5장. 딥러닝 –V

5. 오토인코더

5.1 특징 추출 오토인코더

5.2 잡음 제거 오토인코더

5.3 희소 오토인코더

5.4 변분 오토인코더

5.1 오토인코더

❖ **비지도 학습 신경망 모델**

- **특징 추출에 사용**
- **제한적 볼츠만 머신**(RBM, Restricted Boltzmann Machine)
- **오토인코더**(autoencoder)

❖ **오토인코더**

- 입력 노드의 개수와 출력 노드의 개수가 같은 다층 신경망으로 구성
- 모래 시계와 같은 모양의 계층 구조

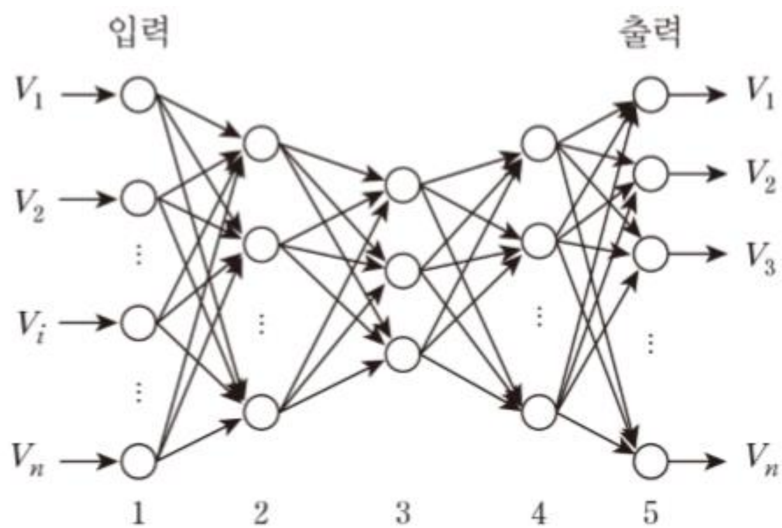
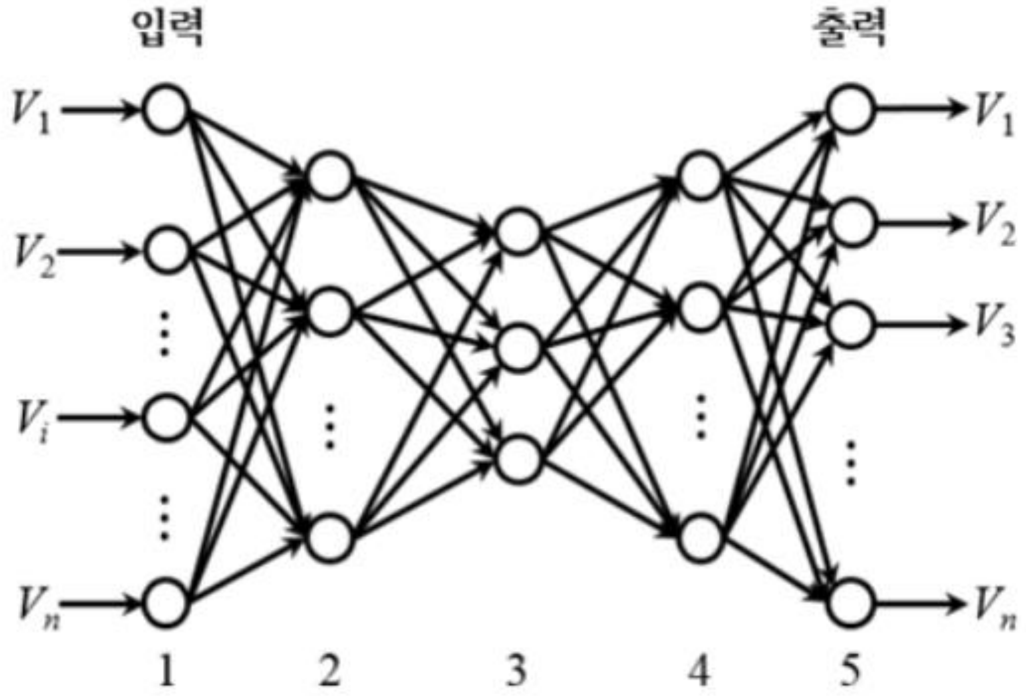


