Clova
Developers

Neural Discrete Representation Learning (VQ-VAE)

BizAI
곽동현

Clova
Developers

Contents

Paper Review 1

Neural Discrete Representation Learning

Aaron van den Oord
DeepMind
avdnoord@google.com

Oriol Vinyals
DeepMind
vinyals@google.com

Koray Kavukcuoglu
DeepMind
korayk@google.com

Contributions

1. VQ-VAE라는 새로운 알고리즘 제안
2. PixelCNN을 이용한 Autoregressive prior 사용
3. 다양한 domain에서 실험(image, audio, video)

Posterior Collapse 문제

- Powerful decoder가 latents를 무시하고 이미지를 reconstruction하는 현상
- 완전히 posterior collapsing이 된 경우, 아래와 같이 posterior가 x와 독립

$$q_\phi(z|x) \simeq q_\phi(z) = \mathcal{N}(a, b)$$

- Posterior collapse가 일어나는 이유는 latent variable z가 too noisy(high variance)해서, 이를 무시하도록 decoder가 generation을 학습하기 때문

Goyal, Anirudh Goyal Alias Parth, et al. "Z-forcing: Training stochastic recurrent networks." Advances in neural information processing systems. 2017.

Vector-Quantized VAE의 장점

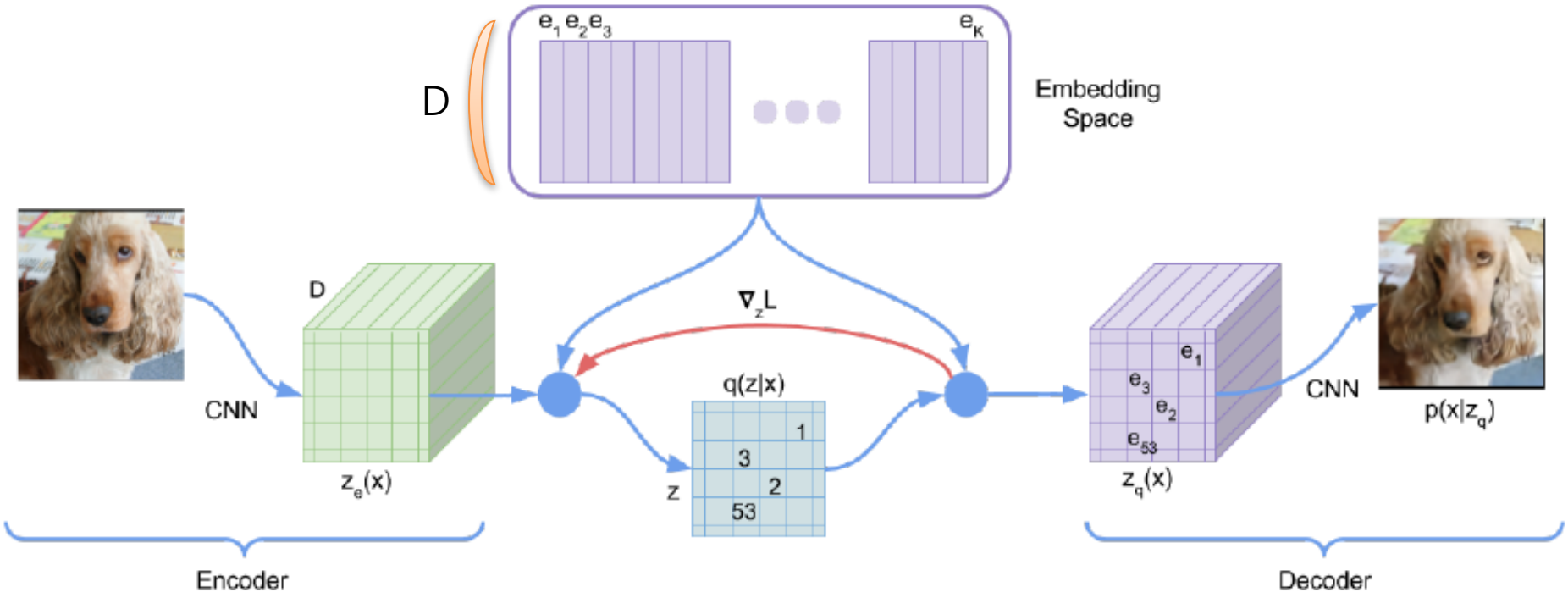
- Discrete variable이므로 high-variance를 겪지 않는다.(Posterior collapse완화)
- 상당한 정보량의 압축이 가능하다.(256*32bits → 1*7bit / 1170배)
- 해석 가능성이 높다.
- 그럼에도 불구하고 reconstruction이 나쁘지 않다.

