Journal of Information Technology and Architecture
Vol. 20. No. 4, December 2023, Pages 341-350
DOI http://doi.org/10.22865/jita.2023.20.4.341

시각화 오토인코더를 사용한 학습된 심층신경망의 잠재공간 조작 시스템

고영민[1] · 이은주[1] · 민정익[2] · 고선우[2]*

[1]전주대학교 문화기술학과 박사과정, 인공지능연구소
[2]전주대학교 인공지능학과, 인공지능연구소

gjtrj55@jj.ac.kr, leeeunju@jj.ac.kr, minji@jj.ac.kr, godfriend@jj.ac.kr

(2023년 11월 13일 접수; 2023년 12월 14일 수정; 2023년 12월 25일 채택)

요약: 최근에 생성모델, 특히 ChatGPT와 DALLE-3과 같은 생성모델들은 사회에 큰 영향을 미치며, 이러한 모델들의 잠재공간을 정교하게 조작하고 이해하는 기술이 중요해지고 있다. 본 논문에서는 학습된 생성모델의 잠재공간을 시각화하여 조작하는 새로운 방법인 “시각화 오토인코더” 방법을 제안한다. 이 방법은 생성모델의 잠재벡터를 입력으로 하여 3차원 이하의 시각화 가능한 차원으로 축소하여 조작하는 것을 목표로 한다. 본 연구에서는 생성모델의 잠재공간 조작을 “관심 있는 샘플 생성”과 “샘플 간 변환 과정 시각화” 두 가지로 정의하고, 이를 위해 필요한 수학적 성질 “일대일 대응”과 “locally smoothness”를 가정한 심층신경망 기반의 생성모델로 제한한다. MNIST 데이터셋을 사용하여 실험한 결과 이러한 성질을 만족하는 심층신경망 모델에 대해 제안된 시각화 오토인코더 방법으로 관심 있는 샘플을 생성, 변환 과정을 시각화할 수 있음을 확인하였다.

주제어: 시각화 오토인코더, 잠재공간 조작, 심층신경망, 생성모델

A Method for Latent Space Control System of a Learned Deep Neural Network Using a Visualization Autoencoder

Young-Min Ko[1], Eun-Ju Lee[1], Jeong-Ik Min[2] and Sun-Woo Ko[2]*

[1]Ph.D Student, Dept. of Culture Technologyand Artificial Intelligence Research Center JeonJu University
[2]Professor, Dept. of Artificial Intelligence and Artificial Intelligence Research Center, JeonJu University

(Received November 13, 2023; Revised December 14, 2023; Accepted December 25, 2023)

Abstract: Recent generative models, especially generative models such as ChatGPT and DALLE-3, have had a great impact on society, and technology to intricately manipulate and understand the latent space of these models is becoming important. In this paper, we propose the “Visualization Autoencoder” method, a new method that visualizes and manipulates the latent space of the learned generative model. This method aims to manipulate the latent vectors of the generative model by reducing them to three or less dimensions that can be visualized. In this study, the latent space manipulation of the generative model is defined as two types: “generating samples of interest” and “visualizing the transformation process between samples”. For this purpose, we limit our-

*Corresponding Author
본 연구는 2023년도 중소벤처기업부의 기술개발사업 지원에 의한 연구임 [00207554]

selves to a deep neural network-based generative model that assumes the necessary mathematical properties “one-to-one correspondence” and “locally smoothness”. As a result of experiments using the MNIST dataset, we confirmed that the proposed visualization autoencoder method for a deep neural network model that satisfies these properties can generate samples of interest and visualize the conversion process.

Keywords: Visualization Auto-Encoder, Latent space control system, Deep Neural Network, Generative Model

1. 서 론

최근에 텍스트를 생성하는 Chat GPT, 이미지를 생성하는 DALLE-3를 비롯한 다양한 생성모델이 사회 전반에 큰 영향을 미치고 있다. 이에 사용자가 원하는 데이터를 더욱 정교하게 생성해주는 조작 시스템이 중요해지고 있으며, 이런 조작을 위해 생성모델의 잠재공간을 이론적으로 이해하고 발전시키는 연구들이 활발히 진행되고 있다[1].

생성모델의 잠재공간을 조작하는 연구는 크게 두 축으로 구분되며 하나는 생성모델의 잠재공간이 특정한 성질을 가지도록 설계하는 것이고[2], 다른 하나는 학습된 생성모델이 주어졌을 때 그 잠재공간을 조작하는 방법이다[1].

일반적으로 생성모델의 잠재공간의 차원은 3차원보다 큰 차원을 가지므로 잠재공간을 시각적으로 다루기 어렵다. 이에 본 연구는 학습된 생성모델에 대한 잠재공간 조작 시스템으로 시각화 오토인코더 방법을 제안한다. 제안된 방법은 학습된 생성모델의 잠재공간 내에 있는 잠재벡터를 3차원 이하의 시각화 가능한 오토인코더로 학습하여 조작하는 방법이다.

이를 설명하기 위해 생성모델의 잠재공간 조작을 두 가지(관심있는 샘플 생성과 샘플 간 변환 과정 시각화)로 정의하고 이 논리를 뒷받침하기 위해 필요한 수학적 성질 두 가지(일대일 대응과 locally smoothness)를 가정한 학습된 심층신경망을 기반으로 하는 생성모델로 제한한다.

