澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

Reinforcement Learning for Generative AI:A Survey

speaker: Xiongyi Li

Dec 23th 2023

1.Background

1.1 Generative Models

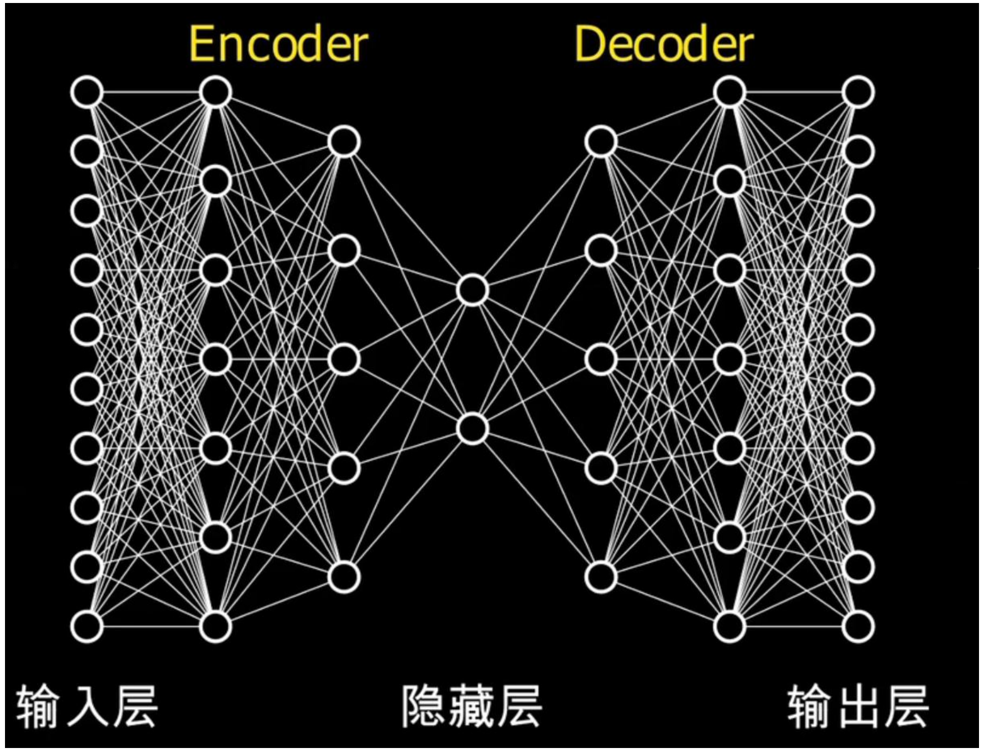
- Variational Autoencoder
- Generative Adversarial Networks
- Energy-Based Models
- Autoregressive Models
- Normalizing Flows