VAE와의 비교

- Continuous latent variable VS Discrete latent variable
- Categorical distribution vs Gaussian distribution
- Embedding space가 명시적으로 존재함
- 미분 불가능한 Regularization loss-term이 있음.

Neural discrete representation learning (NeurIPS 2017, Aaron DeepMind)

Vector-Quantized VAE 아키텍쳐

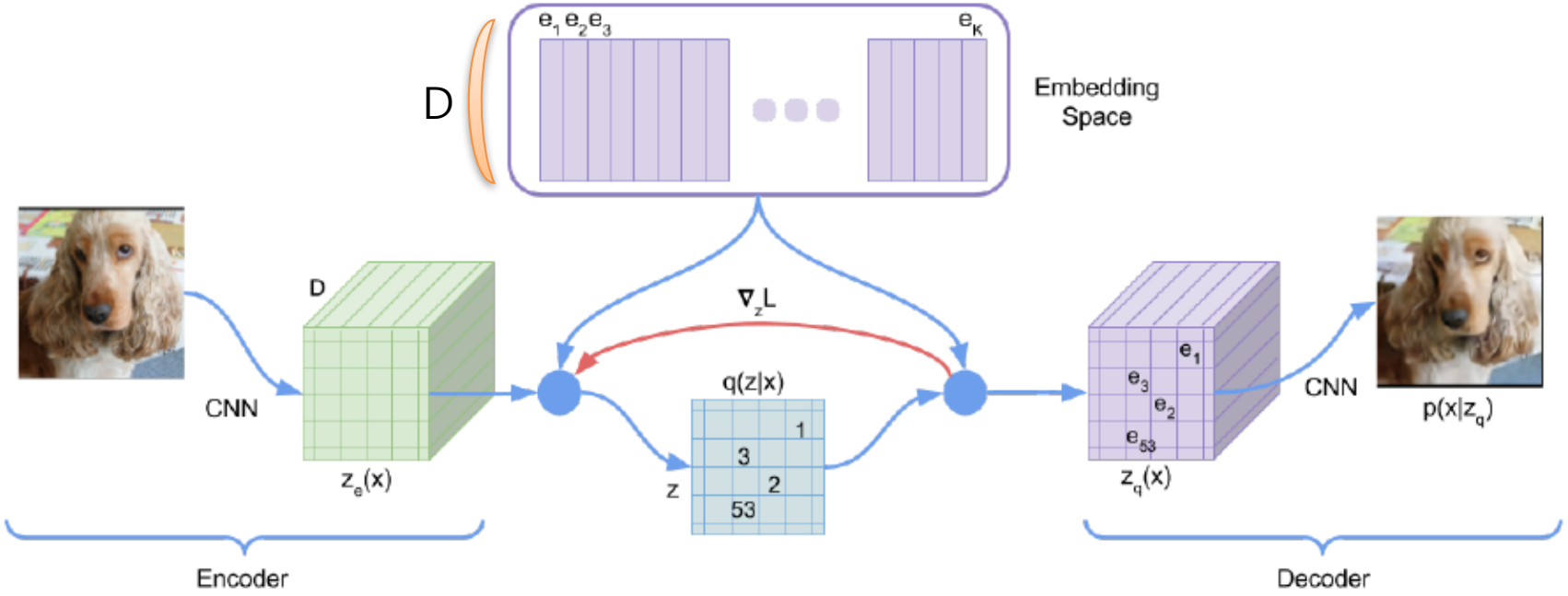


- Learnable Parameter: Encoder / Embedding Space / Decoder

- Input pixel → Encoding vector(z_e(x)) → Referencing embedding space → Discrete latent variables(z) → decoding vector(z_q(x)) → Output pixel

Neural discrete representation learning (NeurIPS 2017, Aaron DeepMind)