2. 관련 연구

학습 과정 중 잠재공간을 조작하고자 시도한 연구들[2,3,4]의 경우 잠재공간을 효과적으로 다루고자 모델의 구조 변형 및 학습 알고리즘을 제안[5,6]하였으며 FactorVAE[7], FlatGAN[8], GLOW[9] 등이 있다. Chi-Hieu Pham et al.[10]의 경우 공분산 손실 항을 사용해 잠재공간을 조직화하고 다양한 속성을 분리하여 이미지의 크기 및 회전과 같은 다양한 기하학적 특성을 조작할 수 있는 의미있는 축을 발견하였다. Irina Higgins et al.[11]는 데이터의 독립적인 생성 요소를 해석 가능한 요소로 분해하고자 VAE에 조절 가능한 하이퍼파라미터 β를 도입하여 더 넓은 범위의 변이 요소 값을 포함하여 더욱 깨끗하게 해체되는 표현을 학습하였다. Tero Karras et al.[12]은 입력 잠재코드를 중간 잠재공간에 내장하여 선형적이고 덜 엉킨 표현을 허용함으로써 이미지의 특성을 조절할 수 있는 방법인 StyleGAN을 제안하였으며, Taesung Park et al.[13]는 이미지의 구조와 텍스처를 분리하고 새로운 입력의 잠재코드를 효과적으로 찾을 수 있도록 조작에 특화된 모델인 Swapping Autoencoder 제안하였다.

학습이 끝난 생성모델에 대해 잠재공간을 조작하고자 시도한 연구들의 경우 사전 생성모델의 특성을 깊게 분석하여 특정 방향을 탐색하여 조작하고자 하였다. 이준하 등[14], LowRankGAN[15], SeFA[16], AdvStyle[17] 등이 제안되었으며, 이미지의 의미있는 변환을 반영하기 위해 시도하였고[18,19], 의미론적 요소를 탐색하고자 하였다[20,21]. Nurit Spingarn Eliezer et al.[22]은 조작경로를 훈련이나 최적화 없이 직접 생성자의 가중치에서 폐쇄된 형태로 계산할 수 있음을 보여주며, 기존 방법보다 더욱 빠르고 의미있는 방향을 찾을 수 있음을 말하였다. Xuanchi Ren et al.[1]은 이미지 변형을 강조하는 첫 번째 통합 프레임 워크를 제시하였고, Erik Härkönen et al.[23]는 PCA를 기반으로 잠재방향을 식별하여 모델의 레이어별로 다양한 해석 가능한 방향들을 발견하였다. Or Patashnik et al.[24]는 CLIP(Contrastive Language-Image Pre-training)를 활용하여 직관적인 텍스트 기반 의미론적 이미지 조작을 가능하게 하는 StyleCLIP을 제안하였다. Weili Nie et al.[25]은 제어 가능한 고품질 생성과 속성들의 합성 생성을 다루기 위해 StyleGAN[12]과 같은 사전 학습된 생성 모델의 잠재 공간에 EBM

(Energy-Based Models)을 도입하여 고품질 이미지 제어를 가능하게 하였다. Zongze Wu et al.[26]은 의미있는 조작을 위해 StyleGAN2[27]의 잠재공간인 StyleSpace를 탐색하고 분석하였고, 특정 의미 영역에서 일관되게 활성화되는 채널을 발견하여 시각적 속성을 조작할 수 있음을 말하였다.

위와 같이 진행된 연구들은 많은 실험들을 근거로 제시하여 생성 모델의 잠재공간을 원하는 목적에 맞게 조작할 수 있음을 보여주고 있다. 기존연구들[15,16,19,21]은 학습된 생성모델의 잠재공간이 인간이 해석 가능한 변환을 가지고 있다는 것을 실험을 통해 발견하여 이를 근거로 연구들이 진행되었다. 하지만 학습된 생성모델의 잠재공간을 좀 더 세밀하게 다루려면 이론적 안정성이 뒷받침되어야 한다.

이에 본 연구는 기존 연구와 다르게 잠재공간을 다루기 위해 필요한 성질인 일대일 대응과 locally smoothness 성질을 만족하는 학습된 생성모델로 제한하여 잠재공간을 조작하기 위한 시각화 오토인코더 방법을 연구하였고 이론적 해석을 제공한다.

3. 시각화 오토인코더를 사용한 학습된 심층신경망의 잠재공간 조작 시스템

3.1 학습된 심층신경망의 잠재공간 조작 정의와 필요한 성질

본 연구에서 학습된 심층신경망의 잠재공간을 조작한다는 것은 다음 두 가지 행위로 정의한다. 첫 번째는 잠재공간의 잠재벡터를 조작하여 학습된 심층신경망의 관심 있는 샘플들 간의 변환 과정을 시각화하는 것이고, 두 번째는 학습된 심층신경망의 잠재공간의 잠재벡터를 조작하여 관심 있는 샘플을 생성하는 것이다. 이와 같은 조작을 하기 위해서 학습된 심층신경망이 필요로 하는 수학적 성질 2가지를 가정한다.

3.1.1 일대일 대응 성질

일반적인 학습된 심층신경망은 n_0개 특징을 가지는 데이터 $z^{(1)} \in R^{n_0}$을 입력으로 하는 입력층과 L개의 은닉층, 그리고 출력층으로 구성되어있다. 구체적으로 임의의 l번째 은닉층의 입력으로 들어가는 잠재벡터 $z^{(l)} \in R^{n_{(l-1)}}$은 선형변환 $W^{(l)}:R^{n_{(l-1)}} \to R^{n_l}$ 과 비선형변환 $\sigma^{(l)}:R^{n_l} \to R^{n_l}$을 거쳐 $z^{(l+1)}$을 출력하고, 마지막 L번째 은닉층의 출력 $z^{(L+1)} \in R^{n_L}$은 출력층의 선형변환을 통해 $z^{(out)} \in R^{n_{(L+1)}}$을 출력한다.

만약 학습된 심층신경망이 일대일 대응 성질을 보장할 수 없는 경우 잠재벡터 $z^{(l)}$이 $z^{(out)}$으로 매핑되었을 때 정확히 추적할 수 없는 위험성이 존재한다. 이 문제를 해결하기 위해 학습된 심층신경망의 임의의 l번째 선형변환의 rank가 $\min[n_{(l-1)}, n_l]$을 가진다고 하자. 그리고 l번째 비선형변환이 일대일 대응 성질을 만족하는 활성 함수, 예를 들어 Tanh 함수 또는 파라미터를 적용한 abTanh[28] 등을 사용하였다고 가정하자.

