Neural Discrete Representation Learning

2022.05.07

김기범

Abstract

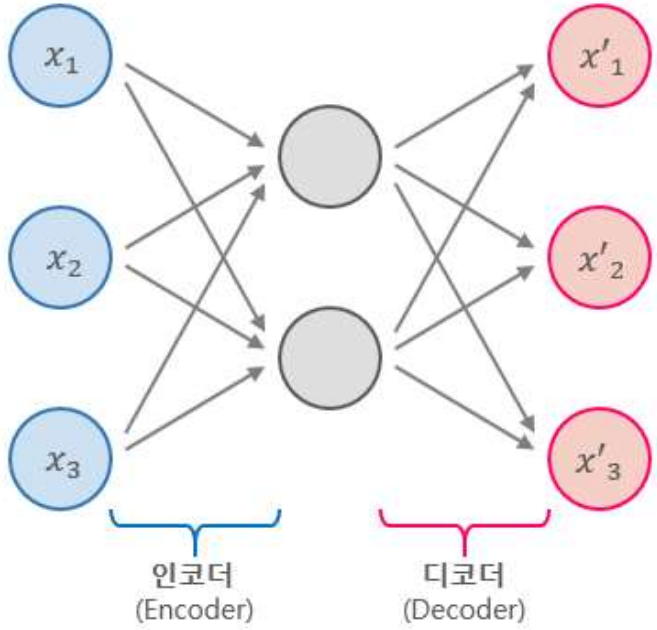
- Vector QuantisedVariational AutoEncoder (VQ-VAE)를 소개한 논문.

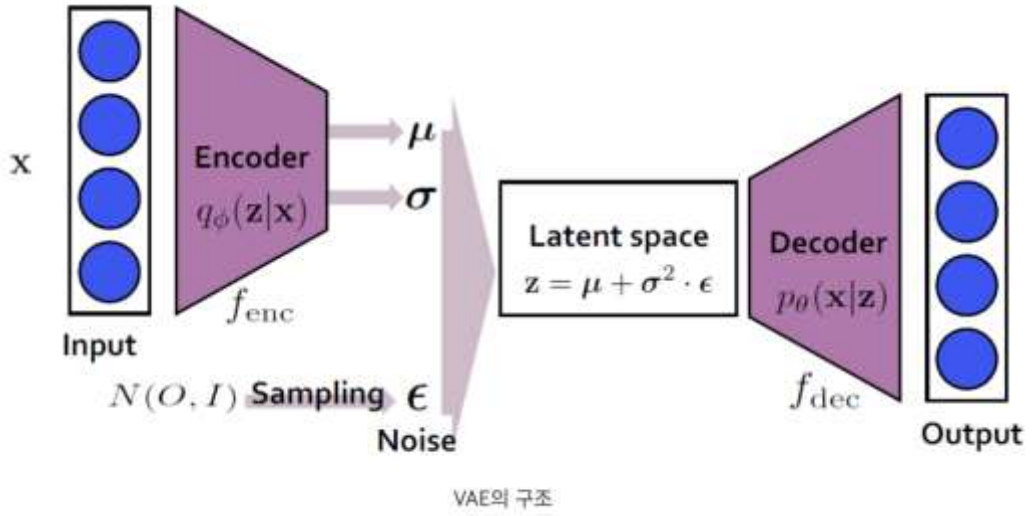
- 기존 VAE과 두 부분의 큰 차이를 보임.
  1. Encoder network를 거친 결과가 discrete하다.
  2. prior은 static하지 않고, 학습이 가능하다.

- 이산 잠재 표현(discrete latent representation)을 학습하기 의해 vector quantization의 아이디어를 사용했다.

- VQ 방법을 사용하면, VAE에서 발생하는 Posterior Collapse 문제를 피할 수 있다.