그림 5.65 오토인코더autoencoder

5.1. 특징 추출 오토인코더

❖ 오토인코더의 역할



n-차원 데이터	m-차원 특징	3-차원 특징	m-차원 복원 특징	n-차원 복원 데이터

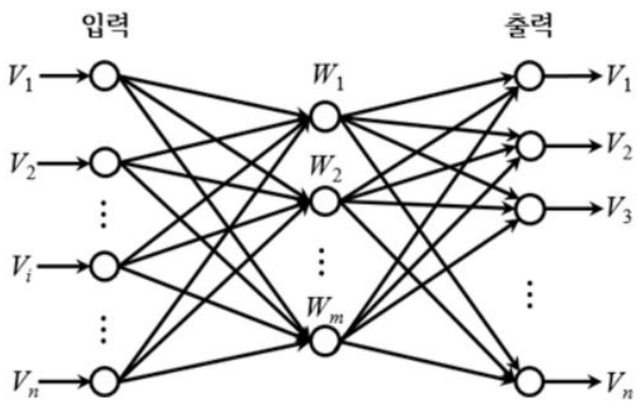
특징 추출기
인코더

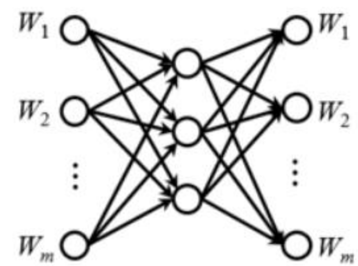
디코더

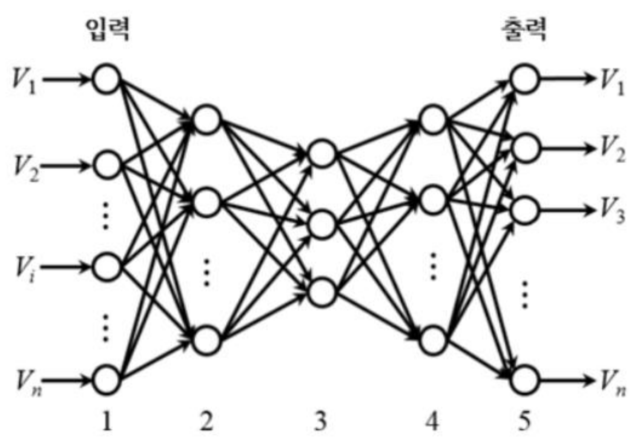
특징 추출 오토인코더

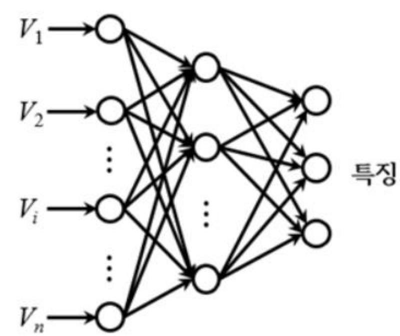
❖ 오토인코더의 학습

- **출력 노드가 입력 값과 동일한 결과를 생성하도록 학습**
- 은닉층의 노드값을 새로운 입력으로 하는 은닉층이 하나인 다층 퍼셉트론 학습후 결함









