Generative Model

2020. 09. 18

방준일

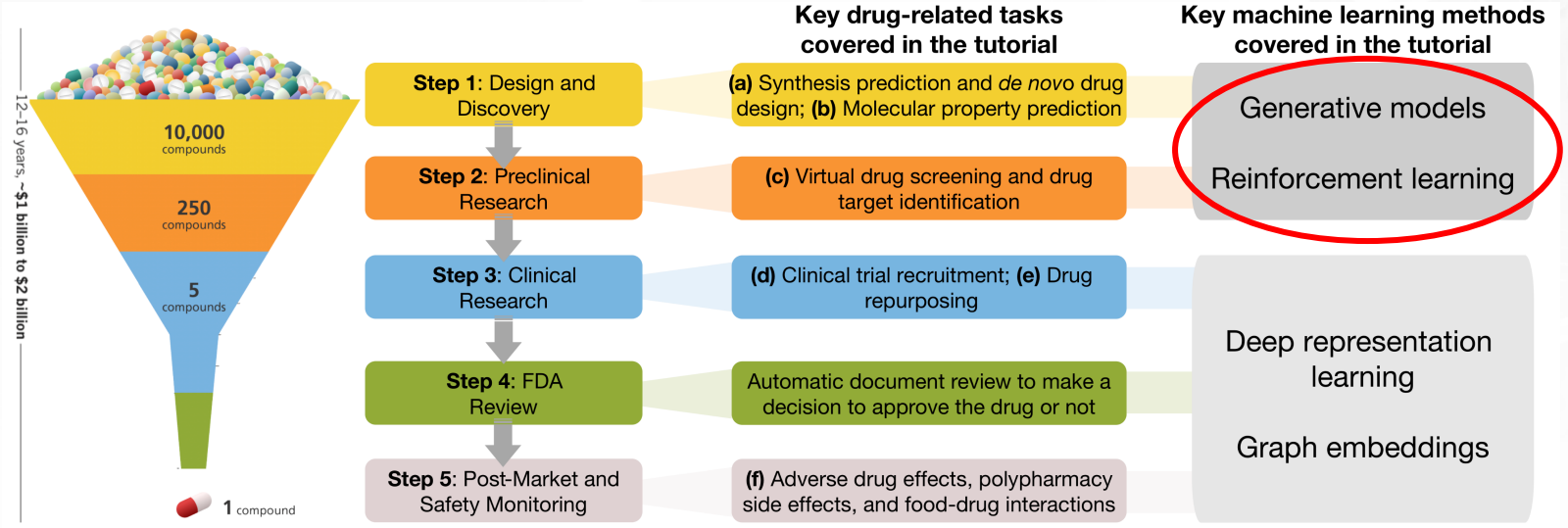
Korea AI Center for Drug Discovery and Development