Variational Autoencoder

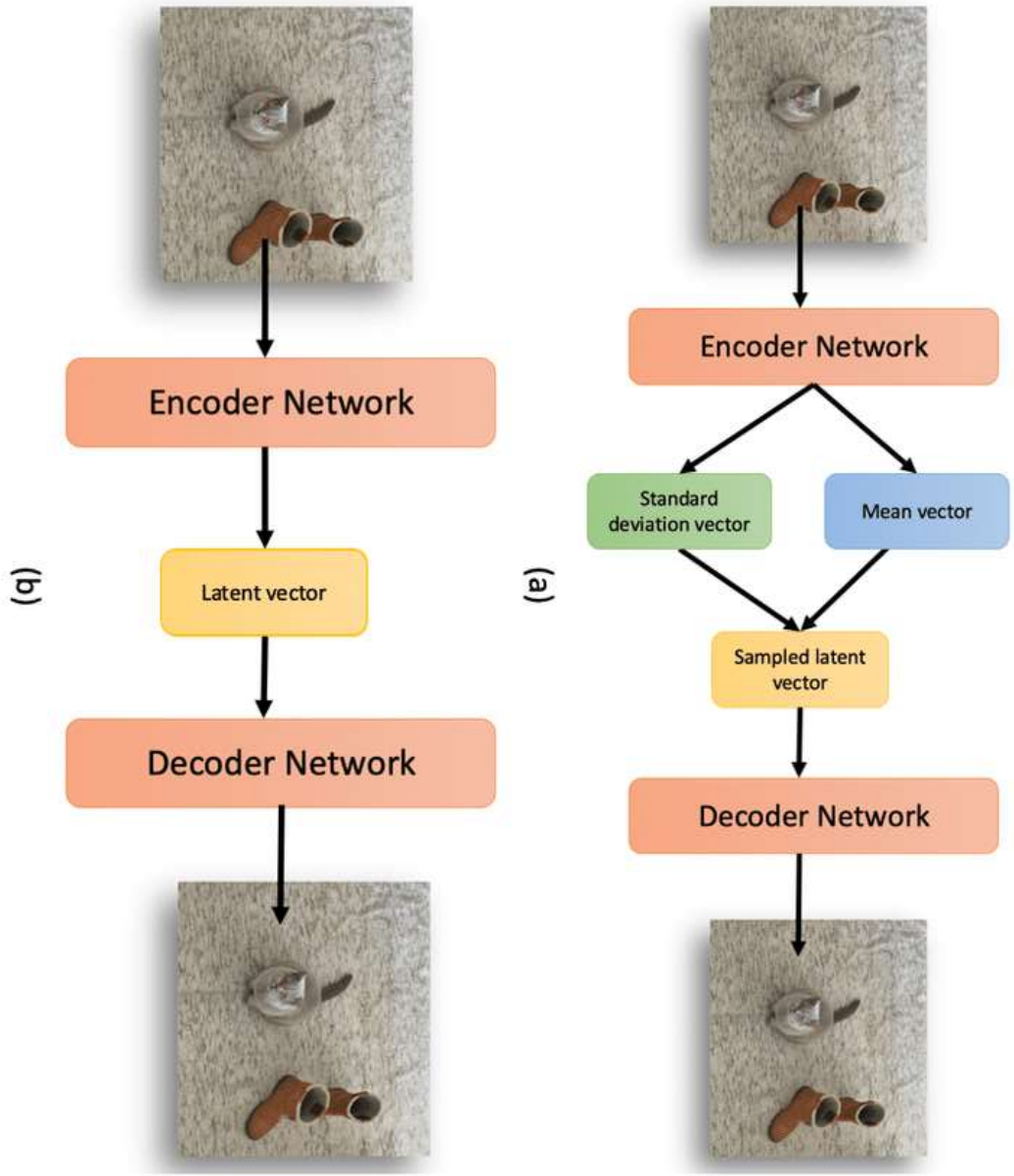




VAE의 구조

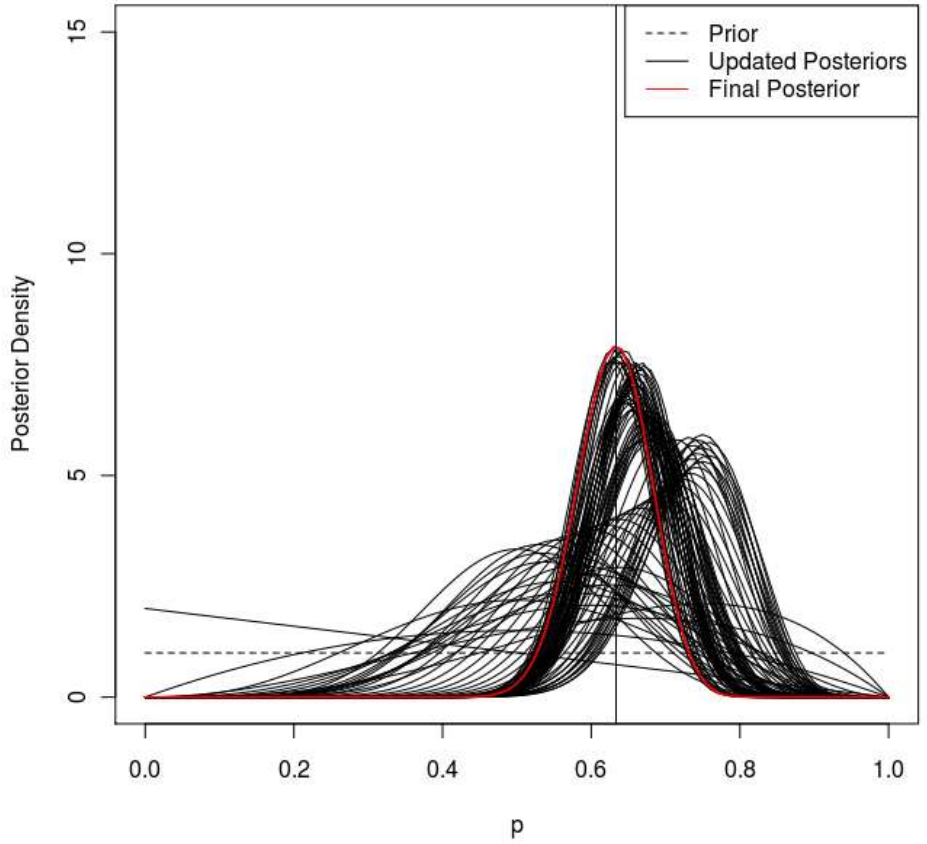
Autoencoder: latent vector z를 잘 추출하자.

VAE: encoding의 distribution이 prior로 주어짐.

