Generative Models

Discriminative vs Generative

Understanding Machine Learning

Archetype: Naïve Bayes

probabilistic model:

$$P(Y|X_1,\cdots,X_p) = \frac{P(Y,X_1,\cdots,X_p)}{P(X_1,\cdots,X_p)} = \frac{P(Y)P(X_1,\cdots,X_p|Y)}{P(X_1,\cdots,X_p)} \propto P(Y)P(X_1,\cdots,X_p|Y)$$

Bayes' rule constant to be estimated

approach:

1. estimate $P(Y,\boldsymbol{X})$ → generative model (can be used to generate new samples)
2. calculate $P(Y|\boldsymbol{X})$ from $P(Y,\boldsymbol{X})$ → used for discriminative task (classification)

Independence Assumption

(naïve) assumption: conditional independence of features given target

$$P\left(X_j | Y, X_1, \cdots, X_{j-1}, X_{j+1}, \cdots, X_p\right) = P\left(X_j | Y\right)$$

$$\Longrightarrow P\left(Y | X_1, \cdots, X_p\right) = \frac{P(Y) \prod_{j=1}^{p} P\left(X_j | Y\right)}{P\left(X_1, \cdots, X_p\right)}$$

→ independent feature contributions (ignoring feature correlations)

→ robust against curse of dimensionality

Estimation of Feature Contributions

separate estimations of $P(X_j|Y)$ for each feature

requires assumption of distributions (e.g., Gaussian naïve Bayes) or non-parametric methods (kernel density estimation)

Gaussian feature likelihoods: $$P(x_{ij}|y) = \frac{1}{\sqrt{2\pi\sigma_{y,j}^2}} \exp\left(-\frac{(x_{ij}-\mu_{y,j})^2}{2\sigma_{y,j}^2}\right)$$

parameter estimation (e.g., mean and variance of Gaussians) can be done with maximum likelihood method (y known in training)

→ no Bayesian methods needed

Maximum a Posteriori Classification

$$\hat{y}_i = \underset{y}{\operatorname{argmax}}\, P(y) \prod_{j=1}^{p} P\left(x_{ij} | y\right)$$

despite potentially inaccurate probability estimates (due to naïve independence assumption), good identification of correct class via maximum probability

→ bad for regression tasks (if independence assumption is too naïve, i.e., features are correlated)

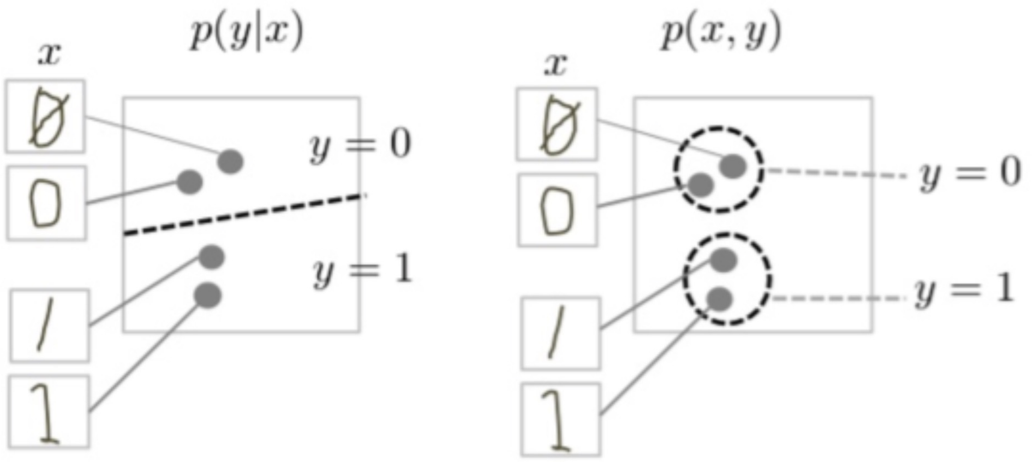
Generative vs Discriminative Models

generative models: predict joint probability $P(Y, \boldsymbol{X})$ (what allows to create new data samples) or directly generates new data samples

discriminative models: predict conditional probability $P(Y|\boldsymbol{X})$ or directly output (label for classification, real value for regression)

task of generative models more difficult: model full data distribution rather than merely find patterns in inputs to distinguish outputs

discriminative model generative model



source

Naïve Bayes and Logistic Regression

generative-discriminative pair of classification algorithms

- binary case: logit of naïve Bayes' outputs, $\log\left(\frac{P(y_i=1|\boldsymbol{x}_i)}{P(y_i=0|\boldsymbol{x}_i)}\right)$, corresponds to output of logistic regression's linear predictor
- for discrete inputs or Gaussian naïve Bayes: naïve Bayes can be reparametrized as linear classifier

for discriminative task: identical in asymptotic limit (infinite training samples) if independence assumption holds (otherwise naïve Bayes less accurate)

naïve Bayes has greater bias but lower variance than logistic regression → to be preferred for scarce training data (if bias, i.e., independence assumption, correct)