Contents

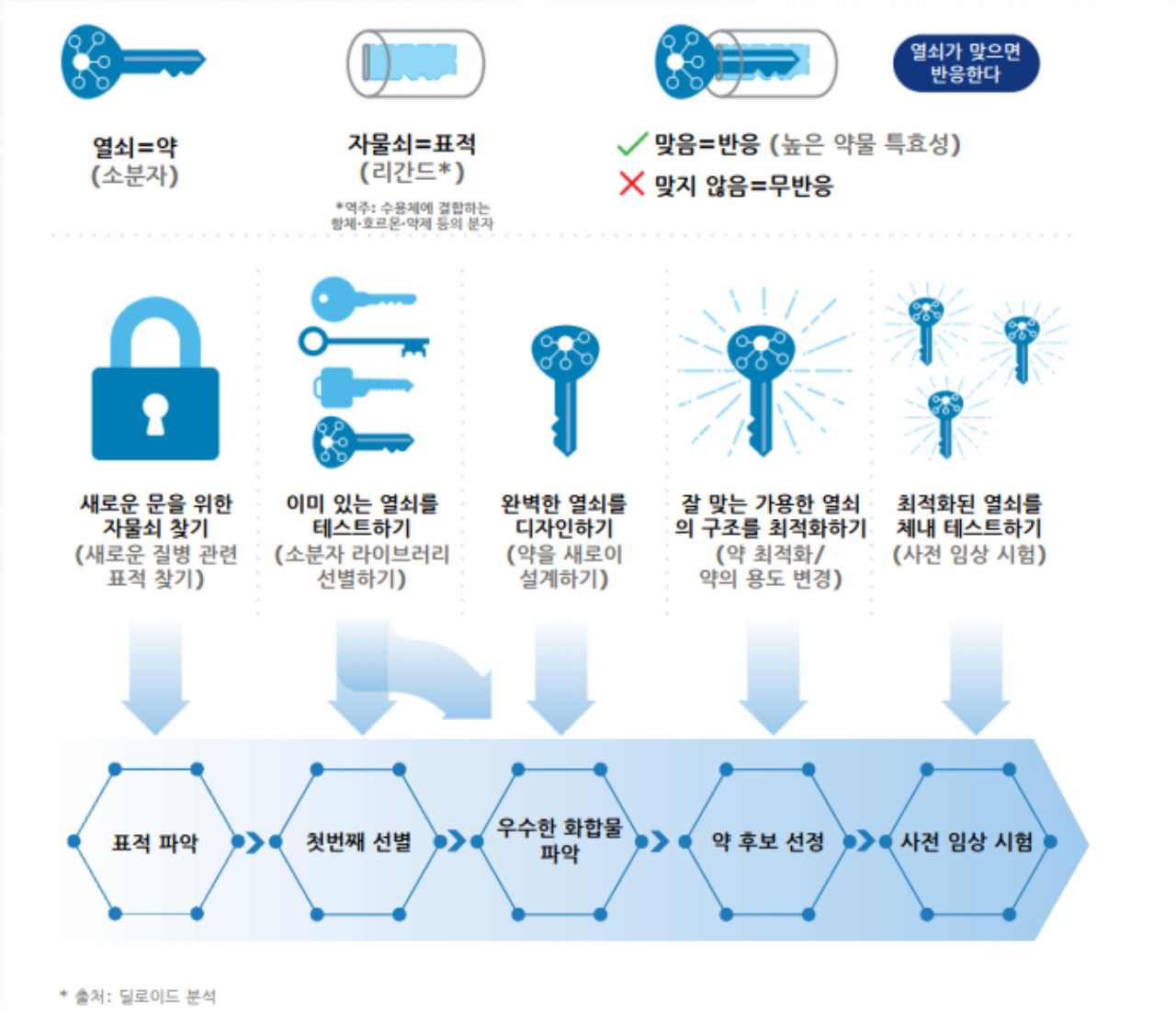
1. AI Based Drug Development

AI Based Drug Development

- a : 합성 예측 및 새로운 **약물 설계**, 복잡한 분자 합성을 지원하기 위해 화학적으로 올바른 구조를 생성
- b : 분자 **특성 예측**, 분자 데이터에서 효능, 생물 활성 및 독성과 같은 특성을 예측하여 분자의 치료 효과를 식별
- c : 가상 약물 스크리닝 및 **약물 표적 식별**, 약물이 표적 단백질에 결합하고 다운 스트림 활동에 영향을 줌으로써 약물이 인체에 미치는 영향을 예측



AI Based Drug Development



2. 개요

Classification

- 분류 문제에서 사용되는 모델은 크게 2가지

- **Discriminative** Model – **판별** 모델

- **Generative** Model – **생성** 모델

Discriminative Classifiers

- Logistic regression
- Scalar Vector Machine
- Traditional neural networks
- Nearest neighbour
- Conditional Random Fields (CRF)s

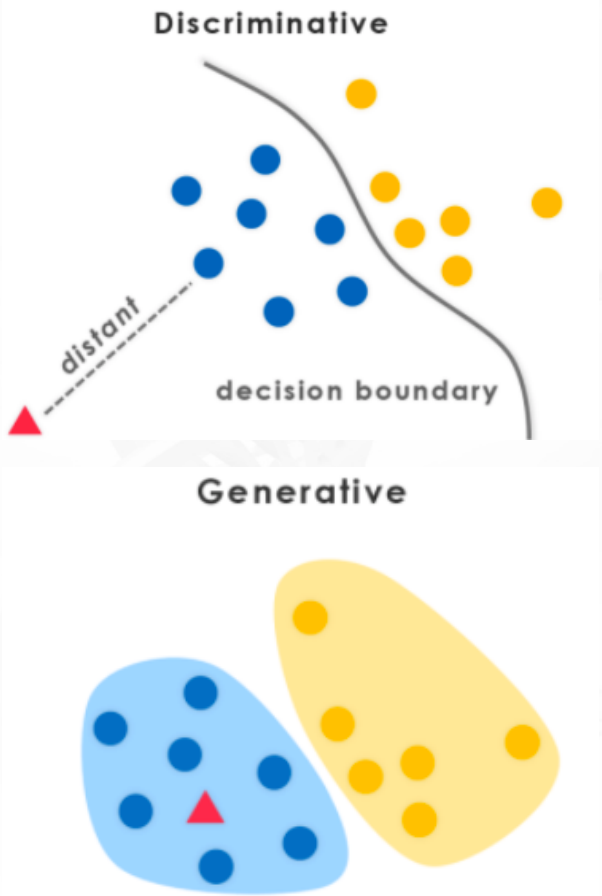
Generative classifiers

- Naïve Bayes
- Bayesian networks
- Markov random fields
- Hidden Markov Models (HMM)

Classification

- Class의 **차이점**에 주목하여 확률적 계산을 통해 판별
  - 데이터가 충분할수록 좋은 성능
  - 구분만이 목적, 본질에 대한 이해가 어려움

- Class의 **분포**(거리, 밀도 등)에 주목
  - 데이터가 충분하지 않아도 성능을 보임
  - 본질 파악 가능
  - **유사 데이터 생성 가능**