3.1.2 locally smoothness 성질

두 번째는 임의의 l번째 은닉층의 입력으로 잠재벡터 $z^{(l)}$에 임의의 작은 실수 벡터 ε만큼 변화하였을 때, 학습된 심층신경망의 출력값 $f(z^{(l)}+\varepsilon)$과 $f(z^{(l)})$의 차이를 나타내는 거리함수 d값이 어떤 양의 상수 δ값보다 큰 경우 잠재벡터의 변화에 대한 안정성을 보장할 수 없다. 이를 방지하기 위해 학습된 심층신경망이 locally smoothness 식 (1)을 만족한다고 가정한다.

$$d\left(f\left(z^{(l)}\right), f\left(z^{(l)}+\epsilon\right)\right) < \delta \tag{1}$$

위 두 가지 가정을 만족하는 학습된 심층신경망을 보다 자세히 이해해보자. 예를 들어, 주어진 손실함수 L에 대해 적당한 양의 상수 η값 보다 작게 학습된 일대일 대응, locally smoothness 성질을 만족하는 심층오토인코더 f가 주어졌다고 해보자. 학습된 심층오토인코더 f는 인코더 함수 f_e와 디코더 함수 f_d로 구성되어 있고, n_0개 특징을 가지는 n개의 입력 데이터 $\{z_i^{(1)}\}_{i=1}^n$을 복원된 데이터 $\{\tilde{z}_i^{(1)}\}_{i=1}^n$로 출력한다(식 (2,3)).

$$L\left(z_i^{(1)}, \tilde{z}_i^{(1)}\right) < \eta, i = 1,...,n \tag{2}$$

$$\tilde{z}_i^{(1)} = f_d\left(f_e\left(z_i^{(1)}\right)\right) \equiv f\left(z_i^{(1)}\right), i = 1,...,n \tag{3}$$

이를 기하학적 관점으로 이해하기 위해 28×28 픽셀을 갖는 숫자 이미지 MNIST 데이터셋을 입력($n_0 = 784$)으로 하는 f를 생각해보자. 모집단 P는 0부터 9까지 n_0 차원에 있는 숫자 이미지이고, 입력 공간인 n_0 차원에서 각 숫자에 대응하는 여러 그룹의 모집단이 있을 수 있다<Figure 1>. f는 입력 공간에서 정의된 모집단 P 중 n개의 관측된 샘플들 $z_i^{(1)}$이 f_e을 거쳐 n_l차원의 잠재공간의 잠재벡터 $z_i^{(l+1)}$로 매핑한 후, f_d을 통해 다시 입력 공간으로 복원하는 $\tilde{z}_i^{(1)}$을 학습하

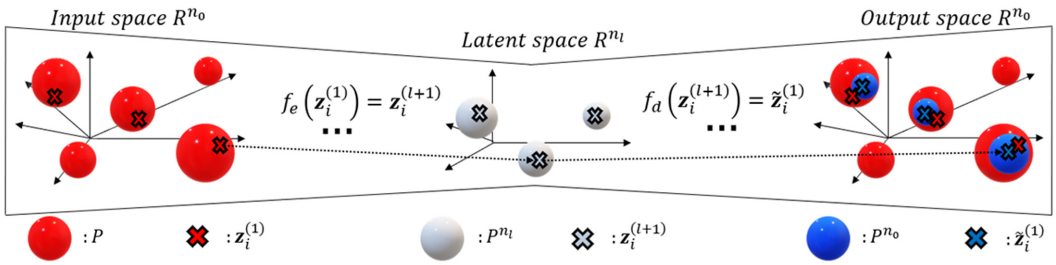


Figure 1. A Learned Deep Autoencoder(f), P : Population, P^{n_0} : A Subset of the Population Generated by f, P^{n_l} : A Subset in the R^{n_l} that Satisfies the One-to-one Correspondence and Locally Smoothness Properties for P^{n_0}, $z_i^{(1)}$: Observed Samples, $\tilde{z}_i^{(1)}$: Reconstructed $z_i^{(1)}$, $z_i^{(l+1)}$:$z_i^{(1)}$ Mapped to R^{n_l} .

였다고 할 수 있다.

이때 식 (2)을 만족한다는 의미가 <Figure 1>에서 $\tilde{z}_i^{(1)}$ 이 P에 속하면서 $z_i^{(1)}$ 와 충분히 가깝게 학습된 경우라고 해보자. 그리고 n_l차원의 잠재벡터 $z_i^{(l+1)}$ 에 임의의 작은 실수 벡터 ε만큼 더해진 벡터들의 집합 P^{n_l} 이 f_d을 통해 n_0차원으로 매핑된 벡터들의 집합 P^{n_0} 을 생각해볼 수 있다. 이때 잠재공간에서의 P^{n_l} 은 일대일 대응 성질로 인해 입력 공간의 P^{n_0} 에 일대일 대응되고, locally smoothness 성질에 의해 P^{n_0} 은 P에 속할 수 있다.

3.2 시각화 오토인코더를 사용한 잠재공간 조작 방법

이제 n차원의 P^{n_l} 을 통해 샘플 $z_i^{(1)}$ 와 비슷한 샘플들 P^{n_0} 을 생성할 수 있지만, $n \gg 3$ 인 경우 P^{n_l} 을 시각적으로 다룰 수 없다. 이를 해결하기 위해 이번 절에서는 시각화 오토인코더를 사용한 잠재공간 조작에 대해 설명한다.

