On Unifying Deep Generative Models

Zhiting Hu\textsuperscript{1,2}, Zichao Yang\textsuperscript{1}, Ruslan Salakhutdinov\textsuperscript{1}, and Eric P. Xing\textsuperscript{1,2}

Carnegie Mellon University\textsuperscript{1}, Petuum Inc.\textsuperscript{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 → sharp yet low-diversity images; VAEs → more blurry
- Enables a perhaps more principled perspective of the broad landscape of generative modeling
- Subsumes the many variants into the unified framework
- Describes 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 VAEs and GANs
- GANs and VAEs involve minimizing KLD of respective posterior and inference distributions, with the generative parameter $\theta$ in opposite directions:
  \begin{align}
  \text{GANs: } & \min_{\theta} \text{KL}(P_\theta||Q_\phi) - JSD_\eta \\
  \text{VAEs: } & \min_{\theta} \text{KL}(Q_\phi||P_\theta) 
  \end{align}
  \tag{1}

  $\phi$: discriminator parameters; $\eta$: encoder parameters
- Interpreting sample generation in GANs as performing posterior inference
- VAEs has a degenerated adversarial mechanism that filters 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 adaptations, 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 → Importance weighted GAN
  - Adversarial mechanism in GANs → Adversary-activated VAE

The Unified View

The HKL of\textsuperscript{3} DGMs:
- $p(x|z) = \left\{ \begin{array}{ll} p_{\theta}(x) & y = 0 \\ p_{\theta}(x) & y = 1 \end{array} \right.$
  \tag{2}

  $y \in \{0,1\}$: domain indicator; $p_{\theta}$: generative distribution
- Discriminator distribution $q_{\phi}(z|x)$; Let $q_{\phi}^*(y|x) = 1 - q_{\phi}(y|x)$
- GANs’ objectives can be written as:
  \begin{align}
  \max_{\theta} \quad & \mathcal{L}_{\text{GAN}} = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log q_{\phi}(y|x)] \\
  \max_{\phi} \quad & \mathcal{L}_{\text{GAN}} = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log q_{\phi}^*(y|x)] 
  \end{align}
  \tag{3}

  Lemma 1. Let
  - $p(y)$ be the uniform distribution
  - $p_{\phi}(x) = \mathbb{E}_{y \sim p(y)}[p_{\phi}(x|y)]$: prior distribution
  - $q_{\phi}(x|y) = q_{\phi}^*(y|x)p_{\phi}(x)$: posterior distribution

  Therefore, the updates of $\theta$ at $\theta_{t+1}$ have
  \begin{align}
  \nabla_{\theta} \left[ \mathbb{E}_{x \sim p_{\text{data}}(x)} \left[ \log q_{\phi}(y|x) \right] \right]_{\theta_{t+1}} = & \mathbb{E}_{x \sim p_{\text{data}}(x)} [KL(p_{\phi}(x|y)||p(y|x))] \\
  \nabla_{\phi} \left[ \mathbb{E}_{x \sim p_{\text{data}}(x)} [KL(p_{\phi}(x|y)||p(y|x))] - JSD(p_{\phi}(x|y) = 0)||p_{\phi}(x|y) = 1) \right]_{\theta_{t+1}} = & 0 
  \end{align}
  \tag{4}

  Variational Inference: $p_{\phi}(x|y)$ variational distribution
- Training dynamics: The KLD essentially pushes $p_{\phi}$ to a mixture of $p_{\phi}$ and $p_{\text{data}}$
- 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:
- $p_{\phi}(x|z, y) = \left\{ \begin{array}{ll} p_{\theta}(x) & y = 0 \\ p_{\theta}(x) & y = 1 \end{array} \right.$
  \tag{5}

  - Inference distribution $q_{\phi}(z|x)$
  - Perfect discriminator $q_{\phi}(y|x)$; $q_{\phi}(y = 1|x \in \text{data}) = q_{\phi}(y = 0|x \in \text{samples}) = 1$
  \begin{align}
  \mathcal{L}_{\text{VAE}} = & \mathbb{E}_{(x,y) \sim p_{\text{data}}(x,y)} \left[ \log q_{\phi}(y|x) + KL(q_{\phi}(z|x, y)||q_{\phi}(z|y)) \right] \\
  = & \mathbb{E}_{(x,y) \sim p_{\text{data}}(x,y)} [KL(q_{\phi}(z|x, y)p_{\phi}(y|x)||p_{\text{data}}(x, y))] 
  \end{align}
  \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:
  \begin{align}
  \nabla_{\theta} \mathcal{L}_{\text{IWGAN}}(y) = & \mathbb{E}_{x \sim p_{\text{data}}(x,y)} \left[ \sum_{i=1}^m \tilde{w}_i \nabla_{\theta} \log q_{\phi}(y|x, z, \theta_i) \right] \end{align}
  \tag{7}

  where $\tilde{w}_i$ is the normalization of $w_i = \frac{q_{\phi}(y|x, z, \theta_i)}{q_{\phi}^*(y|x)}$.
- Adversary Activated VAE (AAVAE) by replacing the perfect discriminator with a parameterized discriminator $q_{\phi}(y|x)$ 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 motivations for many existing GAN-VAE joint models, etc.
- A symmetric view of generation and inference (or, latent variables and visible variables)

ICML2018 Workshop \textit{“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.