Generative Model

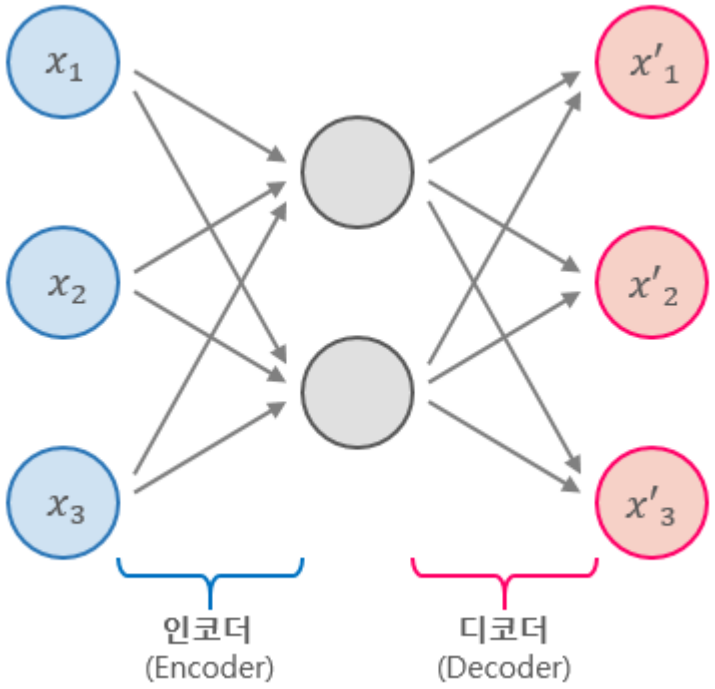
- Loss 값의 계산과 Back Propagation(역전파) 학습이 어려움
- 성능을 명확히 수치화하기 어려워 사용이 많지 않았음
- 여러 방식들의 활용 보고로 인하여 가능성이 조명을 받게 됨