본 논문에서 시각화 오토인코더 v란 일대일 대응과 locally smoothness 성질을 만족하는 시각화 가능한 잠재 차원을 가지는 오토인코더이다. 구체적으로 임의의 n_l차원을 입력으로 하여 인코더 함수 v_e을 통해 3차원 이하의 시각화 가능한 잠재 차원 n_v로 매핑 후, 디코더 함수 v_d을 통해 n_l차원으로 복원하는 오토인코더이다.

이를 기하학적으로 이해하기 위해 $n_v = 2$을 가지는 v을 표현한 그림은 <Figure 2>이다. <Figure 2>에서 $z_i^{(l+1)}$ 을 v_e을 통해 2차원 잠재벡터 z_i로 매핑 후, v_d을 통해 $\tilde{z}_i^{(l+1)}$ 로 복원한 것이다. 이때 z_i을 포함하는 2차원 벡터들의 집합 P^{n_v} 는 v_d을 통해 n_l차원에서 $\tilde{z}_i^{(l+1)}$ 을 지나는 벡터들의 집합 $P_v^{n_l}$ 와 일대일 대응하고 곡면을 이룬다.

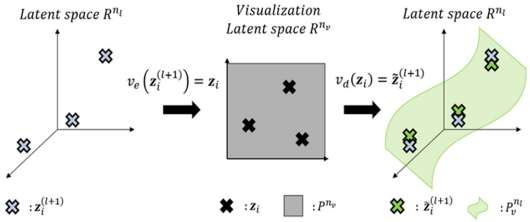


Figure 2. Visualization Autoencoder(v), a Structure that Takes a Latent Space P^{n_l} as Input, Encodes it into Visualizable P^{n_v} Dimensions, and Reconstructs it Into P^{n_l}.

$$z_i = v_e\left(z_i^{(l+1)}\right), i = 1,\ldots,k \tag{4}$$

$$\tilde{z}_i^{(l+1)} = v_d\left(z_i\right), i = 1,\ldots,k \tag{5}$$

$$P_v^{n_l} = v_d\left(P^{n_v}\right) \tag{6}$$

이를 3.1절의 학습된 심층오토인코더 f에 적용하면 다음과 같다. 먼저, f의 n_l차원의 잠재벡터 $z_i^{(l+1)}$ 중 관심 있는 k개($k \ll n$)을 입력으로 하고 $n_v = 2$을 갖는 v에 대해서 적당한 양의 상수 η_2값 보다 작은 손실함수 L값을 갖게 학습시켰다고 하자.

$$L\left(z_i^{(l+1)}, \tilde{z}_i^{(l+1)}\right) < \eta_2, i = 1,\ldots,k \tag{7}$$

k개의 $z_i^{(l+1)}$ 는 v_e을 통해 z_i로 매핑되고(식 (4)), z_i는 v_d을 통해 $\tilde{z}_i^{(l+1)}$ 로 매핑된다(식 (5)). $\tilde{z}_i^{(l+1)}$ 는 f_d을 통해 n_0차원의 $\tilde{z}_{i,v}^{(1)}$ 로 매핑된다(식 (8)).

$$f_d\left(\tilde{z}_i^{(l+1)}\right) = \tilde{z}_{i,v}^{(1)} \tag{8}$$

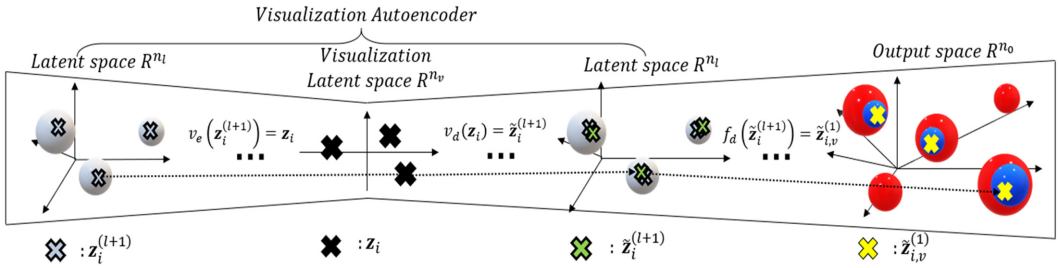


Figure 3. Data Reconstructed with the Trained Visualization Autoencoder(v) and Decoder (f_d).

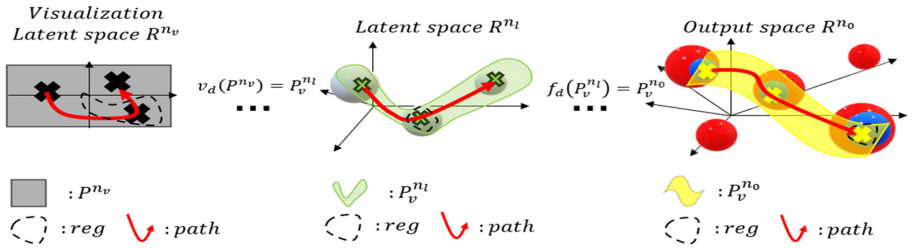


Figure 4. A Geometric Picture in which a Set P^{n_v} Defined within a visualization latent space R^{n_v} is represented in n_0 Dimensions by v_d and f_d.

