VQ-VAE(Neural Discrete Representation Learning)

Vector Quantization이란?

간단한 정의)

: N개의 특징 벡터 집합 x를 K개의 특징 벡터들의 집합 Y로 사상(mapping)하는것.

사상=매핑은 특정 데이터를 특정 카테고리로 맞춰준다는 뜻.

이때, 매핑 하는 방법을 사상함수라고 하며, y=f(x)라고 했을때 f()를 양자화 연산자라고 한다.

벡터 Y의 각 특징(원소)들은 벡터X의 원소가 매핑되는 큰 카테고리이고, 코드워드, 코드벡터, 클러스터라고 부르며, 이 Y는 코드북이라고 부른다.

Neural Discrete Representation Learning

> https://github.com/1Konny/VQ-VAE

본 논문에서는 이러한 이산 표현을 학습하는 간단하면서도 강력한 생성 모델을 제안한다. 제안하는 **VQ-VAE(Vector Quantised-Variational AutoEncoder)**모델은 두 가지 면에서 VAE와 다르다:

1. Encoder network는 연속적인 codes를 생성하나 VQ-VAE는 이산 codes를 출력한다.
2. Prior는 정적(static)이지 않고 대신 학습이 가능하다.

이산 잠재 표현을 학습하기 위해 벡터 양자화(**VQ: Vector Quantisation**)의 아이디어를 통합하였다.

VQ 방법을 사용하면 모델이 강력한 autoregressive decoder와 짝을 이룰 때 latent들이 무시되는 **“Posterior Collapse”**문제를 피할 수 있다.

이러한 표현을 autoregressive prior와 짝지으면 모델은 고품질 image, video, audio을 생성할 수 있을 뿐만 아니라 speaker conversion 및 음소의 비지도 학습을 수행하여 학습된 표현이 유용함을 보일 수 있다.

Motivation

언어는 본질적으로 이산적인 성질을 가지며(단어의 수는 한정적이며, 단어와 단어 사이 중간쯤이 명확히 정의되지 않음을 생각하면 된다), 음성도 비슷한 특성을 가진다. 이미지는 언어로 표현될 수 있다. 이러한 점에서, 이산적인 표현은 이러한 이산적인 domain에 잘 맞을 것이라 생각할 수 있다.

VQ-VAE는 vector quantization이라는 기법으로 latent space를 discrete 하게 만든다.

Vector Quantization

VQ-VAE는 Autoencoder 구조에 discrete 한 codebook을 더한 구조이다.

Codebook은 기본적으로 벡터를 요소로 가지는 리스트입니다. Encoder의 출력으로 어떤 벡터가 나오면, codebook의 모든 벡터들 간 거리를 계산하게 된다.

Codebook의 벡터들 중 encoder의 출력 벡터와 가장 거리가 짧은 벡터를 찾는다. 이후, 그 벡터를 decoder에 넣어 학습한다.

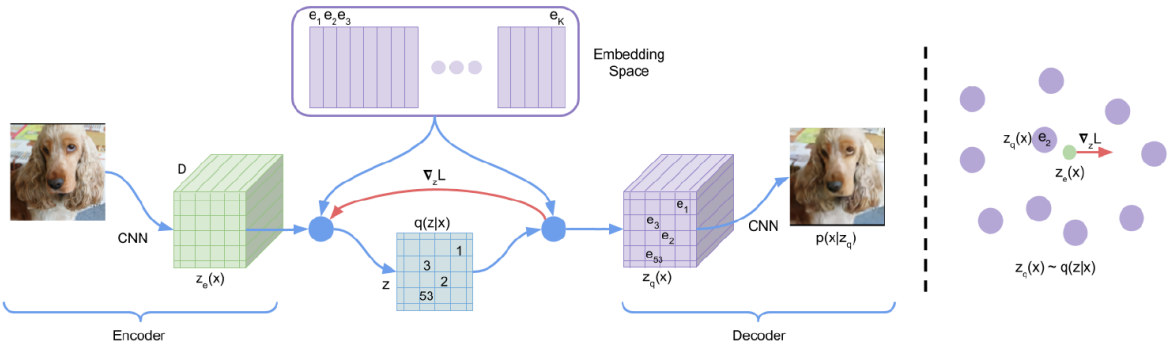
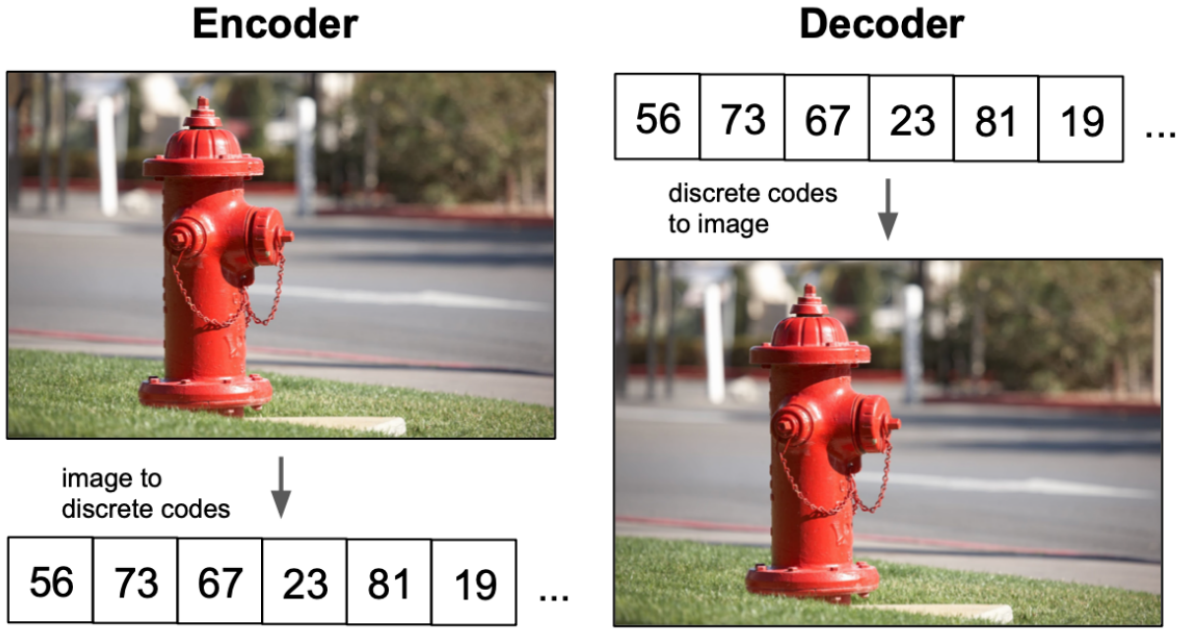


