VQ-VAE (Neural Discrete Representation Learning)

https://arxiv.org/abs/1711.00937

정영석

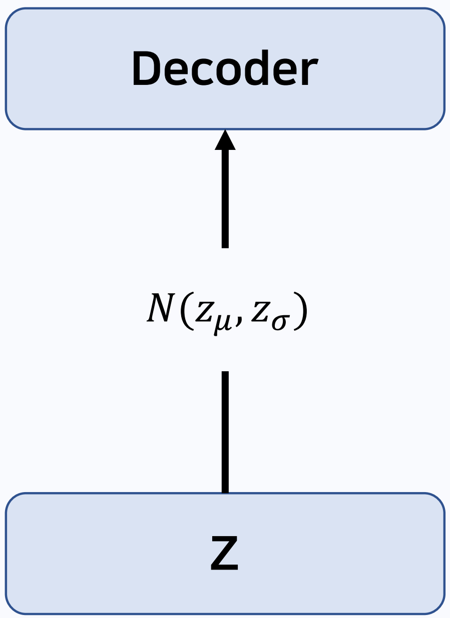
Contents

목 차

Contents

Introduction

- **VAE (Variational AutoEncoder)**



$$\log p(x) = \int \log p(x|z)\, p(z) dz$$

$$\log p(x) \geq \underbrace{\int \log p(x|z)\, q(z|x) dz}_{\text{Reconstruction term}} - \underbrace{D_{kl}(q(z|x)||N(0,1))}_{\text{Regularization term}}$$

Contents

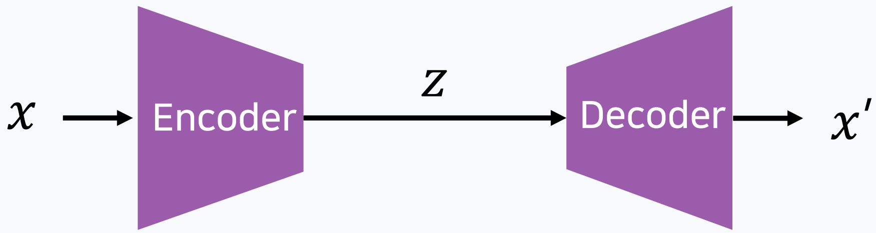
Introduction

- **왜 VQ(Vector Quantized) 이어야 하는가?**

✓ 현실 세계에 적합한 모델링 방법

- 음성의 언어적 정보
- 이미지 역시 Discrete한 표현으로 표현이 가능함

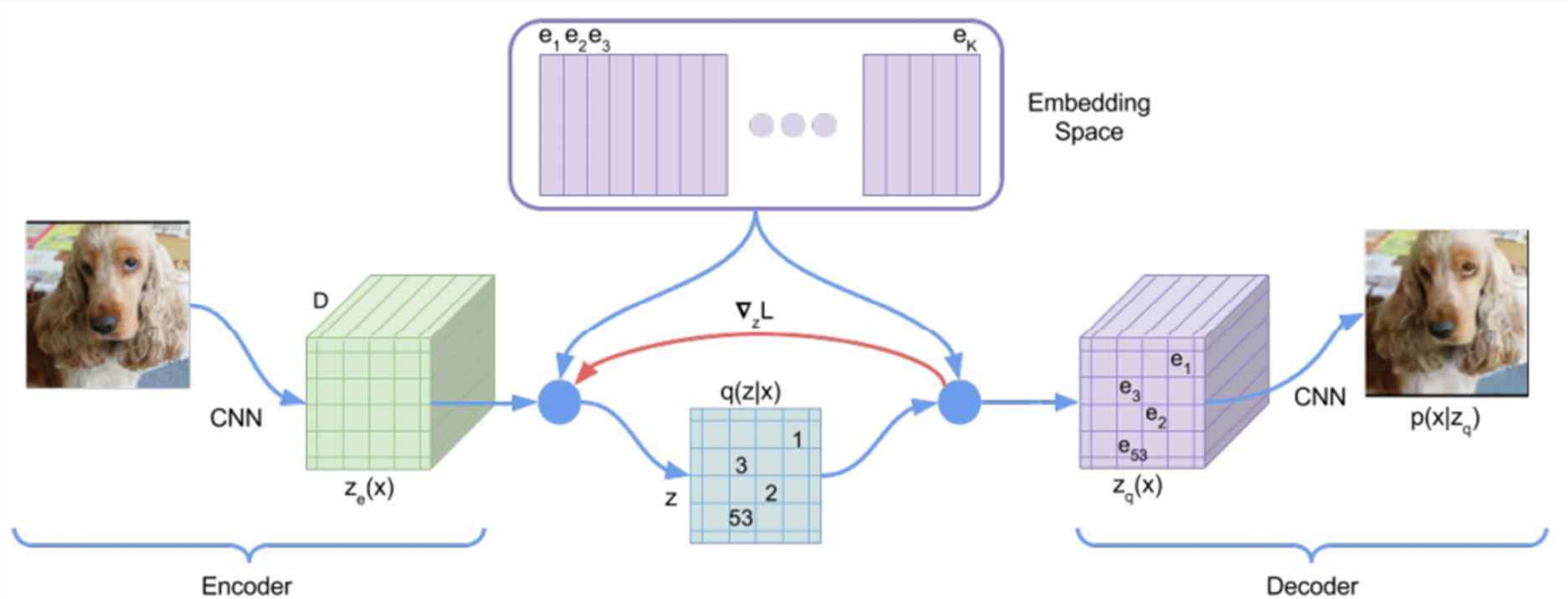
✓ Posterior collapse 를 방지함.