이를 기하학적 관점에서 해석하면 다음 <Figure 3>과 같다. 시각화 가능한 n_v차원의 잠재벡터 z_i는 일대일 대응 성질로 인해 n_0차원에서 $\tilde{z}_{i,v}^{(1)}$에 대응한다. 이때 $\tilde{z}_{i,v}^{(1)}$는 v_d와 f_d의 locally smoothness 성질로 인해 샘플 $z_i^{(1)}$와 가까우면서 P에 속할 수 있다.

n_v차원에서 z_i을 포함하는 벡터들의 집합 P^{n_v}을 v_d을 통해 n_l차원의 $P_0^{n_l}$로 매핑하고(식 (6)), 다시 f_d을 통해 n_0차원의 $P_v^{n_0}$로 매핑한 것(식 (9))을 표현하면 <Figure 4>와 같다.

$$P_v^{n_0} = f_d\left(P_v^{n_l}\right) \tag{9}$$

<Figure 4>에서 P^{n_v}는 우리가 시각화하여 다룰 수 있는 z_i을 포함하는 집합으로써 일대일 대응 성질로 인해 n_0차원의 $P_v^{n_0}$에 일대일 대응한다. 이때 $P_v^{n_0}$는 n_0차원에서 곡면을 이루고 P^{n_v}는 다음과 같은 성질을 가진다. 첫 번째로 P^{n_v} 내에 z_i에 임의의 작은 실수 벡터 ε만큼 더해진 집합(<Figure 4>의 reg)은 일대일 대응과 locally smoothness 성질에 의해 n_0차원에서 $z_i^{(1)}$와 비슷한 샘플을 생성할 수 있다. 두 번째는 P^{n_v} 내에 z_i와 z_j 사이를 잇는 집합(<Figure 4>의 $path$)은 일대일 대응과 locally smoothness 성질에 의해 n_0차원에서 $\tilde{z}_{i,v}^{(1)}$와 $\tilde{z}_{j,v}^{(1)}$ 사이의 변환 과정을 시각화할 수 있다.

4. 실험

4.1 시각화 오토인코더의 기하학적 관점

시각화 오토인코더의 비선형 차원 축소 의미를 기하적으로 보기 위해 3차원에서 4개의 데이터를 입력으로 각각 2차원과 1차원차원으로 매핑한 후 복원하는 시각화 오토인코더를 실험하였다.

각각의 시각화 오토인코더는 손실함수 값이 0에 근사하도록 구조를 적절히 선택하여 학습하였고 결과는

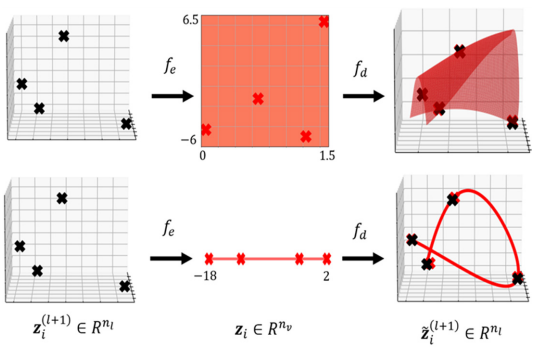


Figure 5. A Visualization Autoencoder for four Input Data in Three Dimensions, Visualized in Two and One Dimension Respectively.

<Figure 5>에 제시하였다. <Figure 5>의 위에 있는 그림은 2차원 잠재공간을 가지는 시각화 오토인코더로써 2차원에 매핑된 데이터가 각 축으로 [0,1.5], [6.5,-6] 범위 안에 매핑된 것을 볼 수 있다. 이때 이 범위를 f_d을 통해 3차원으로 복원하면 3차원 공간에서 휘어진 2차원 곡면을 만들어 내는 것을 볼 수 있다. 마찬가지로 <Figure 5>의 아래 그림은 1차원으로 축소한 것을 나타내며 [-18,2] 범위를 3차원에 복원하면 3차원 공간에서 1차원 곡선이 만들어진다.

4.2 MNIST 데이터를 사용한 시각화 오토인코더의 잠재공간 조작

시각화 오토인코더를 사용하여 잠재공간 조작 실험을 하기 위해 MNIST 숫자 데이터셋을 사용하였고 먼저, 학습된 심층오토인코더는 <Table 1>에 제시하였다.

손실함수는 Binary cross-entropy, Adam optimizer를 사용하여 50epoch 후 손실함수 값은 69.48을 가졌으며 원래 이미지와 학습 후 복원된 이미지를 <Figure 6>에 나타냈다. <Table 1>에서 필터 F, 커널사이즈 K, 스트라이드 S, 패드 pad을 가지는 합성곱과 노드 수 N을 가지는 Dense 구조를 나타내며 마지막 Sigmoid를 제외한 모든 층에 abTanh 함수를 사용하였다. <Table 1>의 W rank는 각 층에서 학습된 선형변환 파라미터 W의 min rank를 나타낸다.

학습된 심층오토인코더의 locally smoothness 성질을 보기 위해 10차원 잠재공간에 매핑된 임의의 학습 데이터 1개 $z_i^{(l+1)}$에 대해 3종류의 스케일링된 노이즈

Table 1. Learned autoencoder architecture

Layers	Learned autoencoder(f)	W rank
Encoder(f_e)		
Conv1	F=64, K=3, S=2, pad='valid'	9
Conv2	F=64, K=3, S=2, pad='valid'	64
Dense1	128	128
Dense2	10	10
Latent space=10		
Decoder(f_d)		
Dense1	128	10
Dense2	7×7×32	128
Dconv1	F=64, K=3, S=2, pad='same'	64
Dconv2	F=64, K=3, S=2, pad='same'	64
Dconv3	F=1, K=3, S=1, pad='same', Sigmoid	1

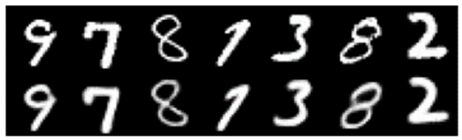


Figure 6. Top: Original Images, Bottom: Reconstructed Image.

