조건부 변분 오토인코더(CVAE) 구현 및 실험 보고서

나노전자물리학과 20201914 박규민

April 9, 2025

1 서론

본 과제의 목표는 MNIST 숫자 이미지 데이터셋을 기반으로 조건부 변분 오토인코더(Conditional Variational AutoEncoder, CVAE)를 구현하고, 이를 통해 class label(숫자 0부터 9까지)에 조건을 부여한 이미지 생성 모델을 학습하는 것이다. CVAE는 입력 이미지와 조건 정보를 함께 인코더에 전달하여 잠재 공간(latent space) 상에서의 분포를 학습하며, 이후 디코더는 해당 조건에 맞는 이미지를 생성할 수 있도록 훈련된다.

학습이 완료된 후, 모델의 생성 성능을 확인하기 위해 조건(label)을 입력으로 하여 다양한 숫자 이미지를 생성하는 실험을 진행하였다. 본 보고서에서는 CVAE의 전체 구조 설계, 학습 과정, 그리고 생성 결과에 대한 분석을 중심으로 과제 수행 내용을 정리한다.

2 모델 설계

본 과제에서 구현한 CVAE 모델은 조건 정보(숫자 클래스)를 입력으로 받아, 해당 조건에 맞는 MNIST 숫자 이미지를 생성하는 구조로 설계되었다. 전체 네트워크는 인코더, 잠재 공간 샘플링 과정, 그리고 디코더의 세 부분으로 구성된다.

인코더 구조

인코더는 입력 이미지와 조건 벡터를 결합하여 잠재 벡터의 평균(mean)과 로그 분산(log variance)을 추정하는 역할을 한다. 입력 이미지 $x \in R^{28\times 28}$는 일렬로 펼쳐져 784차원 벡터로 변환되고, 조건 라벨 $c \in \{0, 1, \ldots, 9\}$는 원-핫 인코딩으로 변환되어 10차원 벡터로 표현된다. 이 둘을 concat하여 총 794차원의 입력으로 구성된 후 MLP에 전달된다.

원-핫 인코딩된 조건 벡터를 입력 이미지와 함께 concat하는 방식은, 각 클래스 간의 명확한 구분을 유지하면서도 조건 정보를 모델에 명시적으로 전달할 수 있는 간단하고 효과적인 방법이다. 특히 MLP 구조에서는 채널 기반의 조건 주입이 어려우므로, 이처럼 벡터 차원에서 조건을 결합하는 방식이 적합하다.

인코더 MLP는 총 3개의 선형 계층으로 구성되며, 각 층마다 Layer Normalization, ReLU 활성화 함수, 그리고 Dropout(0.25)이 적용되었다. 최종적으로 128차원 중간 표현을 얻은 후, 각각 독립적인 선형 계층을 통해 평균 벡터 μ와 로그 분산 벡터 $\log \sigma^2$를 생성한다. 이들은 latent 공간 차원인 2차원으로 투영된다.

잠재 공간 샘플링

추출된 평균과 로그 분산을 이용해 reparameterization trick을 적용하여 $z = \mu + \sigma \cdot \epsilon$ 형태로 샘플링을 수행한다. 여기서 $\epsilon \sim \mathcal{N}(0, 1)$이고, $\sigma = \exp(0.5 \cdot \log \sigma^2)$이다. latent 공간의 차원을 2로 설정하여 시각적 조작 및 해석이 용이하도록 하였다.

디코더 구조

디코더는 샘플링된 z 벡터와 조건 c를 결합하여 원본 이미지 x를 복원 또는 생성하는 역할을 한다. z와 c를 concat한 후, 총 12차원의 입력이 4단계의 선형 계층을 거치며 다시 784차원의 출력으로 확장된다. 각 계층마다 Layer Normalization, ReLU, Dropout(0.25)이 적용되며, 마지막 출력 계층에는 시그모이드 함수를 사용하여 픽셀 단위의 확률 분포를 모델링하였다.

전체 모델은 PyTorch Lightning 기반으로 구현되어 학습 과정의 모듈화와 자동 로깅, 모델 저장이 용이하도록 구성하였다.

3 학습 설정 및 진행

본 과제에서는 PyTorch Lightning의 Trainer 클래스를 이용하여 모델 학습을 수행하였다. 학습 데이터는 torchvision 패키지의 MNIST 숫자 이미지 데이터셋을 사용하였고, 전처리는 `ToTensor()`만을 적용하였다. 학습 설정은 아래와 같다.