Data Generation

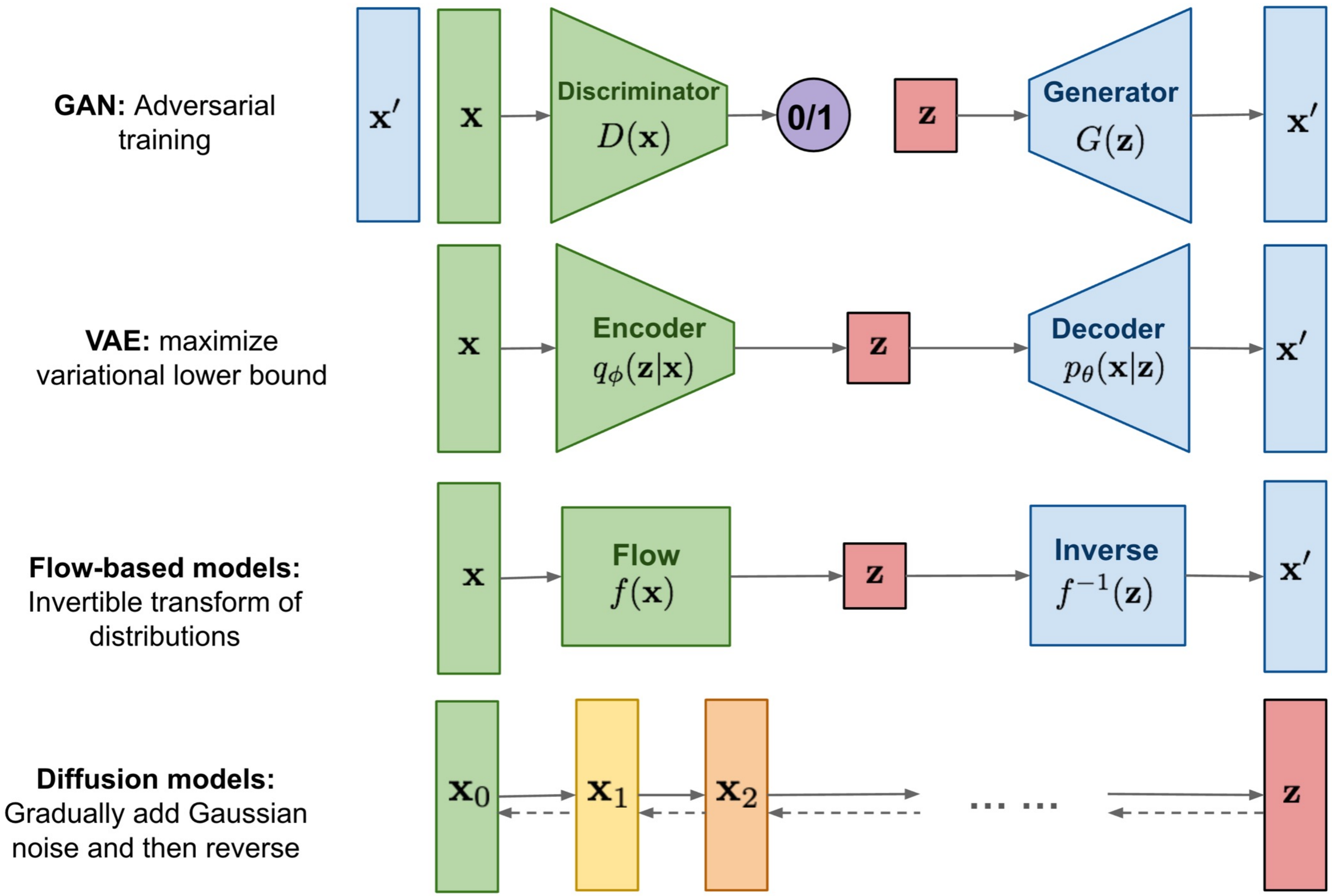
generative models can be used for discriminative tasks (although potentially inferior to direct discriminative methods)

but generative methods do more than discriminative ones: model full data distribution

→ allows generation of new data samples (can be images, text, video, audio, proteins, materials, time series, structured data, …)

large (auto-regressive) language models examples of generative models

Different Types of Generative Models



source

Variational Autoencoder (VAE)

