

NYCU DL

Lab5 - Conditional VAE for Video Prediction

謝宏笙

April 25, 2023

Outline

- Lab Objective
- Important Date
- Lab Description
- Scoring Criteria
- Implement Hints

Lab Objective

- In this lab, you need to implement a conditional Variational Autoencoder (VAE) for video prediction.
- Example:



Important Date

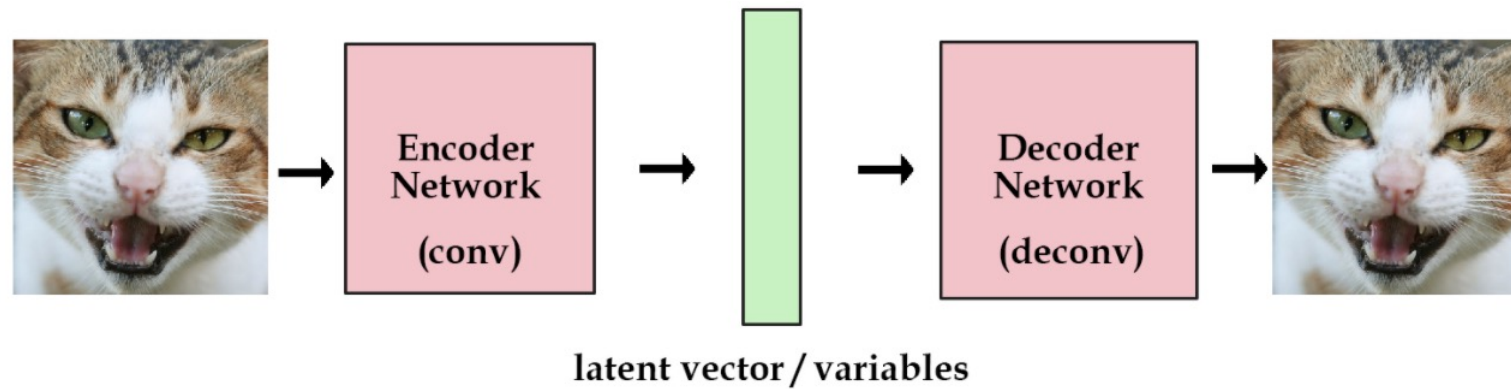
- Report Submission Deadline: 5/16 (Tue.) 11:59 a.m.
- Demo date: 5/16 (Tue.)
- Zip all files into one file
 - Report (.pdf)
 - Source code
- Name it like 「 DL_LAB5_yourstudentID_name.zip 」
 - ex: 「 DL_LAB5_310551109_謝宏笙.zip 」

Lab Description

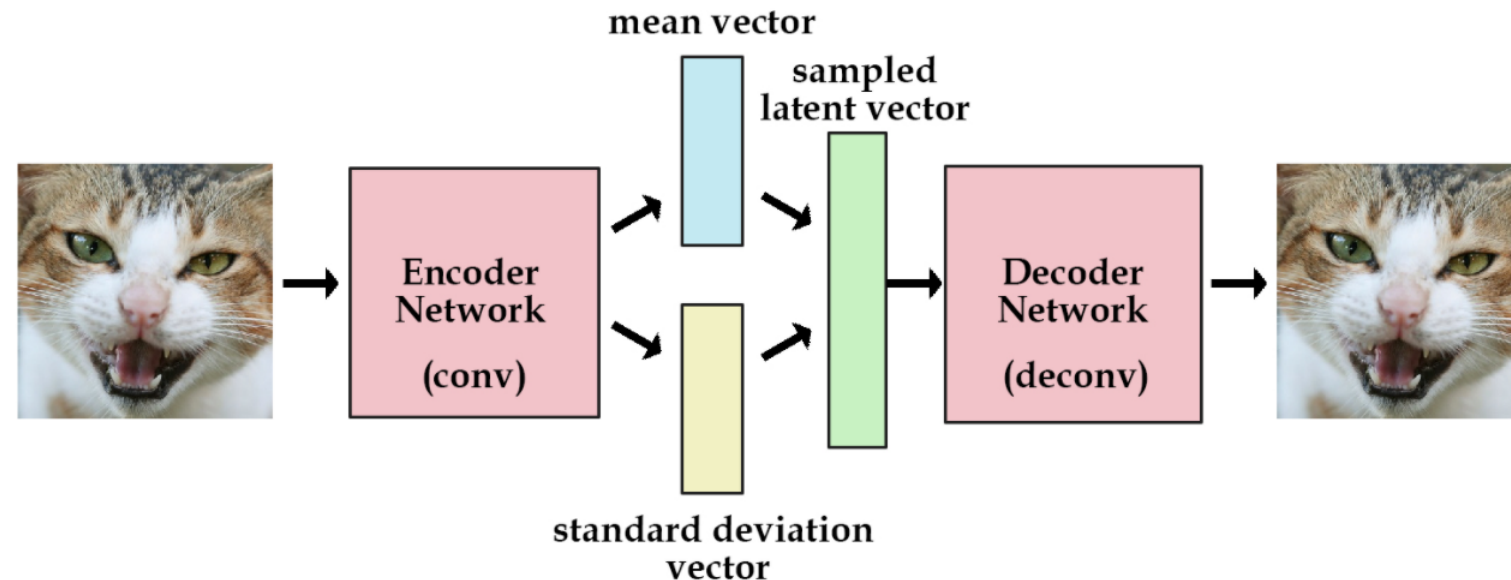
- Variational Autoencoder
- Reparameterization Trick
- Overall architecture
- KL Cost Annealing
- Dataset

