는 1차원 점수 기반으로 단순히 합성 이미지만의 충실도와 다양성만을 고려하기 때문에 실제 이미지와의 정확한 비교가 불가능하다는 한계점이 있다. 그리고 클래스당 한 이미지만 생성하는 경우, 다양성이 낮음에도 불구하고 $p(y)$가 균등하여 IS 값이 높게 나오는 문제점이 있다.

2. FID(Fréchet Inception Distance)

FID는 실제 데이터의 분포를 고려하지 않는 IS의 단점을 개선하였으며, 실제 데이터와 합성 이미지 데이터 간 차이를 측정하는 방식으로 평가한다. 먼저 Inception 네트워크의 convolutional layer를 사용하여 합성 이미지와 실제 이미지의 특징(feature) 추출 작업을 수행한다. 이후, 추출된 특징들을 사용하여 생성된 이미지 집합과 실제 이미지 집합의 특징 분포를 나타내는 평균 벡터(μ)와 공분산 행렬($\sum$)을 계산하고 이를 다변량 가우시안으로 모델링한다. 마지막으로 계산된 합성 이미지의 다변량 가우시안 분포와 실제 이미지의 다변량 가우시안 분포 간 거리를 Fréchet Distance로 계산한다. 이때, FID 값이 낮을수록 실제에 가까운 양질의 합성 이미지가 생성되었다고 판단한다. 아래 산식의 첫 번째 항(실제 이미지의 평균 벡터와 합성 이미지의 평균 벡터를 비교한 값)이 작을수록 충실도가 높고, 두 번째 항(두 이미지 집합의 공분산 행렬 간의 거리)이 클수록 다양성이 높다.[3]

$$FID = |\mu_r - \mu_g|^2 + T_r(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{\frac{1}{2}})$$

r = 실제 이미지, g = 합성 이미지, T_r = 행렬의 Trace 연산

그러나 FID는 IS와 마찬가지로 1차원 점수 지표이기 때문에 충실도와

다양성을 분리하여 평가가 불가하다는 단점이 있다. 그리고 FID는 실제 이미지와 합성 이미지의 특징 벡터를 기반으로 각 이미지 집합을 다변량 정규분포로 모델링하여 비교하는 지표이기 때문에 특징 분포가 다변량 정규분포를 따르지 않는 경우에는 FID 값을 설명하기 어려운 한계점이 있다.

3. Precision 및 Recall

합성 이미지 데이터의 품질을 평가할 때는 충실도와 다양성 간의 균형이 중요하다. 앞서 살펴본 IS, FID는 충실도와 다양성을 분리하여 평가가 불가하다는 한계점이 있었으나, Precision 및 Recall은 2차원 점수 기반의 품질평가 지표로서 충실도와 다양성을 각각 평가 가능하다는 장점이 있다. 그리고 두 지표는 상호 반비례적인 관계를 갖기 때문에 한 지표를 높이면 다른 지표는 감소할 수 밖에 없어 단독으로 해석하기보다는 두 지표가 함께 고려되어야 한다. Precision은 합성 이미지가 얼마나 정밀하게 실제 이미지를 묘사하였는지 나타내는 충실도 평가를 위한 지표로, 합성 이미지 중에서 실제 이미지 샘플 분포에 속한 비율을 의미한다. 이때, Precision의 값이 높을수록 합성 이미지의 충실도가 높다고 판단한다. Recall은 실제 이미지가 얼마나 많이 합성 이미지를 통해 재현되었는지 나타내는 다양성 평가를 위한 지표로, 실제 이미지 중에서 실제 이미지 샘플 분포에 속한 비율을 의미한다. Recall의 값이 클수록 합성 이미지의 다양성이 높다고 판단한다.

$$Precision = \frac{TP}{TP+FP} = \frac{\text{실제 이미지 분포에 포함된 합성 이미지 수}}{\text{전체 합성 이미지 수}}$$

$$Recall = \frac{TP}{TP+FN} = \frac{\text{합성 이미지로 재현된 실제 이미지 수}}{\text{전체 실제 이미지 수}}$$

반면에 Precision 및 Recall은 특정 클래스에 대한 모델의 예측 정확성을 측정하는 지표로서 데이터 분포를 비교하는 데는 적합하지 않고, 이상치 및 클래스 불균형에 취약하다. 뿐만 아니라, 데이터가 균일하게 밀집되어 있다고 가정하여 그렇지 않은 경우를 고려하지 않은 한계가 있으며 초기 중심점 선택에 따라 결과값이 크게 달라질 수 있는 K-means 알고리즘을 사용했다는 문제점이 있다.