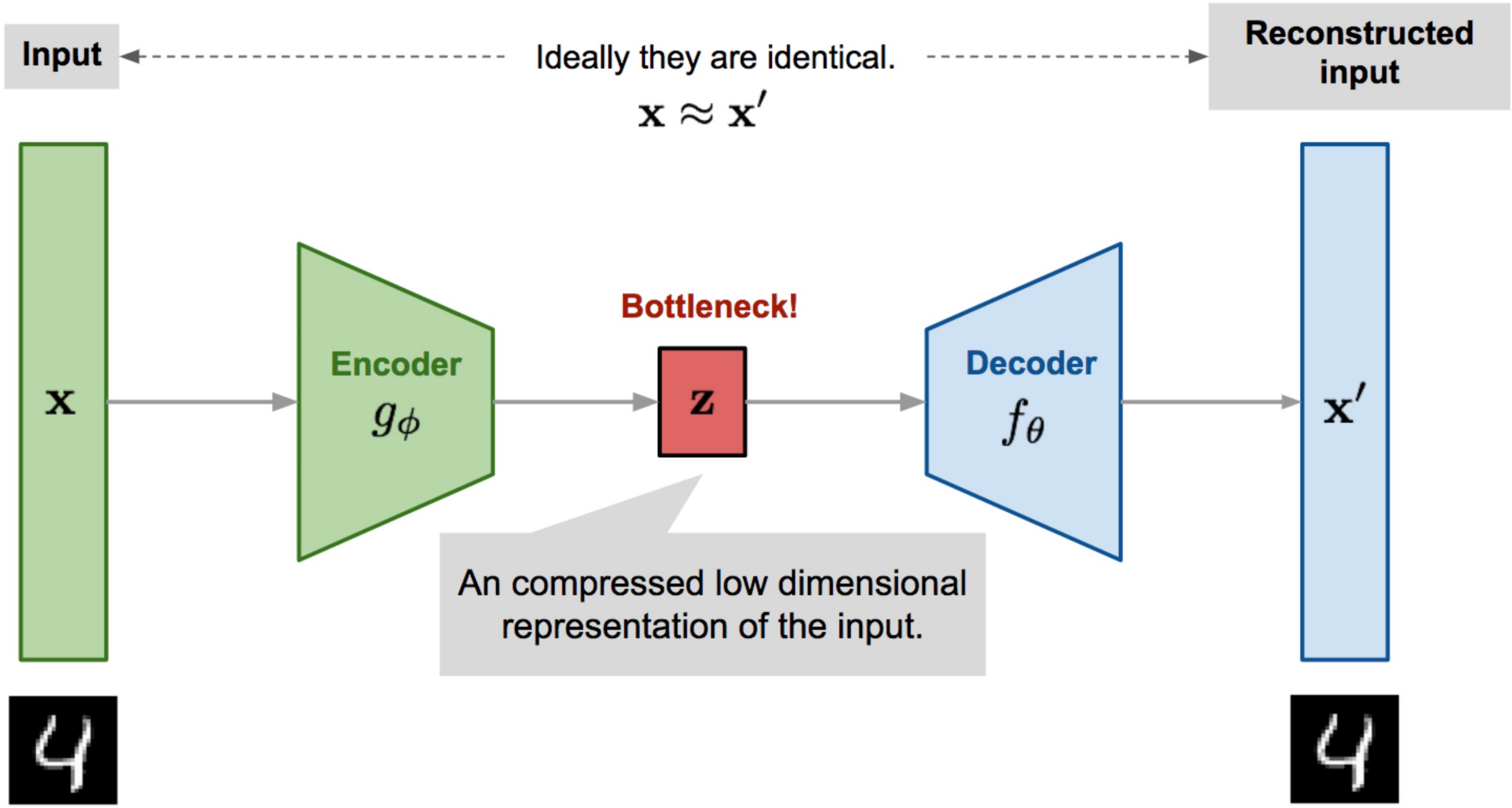
Recap: Autoencoder

(deep) encoder network

(deep) decoder network

learned together by minimizing differences between original input and reconstructed input (expressed as losses)

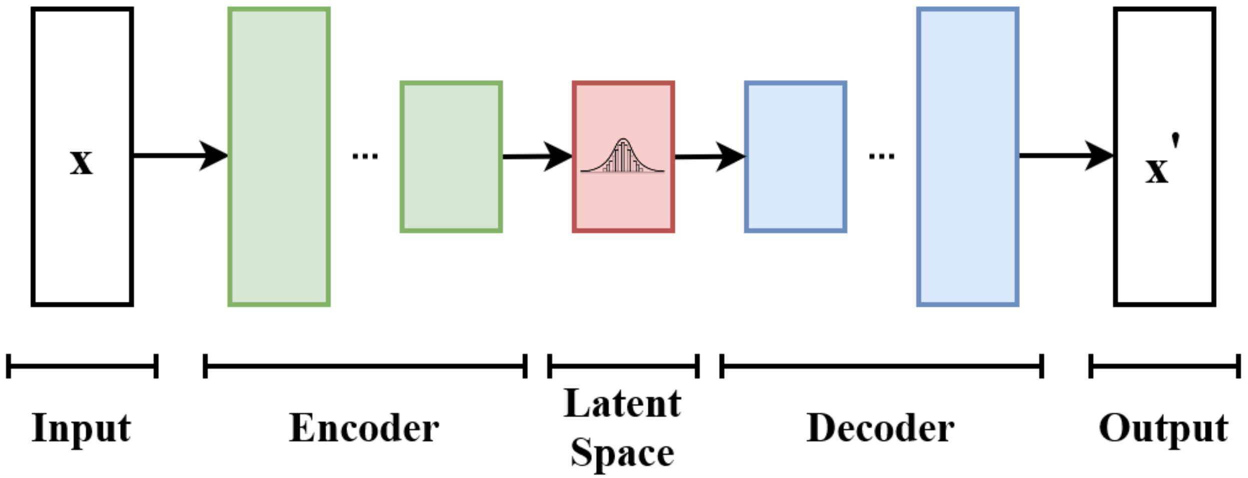
compressed intermediate representation: dimensionality reduction