Table 2. Table Showing the Locally Smoothness Properties

$d=l_2norm, \varepsilon \sim N(0,I)$	Mean	Std
$d(f_d(z^{(l+1)}),\alpha_1)$ $\alpha_1=f_d(z^{(l+1)}+(1/\|z^{(l+1)}\|_2)\varepsilon)$	0.002	0.001
$d(f_d(z^{(l+1)}),\alpha_2)$ $\alpha_2=f_d(z^{(l+1)}+(0.5/\|z^{(l+1)}\|_2)\varepsilon)$	0.016	0.004
$d(f_d(z^{(l+1)}),\alpha_3)$ $\alpha_3=f_d(z^{(l+1)}+\varepsilon)$	0.05	0.016

를 더하여 나온 이미지와 그 l_2norm 거리함수 값을 <Table 2>와 <Figure 7>에 나타내었다.

$z_i^{(l+1)}$에 노이즈를 더하여 f_d를 통해 나온 α_1, α_2, α_3는 표준정규분포에서 샘플링한 노이즈 ε에 $z_i^{(l+1)}$ 벡터 크기를 반영한 스케일링을 곱한 것이다. 각 α_i마다 6회 실시하였고 <Table 2>에 거리함수 값의 평균(Mean)과 표준편차(Std)를 나타낸다.

복원된 이미지 α_1, α_2을 보면 f_d가 locally smoothness 성질을 만족하는 것을 알 수 있다. 좀 더 큰 스케일링된 노이즈가 반영된 α_3 같은 경우는 $f_d(z_i^{(l+1)})$ 와 다른 이미지를 형성하는 것을 볼 수 있다.

이제 학습된 잠재공간을 조작하기 위해 잠재 차원이 각각 1차원, 2차원을 갖는 시각화오토인코더를 실험에 사용하였고 <Figure 8>은 시각화 오토인코더를 통해 복원된 n_l차원 데이터를 f_d을 통해 이미지로 매핑하는 과정을 나타낸다.

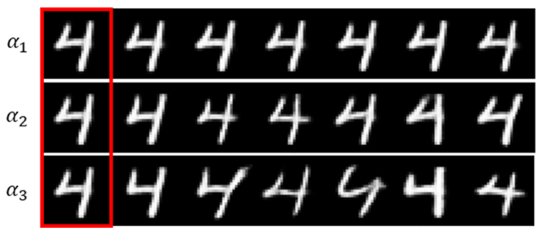


Figure 7. Image Reconstruction of Latent Vectors with Scaled Noise Added, Red Box: $f_d(z_i^{(l+1)})$.

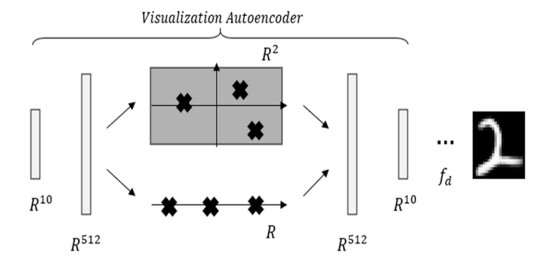


Figure 8. Visualization Autoencoder and Learned Decoder f_d.

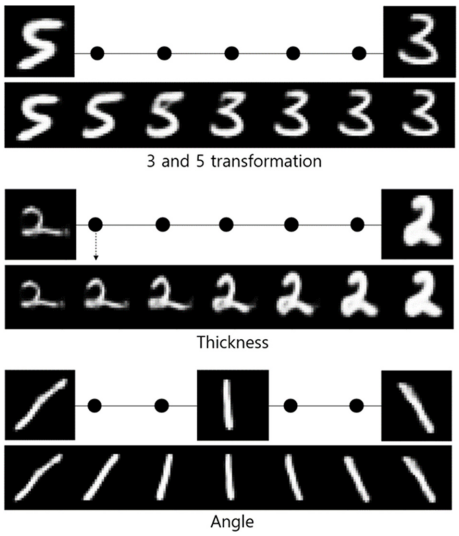


Figure 9. Visualization of the Transformation Process of Various Samples.

먼저 1차원 시각화 오토인코더를 사용하여 관심 있는 샘플들의 변환 과정을 <Figure 9>에 제시하였다. 두께, 각도 그리고 다른 숫자 이미지 변환 등 다양한 변환 과정을 시각화할 수 있다. 구체적으로 <Figure 9>의 두께 변환 과정을 나타내는 것은 얇은 2와 굵은 2, 두 개를 골라 1차원 시각화 오토인코더에 학습한 것이다. <Figure 9>의 학습된 모든 데이터는 손실함수 값이 0에 가깝게 수렴했으며 이 의미는 <Figure 3>의 $\tilde{z}_{i,v}^{(1)}$가 $z_i^{(1)}$와 매우 가까움을 의미한다.

다음은 2차원 시각화 오토인코더를 통해 관심있는 샘플들을 생성한 결과를 <Figure 10>에 제시하였다. 숫자 5 샘플을 생성하기 위해 각각 5를 나타내는 $z_i^{(l+1)}$을 4000개, 2000개, 200개, 20개를 2차원 시각화 오토인코더에 학습하였다.