3. 기초

Auto Encoder

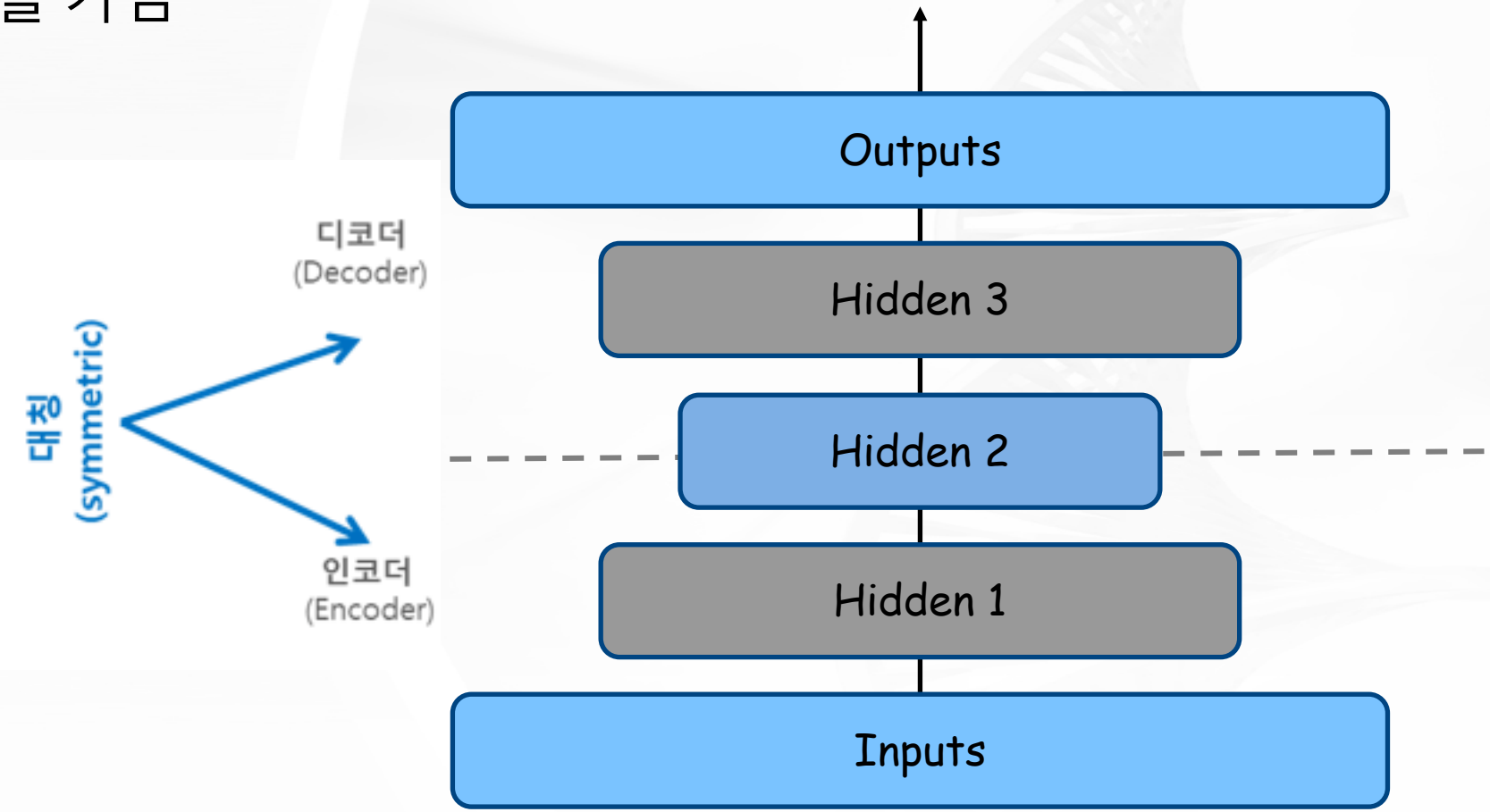
- 자기 자신을 **재생산**하는 모델

- Recognition network – encoder
- Generative network – decoder



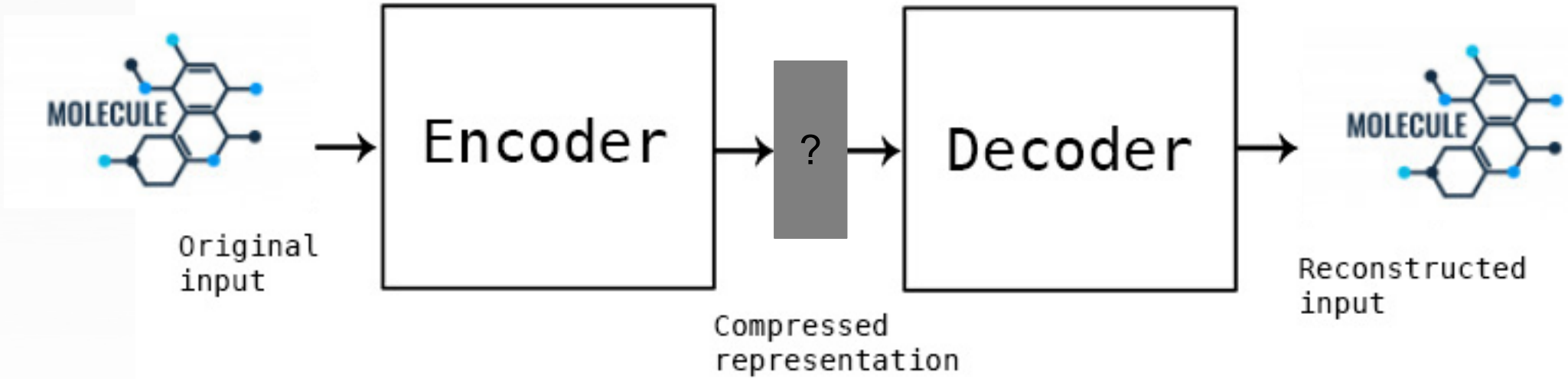
Auto Encoder

- Stacked Auto Encoder
- 여러 개의 히든 레이어를 가짐
- 보다 복잡한 학습 가능



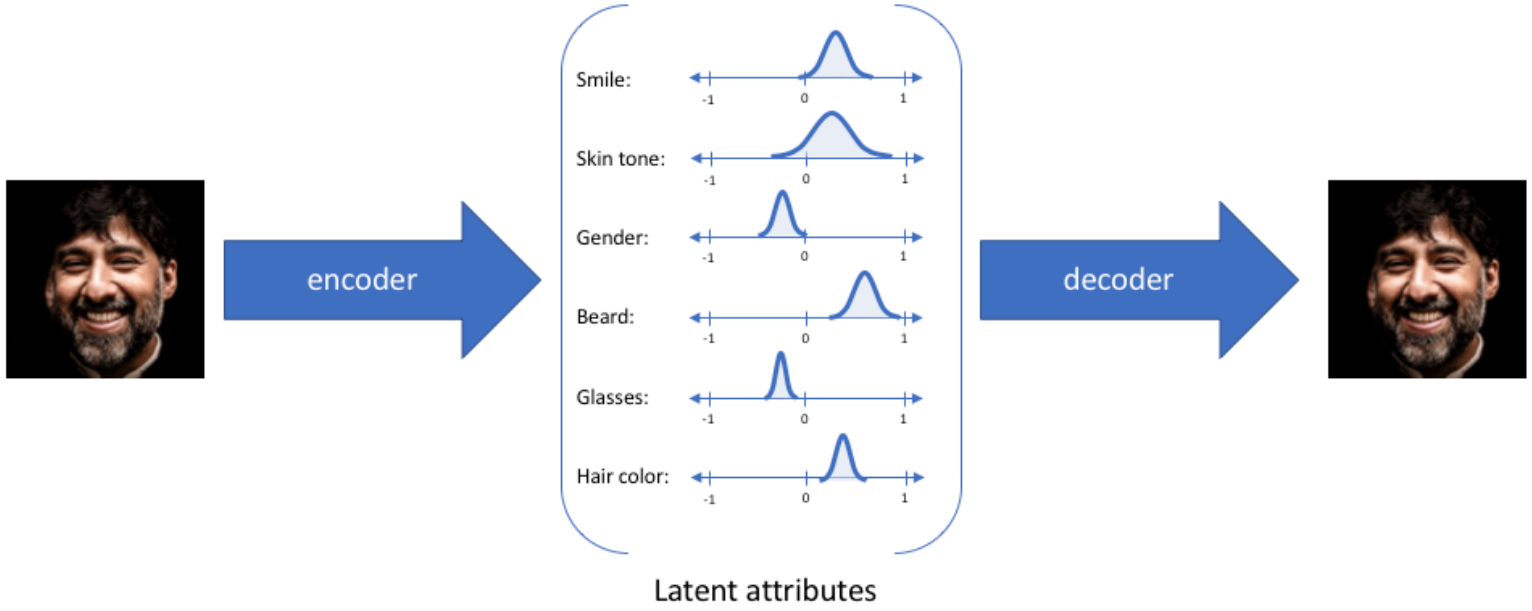
Auto Encoder

- 압축된 Latent Space는 input 데이터의 특징을 가지고 있음
- 해당 특징을 보존하여 output 생성



Auto Encoder

- 그림 예시



참조 : https://www.jeremyjordan.me/variational-autoencoders/

Auto Encoder

- Data- Specific :
  훈련된 데이터와 비슷한 데이터로만 압축 및 생성 가능