source

Idea

- …



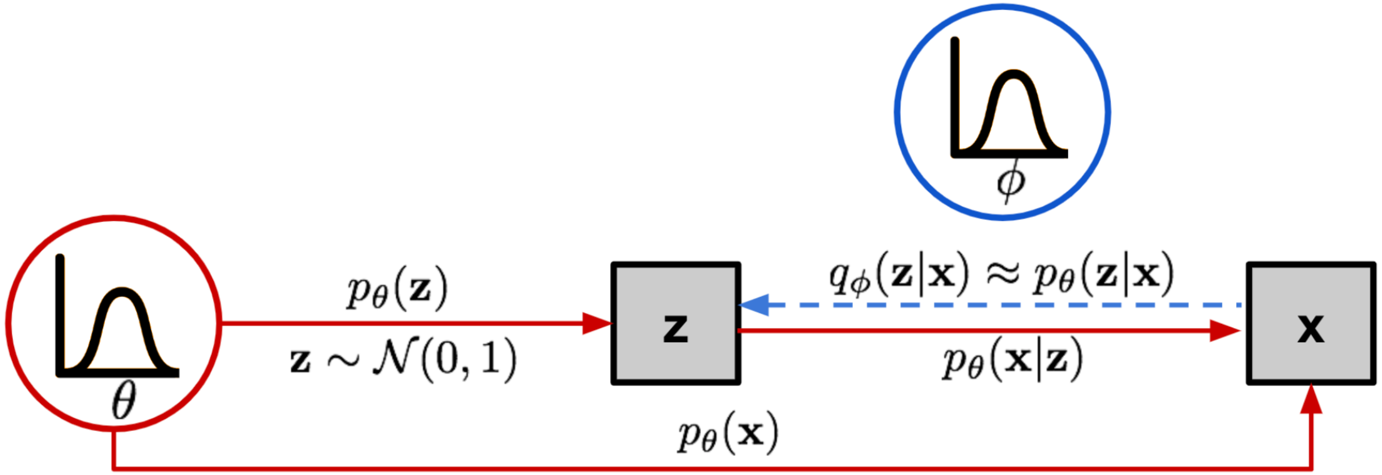
from wikipedia

Variational Inference

- Bayesian …

Surrogate Loss

- …