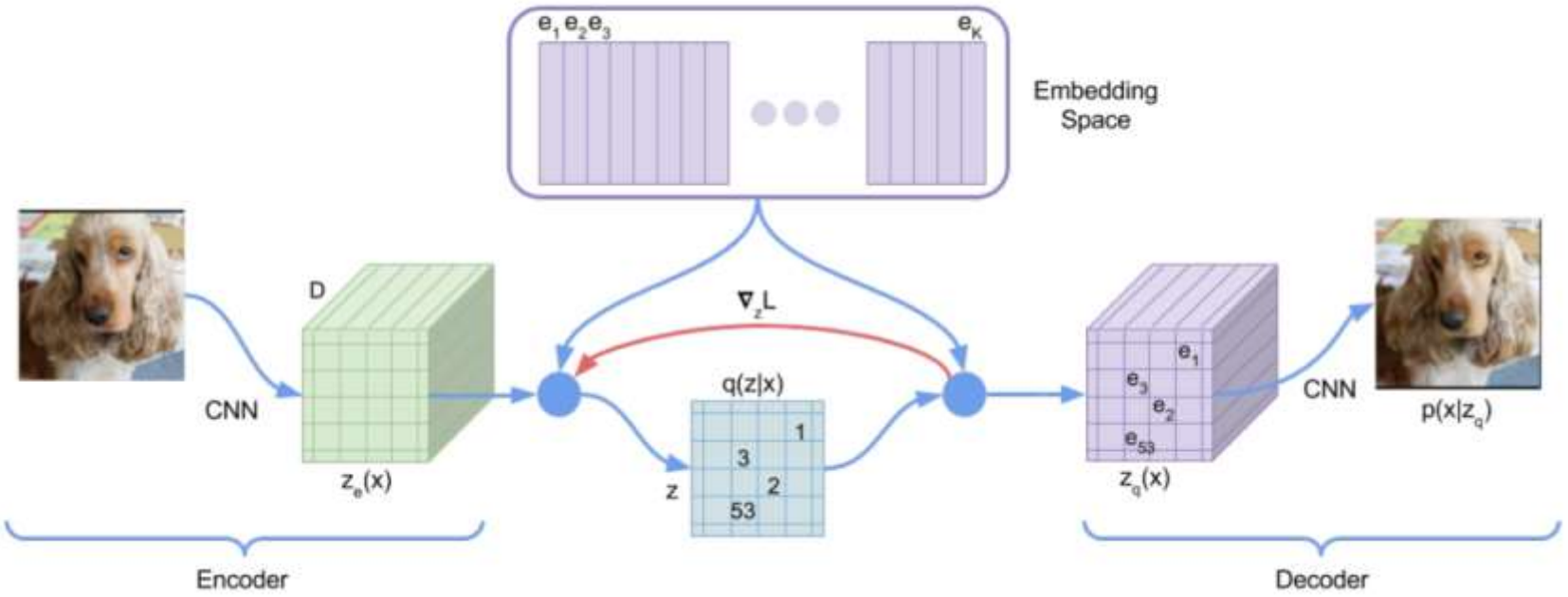

(b)
Encoder Network
Latent vector
Decoder Network
(a)
Encoder Network
Standard deviation vector
Mean vector
Sampled latent vector
Decoder Network

Posterior collapse

- 일종의 local minima state
- Approximate posterior가 prior을 그대로 따라함.
- Autoencoder가 latent variable을 무시한 상태에서 학습이 진행됨.