Nearest-neighbour look-up



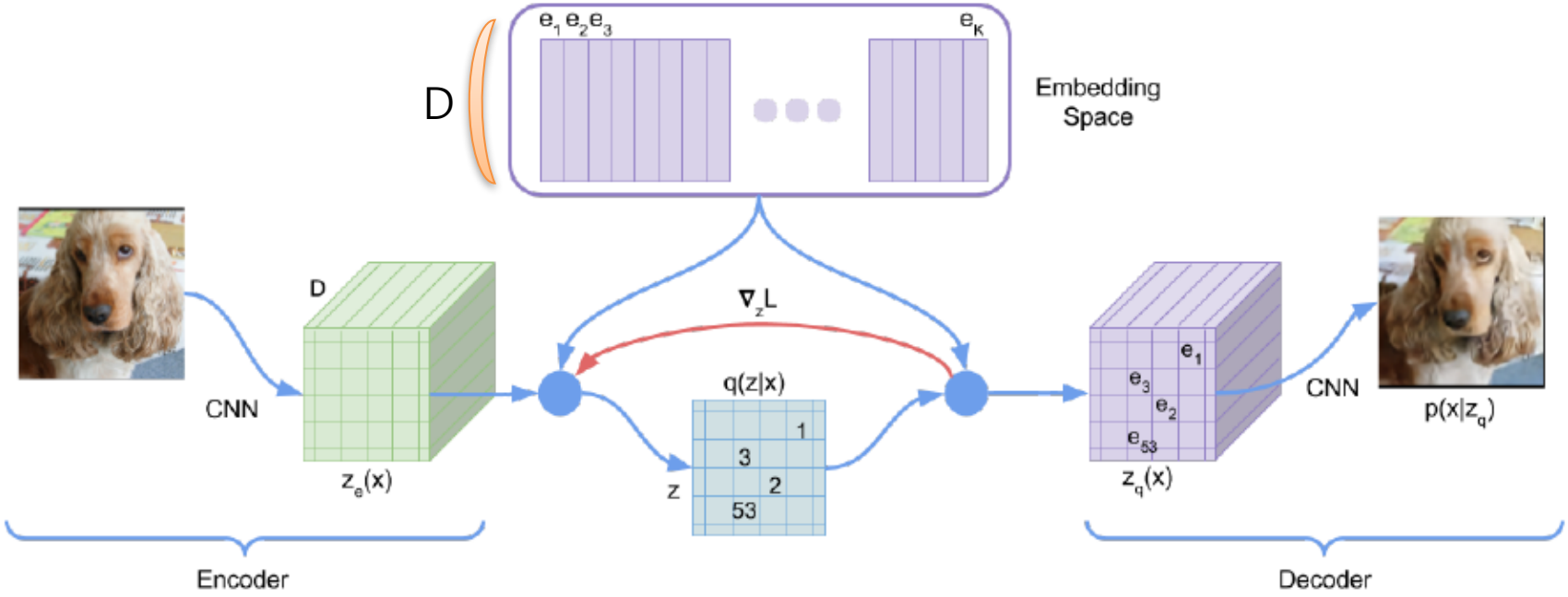
- Categorical distribution q(z|x)의 분포는 다음과 같이 정의

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

- 위의 분포를 one-hot encoding으로써 이용해, **vector to index mapping**

Neural discrete representation learning (NeurIPS 2017, Aaron DeepMind)