- 자동적 학습 :
  특정 입력값에 대하여 잘 작동하는 모델로 훈련

- 손실 값 존재 :
  압축 해제, 즉, 디코딩되어 생성된 값은 원본 보다 손실된 결과

Auto Encoder

- Auto Encoder는 타 알고리즘들의 활용 빈도수 보다 낮았으며, 실제 응용이 드물었음

- 다만, 비지도 학습의 문제를 풀어낼 **KEY**로 생각되었음

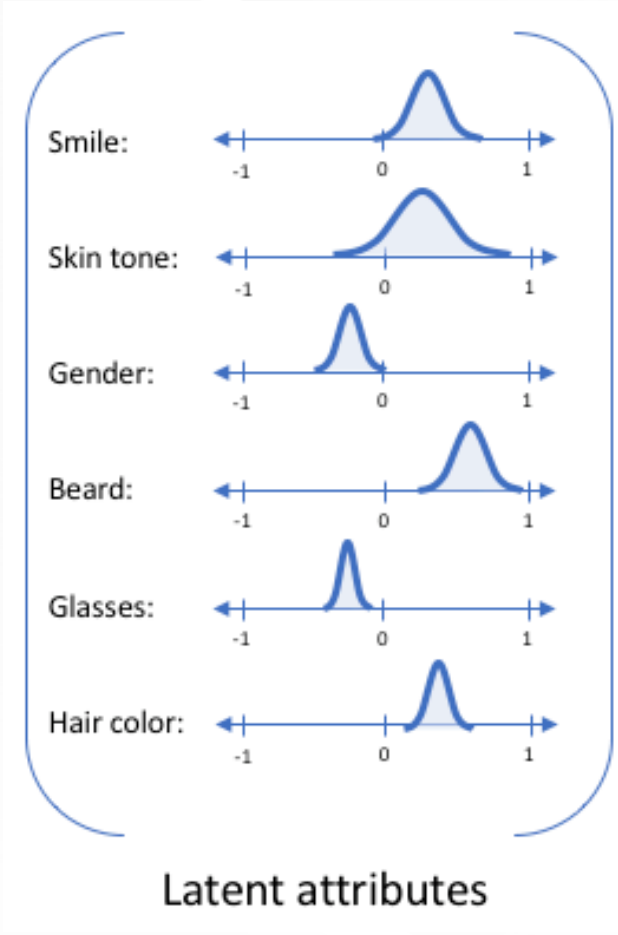
- Deep Auto Encoder, Convolutional Auto Encoder, Sequence-to-sequence Auto Encoder(used RNN) 등의 방향으로 발전

4. VAE

Variational Auto Encoder

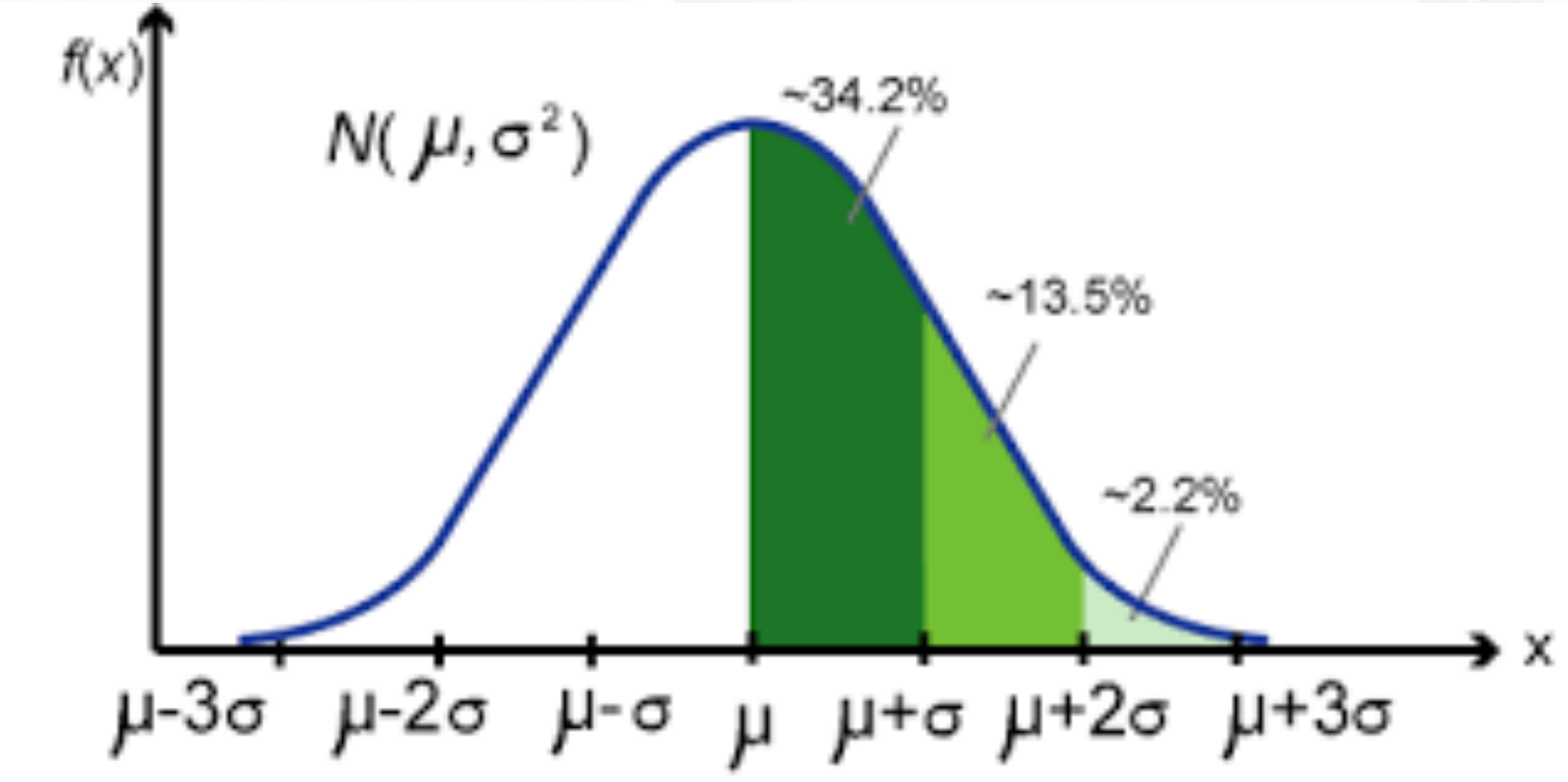
- 제약 조건이 추가된 Auto Encoder
- 입력 데이터에 대한 Latent Variable Model을 학습
- 임의의 함수 학습 대신, 데이터를 모델링하는 확률 분포의 매개변수를 학습