- Optimizer: Adam
- Learning rate: 0.001
- Epochs: 10
- Loss function: MSE Loss
- Batch size: 100

모델의 학습은 입력 이미지 x와 레이블 c를 조건으로 하여 진행되며, `forward()` 함수 내부에서 인코더를 통해 잠재 변수의 평균과 로그 분산을 구하고, reparameterization trick을 이용하여 잠재 변수 z를 샘플링한다. 이후, z와 조건 c를 다시 디코더에 전달하여 복원 이미지 $\hat{x}$를 생성하며, 손실 함수는 원본 이미지 x와 $\hat{x}$ 간의 평균 제곱 오차로 계산된다.

전체 학습 파라미터 수는 약 1.1M이며, 학습 완료 후 모델의 가중치는 `cvae.ckpt` 파일로 저장되었다.

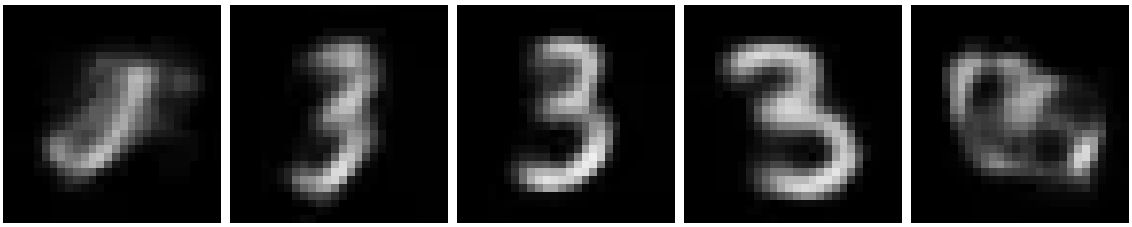
4 결과 분석

4.1 Gradio를 통한 이미지 생성

학습된 CVAE 모델을 기반으로 Gradio 인터페이스를 구성하여, 사용자로부터 class label과 latent vector를 입력받아 실시간으로 이미지를 생성할 수 있도록 구현하였다. 사용자는 슬라이더를 통해 latent vector의 두 차원(z_0, z_1)을 조절할 수 있으며, 그에 따라 출력 이미지가 동적으로 변화하는 것을 확인할 수 있다.

4.2 생성 이미지 예시

Figure 1은 label=3으로 고정한 후, z_0, z_1 값을 변화시키며 생성한 이미지 예시이다. 연속적인 latent 공간 내 이동을 통해 출력 이미지가 어떻게 달라지는지를 확인할 수 있다.



$z_0 = -5,\ z_1 = -5$ $z_0 = -1,\ z_1 = -1$ $z_0 = 0,\ z_1 = 0$ $z_0 = 1,\ z_1 = 1$ $z_0 = 5,\ z_1 = 5$

Figure 1: Latent vector 변화(z_0, z_1)에 따른 이미지 생성 예시

위 그림은 label을 3으로 고정한 상태에서 latent vector의 값을 점진적으로 변화시키며 생성한 이미지들이다. 왼쪽부터 오른쪽으로 갈수록 (z_0, z_1) 값이 $(-5, -5)$에서 $(5, 5)$로 변화하며, 생성된 이미지의 형태 또한 함께 변화하는 것을 확인할 수 있다.

특히 $(z_0, z_1) = (-5, -5)$나 $(5, 5)$와 같이 latent 공간의 극단에 위치한 경우에는 숫자의 형태가 왜곡되거나 불분명해지는 경향을 보였다. 반면, $(z_0, z_1) = (0, 0)$ 주변의 중앙 영역에서는 가장 명확하고 전형적인 숫자 '3'이 생성되었다.

이러한 결과는 CVAE 모델이 latent 공간 내에서 의미 있는 표현 구조를 학습하고 있음을 보여준다. latent vector의 연속적인 변화가 이미지의 스타일, 획의 굵기, 기울기 등 다양한 표현적 특성에 영향을 미치며, 이는 생성된 이미지가 단순한 복제물이 아니라 잠재 표현에 기반한 결과임을 시사한다.