특징 추출 오토인코더

❖ **적층 제한적 볼츠만 머신(stacked RBM)을 이용한 학습**

- 점진적 RBM 학습
- 학습된 RBM의 구조를 대칭적으로 복사하여 오토인코더 구조 구성

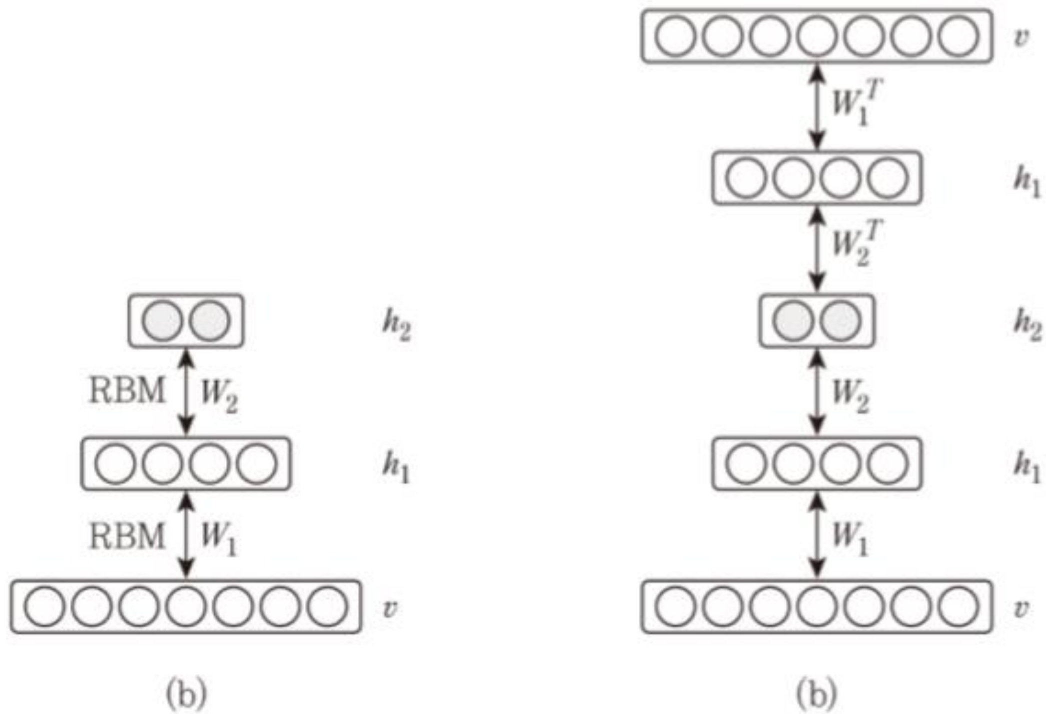


그림 5.67 RBM을 이용한 오토인코더의 학습

- 오차 역전파(backpropagation) 알고리즘 적용

5.2 잡음제거 오토인코더

❖ 잡음제거 오토인코더 (Denoising Autoencoder)

- 학습 데이터
  - 입력 : $\hat{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{r}$: $\boldsymbol{x}$ (원본 데이터), $\boldsymbol{r}$ (무작위 잡음)
  - 출력 : $\boldsymbol{x}$
- 입력된 데이터의 정보를 유지하면서 보다 좋은 특징 을 추출할 뿐만 아니라, 입력에 포함된 잡음을 제거하는 역할

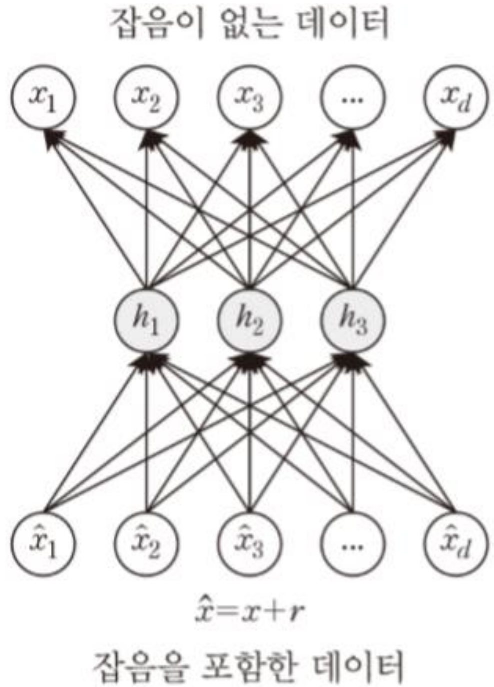
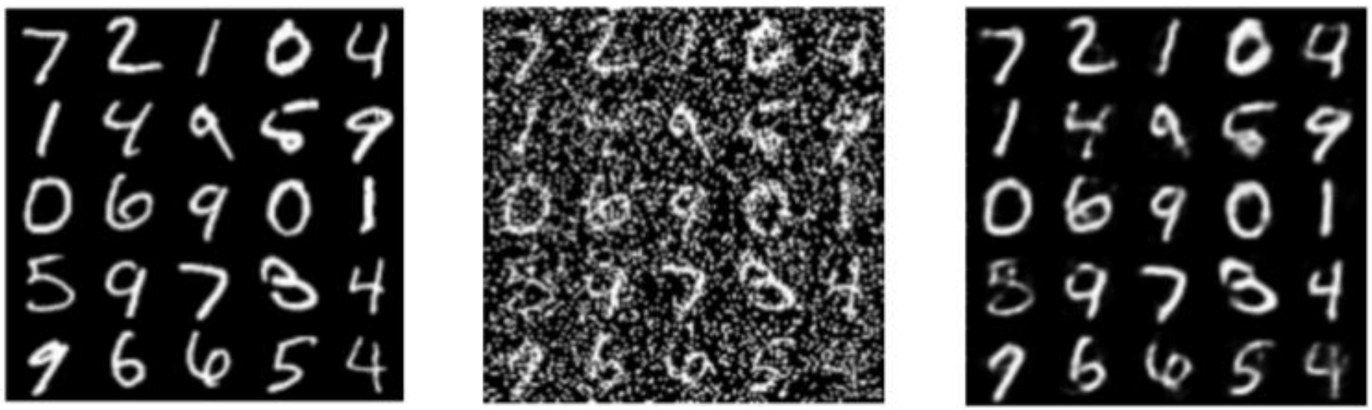