Variational Auto Encoder



Variational Auto Encoder

$$z=\mu(x)+\sigma(x)\times\epsilon,\epsilon\sim N(0,1)$$

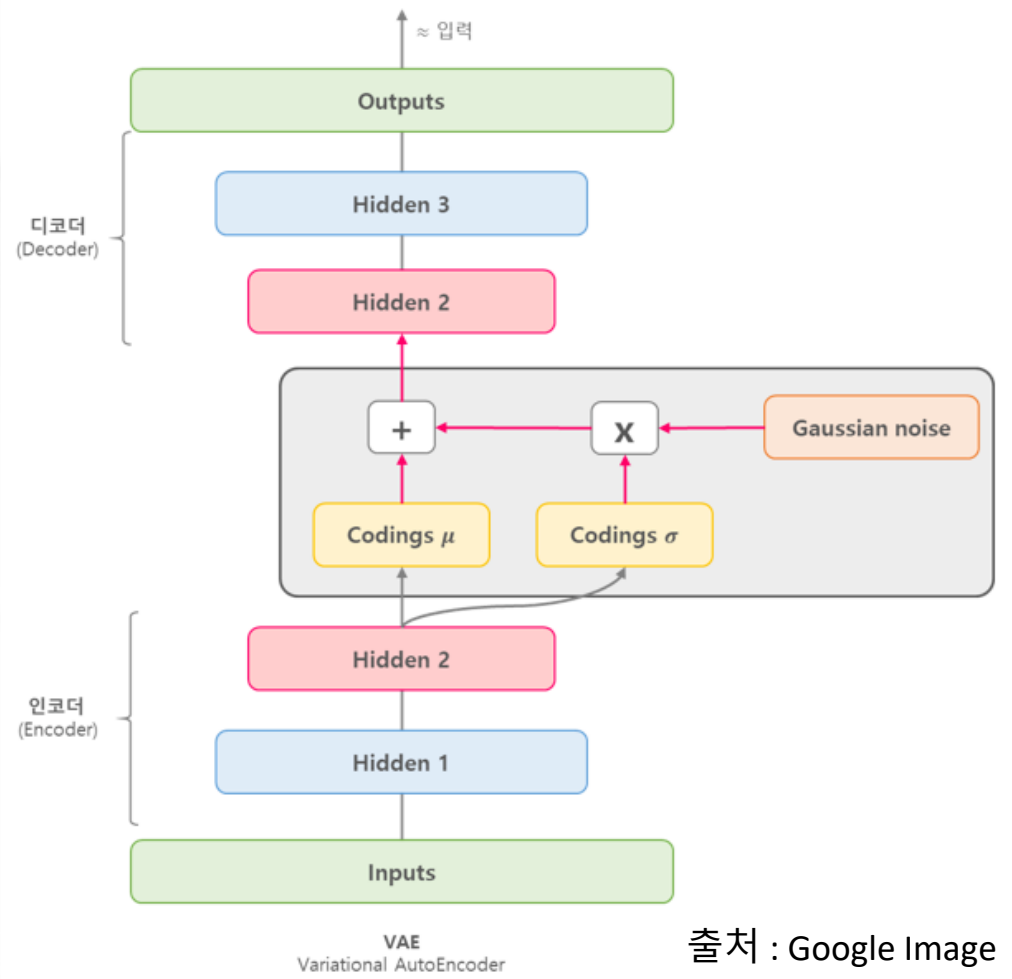


Variational Auto Encoder

- 확률적 오토인코더 :
  학습이 끝난 후에도 출력이 부분적으로
  우연에 의해 결정 – 가우시안 노이즈 영향

- 생성 오토인코더 :
  데이터셋에서 샘플링 된 것과 같은
  새로운 샘플을 생성 가능 – 가우시안 분포 영향

- 2014년 D.Kingma와 M.Welling
  Auto-Encoding Variational Bayes



출처 : Google Image

Variational Auto Encoder

- 보통의 Auto Encoder와 다르게, 주어진 입력에 추가적 연산

- 평균이 μ이고 표준편차가 σ 인 가우시안 분포에서 랜덤하게 샘플링 후, 디코더가 원본 입력으로 재구성
  -> 가우시안 분포(정규분포)에서 샘플링 된 것처럼 보이는 연산
  $z=\mu(x)+\sigma(x)\times\epsilon, \epsilon\sim N(0,1)$