4.3 생성 이미지 분석

Latent space 상의 연속적인 변화를 주었을 때, 출력되는 이미지의 스타일, 획의 두께, 숫자의 기울기 등 세부적인 표현이 자연스럽게 달라지는 것을 확인할 수 있었다. 이는 단순한 숫자 모양 생성뿐만 아니라, CVAE가 latent 변수에 따른 다양한 표현 양상을 학습하고 있음을 의미한다. 또한, 동일한 label을 유지하더라도 latent vector의 값에 따라 시각적 표현이 다채롭게 나타났다는 점도 관찰되었다.

Figure 2는 이러한 관찰을 보다 정량적, 시각적으로 뒷받침하기 위해 이전 과제의 오토인코더로부터 추출한 3차원 latent vector를 t-SNE를 통해 2차원 공간에 시각화한 것이다. 그 결과, label 4와 label 9의 latent 표현이 서로 밀접하게 분포되어 있는 것을 확인할 수 있었다. 이는 두 숫자의 형태적 유사성이 모델의 latent 공간에서도 반영되었음을 시사하며, 두 클래스 간의 잠재적 모호성이 존재할 가능성을 제기한다.

이를 바탕으로 해당 절에서는 label 4와 9를 조건으로 한 생성 이미지 비교를 통해, latent 공간에서의 거리 기반 유사성이 실제 생성 이미지의 시각적 유사성과 어떻게 연결되는지를 분석하고자 한다.

먼저, 동일한 latent vector z를 기준으로 label이 4인 경우와 9인 경우를 각각 조건으로 주어 생성된 이미지를 비교하면, 전반적으로 유사한 형태적 특징을 공유함을 확인할 수 있다. 특히 z = (0, 0), (-1, -1), (1, 1) 등의 중심 근처 좌표에서는 생성된 숫자들이 모두 비교적 명확하고 전형적인 형태로 나타난다. 이 영역은 latent space에서 데이터가 밀집된 부분으로, CVAE 모델이 보다 안정적인 표현을 학습한 것으로 해석된다.