- VAE relies on a surrogate loss

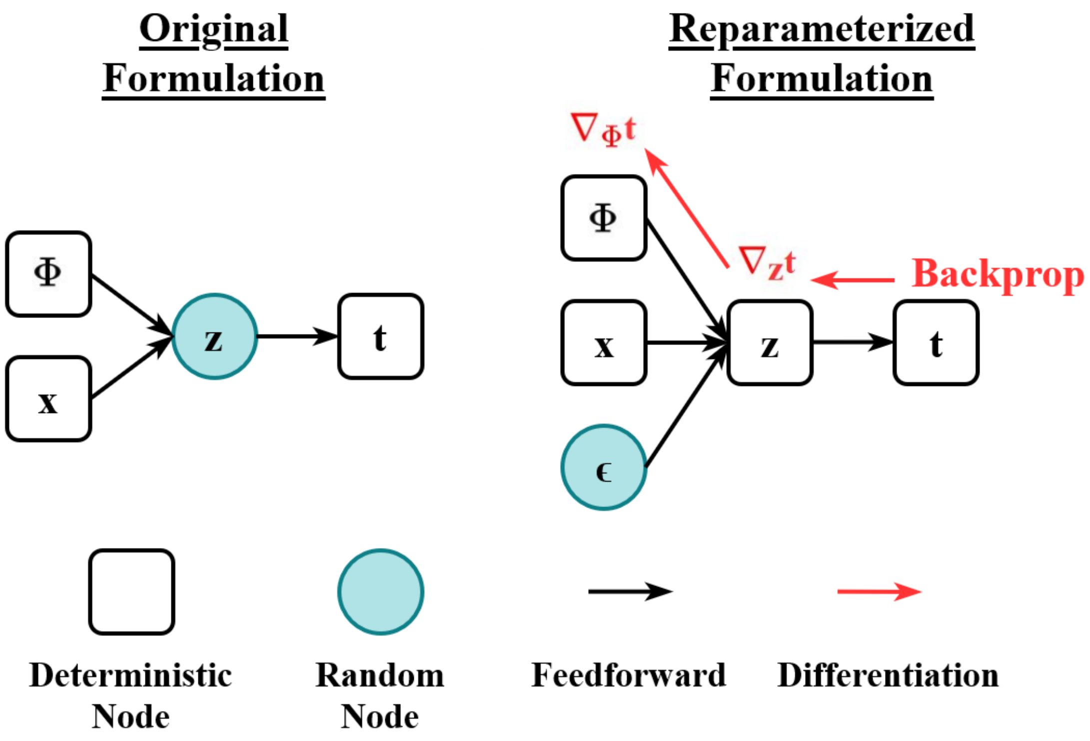


source

ELBO

Reparameterization Trick

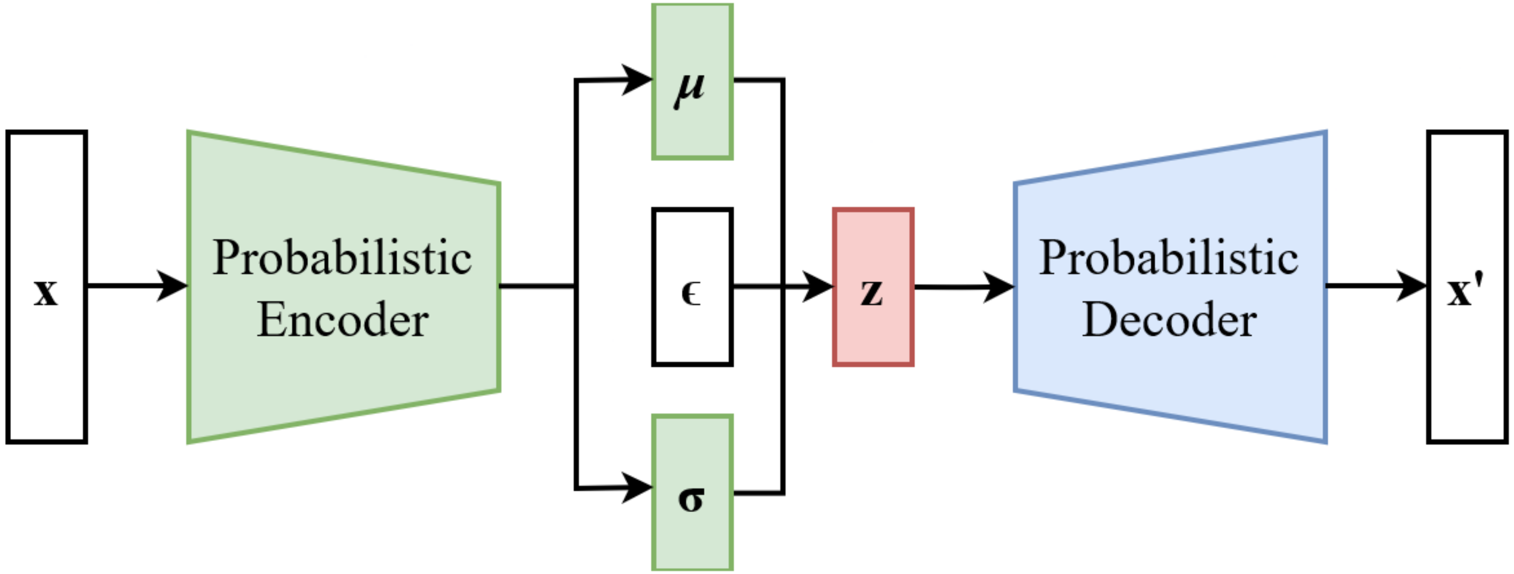
- ...