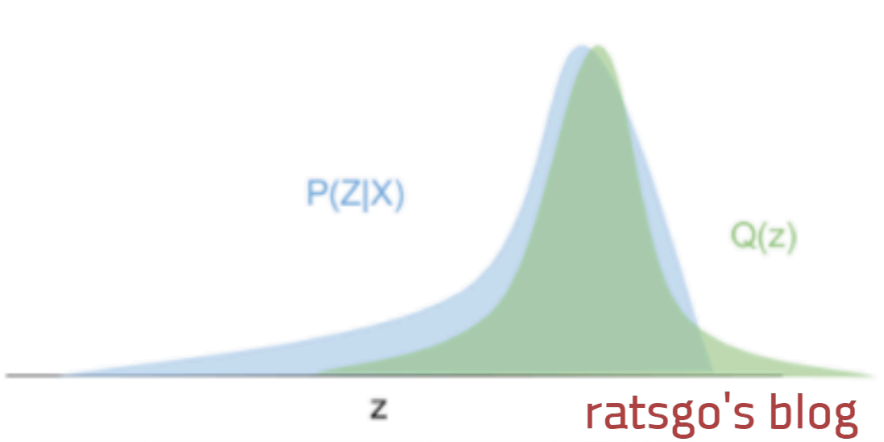
- 학습하는 동안 손실함수가 가우시안 분포와 유사한 형태의 Latent Sapce(잠재변수 공간)을 따르도록 함

Variational Auto Encoder

- KL Divergence - 쿨백-라이블러 발산(Kullback-Leibler divergence, KLD)
  - 두 확률분포의 차이를 계산하는 데 사용하는 함수

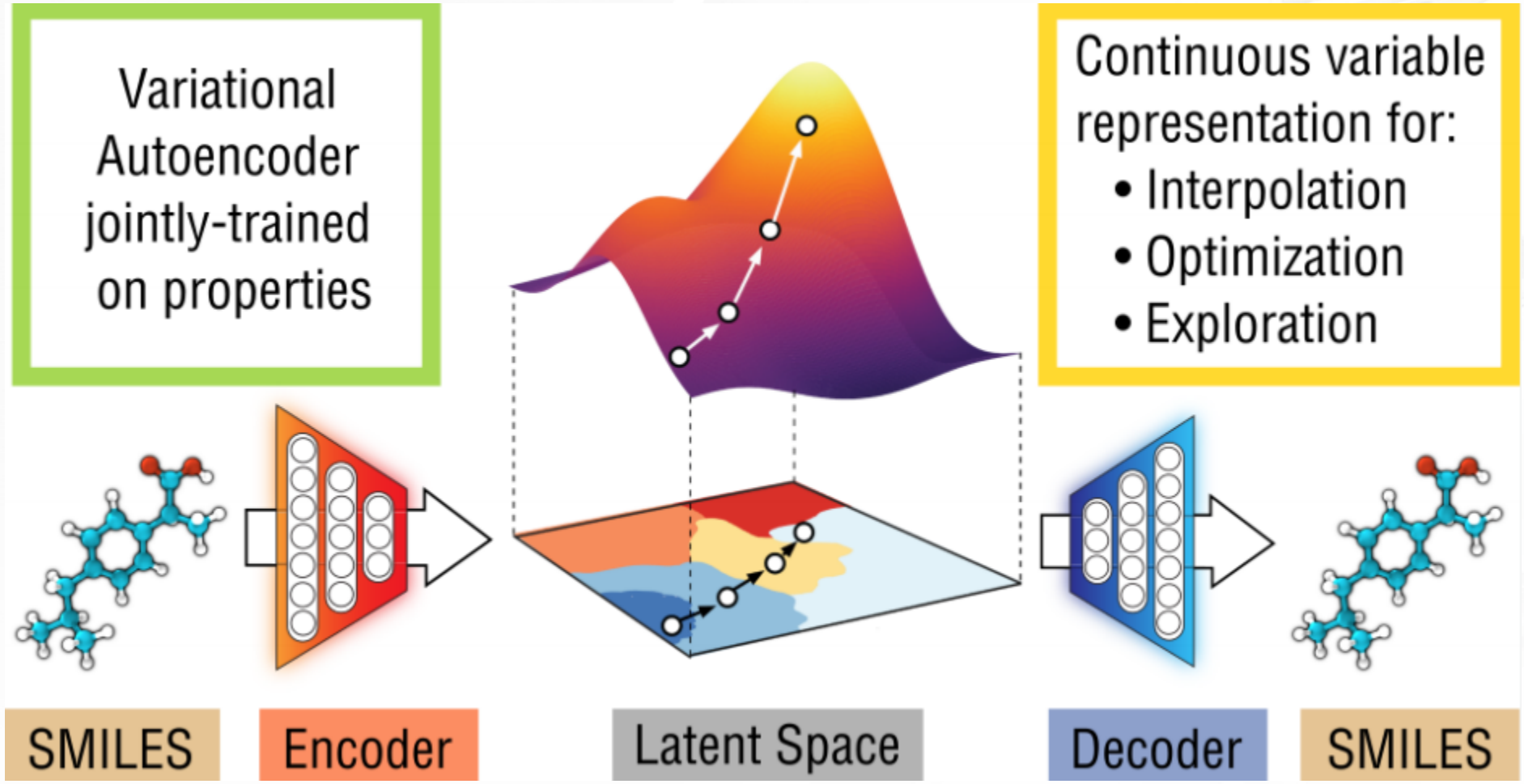
$$D_{KL}(P||Q) = E_{X\sim P}\left[\log \frac{P(x)}{Q(x)}\right] = E_{X\sim P}\left[-\log \frac{Q(x)}{P(x)}\right]$$