Lab Description - VAE

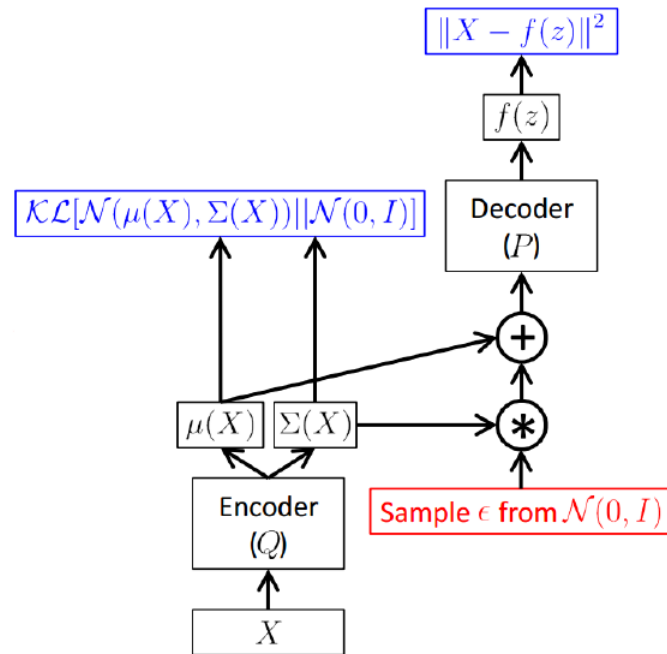
AE



VAE



Lab Description – Reparameterization Trick



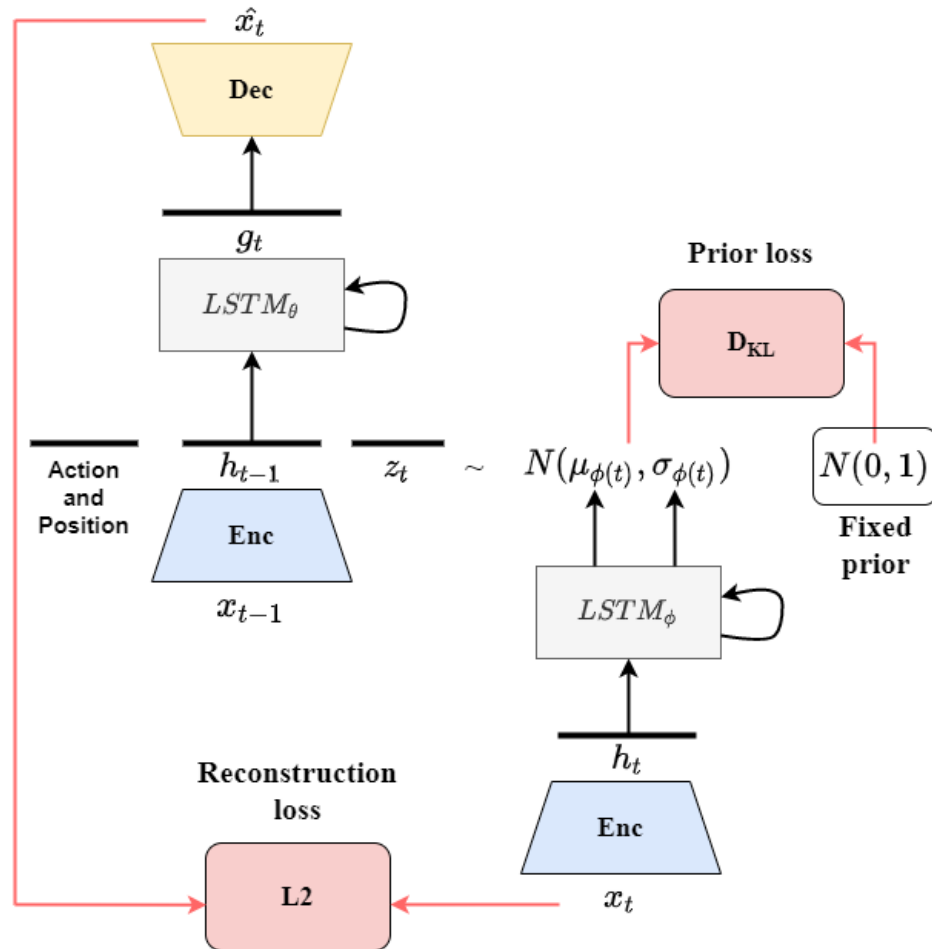
$$\mathcal{L}(X, q, \theta) = E_{Z \sim q(Z|X; \phi)} \log p(X|Z; \theta) - KL(q(Z|X; \phi) || p(Z))$$