from wikipedia

Putting Everything Together

- …



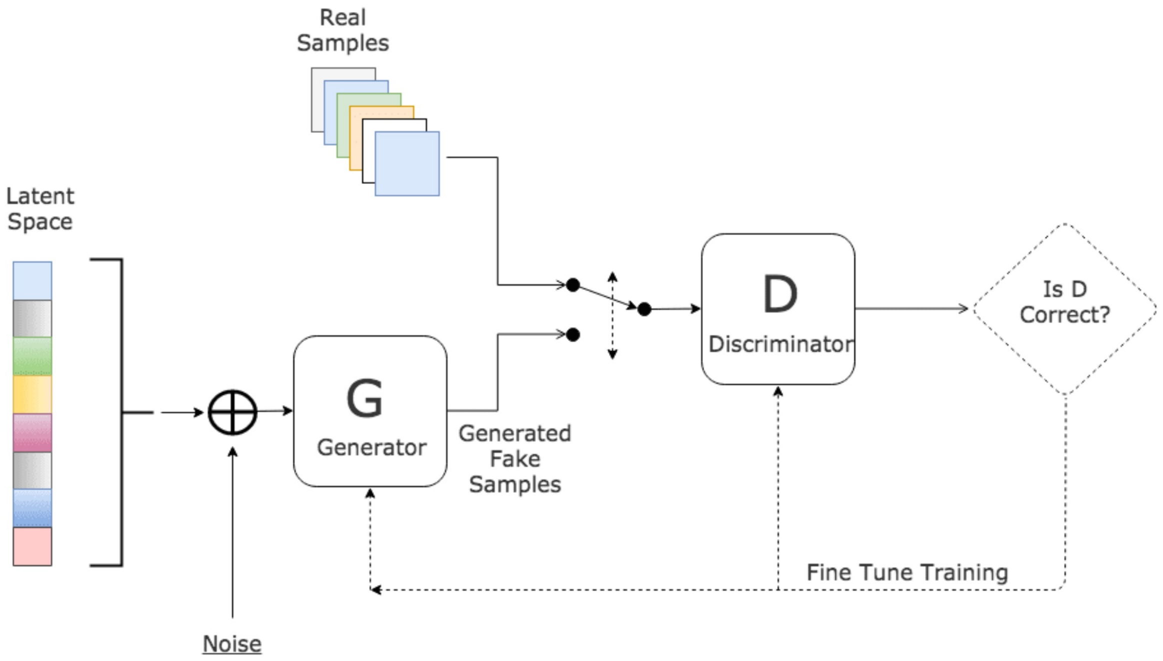
from wikipedia

Generative Adversarial Networks (GAN)

Zero-Sum Game

- …
- two neural networks playing a zero-sum game …

- …
- … noise



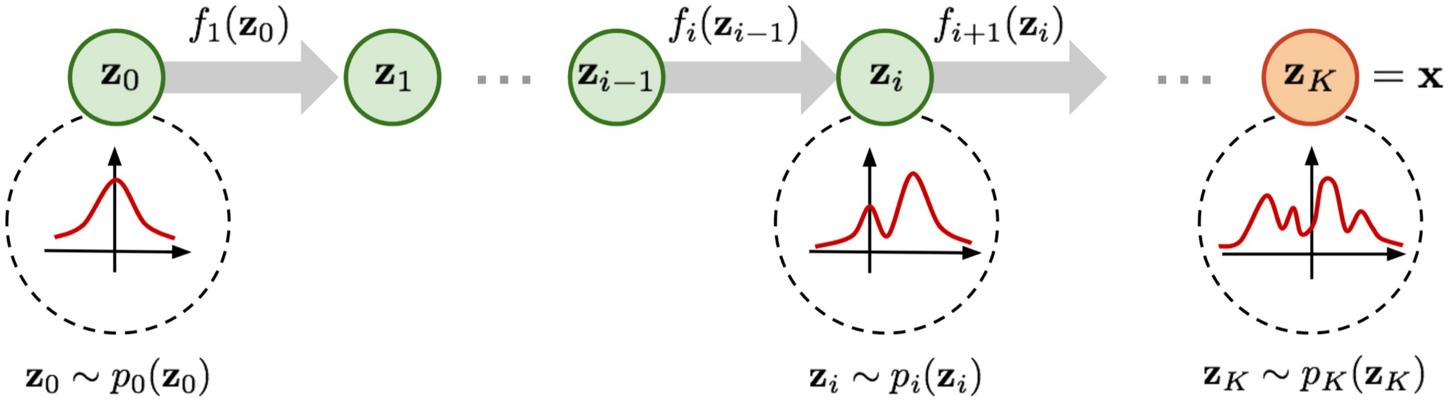
[source]

Issues in GANs

- …
- …potentially unstable training and less diversity in generation

Flow-Based Methods

- ...


$f_1(\mathbf{z}_0)$
$\mathbf{z}_0$
$\mathbf{z}_1$
$\mathbf{z}_{i-1}$
$f_i(\mathbf{z}_{i-1})$
$\mathbf{z}_i$
$f_{i+1}(\mathbf{z}_i)$
$\mathbf{z}_K = \mathbf{x}$
$\mathbf{z}_0 \sim p_0(\mathbf{z}_0)$
$\mathbf{z}_i \sim p_i(\mathbf{z}_i)$
$\mathbf{z}_K \sim p_K(\mathbf{z}_K)$