- $q_\phi(z \mid x) = p(z)$

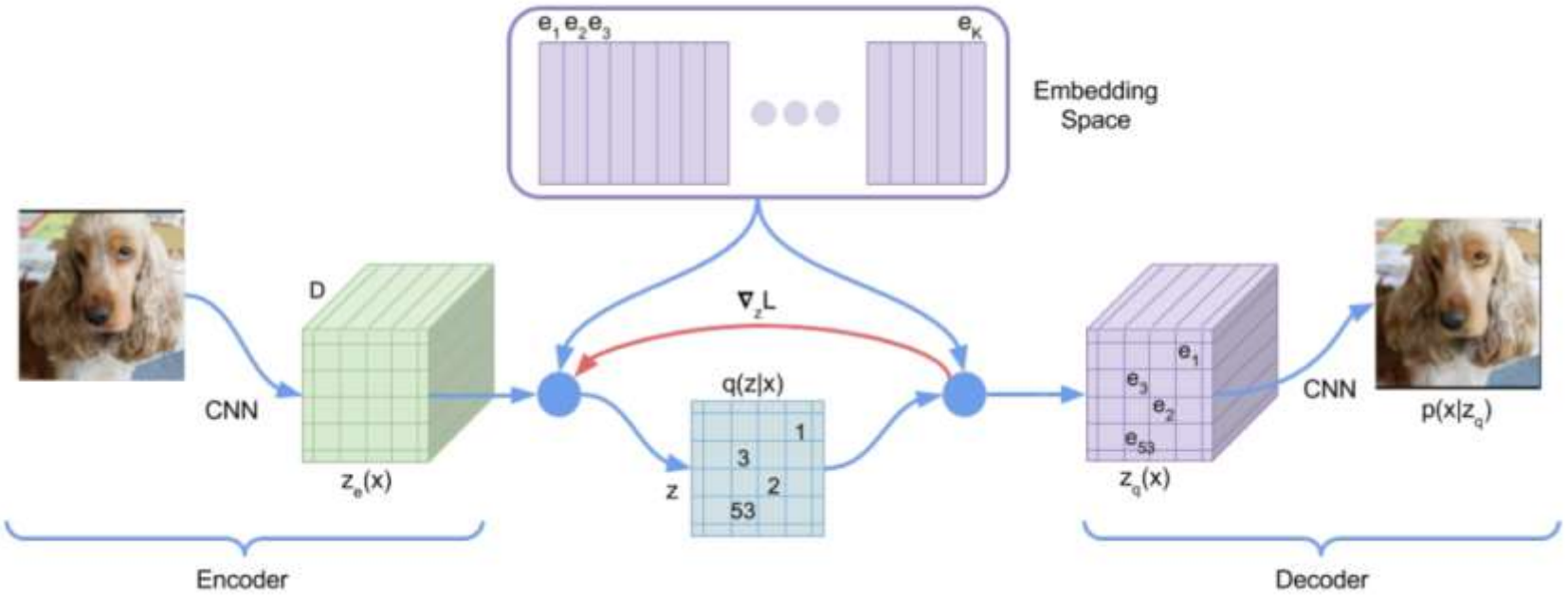


VQ-VAE



- VQ-VAE는 Vector quantization(VQ)을 이용하여 이산 표현을 다룸.
- VQ를 사용할 때, posterior과 prior distribution은 categorical distribution이다.
- 이 분포로부터 생성된 sample은 embedding table을 indexing한다. 이 embedding는 decoder의 입력으로 들어간다.

