On Unifying Deep Generative Models

Zhiting Hu[1,2], Zichao Yang[1], Ruslan Salakhutdinov[1], and Eric P. Xing[1,2]

Carnegie Mellon University[1], Petuum Inc.[2]

Overview

Conventional separate view of DGMs:

- Various deep generative models (DGMs) have been viewed as distinct learning paradigms, e.g.,
  - VAEs: Maximize lower bound of the data likelihood
  - GANs: Seek an equilibrium between generator and discriminator
- Extensive research in each of the lines; Many model variants.

Benefits of a unified statistical view of DGMs:

- Provides new theoretical understanding of different model behaviors
  - E.g., GANs $\to$ sharp yet low-diversity images; VAEs $\to$ more blurry
- Enables a perhaps more principled perpective of the broad landscape of generative modeling
  - Subsumes the many variants into the unified framework
  - Depicts a consistent roadmap of the advances in the field
- Enables transfer of technique across research lines in a principled way

This work attempts to compile such a unified view:

- Develops a new formulation of GANs and VAEs
  - GANs and VAEs involve minimizing KLD of respective posterior and inference distributions, with the generative parameter $\boldsymbol{\theta}$ in *opposite* directions:

$$\begin{aligned} &\text{GANs: } \min_\theta \mathrm{KL}(P_\theta \| Q_{\phi_0}) - \mathrm{JSD}_\theta \\ &\text{VAEs: } \min_\theta \mathrm{KL}(Q'_\eta \| P'_\theta) \end{aligned} \tag{1}$$

$\boldsymbol{\phi}$: discriminator parameters; $\boldsymbol{\eta}$: encoder parameters

  - Interpreting sample generation in GANs as performing posterior inference
  - VAEs has a *degenerated* adversarial mechanism that filers out generated samples and only uses real examples for model training
- Links back to the classic variational inference and the wake-sleep algorithms
- Extends easily to InfoGAN, VAE/GAN joint models, CycleGAN, AAE, adversarial domain adaptions, etc
  - All these methods can be easily formulated as instances or approximations of a loss-augmented variational posterior inference problem of latent variable graphical models
- Transfers techniques between VAE- and GAN-families:
  - Importance weighted VAE $\to$ Importance weighted GAN
  - Adversarial mechanism in GANs $\to$ Adversary-activated VAE