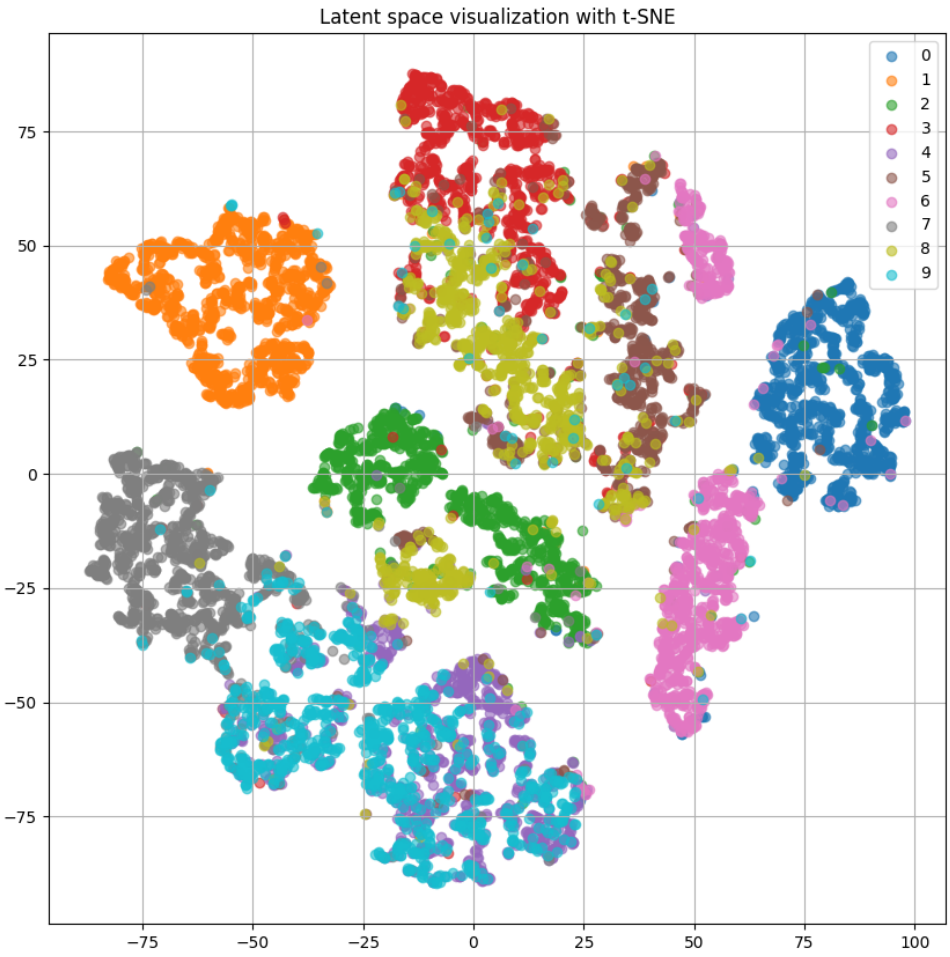
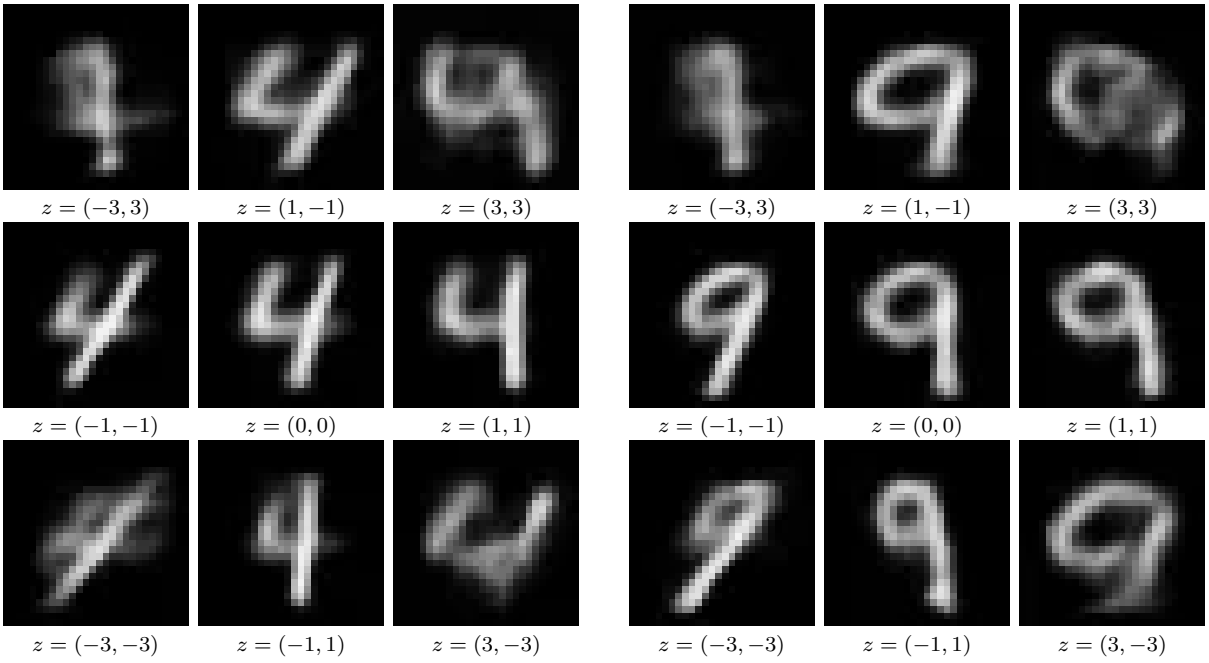


Figure 2: Autoencoder의 latent vector 시각화 결과 — label 4와 9가 근접한 위치에 분포함



Label 4 생성 결과 (9개 latent vector)

Label 9 생성 결과 (9개 latent vector)

Figure 3: Label 4와 Label 9의 동일 latent vector에 대한 생성 이미지 비교

반면, z = (-3, -3)이나 (3, 3), (3, -3), (-3, 3)과 같이 값이 크고 주변부에 위치한 latent vector에서는 label에 상관없이 생성 이미지의 품질이 저하되며, 형태가 서로 유사하게 뭉개지거나 기울어지는 현상이 관찰된다.