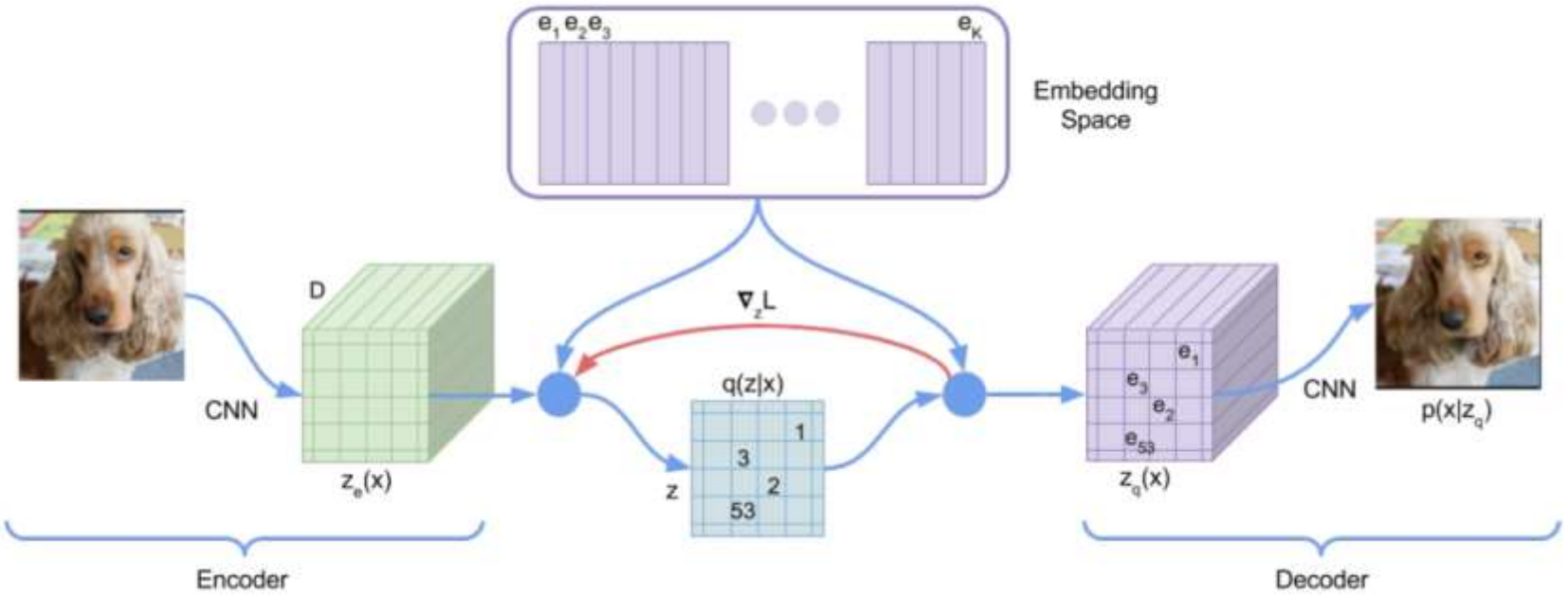
VQ-VAE: Discrete latent variable



- $e \in R^{R*D}$: latent embedding space(codebook),
  - K: discrete latent space의 크기
  - D: e_i의 차원
- 모델의 encoder는 입력 x 를 받아 discrete latent variable $z_e(x)$ 를 출력
- posterior categorical cistribution q(z|x)의 확률은 크기 K의 one-hot encoding으로 정의됨.

$$q(z=k|x) = \begin{cases} 1 & \text{for } k = \text{argmin}_j \|z_e(x) - e_j\|_2, \\ 0 & \text{otherwise} \end{cases}$$