Contents

Methodology

- **모델의 구조**



Contents

Methodology

- **모델의 구조**

✓ Encoder

- 입력된 x로부터 $z_e(x)$를 생성함.

✓ Embedding Space

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

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

Contents

Methodology

- **Objective function**

  ✓ VQ-VAE log-likelihood

$$\log p(x) \geq \int \log p(x|z)\, q(z|x) dz - D_{kl}(q(z|x)||p(z))$$

$$\log p(x) \geq \int \log p(x|z)\, q(z|x) dz - D_{kl}(q(z|x)||\frac{1}{k}), (k = \#\ of\ embedding\ )$$

$$D_{kl}(q(z|x)||1/k) = \sum q(z|x) \log\left(\frac{q(z|x)}{p(z)}\right)$$

$$= q(k|x) \log\left(\frac{q(z|x)}{p(z)}\right)$$

$$= 1 * \log\left(\frac{1}{(\frac{1}{k})}\right) = \log k$$

Contents

Methodology

- **Objective function**

  ✓ VQ-VAE 의 likelihood

$$p(x) = \int \log p(x|z)\, p(z) dz$$

$$p(x) \approx \log p(x|z_k)\, p(z_k)$$

  ✓ VQ-VAE 의 objective function

$$L = \underbrace{\log p\big(x \big| z_q(x)\big)}_{\text{Reconstruction loss}} + \underbrace{\big|\big|sg\ [z_e(x)]\ - e\big|\big|_2^2 - \beta\big|\big|z_e(x) - sg\ [e]\big|\big|_2^2}_{\text{Commitment loss}}$$

Contents

experiments

- 실험 결과 분석(Images)

Contents

experiments

- **실험 결과 분석(Images – pixelCNN prior + VQ-VAE Decoding)**

kit fox, gray whale, brown bear, admiral (butterfly), coral reef, alp, microwave, pickup

Contents

experiments

- **실험 결과 분석(Audio)**

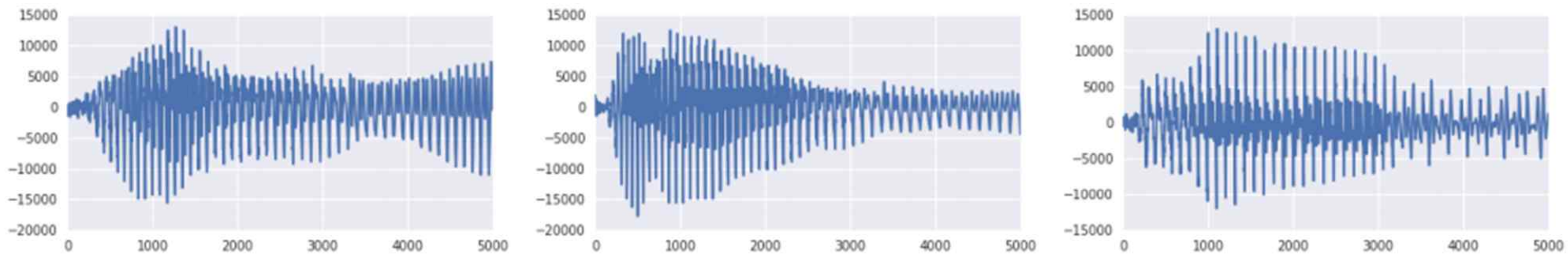


Figure 6: Left: original waveform, middle: reconstructed with same speaker-id, right: reconstructed with different speaker-id. The contents of the three waveforms are the same.