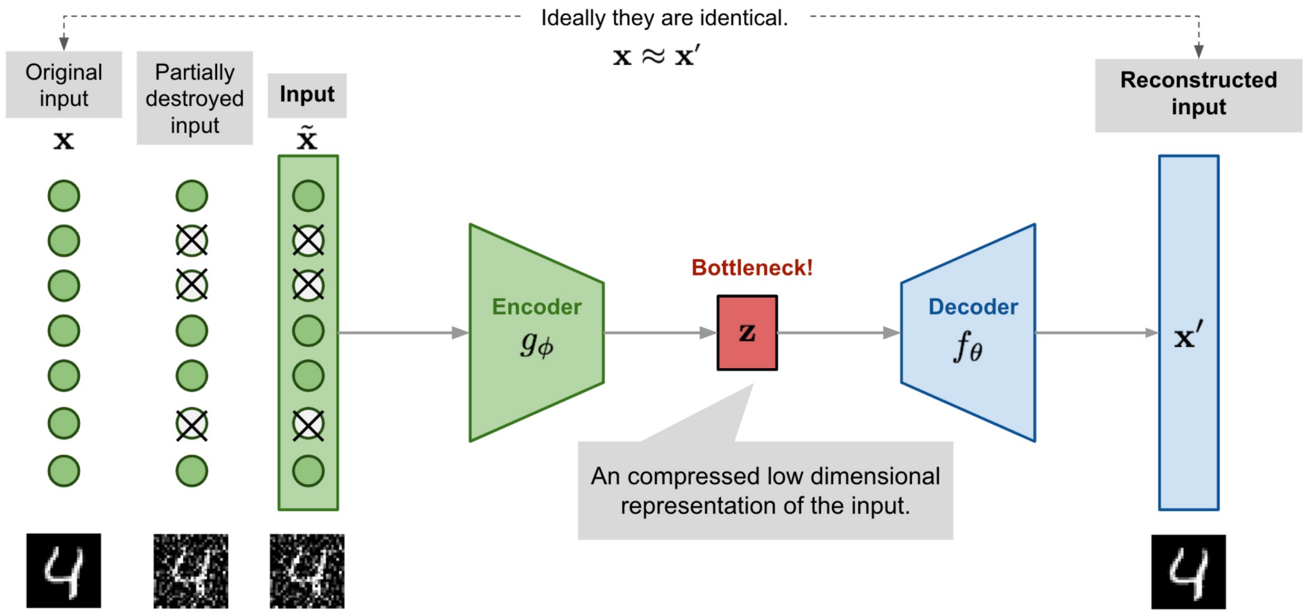
[source]

- specialized architectures to construct reversible transform

Diffusion Models

Denoising Autoencoder

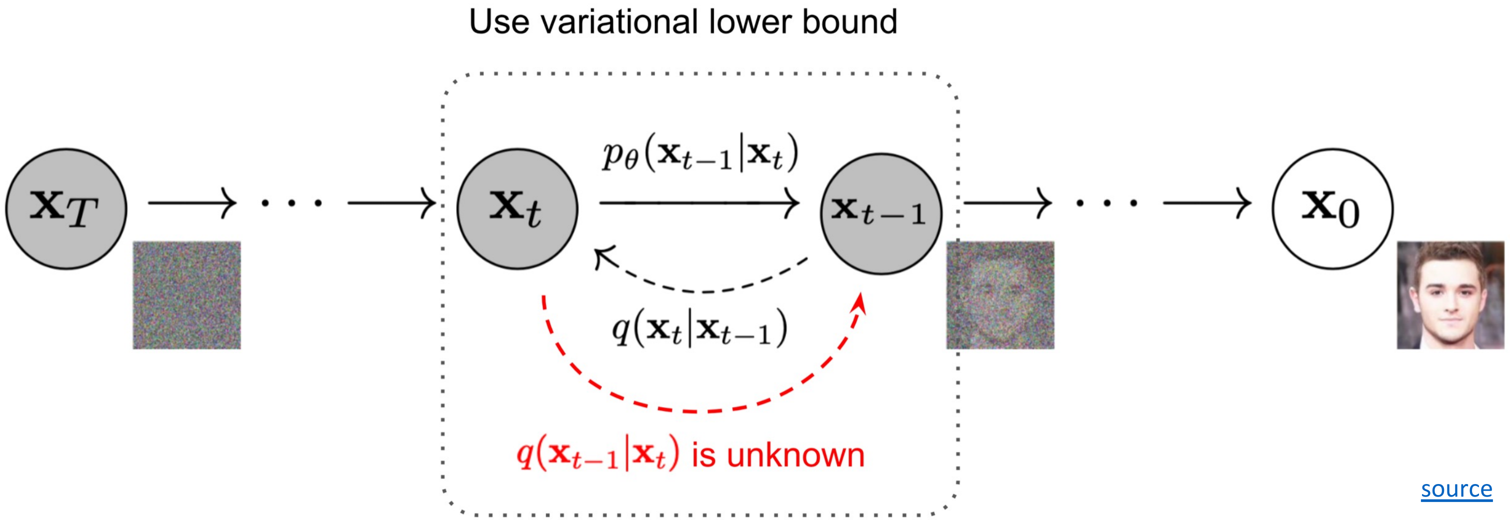
- …



source

- …chain of denoising autoencoders…
- inspired by non-equilibrium thermodynamics
- Markov chain of diffusion steps to slowly add random noise to data
- then learn to reverse the diffusion process to construct desired data samples from the noise
- Unlike VAE or flow models, diffusion models are learned with a fixed procedure and the latent variable has high dimensionality (same as the original data).