그림 5.68 잡음제거 오토인코더의 학습

잡음제거 오토인코더



(a) (b) (c)

그림 5.69 잡음제거 오토인코더의 잡음제거 효과.

(a) 학습에 사용된 데이터 (b) 잡음이 포함된 입력 데이터 (b) 오토인코더의 출력

5.3 희소 오토인코더

❖ 희소 오토인코더 (Sparse Autoencoder)

- 가능하면 코딩 층에서 출력값이 0이 아닌 노드의 개수가 적어지도록 만드는 오토인코더
- 목적 함수

$$E = \sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{y}_i)^2 + \beta \sum_{j=1}^{M} KL(\rho \| \hat{\rho}_j)$$

기대출력과 실제 출력의 차이

임계값보다 큰 활성도 정도를 갖는 코딩 층의 노드에 벌점을 주는 규제항

ρ : 코딩층 노드들의 평균 활성도 정도에 대한 목표값

ρ_j: 현재 미니배치에 대한 j번째 노드의 평균 활성화 값

$$\hat{\rho}_j = \frac{1}{N}\sum_{i=1}^{N} f_j(\boldsymbol{x}_i)$$

j번째 노드의 평균 활성화 값

희소 오토인코더

❖ **쿨벡-라이블러 발산** (Kullback-Leibler divergence, KL-發散)

$$KL(\rho \| \hat{\rho}_j)$$

- 확률 분포 ρ와 $\hat{\rho}$ 사이의 차이 측정
- ρ와 $\hat{\rho}$가 서로 비슷할 수록 0에 접근

$$KL(\rho \| \hat{\rho}_j) = \rho \log\left(\frac{\rho}{\hat{\rho}_j}\right) + (1-\rho)\log\left(\frac{1-\rho}{1-\hat{\rho}_j}\right)$$

❖ **희소 오토인코더의 학습**

- 경사 하강법 적용

5.4 변분 오토인코더

- **변분 오토인코더**(Variational Autoencoder, VAE)
  - 학습 데이터의 분포를 따르는 새로운 데이터를 만드는 오토인코더 기반의 생성 모델
  - 변분법 사용

- **변분법**(Variational method)
  - 어떤 함수 $p(\boldsymbol{x})$의 극점을 찾는 문제에서 해당 함수를 직접 다루는 것이 쉽지 않을 때, 쉽게 다룰 수 있는 다른 함수 $q(\boldsymbol{x})$로 대체해 이를 최적화하여, $p(\boldsymbol{x})$에 대한 근사적인 해를 구하는 방법