<Figure 10>은 2차원으로 매핑된 z_i을 중심으로 빨간박스 구역을 그려 그 구역에서 격자로 100개 샘플을

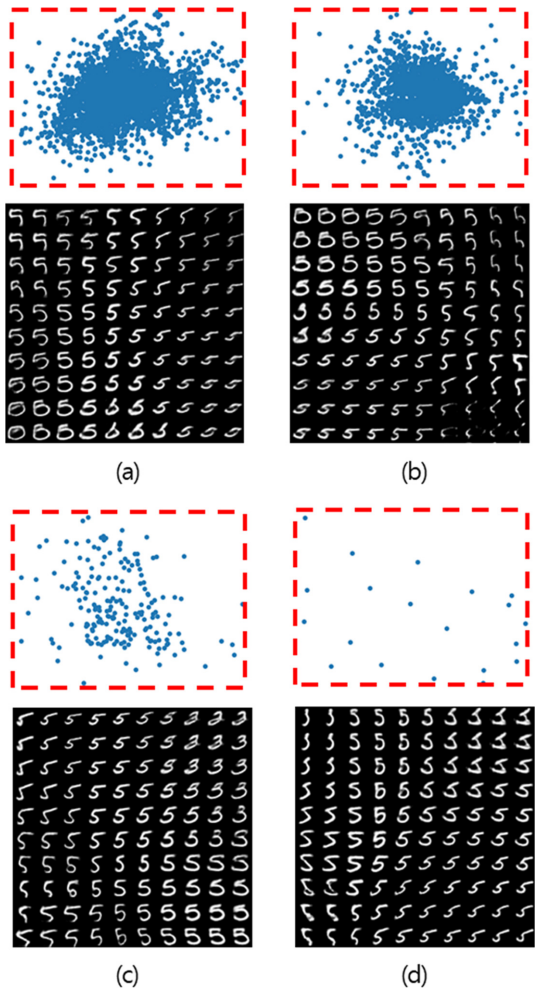


Figure 10. Generation of 5 Image Samples Using 2D Visualization Autoencoder, the Number of Data Used for Training (a) : 4000, (B) : 2000, (C) : 200, (D) : 20.

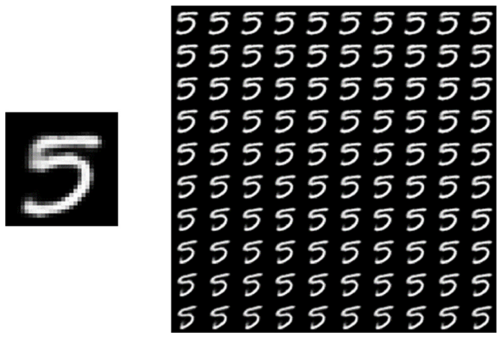


Figure 11. Left : Data Used for Training, Right : Generated Samples.

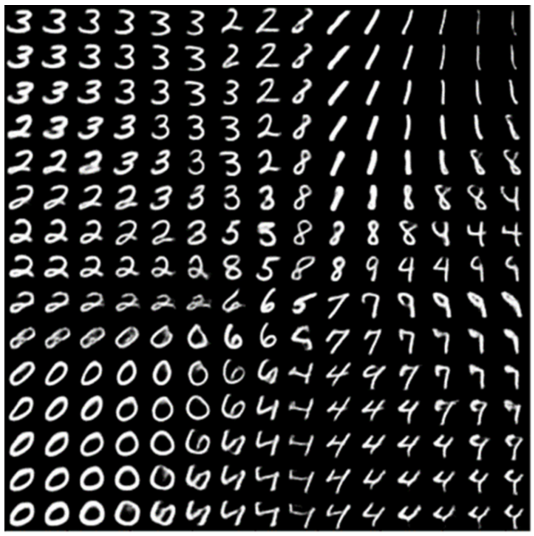


Figure 12. Samples generated by training 200 randomly selected $z_i^{(l+1)}$.

생성한 것이다. (A)~(D) 모두 다양한 숫자 5 샘플을 생성하고 변환 과정을 볼 수 있으며 (A)~(D)의 빨간 박스 구역은 <Figure 4>의 P^{n_v} 에 대응하고 생성된 샘플들은 $P_v^{n_0}$ 에 대응한다.

특히 <Figure 11>처럼 특정 데이터 1개만 2차원 시각화 오토인코더에 학습하여 비슷한 샘플을 생성하는 것도 가능하다.

또한 <Figure 12>처럼 랜덤으로 200개 $z_i^{(l+1)}$ 을 학습하여 다양한 숫자 이미지를 생성하고 변환 과정을 그려볼 수 있다.

5. 결 론

최근의 생성 모델은 다양한 분야에서 두각을 나타내고 있으며, 이러한 모델의 잠재공간 조작 및 이해는 중요한 연구 주제이다. 본 논문에서는 학습된 생성모델의 잠재공간을 조작하고 시각화하는 새로운 방법으로 “시각화 오토인코더”를 제안하였다.

MNIST 데이터셋에 대해 실험한 결과, 조작을 하기 위해 필요한 수학적 성질 두 가지(일대일 대응, locally smoothness)를 만족하는 경우에 대해 시각화 오토인코더를 사용하여 원하는 샘플 이미지 생성과 안정적인 이미지 변환이 가능함을 보았다. 특히, 더 적은 학습 데이터만을 사용하여도 의미 있는 결과를 얻을 수 있음을 확인하였다.

이러한 연구 결과는 생성모델의 잠재공간을 보다 깊게 이해하고 효율적으로 활용하기 위한 기초를 제공한다. 앞으로의 연구에서는 본 논문에서 제안된 방법을 다양한 데이터셋 및 환경에서 적용하여 그 범용성을 검증하고, 잠재공간의 다른 특성들에 대한 깊은 이해를 위한 추가적인 연구가 필요하다고 판단된다.

REFERENCES

[1] X. Ren, T. Yang, Y. Wang, and W. Zeng, Learning Disentangled Representation by Exploiting Pretrained Generative Models: A Contrastive Learning View, arXiv preprint arXiv:2102.10543, 2021.

[2] J. Lezama, Overcoming the Disentanglement vs Reconstruction Trade-off via Jacobian Supervision, International Conference on Learning Representations, 2019.

[3] W. Peebles, J. Peebles, J.-Y. Zhu, A. Efros, and A. Torralba, The Hessian Penalty: A Weak Prior for Unsupervised Disentanglement, Computer Vision–ECCV 2020: 16th European Conference, Vol. 12351, pp. 581-597, 2020.

[4] M. Connor and C. Rozell, Representing Closed Transformation Paths in Encoded Network Latent Space, Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 04, 2020.

[5] K. H. Min, J. Y. Kim and S. B. Cho, Algorithmic Bias Mitigation Method via Controlling Latent Space regards to Protected Feature, Proceedings of the Korea Conference on Software Engineering, pp. 491-493, 2019.