where $q(Z|X; \phi)$ is considered as encoder and $p(X|Z; \theta)$ as decoder.

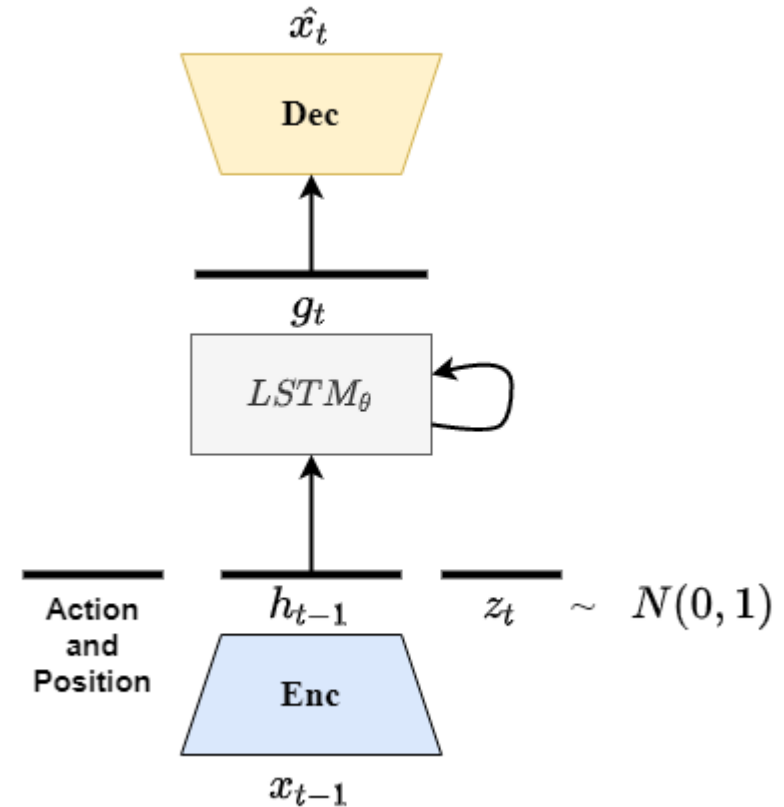
$$\underbrace{E_{Z \sim q(Z|X; \theta')} \log p(X|Z; \theta)}_{\text{Re-parameterization for end-to-end training}} - KL(q(Z|X; \theta') || p(Z))$$

- Log variance
 - Output should be log variance (not variance)