The Unified View

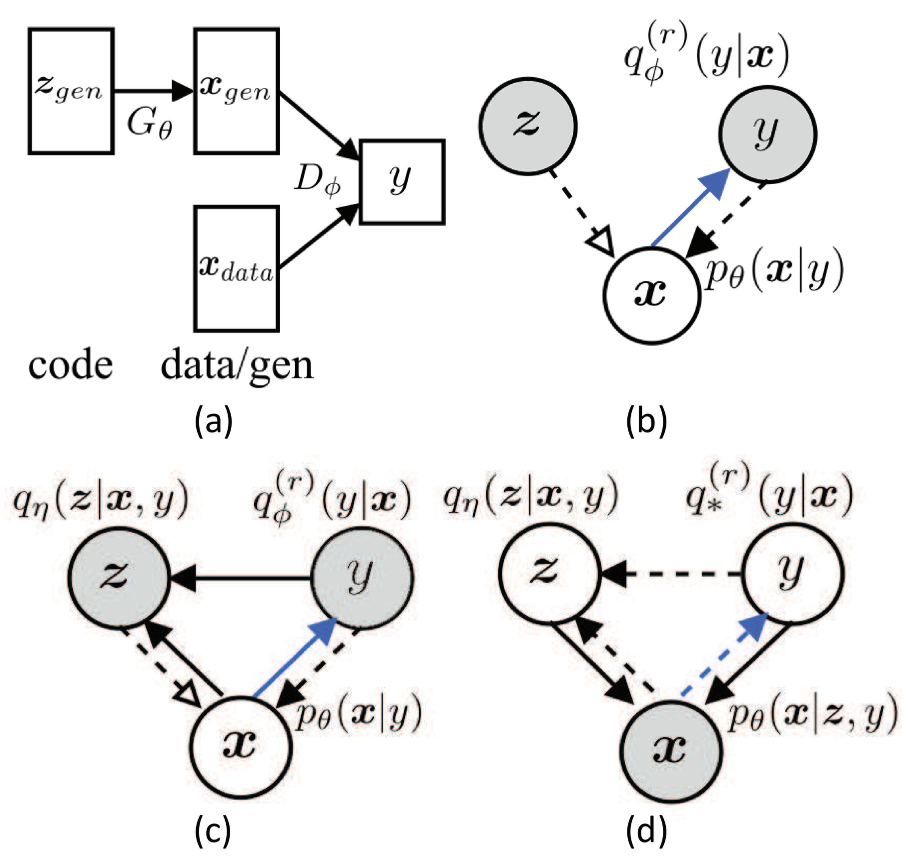

FIG 1: **(a)**: Conventional view of GANs. **(b)**: Schematic graphical model of GANs. **(c)**: InfoGAN. **(d)**: VAEs.

GANs:

- $$p_\theta(\boldsymbol{x}|y) = \begin{cases} p_{g_\theta}(\boldsymbol{x}) & y = 0 \\ p_{data}(\boldsymbol{x}) & y = 1. \end{cases} \tag{2}$$

  $y \in \{0, 1\}$: domain indicator; p_{g_θ}: generative distribution.
- Discriminator distribution $q_\phi(y|\boldsymbol{x})$; Let $q^r_\phi(y|\boldsymbol{x}) = 1 - q_\phi(y|\boldsymbol{x})$
- GANs' objectives can be written as:

$$\begin{aligned} &\max_{\boldsymbol{\phi}} \mathcal{L}_\phi = \mathbb{E}_{p_\theta(\boldsymbol{x}|y)p(y)} \left[\log q_\phi(y|\boldsymbol{x})\right] \\ &\max_{\boldsymbol{\theta}} \mathcal{L}_\theta = \mathbb{E}_{p_\theta(\boldsymbol{x}|y)p(y)} \left[\log q^r_\phi(y|\boldsymbol{x})\right]. \end{aligned} \tag{3}$$

Lemma 1. *Let*
- *$p(y)$ be the uniform distribution*
- $p_{\theta_0}(\boldsymbol{x}) = \mathbb{E}_{p(y)}[p_{\theta_0}(\boldsymbol{x}|y)]$: *prior distribution*
- $q^r(\boldsymbol{x}|y) \propto q^r_{\phi_0}(y|\boldsymbol{x})p_{\theta_0}(\boldsymbol{x})$: *posterior distribution*

Therefore, the updates of $\boldsymbol{\theta}$ at $\boldsymbol{\theta}_0$ have

$$\begin{aligned} &\nabla_\theta\Big[-\mathbb{E}_{p_\theta(\boldsymbol{x}|y)p(y)}\left[\log q^r_{\phi_0}(y|\boldsymbol{x})\right]\Big]\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \\ &\nabla_\theta\Big[\mathbb{E}_{p(y)}\left[KL\left(p_\theta(\boldsymbol{x}|y) \| q^r(\boldsymbol{x}|y)\right)\right] - JSD\left(p_\theta(\boldsymbol{x}|y=0) \| p_\theta(\boldsymbol{x}|y=1)\right)\Big]\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}, \end{aligned} \tag{4}$$

- **Variational Inference:** $p_\theta(\boldsymbol{x}|y)$ variational distribution
- **Training dynamics:** The KLD essentially pushes p_{g_θ} to a mixture of p_{g_θ} and p_{data}
- **The JSD term** is upper bounded by the KLD term.
- **Missing mode issue** due to the asymmetry of KLD and symmetry of JSD.
- **No optimality assumption on discriminator:** a generalization of previous results.
- Easily extends to InfoGAN, Adversarial Autoencoder, Prediction Minimization, and cycleGAN, etc.