[6] Y. G. Kim, Variational Auto Encoder Distributed Restrictions for Image Generation, The Journal of the Institute of Internet, Broadcasting and Communication, Vol. 23, No. 03, pp. 91-97, 2023.

[7] H. Kim and A. Mnih, Disentangling by Factorising, nternational Conference on Machine Learning. PMLR, Vol. 80, pp. 2649-2658, 2018.

[8] T. Möllenhoff and D. Cremers, Flat Metric Minimization with Applications in Generative Modeling, International Conference on Machine Learning. PMLR, Vol. 97, pp. 4626-4635, 2019.

[9] D. P. Kingma and P. Dhariwal, Glow: Generative Flow with Invertible 1x1 Convolutions, Advances in Neural Information Processing Systems, Vol. 31, 2018.

[10] C-H. Pham, S. Ladjal, and A. Newson, PCA-AE: Principal Component Analysis Autoencoder for

Organising the Latent Space of Generative Networks, Journal of Mathematical Imaging and Vision, Vol. 64, pp.569-585, 2022.

[11] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework, International Conference on Learning Representations, 2017.

[12] T. Karras, S. Laine, and T. Aila, A Style-Based Generator Architecture for Generative Adversarial Networks, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401-4410, 2019.

[13] T. Park, J. Y. Zhu, O. Wang, J. Lu, E. Shechtman, A. Efros, and R. Zhang, Swapping autoencoder for deep image manipulation, Advances in Neural Information Processing Systems Vol. 33, pp. 7198-7211, 2020.

[14] J. H. Lee, C. S. Lee, and S. J. You, Attribute Preserving Face De-identification Method Using Latent Space Manipulation, Proceedings of the Korean Society of Precision Engineering Conference, pp. 596-596, 2022.

[15] J. Zhu, R. Feng, Y. Shen, D. Zhao, Z-J. Zha, J. Zhou, and Q. Chen, Low-RankSubspaces in GANs, Advances in Neural Information Processing Systems, Vol. 34, 2021.

[16] Y. Shen and B. Zhou, Closed-Form Factorization of Latent Semantics in GANs, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1532-1540, 2021.

[17] H. Yang, L. Chai, Q. Wen, S. Zhao, Z. Sun, and S. He, Discovering Interpretable Latent Space Directions of GANs Beyond Binary Attributes, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 12177-12185, 2021.

[18] A. Jahanian, L. Chai and P. Isola, On the "steerability" of generative adversarial networks, arXiv preprint arXiv:1907.07171, 2019.

[19] O. K. Yüksel, E. Simsar, E. G. Er, and P. Yanardag, LatentCLR: A Contrastive Learning Approach for Unsupervised Discovery of Interpretable Directions, Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 14263-14272, 2021.

[20] J. Choi, J. Lee, C. Yoon, J. H. Park, G. Hwang, and M. Kang, Do Not Escape From the Manifold: Discovering the Local Coordinates on the Latent Space of GANs, arXiv preprint arXiv:2106.06959, 2021

[21] A. Voynov and A. Babenko, Unsupervised Discovery of Interpretable Directions in the GAN Latent Space, International conference on machine learning. PMLR, Vol. 119, pp.9786-9796, 2020.

[22] N. Spingarn, R. Banner, and T. Michaeli, GAN "Steerability" without optimization, International Conference on Learning Representations. 2020.

[23] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris, GANSpace: Discovering Interpretable GAN Controls, Advances in Neural Information Processing Systems, Vol. 33, 2020.

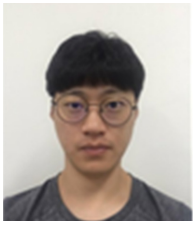
[24] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery, Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2085-2094, 2021.

[25] W. Nie, A. Vahdat and A. Anandkumar, Controllable and Compositional Generation with Latent-Space Energy-Based Models, Advances in Neural Information Processing Systems, Vol. 34, 2021.

[26] Z. Wu, D. Lischinski, and E. Shechtman, StyleSpace Analysis: Disentangled Controls for StyleGAN Image Generation, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12863-12872, 2021.

[27] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, Analyzing and Improving the Image Quality of StyleGAN, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8110-8119, 2020.

[28] Y. M. Ko and S. W. Ko, Alleviation of Vanishing Gradient Problem Using Parametric Activation Functions, KIPS Transactions on Softward and Data Engineering, Vol.10, No.10, pp.407-420, 2021.



고영민(Young-Min Ko)

2022년 전주대학교 인공지능학과에서 공학석사를 취득하였다. 현재는 전주대학교 문화기술학과 인공지능 전공으로 박사과정 중에 있다. 관심분야는 AI, 생성모델, VAE, Manifold learning 등이다.

이은주(Eun-Ju Lee)

2022년 전주대학교 스마트 Agro ICT 융합학과에서 공학석사를 취득하였다. 현재 전주대학교 문화기술학과 인공지능전공 박사과정 중에 있다. 관심분야는 인공지능, 데이터분석 등이다.

민정익(Jeong-Ik Min)

2000년 한국과학기술원 산업공학과에서 공학박사를 취득하였다. 1989년부터 LG-CNS, 한국IBM, SAP Korea, KT등에서 근무한 바 있으며, 현재는 전주대학교 인공지능학과, 전주대학교 일반대학원 Agro AI 학과, 전주대학교 인공지능연구소에 재직 중에 있다. 관심분야는 스마트팜, 인공지능 등이다.

고선우(Sun-Woo Ko)

1985년 고려대학교 산업공학과를 졸업하고, 1988년 한국과학기술원 산업공학과 석사, 1992년 한국과학기술원 산업공학과 공학박사를 취득하였다. 현재는 전주대학교 인공지능학과 교수, 전주대학교 인공지능연구소 연구소장으로 재직 중에 있다.