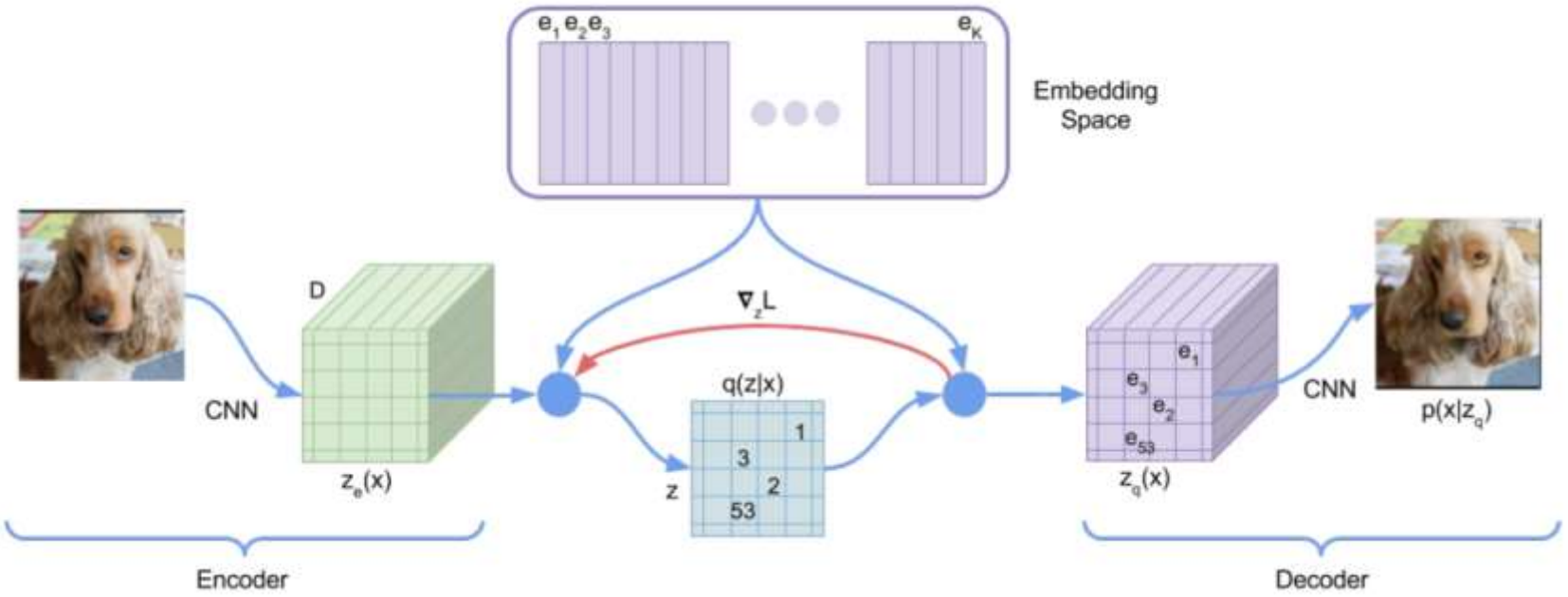
VQ-VAE: Discrete latent variable



- decoder의 input $z_q(x)$를 얻기 위해 아래 식과 같이 embedding(codebook) space e에서 $z_e(x)$와 가장 가까운 원소 e_j를 찾는다.

$$z_q(x) = e_k, \quad \text{where} \quad k = \text{argmin}_j \|z_e(x) - e_j\|_2$$

VQ-VAE: Learning



- z_q(x)를 계산하는 식(*argmin*)에서는 gradient를 정의할 수 없음.
- z_q로 들어온 gradient를 그대로 z_e에 복사하는 형태로 gradient($\nabla_z L$)를 encoder쪽으로 보내는 방식을 사용했다.

VQ-VAE: Learning

- z_q(x)를 계산하는 식(*argmin*)에서는 gradient를 정의할 수 없음.
- z_q로 들어온 gradient를 그대로 z_e에 복사하는 형태로 gradient($\nabla_z L$)를 encoder쪽으로 보내는 방식을 사용했다.

- Loss Function은 아래와 같다.

$$L = \log p(x|z_q(x)) + \|\mathrm{sg}[z_e(x)] - e\|_2^2 + \beta\|z_e(x) - \mathrm{sg}[e]\|_2^2,$$

  - 첫번째 항: reconstruction loss로 encoder, decoder 모두를 최적화
  - 두번째 항: 첫번째 항에서 gradient가 z_q에서 z_e로 바로 넘어가기 때문에 embedding e를 학습시키지 못한다. Encoder로부터 뽑혀 나온 z_e 와 비슷해지도록 e를 업데이트
  - 세번째 항: embedding e_i는 encoder parameter만큼 빠르게 학습되지 못한다. 그래서 e와 z_e가 비슷한 속도로 학습될 수 있도록 commitment loss를 넣었다.