VAEs:

- $$p_\theta(\boldsymbol{x}|\boldsymbol{z}, y) = \begin{cases} p_\theta(\boldsymbol{x}|\boldsymbol{z}) & y = 0 \\ p_{data}(\boldsymbol{x}) & y = 1. \end{cases} \tag{5}$$
- Inference distribution $q_\eta(\boldsymbol{z}|\boldsymbol{x})$
- *Perfect* discriminator

  $q_*(y|\boldsymbol{x})$: $q_*(y = 1|\boldsymbol{x} \in \text{data}) = q_*(y = 0|\boldsymbol{x} \in \text{samples}) = 1$

Lemma 2. *Let $p_\theta(\boldsymbol{z}, y|\boldsymbol{x}) \propto p_\theta(\boldsymbol{x}|\boldsymbol{z}, y)p(\boldsymbol{z}|y)p(y)$. The VAE objective can be written as:*

$$\begin{aligned} \mathcal{L}^{vae}_{\theta,\eta} &= \mathbb{E}_{p_{\theta_0}(\boldsymbol{x})}\Big[-\mathbb{E}_{q_\eta(\boldsymbol{z}|\boldsymbol{x},y)q^r_*(y|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z}, y)\right] + KL\left(q_\eta(\boldsymbol{z}|\boldsymbol{x}, y)q^r_*(y|\boldsymbol{x}) \| p(\boldsymbol{z}|y)p(y)\right)\Big] \\ &= \mathbb{E}_{p_{\theta_0}(\boldsymbol{x})}\Big[KL\left(q_\eta(\boldsymbol{z}|\boldsymbol{x}, y)q^r_*(y|\boldsymbol{x}) \| p_\theta(\boldsymbol{z}, y|\boldsymbol{x})\right)\Big]. \end{aligned} \tag{6}$$

- **VAEs contains an adversarial mechanism** which is degenerated due to the perfect discriminator. Generated samples are filtered out during training.
- **Covering mode issue** due to the asymmetry of KLD.

Technique Transfer:

- **Importance Weighted GAN (IWGAN)** by simply *copying* the derivations of Importance Weighted Autoencoder with little adaptions:

$$\nabla_\theta \mathcal{L}_k(y) = \mathbb{E}_{\boldsymbol{z}_1,\dots,\boldsymbol{z}_k \sim p(\boldsymbol{z}|y)}\left[\sum\nolimits_{i=1}^k \widetilde{w}_i \nabla_\theta \log q^r_{\phi_0}(y|\boldsymbol{x}(\boldsymbol{z}_i, \boldsymbol{\theta}))\right]. \tag{7}$$

  where $\widetilde{w_i}$ is the normalization of $w_i = \frac{q^r_{\phi_0}(y|\boldsymbol{x}_i)}{q_{\phi_0}(y|\boldsymbol{x}_i)}$
- **Adversary Activated VAE (AAVAE)** by replacing the perfect discriminator with a parameterized discriminator $q_\phi(y|\boldsymbol{x})$ learned jointly with other parts.
  - An adaptive data weighting mechanism that selects high-quality generated samples for model training.
- Improved performance on MNIST and SVHN.

More in the paper:

- GANs and VAEs extend the two learning phases of the wake-sleep algorithm, respectively.
- Provides alternative motivtaions for many existing GAN-VAE joint models, etc.
- A symmetric view of generation and inference (or, latent variables and visible variables)

- **ICML2018 Workshop** "Theoretical Foundations and Applications of Deep Generative Models". Please consider submitting your work and participating!
- **A DGM toolbox** for text generation will be open-sourced soon!
  - Supporting a large variety of DGMs for many text and sequence generation tasks.
  - Highly modularized and extensible for research and industry use.