- ...



Use variational lower bound
$\mathbf{x}_T$
$\mathbf{x}_t$
$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$
$\mathbf{x}_{t-1}$
$\mathbf{x}_0$
$q(\mathbf{x}_t|\mathbf{x}_{t-1})$
$q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is unknown

source

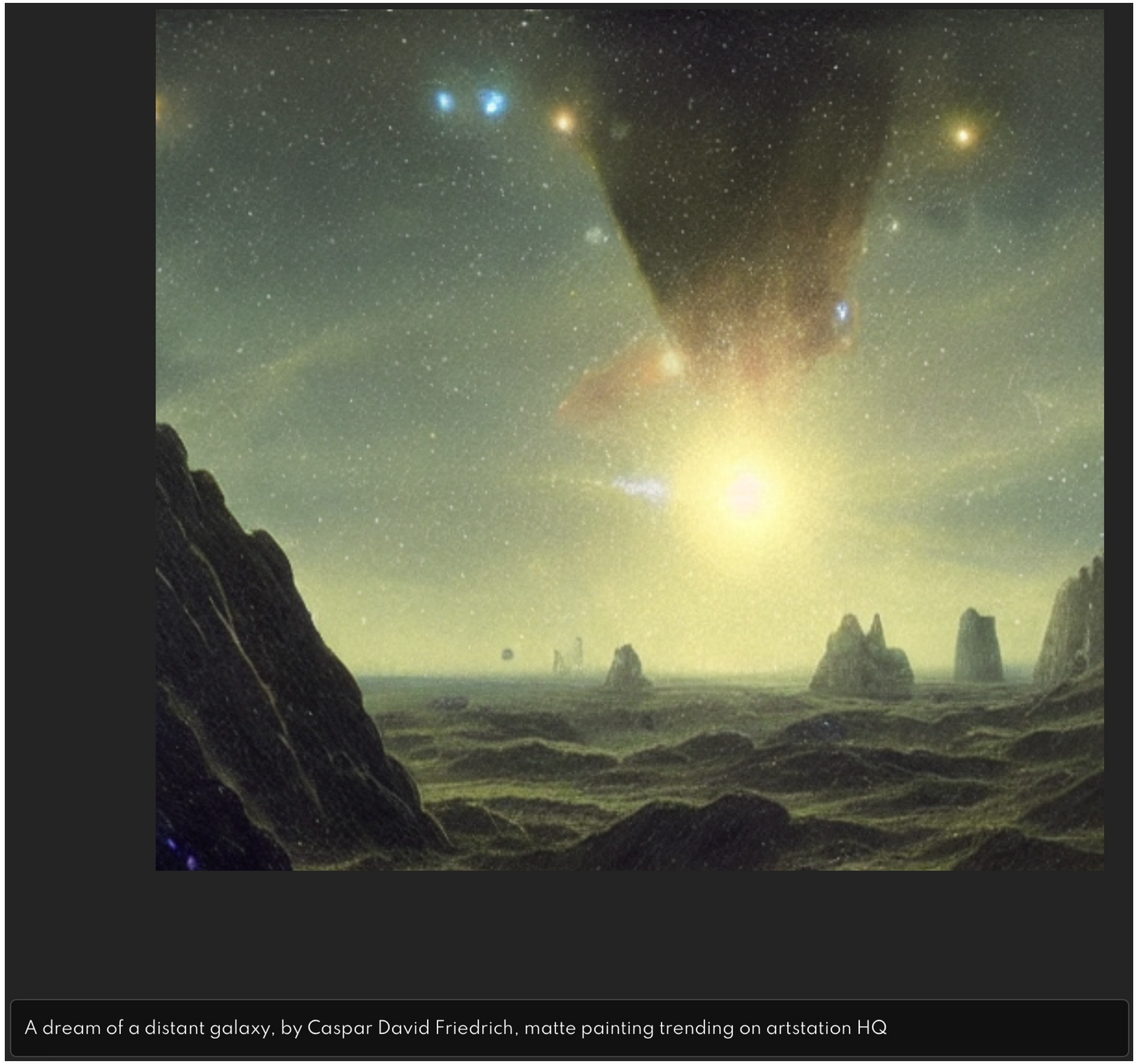
Image Generation

DALL-E 2

...

[Stable Diffusion]

[DreamStudio]



Literature

papers:

- variational autoencoder
- GAN
- normalizing flows
- latent diffusion

Movie-like Intelligence

emergent capabilities of complex systems almost impossible to foresee

mini examples in contemporary ML:

- [large language models]
- [multi-agent reinforcement learning]

one idea: [reward is enough]

philosophical: emotions or consciousness might also occur as emergent capabilities