변분 오토인코더 VAE

- ❖ **변분 오토인코더의 구조**
  - **인코더**
    - 입력 공간의 데이터 $\boldsymbol{x}$를 은닉 공간의 데이터 $\boldsymbol{z}$로 변환
    - 인코더의 확률 모델

$$p_\phi(z|\boldsymbol{x})$$

  - **디코더**
    - 은닉 공간의 데이터 $\boldsymbol{z}$를 입력 공간의 데이터 $\boldsymbol{x}$로 변환
    - 디코더의 확률 모델 probability model of decoder

$$p_\theta(\boldsymbol{x}|z)$$

- ❖ **학습의 목표**
  - 디코더가 원본 데이터를 복원하도록 하는 것
  - $p_\theta(\boldsymbol{x})$를 최대화하는 것

변분 오토인코더 VAE

❖ 목적 함수

- $p_\theta(\boldsymbol{x})$의 로그 가능도 $\boxed{p_\theta(\boldsymbol{x}|z)}$
  - $p_\phi(\boldsymbol{z}|\boldsymbol{x})$ 대신에 다른 확률 분포 $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ 사용

$$\log p_\theta(\boldsymbol{x}) = \sum_{\boldsymbol{z}} q_\phi(z|\boldsymbol{x}) \log p_\theta(\boldsymbol{x}) = \sum_{\boldsymbol{z}} q_\phi(z|\boldsymbol{x}) \log \frac{p_\theta(\boldsymbol{x}, z)}{p_\theta(z|\boldsymbol{x})}$$

$$= \sum_{\boldsymbol{z}} q_\phi(z|\boldsymbol{x}) \log \frac{p_\theta(\boldsymbol{x}, z)}{p_\theta(z|\boldsymbol{x})} \frac{q_\phi(z|\boldsymbol{x})}{q_\phi(z|\boldsymbol{x})}$$

$$= \sum_{\boldsymbol{z}} q_\phi(z|\boldsymbol{x}) \log \frac{q_\phi(z|\boldsymbol{x})}{p_\theta(z|\boldsymbol{x})} \frac{p_\theta(\boldsymbol{x}, z)}{q_\phi(z|\boldsymbol{x})}$$

$$= \sum_{\boldsymbol{z}} q_\phi(z|\boldsymbol{x}) \log \frac{q_\phi(z|\boldsymbol{x})}{p_\theta(z|\boldsymbol{x})} + \sum_{\boldsymbol{z}} q_\phi(z|\boldsymbol{x}) \log \frac{p_\theta(\boldsymbol{x}, z)}{q_\phi(z|\boldsymbol{x})}$$

변분 오토인코더 VAE

❖ 목적 함수 – cont.

$$\log p_\theta(\boldsymbol{x}) = \sum_{\boldsymbol{z}} q_\phi(z|\boldsymbol{x}) \log \frac{q_\phi(z|\boldsymbol{x})}{p_\theta(z|\boldsymbol{x})} + \sum_{\boldsymbol{z}} q_\phi(z|\boldsymbol{x}) \log \frac{p_\theta(\boldsymbol{x}, z)}{q_\phi(z|\boldsymbol{x})}$$

$$= KL(q_\phi(z|\boldsymbol{x}) \| p_\theta(z|\boldsymbol{x})) + L(\theta, \phi, \boldsymbol{x})$$

$$\geq L(\theta, \phi, \boldsymbol{x})$$ 하한(lower bound)

변분 오토인코더 VAE

❖ 목적 함수 – cont.

$$\log p_\theta(\boldsymbol{x}) \geq L(\theta, \phi, \boldsymbol{x})$$

- 하한

$$L(\theta, \phi, \boldsymbol{x}) = \sum_{\boldsymbol{z}} q_\phi(z|\boldsymbol{x}) \log \frac{p_\theta(\boldsymbol{x}, z)}{q_\phi(z|\boldsymbol{x})}$$

$$= E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}, z)\right] - E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log q_\phi(z|\boldsymbol{x})\right]$$

$p_\theta(x, z) = p_\theta(x|z)p_\theta(z)$

$$= E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|z) + \log p_\theta(z)\right] - E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log q_\phi(z|\boldsymbol{x})\right]$$

$$= -KL(q_\phi(z|\boldsymbol{x}) || p_\theta(z)) + E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|z)\right]$$

규제화 항 **복원 손실**(reconstruction loss)