4. Improved Precision 및 Recall

위에서 언급한 문제점을 K-NN(K-Nearest Neighbor) 알고리즘을 사용하여 해결한 평가 지표가 Improved Precision 및 Recall이다. 먼저 Inception 네트워크의 convolutional layer를 사용하여 실제 이미지와 합성 이미지의 특징을 추출한 후 K-NN 알고리즘을 사용해서 실제 이미지 집합과 생성 이미지 집합의 영역을 근사한다. Improved Precision은 실제 이미지를 기준으로 K번째로 가까운 실제 이미지의 영역에 합성 이미지가 존재하는지를 이진값으로 구하여 평균을 계산하고, Improved Recall은 합성 이미지를 기준으로 K번째로 가까운 합성 이미지 영역에 실제 이미지가 존재하는지를 이진값으로 구하여 평균을 계산한다.[4]

$$Improved\ Precision(\Phi_r,\Phi_g) = \frac{1}{|\Phi_g|}\sum_{\phi_g\in\Phi_g} f(\phi_g,\Phi_r)$$

$$Improved\ Recall(\Phi_r,\Phi_g) = \frac{1}{|\Phi_r|}\sum_{\phi_r\in\Phi_r} f(\phi_r,\Phi_g)$$

Φ = 특징들의 집합, ϕ = 집합 Φ의 특징 원소

그러나 Precision 및 Recall 지표와 동일하게 이상치가 존재할 경우 평가 결과를 왜곡시킬 수 있다는 한계점이 있다.

5. Density 및 Coverage

Density는 Precision 지표의 문제점인 이상치로 인해 실제 데이터의 manifold가 과대평가되는 현상을 보정하는 평가 지표이다. Precision은 합성 이미지 데이터가 실제 데이터 경계에 포함되는지를 이진 분류로 평균 내어 계산했지만, Density는 실제 데이터 경계에 포함된 합성 이미지 데이터의 수를 측정하여 Precision보다 세밀하게 평가할 수 있다.

Coverage는 위와 마찬가지로 Recall 지표의 문제점인 합성 이미지 데이터의 manifold가 과대평가되는 현상을 보정하는 평가 지표이다. Recall은 합성 이미지 데이터의 manifold를 실제 데이터의 근처에 구성하여 측정했기 때문에 합성 이미지 데이터의 manifold가 실제 manifold와 다를 경우 정확한 결과를 도출하지 못하는 문제점이 있었다. 이를 개선하기 위해 실제 데이터의 manifold를 기준으로 합성 이미지 데이터의 manifold를 구성하여 기존 Recall 지표의 한계를 극복하였다.

$$Density = \frac{1}{kM}\sum_{j=1}^{M}\sum_{i=1}^{N} 1_{Y_j\in B(X_i,NND_k(X_i))}$$

$$Coverage = \frac{1}{N}\sum_{i=1}^{N} 1_{\exists\, js.t.\ Y_j\in B(X_i,NND_k(X_i))}$$

k = 이웃의 수

Coverage의 경우 0에서 1 사이의 값을 가질 수 있으며, 값이 1에 가까울수록 다양성이 높음을 의미한다. 하지만 Density의 경우 값이 100을 넘을 수가 있어 절대적인 측정값으로 표현이 불가하다는 단점이 있다.[5]

Ⅲ. 결론

본 논문에서는 합성 이미지 데이터의 충실도와 다양성 측면에서 품질평가 지표별 특징 및 한계점에 대해 살펴보았다. 해당 지표들은 각기 다른 제한점이 있으므로 여러 지표를 혼합하여 합성 이미지 데이터의 품질을 파악하고 관리하는 것이 좋다. 또한 실제 합성 이미지 데이터에 적합한 평가 지표를 적용하여 신뢰할 수 있는 평가를 수행하고 해당 결과를 활용한 품질관리를 기대한다. 향후 연구에서는 지표별 한계점을 개선한 평가 지표를 탐색하여 고품질의 합성 데이터 구축에 기여하고자 한다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 초거대AI 확산 생태계 조성 사업(2100-2131-305, 2024년 초거대AI 확산 생태계 조성 사업)에 의해서 수행되었습니다.

참 고 문 헌

[1] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. “Improved Techniques for Training GANs,” Proc.NIPS ‘16, Jun. 2016.

[2] 장유진, 유재준, 홍헬렌, “GAN 기반 의료영상 생성 모델에 대한 품질 및 다양성 평가 및 분석”, 한국컴퓨터그래픽스학회, pp. 11-19, May 2022.

[3] Heusel, M., Ramasuer, H., Unterthiner, T., and Nessler, B. “GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium,” Proc.NIPS ‘17, Jan. 2018.

[4] Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., and Aila, T. “Improved Precision and Recall Metric for Assessing Generative Models,” Proc.NIPS ‘19, Oct. 2019.

[5] Naeem, M., Oh, S., Uh, Y., Choi, Y., and Yoo, J. “Reliable Fidelity and Diversity Metrics for Generative Models,” Proc.ICML ‘20, Jun. 2020.