Figure 1: Left: A figure describing the VQ-VAE. Right: Visualisation of the embedding space. The output of the encoder $z(x)$ is mapped to the nearest point e_2. The gradient $\nabla_z L$ (in red) will push the encoder to change its output, which could alter the configuration in the next forward pass.

위 그림에서 embedding e ∈ RK×D가 이산 표현을 나타내고, 이를 codebook이라 한다. K는 이산 표현 공간의 크기(K way categorical과 같음), D는 각 embedding vector ei의 차원임.

ei ∈ RD, i ∈ 1,2,…,K, 즉 embedding vector K개가 존재하는 상황인 것이며, 벡터의 개수는 제한되는 상황인 것이다.

- Codebook 안의 벡터들의 개수가 제한되면 다양한 이미지를 생성하는 것이 가능한지에 대한 의문이 들 수 있는데, 사실 일반적으로 encoder의 출력은 하나의 벡터가 아니다.
- 예를 들어 이미지를 학습하는 경우 32x32x1 벡터가 출력되도록 encoder를 디자인할 수 있다. 그럼 각 grid는 codebook의 벡터들 중 가장 가까운 벡터로 변환될 것이고, 만약 codebook list의 크기가 512라면 512^(32x32) 만큼의 distinct 한 이미지들을 생성할 수 있게 된다.
- 32x32 feature map의 각 요소 값은 가장 가까운 벡터의 index로 사상되므로, discrete한 vector space를 얻었다고 할 수 있는 것이다.



VQ-VQE의 loss function은 세 term으로 이루어짐.

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

첫 번째 항은 reconstruction loss, 두 번째 항은 codebook alignment loss, 그리고 세 번째 항은 codebook commitment loss.

- sg는 해당 term에는 weight가 업데이트되지 않는 "stop gradient"를 의미.
- Codebook alignment loss는 encoder의 출력인 z_e(x)와 가장 가까운 codebook 내 vector e_i가 z_e(x)와 더 가까워지도록 함.
- Codebook commitment loss는 그 반대로, encoder의 출력 z_e(x)와 가장 가까운 codebook 내 vector e_i가 z_e(x)와 더 멀어지도록 함.

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

Learning priors

- VQ-VAE를 학습하는 과정에서는 prior를 uniform distribution으로 두게 된다.
- VQ-VAE로 discrete 한 latent space를 얻어 학습이 종료되면, prior를 autoregressive model를 활용하여 latent space와 닮아가도록 학습을 진행시킬 수 있다.

Contribution

- VAE와 discrete latent 표현을 위한 VQ를 결합하여 새로운 생성 모델을 제안

- 이러한 latent는 discrete하며, continuous latent와 비교하여 성능이 나쁘지 않다.
- Image, audio, video 모두에 대해서 잘 모델링 및 압축한 후 중요한 내용을 잘 보존하면서 복원이 가능하다.
- Continuous한 latent representation을 위주로 다룬 vanilla VAE에 비해, discrete latent representation을 다루는 방법론이다.