VQ-VAE: Prior

- 이 모델에서 prior distribution p(z)는 categorical distribution이며 z에 대해 autoregressive하게 만들어 질 수 있다. VQ-VAE를 학습할 때엔 constant, uniform하게 유지된다.
- 학습 이후, z에 대한 autoregressive분포에 맞추어 ancestral sampling을 통해 새로운 x를 생성 가능.
- Image의 discrete latent를 학습하기 위해 PixelCNN을 사용했고, raw audio 데이터에 대해서는 WaveNet 모델을 사용했다.

Experiments

- Continuous variable을 사용한 모델과 비교:
  - CIFAR10을 사용하여 VQ-VAE, VAE, VIMCO를 비교. reconstruction error가 continuous variable model과 비슷하게 나왔다.
  - VAE : VQ-VAE : VIMCO = 4.51 : 4.67 : 5.14 bits/dim, negative log likelihood

- Image
  - ImageNet의 128 *128*3 크기 이미지를 purely deconvolutional $p(x|z)$를 통해 z = 32*32 *1 의 discrete space로 압축했고 그 결과로 생성된 z에 대해 prior인 PixelCNN을 학습했다. 이를 통해 학습 속도를 높이고 이미지의 전체적인 특성을 살릴 수 있었다.

Figure 2: Left: ImageNet 128x128x3 images, right: reconstructions from a VQ-VAE with a 32x32x1 latent space, with K=512.

Experiments: Image

- PixelCNN prior으로 discretised 32x32x1 latent space를 학습한 후(spatial masking in the PixelCNN) 생성된 z로 Image generation을 수행.

Figure 3: Samples (128x128) from a VQ-VAE with a PixelCNN prior trained on ImageNet images. From left to right: kit fox, gray whale, brown bear, admiral (butterfly), coral reef, alp, microwave, pickup.

Experiments: Audio

- 109명의 speaker의 음성 녹음본을 담은 VCTK set을 학습에 사용
- 6 strided convolution, stride 2, window-size 4를 사용하여 원본 파일보다 64배 압축, long-term의 정보만 보존하도록 latent space를 만듦.
- 본래 WaveNet에서는 시끄러운 소리가 생성된 반면 VQ-VAE에서는 깨끗한 소리를 생성함.

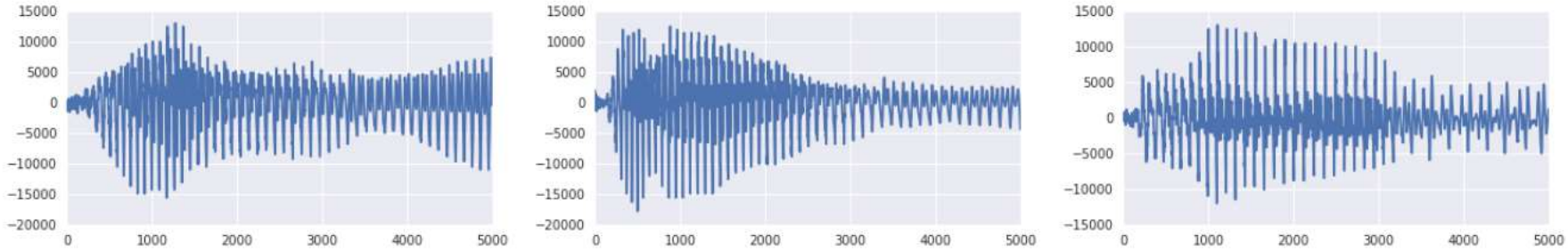


Figure 6: Left: original waveform, middle: reconstructed with same speaker-id, right: reconstructed with different speaker-id. The contents of the three waveforms are the same.

Conclusion

- VAE와 discrete latent 표현을 위한 Vector quantization를 결합하여 새로운 생성 모델을 만들었고, continuous latent와 비슷한 성능을 낸다.
- VQ-VAE는 원본을 작은 latent값으로 잘 압축할 수 있으며, long-term dependency를 잘 모델링한다.