  - 두 확률 분포의 차이를 줄이는 것으로 학습



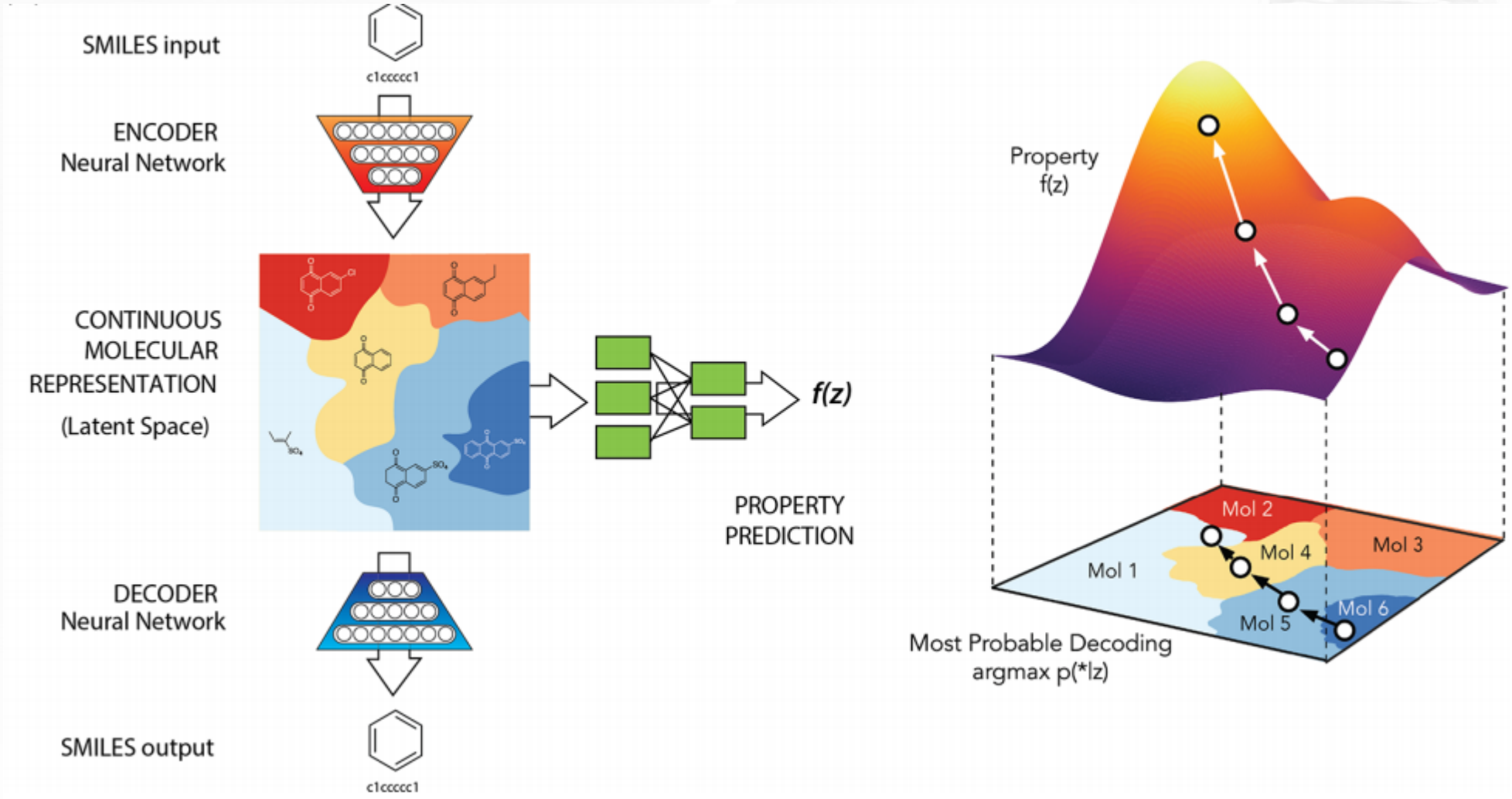
참조 :ratsgo's blog

Variational Auto Encoder



참조 : Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules

Variational Auto Encoder



참조 : Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules

Variational Auto Encoder

- 복잡한 데이터 생성 모델을 설계하고 대규모 set에 적응 할 수 있게 해줌

- VAE는 data 분포가 잘 학습되기만 하면 sampling (=data generation)이 자연스럽게 따라옴

Variational Auto Encoder vs GAN

- Generative Adversarial Networks (GAN) 의 특성 :

- Generator Model의 목적 자체가 어떤 data의 분포를 학습하는 것이 아님

- 진짜 같은 Sample을 Generate하는 것이 목적

5. 실습

예제

- VAE 실습
- **Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules -** Rafael Gomez-Bombarelli, et al
- **https://nextjournal.com/a/MyupVXGXaCQxLpJfKKe1H/**