예를 들어 z = (-3, -3)에서 생성된 label 4와 label 9의 이미지는 모두 기울어진 형태를 보이며, 숫자 고유의 구조보다는 공통적인 스타일이 강조되는 양상을 띤다.

z = (-3, 3)의 경우, 두 숫자 모두 전체적으로 윤곽이 뭉개지면서도 숫자의 형태가 압축되어 좁아지는 느낌을 준다. 마치 외곽의 선이 흘러내리듯 번지며 형태 구분이 어려워진다. 이 좌표에서는 label 4와 label 9의 생성 이미지가 거의 동일하게 나타나며, 실제로 구분이 어려울 정도로 유사한 외형을 가진다. 이는 해당 latent vector가 두 label 간의 구별 정보를 충분히 반영하지 못한 채, 스타일 정보만을 반영하여 이미지를 생성했음을 시사한다.

z = (3, 3)에서는 전체적으로 선이 흐려지고 곡선의 왜곡이 심화되며, 숫자 형태가 불분명해지는 특징을 보인다. z = (3, -3)에서는 두 숫자 모두 글자의 형태가 전체적으로 퍼지고 번져 흐려지는 경향을 보인다. 선의 두께나 형태가 명확하게 닿지 않고 뭉개지며, 숫자 구조 자체가 흐릿해지는 양상이 나타난다. 이는 CVAE가 이러한 영역에서 학습한 표현이 불안정함을 시사한다.

이러한 결과는 CVAE가 latent 공간의 위치에 따라 공통적인 시각적 특성을 생성하고 있으며, 조건(label)에 따라 그 위에 각 숫자에 특화된 구조를 추가적으로 형성하고 있음을 시사한다. 즉, latent vector는 주로 글자의 굵기, 기울기, 획의 방향과 같은 시각적 스타일을 결정하고, 조건 정보는 숫자의 의미를 결정하는 역할을 수행하는 것이다.

결과적으로 동일한 z에 대해 label만 바꾸어 생성한 이미지들을 비교한 결과는, CVAE가 latent 공간의 연속성과 조건 기반 표현 모두를 일정 수준 이상 학습했음을 보여준다. 이러한 구조는 향후 특정 스타일을 유지한 채 숫자만 바꾸거나, 반대로 특정 숫자에서 다양한 스타일을 생성하는 응용 가능성을 열어준다.

5 결론 및 소감

본 과제에서는 조건부 변분 오토인코더(CVAE)를 직접 구현하고, MNIST 숫자 이미지에 조건(label)을 부여하여 다양한 이미지를 생성하는 실험을 수행하였다. 모델은 MLP 기반의 인코더와 디코더로 구성되었으며, latent 공간의 차원을 2로 설정함으로써 시각적인 해석이 가능하도록 하였다.

실험 결과, CVAE는 주어진 조건에 따라 해당 숫자의 전형적인 형태를 생성하는 데 성공하였고, latent vector의 연속적인 변화에 따라 이미지의 획 두께, 기울기, 윤곽 등 이미지의 외형이 변화하는 양상을 보였다. 특히 동일한 latent vector에 대해 label만 바꾸었을 때 생성 이미지 간 유사성이 관찰되었으며, 이를 통해 latent vector는 이미지의 시각적 스타일을, 조건(label)은 숫자의 구조적 의미를 결정한다는 구조적 특성이 확인되었다.

또한, latent 공간의 중심부에서는 비교적 뚜렷한 숫자 형태가 생성되었으나, 주변부에서는 label에 관계없이 불분명하거나 유사한 이미지가 생성되는 경향이 나타났다. 이는 모델이 중심 영역에서 더 안정적인 표현을 학습하는 반면, 주변부에서는 조건 정보가 충분히 반영되지 못할 수 있음을 시사한다.

더 나아간다면 CNN 기반 네트워크로 모델 구조를 확장하거나, KL divergence 항을 추가하여 정식 VAE 구조로의 발전을 시도할 수 있을 것이다. 더불어 latent 공간에 의미 있는 제약을 추가하거나, 다양한 조건 정보를 활용하여 생성 모델의 표현력을 높이는 방향으로 실험을 확장할 수 있을 것이다.

이번 과제를 통해 단순한 이론 이해를 넘어, 실제로 조건부 생성 모델을 설계하고 구현하는 전 과정을 경험할 수 있었다. 특히 latent 공간의 구조가 생성 이미지에 어떻게 반영되는지를 시각적으로 확인하는 과정은 매우 흥미로웠고, 생성 모델이 갖는 표현력의 특징을 보다 깊이 이해할 수 있었다. 향후 생성 모델에 대한 다양한 관심과 이해를 가질 수 있는 의미 있는 과제였다.