Lab Description – Overall architecture



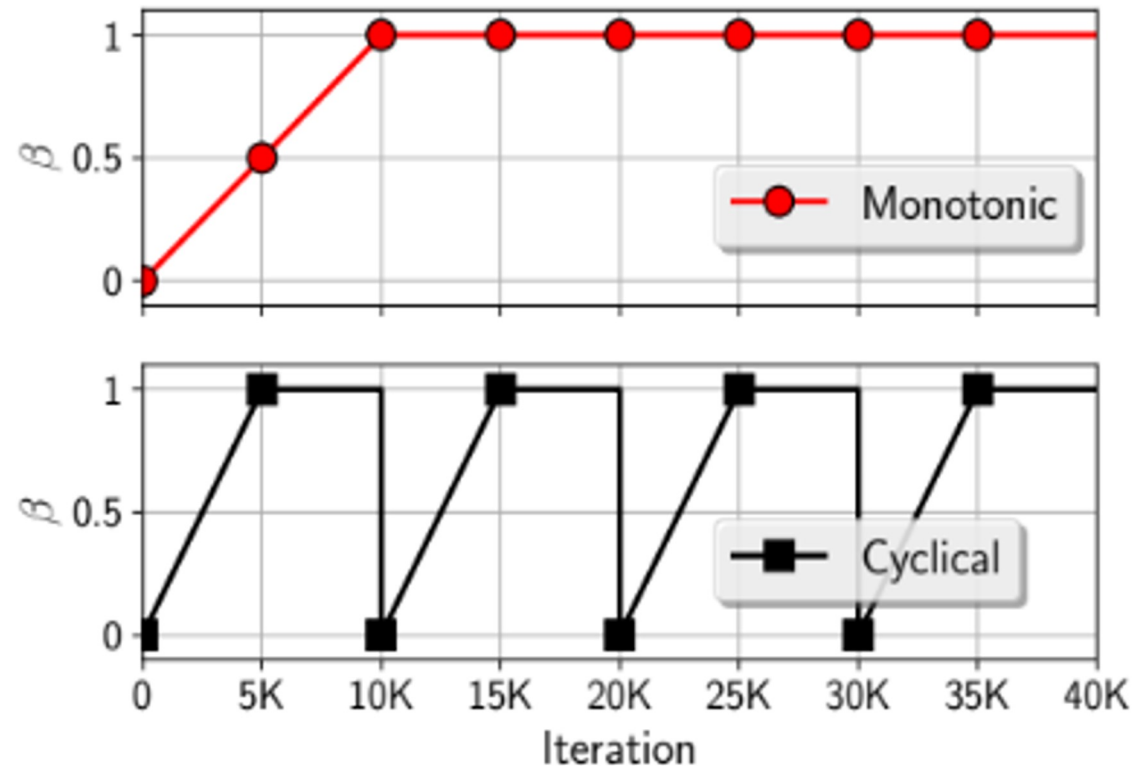
(a) Training procedure



(b) generating procedure

Lab Description - KL Cost Annealing

- Initially set your KL weight to 0
- Maximum value is 1



Lab Description - Dataset

- We use bair robot pushing small dataset to train CVAE
 - This data set contains roughly 44,000 sequences of robot pushing motions, and each sequence include 30 frames