가능한 작게 만들기

$q_\phi(\boldsymbol{z}|\boldsymbol{x})$와 $p_\theta(\boldsymbol{z})$가 서로 비슷하게

변분 오토인코더 VAE

❖ **변분 오토인코더의 학습**

- $q_\phi\,(\boldsymbol{z}|\boldsymbol{x})$의 확률 분포
  - 데이터를 확률 분포로 인코딩
  - 가우시안 분포(Gaussian distribution)를 가정
  - 평균 벡터 $\boldsymbol{\mu}$와 표준편차 행렬 Σ 학습

$$\epsilon \sim N(0,1)$$

$$z = \boldsymbol{\mu} + \epsilon \boldsymbol{\Sigma}$$

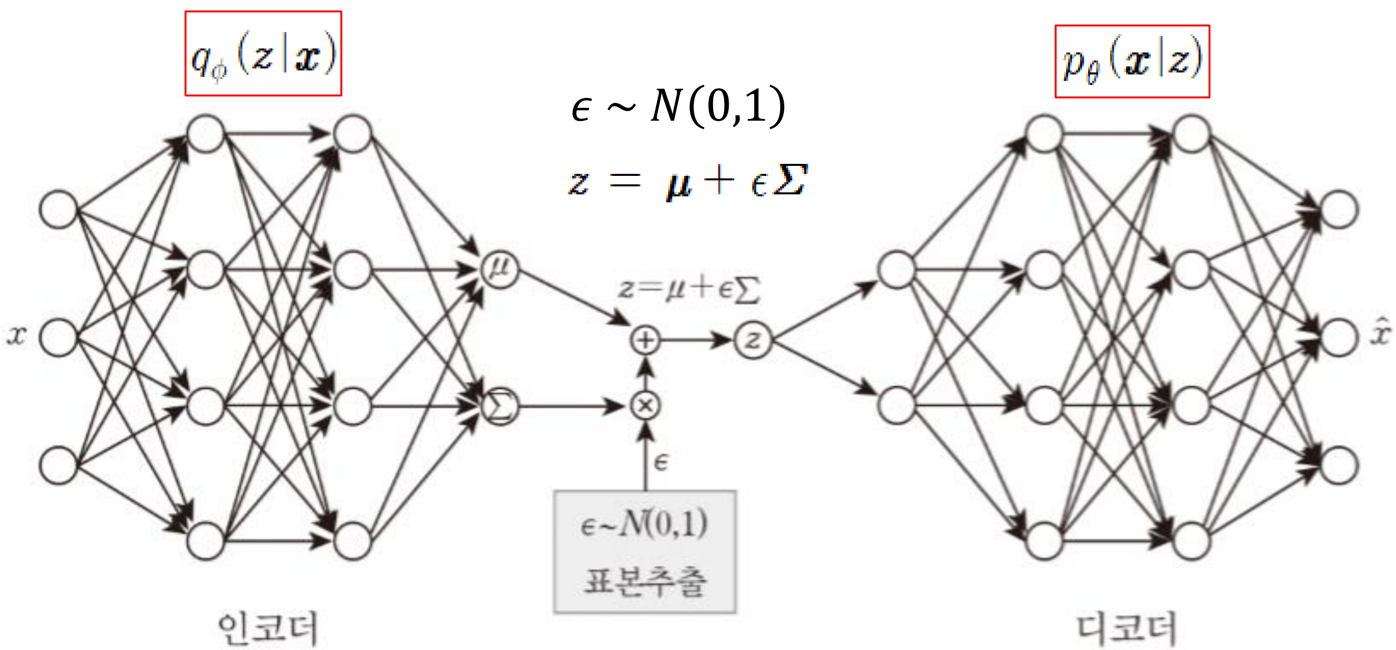


그림 5.70 변분 오토인코더

변분 오토인코더 VAE

❖ 변분 오토인코더의 학습

$$\log p_\theta(\boldsymbol{x}) \geq L(\theta, \phi, \boldsymbol{x})$$

$$\boxed{L(\theta, \phi, \boldsymbol{x}) = -KL(q_\phi(z|\boldsymbol{x}) \| p_\theta(z)) + E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|z)}$$

- $L(\theta, \phi, \boldsymbol{x})$의 최대화를 위해 **경사 상승법**(gradient-ascent method) 적용