Nearest-neighbour look-up

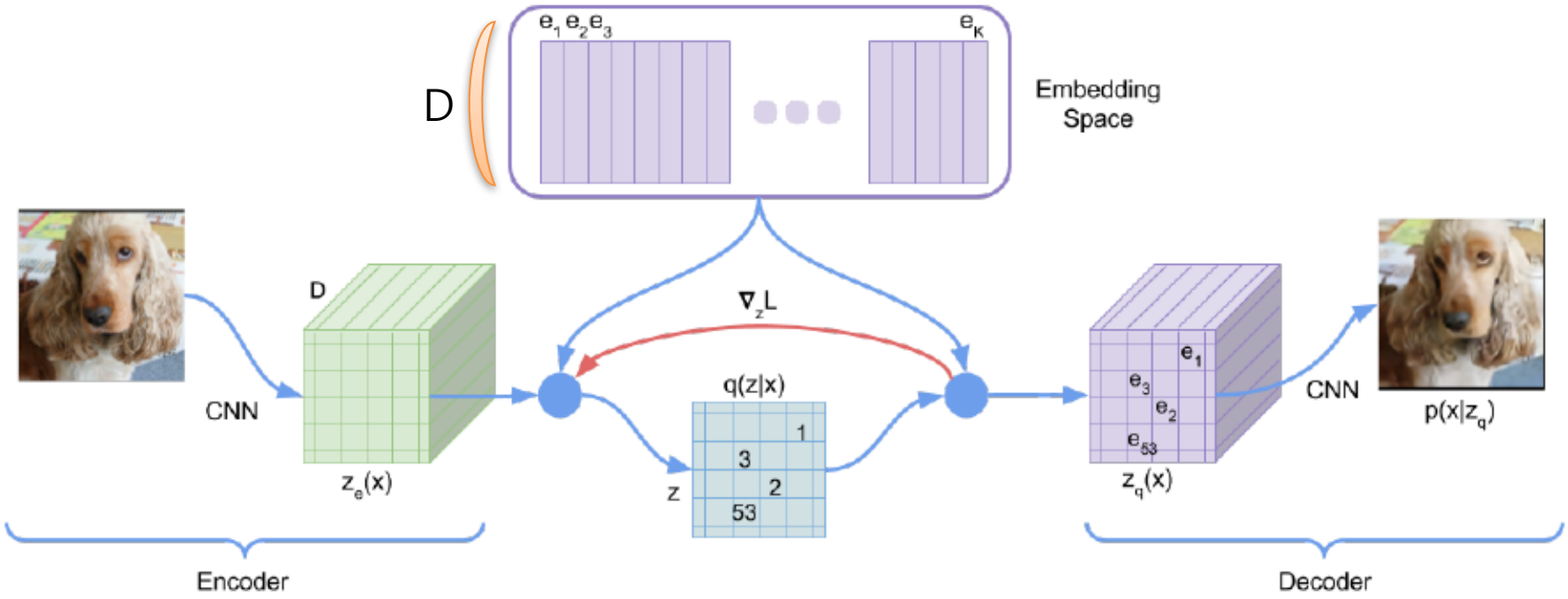


- 동일한 Nearest-neighbour look-up으로 **index to vector mapping**

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

Neural discrete representation learning (NeurIPS 2017, Aaron DeepMind)

Learning



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

Reconstruction loss (encoder, decoder)

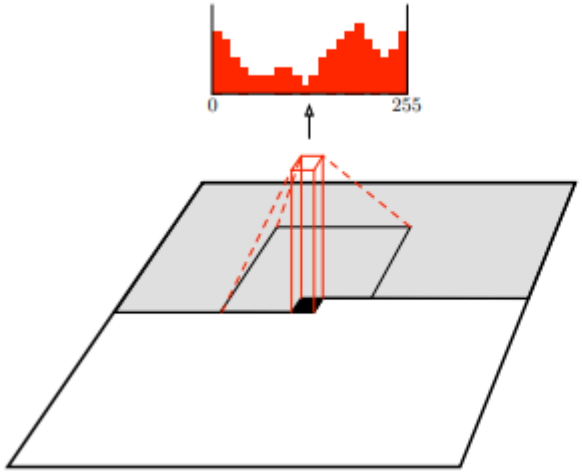
Dictionary learning loss (embedding space)

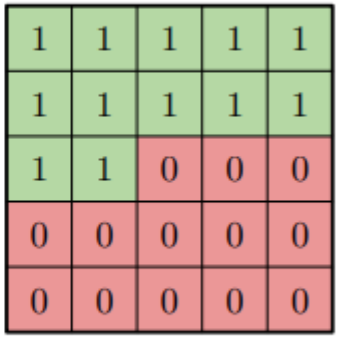
Commitment loss (encoder)

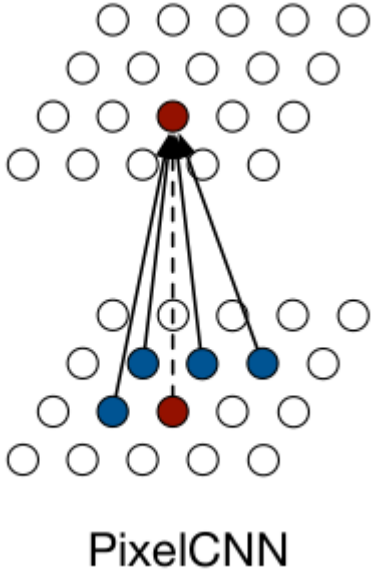
- **sg: stopgradient operator,** to be a non-updated constant.(consider as a GT)

Neural discrete representation learning (NeurIPS 2017, Aaron DeepMind)