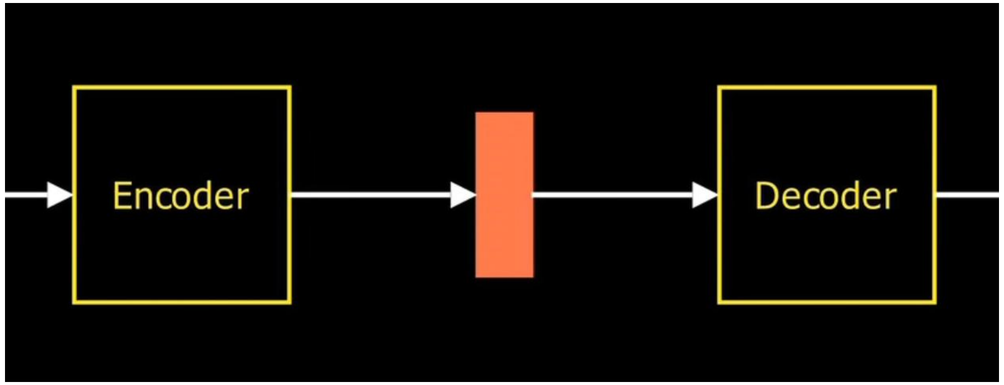
1.2 Reinforcement Learning Methods

- Markov Decision Process
- Model-free Methods
- Model-based Methods

1.1 Generative Models

- Variational Autoencoder
- Auto-encoder



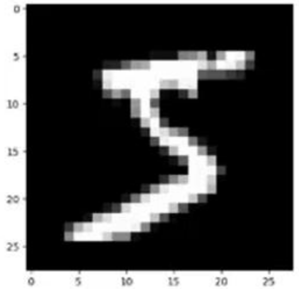


1.1 Generative Models

- Generative Adversarial Networks

1.1 Generative Models

- Models
- $y = f_w(x)$
- x = 784 vectors(28*28)
- y = 10 vectors
- w = parameter
- f = neural networks



28x28=784 -> (0,0,0,0,0,1,0,0,0,0)

- Energy-Based Models
- $E = f_w(x,y)$
- x = 784 vectors(28*28)
- y = 10 vectors
- x,y = 794 vectors
- when (x,y), E get small
- w = parameter
- f = neural networks

1.1 Generative Models

- Autoregressive Models

$$p(x) = p(x_1, x_2, \ldots, x_n) = \prod_{i}^{n} p(x_i | x_1, \ldots, x_{i-1})$$

1.1 Generative Models

- Normalizing Flows

$$\ln p(x_n) = \ln p(x_1) - \sum_i ln \left|\det \frac{f_i}{x_{i-1}}\right|$$

1.2 Reinforcement Learning Methods

- Markov Decision Process

1.2 Reinforcement Learning Methods

- Model-free Methods

$$V^{\pi}(s_t) = \mathbb{E}_{s_{t+1}}[r_{t+1} + \gamma V^{\pi}(s_{t+1})] \text{ and } Q^{\pi}(s_t, a_t) = \mathbb{E}_{s_{t+1}}[r_{t+1} + \gamma Q^{\pi}(s_{t+1}, \pi(s_{t+1}))]$$

1.2 Reinforcement Learning Methods

- Model-based Methods
- AlphaGo

2. Benefits of RL-Based Generative Models

2.1 Solving the Non-dfferentiable Learning Problems

- The genrated variable is non-differentiable
- The training objective is non-differentiable

2.2 Introducing New Training Signal

- Reward by Discriminator
- Reward by Hand-designed rules
- Reward and Divergence
- Reward by data-driven model

2.3 Neural Architecture Search

- State and Action Design
- Sample effiency

um 澳大

2.1 Solving the Non-dfferentiable Learning Problems

- The genrated variable is non-differentiable
- This is achieved by establishing a connection between policy gradient and the GAN objective. The data distribution is optimized by Monte-Carlo estimation,thereby mitigating variance in the policy gradient.

2.1 Solving the Non-dfferentiable Learning Problems

- The training objective is non-differentiable
- BLEU:A Method for Automatic Evaluation of Machine Translation
- $Bleu_n$= m/n (m: both in C and R; n: the length of C)
- Candidate: the cat sat on the mat
- Reference: the cat is on the mat
- $Bleu_1$ = 5/6
- $Bleu_2$ = 3/5

2.2 Introducing New Training Signal

- Reward by Hand-designed rules
- BLEU

2.2 Introducing New Training Signal

- Reward and Divergence
  - KL divergence

2.2 Introducing New Training Signal

- Reward by data-driven model
  - Reward function with a entropy term
  - Learn a model for human preference

2.3 Neural Architecture Search

- Construction of search spaces,
- Optimization algorithms
- Model evaluation

3. Challenges

3.1 Peaked Distribution

3.2 Exploration and Exploitation

3.3 Sparse Reward

3.4 Long-term Credit Assignment

3.5 Generalization

4. Applications

- Natural Language Processing
- Code Generation
- Computer Vision
- Speech and Music Generation
- AI for Science
- Recommender System and Information Retrieval
- Robotics

Thank You!

Avenida da Universidade, Taipa, Macau, China
Tel : (853) 8822 8833 Fax : (853) 8822 8822
Email : mc35289@um.edu.mo Website : www.um.edu.mo