- $KL(q_\phi(z|\boldsymbol{x}) \| p_\theta(z)) = -E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|z) + \log p_\theta(z)]$
  - $p_\theta(\boldsymbol{z})$ ⇦ 표준 정규분포 $N(0,1)$ 가정
  - $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ ⇦ 정규 분포 $N(\mu, \Sigma)$ 가정

$$KL(q_\phi(z|\boldsymbol{x}) \| p_\theta(z)) = -\frac{1}{2}\sum_{j=1}^{J}\left(1 + \log 1(\sigma_j^2) - \mu_j^2 + \sigma_j^2\right)$$

- 복원손실 관련 항 $E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|z)]$: 미니배치 데이터에 대해 계산

$$E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|z)] \approx \frac{1}{L}\sum_{i=1}^{L}\log p_\theta(\boldsymbol{x}|z)$$

변분 오토인코더 VAE

❖ **변분 오토인코더의 동작**

- 입력 $\boldsymbol{x}$ $\rightarrow$ 은닉 변수 $\boldsymbol{z}$ $\rightarrow$ 출력 $\boldsymbol{y}$

- 각 원소의 값이 구간 [0,1]의 값인 경우
  - $\log p_\theta(\boldsymbol{x}|\boldsymbol{z})$의 값

$$\log p_\theta(\boldsymbol{x}|\boldsymbol{z}) = \sum_{i=1}^{D}[x_i \log y_i + (1-x_i)\log(1-y_i)]$$

- 중간 층에 확률적 요소 추가
- 입력과 약간 다른 데이터 생성 가능
  - 학습 데이터의 분포를 따르는 새로운 데이터 생성

  $\Rightarrow$ 생성 모델의 역할

변분 오토인코더 VAE

❖ 은닉 공간과 데이터 공간의 대응 관계

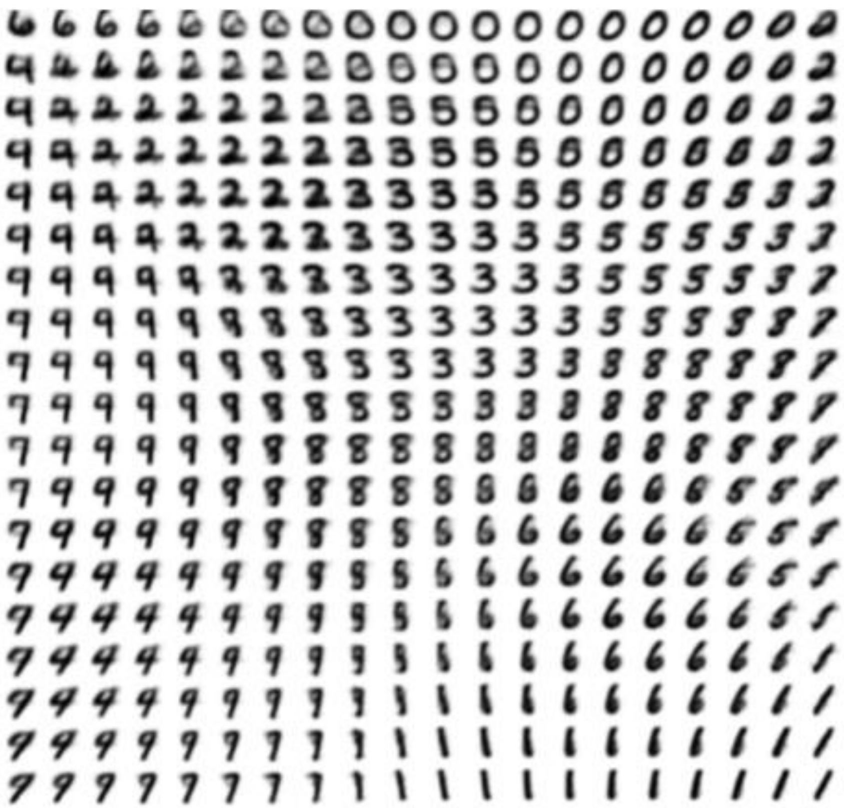


그림 5.71 MNIST 데이터를 학습한 변분 오토인코더의 은닉변수 공간에 코딩된 데이터[출처: Kingma 등 2014]