PixelCNN Prior





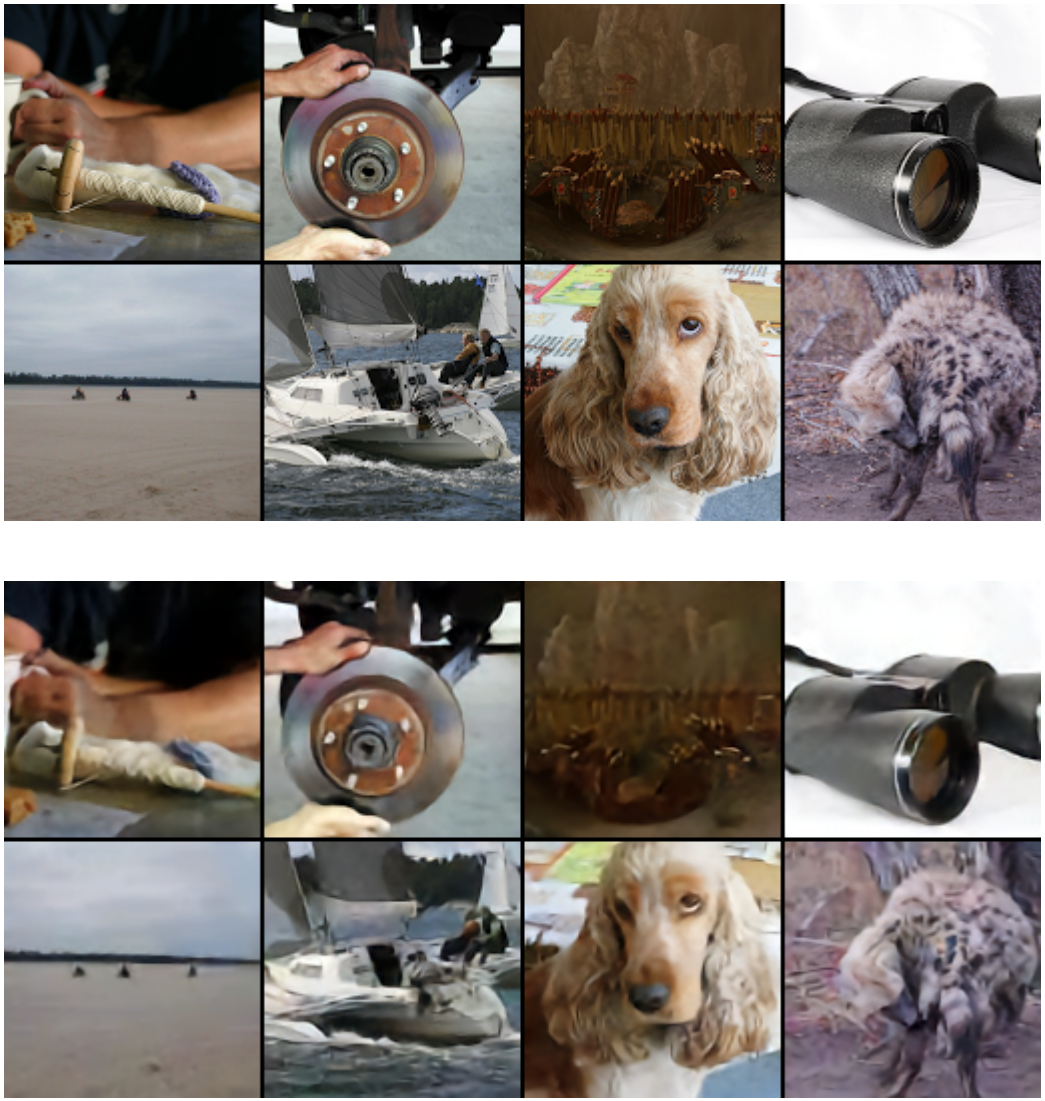


- First training: The VQ-VAE, the prior(p(z)) is kept constant and uniform.
- Second training: We fit an autoregressive distribution over z, p(z), so that we can generate x via ancestral sampling.

Van den Oord, Aaron, et al. "Conditional image generation with pixelcnn decoders." *Advances in neural information processing systems*. 2016.