Lab Description – Get dataset

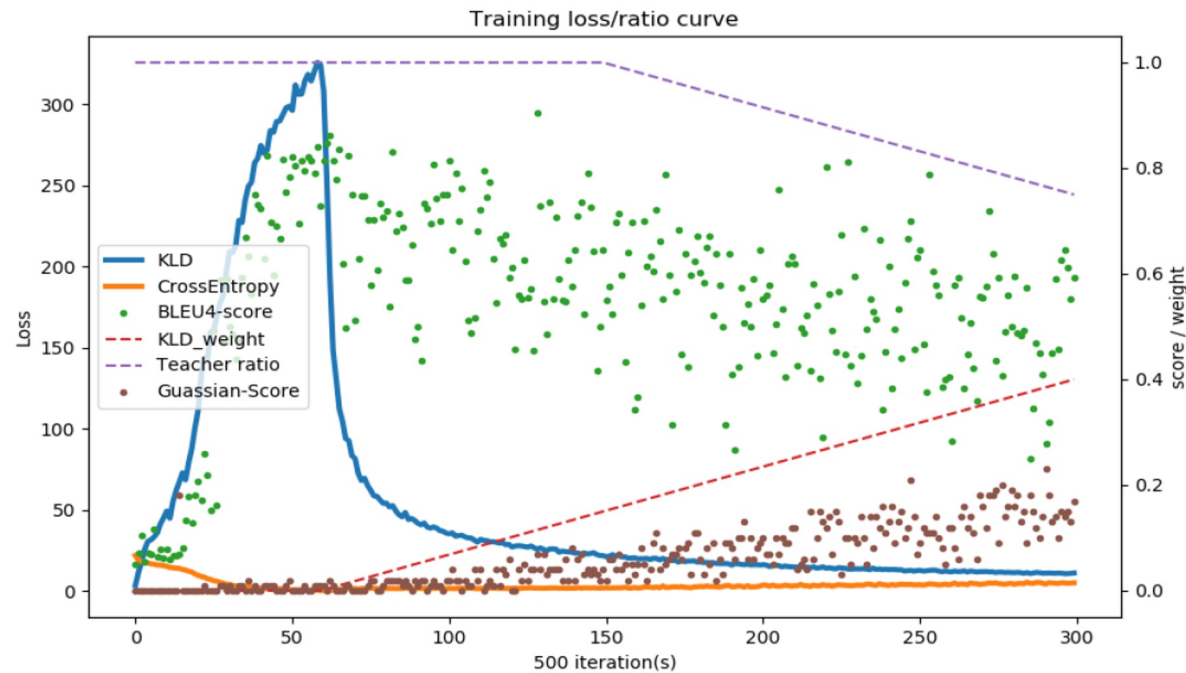
- For Linux
 - Please refer to the pdf file
- For Windows
 - Click the following link and download it: <https://reurl.cc/lvvZxq>

Lab Description – Other details

- Adopt PSNR to evaluate the perceptual quality.
 - Provide in the sample (finn_eval_seq)
 - Given two past frame to prediction next ten frames (average PSNR)

Lab Description – Requirements

- Modify encoder, decoder, and training functions
- Implement dataloader, and reparameterization trick.
- Adopt teacher-forcing and KL loss annealing in your training processing.
- Plot the losses, average PSNR and ratios.



Lab Description – Requirements

- Make videos or gif images for test result (**select one sequence**)



- Output the prediction at each time step (**select one sequence**)



Lab Description – Hints

- **Warning: You need to do the lab as early as possible** because you may take more than two days to train the model
- Model weights
 - Strongly recommend you save your model weights during training
- Teacher forcing ratio and KL weight
 - Influential to the performance of model
 - You can first set your KL weight to 0 to see whether your model works.

Scoring Criteria - Report (60%)

- Introduction(5%)
- Derivation of CVAE(10%)
- Implementation details(15%)
 - Describe how you implement your model. (e.g. dataloader, encoder, decoder, etc)
 - Describe the teacher forcing (including main idea, benefits and drawbacks)

Notice: You must prove that you use previous predicted frame to predict next frame, i.e. teacher forcing ratio = 0 when testing (paste/screenshot your code)

Scoring Criteria - Report (60%)

- Results and discussion(30%)
 - Show your results of video prediction
 - Make videos or gif images for test result (5%)
 - Output the prediction at each time step (5%)
 - Plot the losses, average PSNR and ratios. (5%)
 - Discuss the results according to your settings. (15%)

Notice: This part mainly focuses on your discussion, if you simply just paste your results, you will get a low score

Scoring Criteria - Demo(50%)

- Capability of video prediction.(20%)
 - Your model should use two past frames to predict the next ten frames
 - PSNR ≥ 25 ----- 100%
 - $25 > \text{PSNR} \geq 24$ ----- 90%
 - $24 > \text{PSNR} \geq 23$ ----- 80%
 - $23 > \text{PSNR} \geq 22$ ----- 70%
 - $22 > \text{PSNR} \geq 21$ ----- 60%
 - $21 > \text{PSNR} \geq 20$ ----- 50%
 - PSNR < 20 ----- 0%
- Questions (20%)

Scoring Criteria – Extra(30%)

- Implement learned prior(10%)
- Implement hierarchical structure(10%)
- Implement conditional convolution(10%)

Reference

- Yoojin Choi, Mostafa El-Khamy, and Jungwon Lee. Variable Rate Deep Image Compression With a Conditional Autoencoder. arXiv e-prints, page arXiv:1909.04802, Sept. 2019.
- Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. CoRR, abs/1802.07687, 2018.
- Zhihui Lin, Chun Yuan, and Maomao Li. Haf-svg: Hierarchical stochastic video generation with aligned features. In Christian Bessiere, editor, Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, pages 991–997. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track.