Experiments

Image Reconstruction



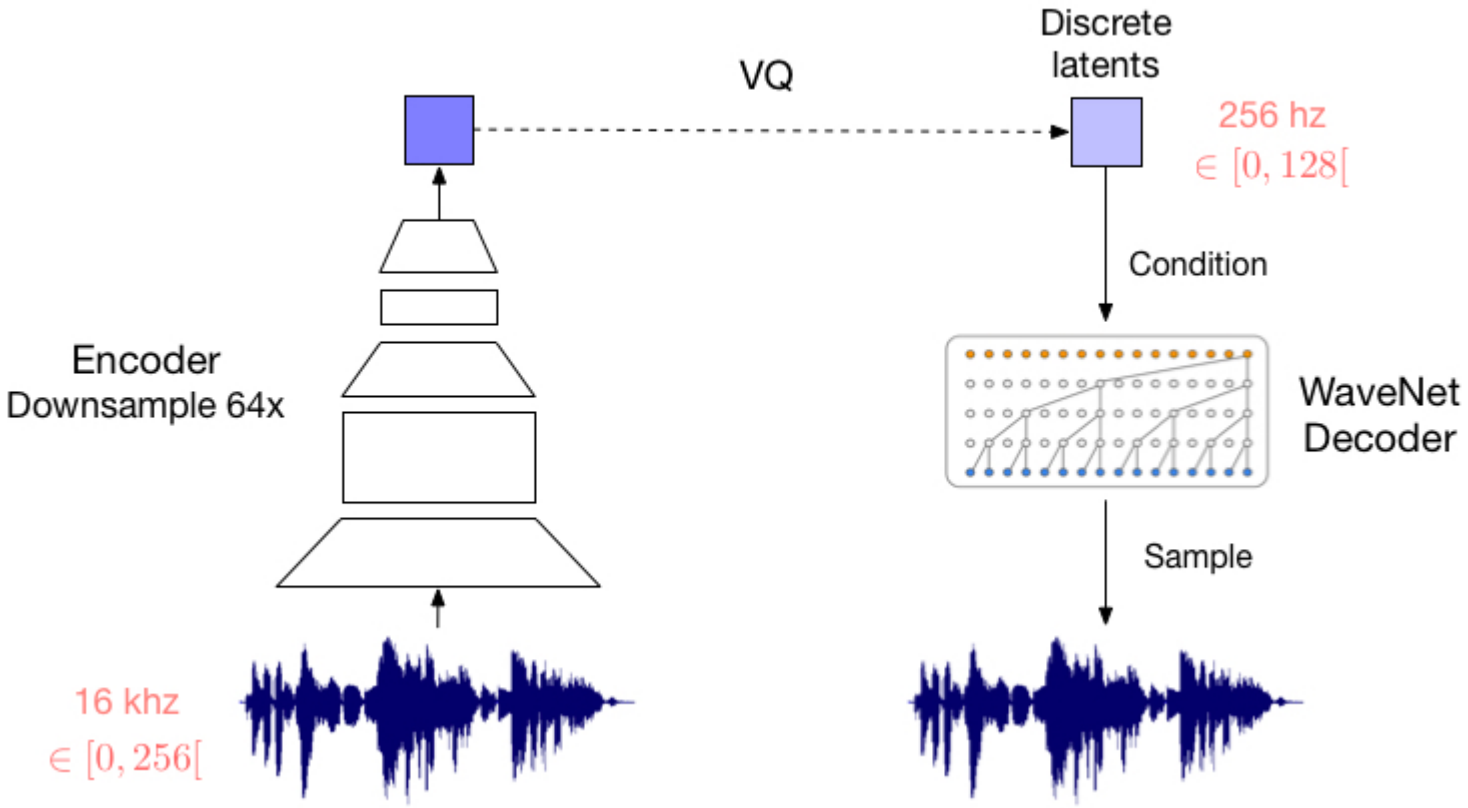
Experiments

Image Sampling with PixelCNN prior

- Samples (128x128) from a VQ-VAE with a PixelCNN prior trained on ImageNet images.

Experiments

Audio Reconstruction / Sampling / Style-Transfer



- https://avdnoord.github.io/homepage/vqvae/

Interpretability as a phoneme

- Speech데이터에 대해 학습이 끝난 VQ-VAE로 128-K 사이즈의 discrete latent variable을 매 time-step마다 뽑는다.
- 동일한 time-step에 대해 이미 알려진 41가지 ground-truth phoneme을 1:1대응 시킨다.(training에서는 gt-phoneme을 알 수 없음)

다음은 추측임.
- 이를 아마도 단순히 count-based로 argmax p(phoneme | code_book seq)을 training에서 구해서 test에 대해 적용한 classification 성능이 49.3% 였다.
- 그리고 단순히 argmax p(phoneme | random codebook)을 한 결과는 7.2%였다.

Code collapse

- VQ-VAE의 대표적 문제 중 한가지로, code book에 모든 index가 사용되지 않는 문제.
- 후속논문에서 이를 해결하고자 2가지 codebook의 곱셈형태로 discrete latent를 표현함

Thank you