A Unified View of Deep Generative Models

Zhiting Hu and Eric Xing
Petuum Inc.
Carnegie Mellon University
Deep generative models
Deep generative models

- Define probabilistic distributions over a set of variables
- "Deep" means multiple layers of hidden variables!
Early forms of deep generative models

- Hierarchical Bayesian models
- Sigmoid brief nets [Neal 1992]

\[
p \left( x_{kn} = 1 \mid \theta_k, z_n^{(1)} \right) = \sigma \left( \theta_k^T z_n^{(1)} \right)
\]

\[
p \left( z_{in}^{(1)} = 1 \mid \theta_i, z_n^{(2)} \right) = \sigma \left( \theta_i^T z_n^{(2)} \right)
\]
Early forms of deep generative models

• Hierarchical Bayesian models
  • Sigmoid brief nets [Neal 1992]

• Neural network models
  • Helmholtz machines [Dayan et al., 1995]
Early forms of deep generative models

• Hierarchical Bayesian models
  • Sigmoid brief nets [Neal 1992]

• Neural network models
  • Helmholtz machines [Dayan et al., 1995]
  • Predictability minimization [Schmidhuber 1995]

Figure courtesy: Schmidhuber 1996
Early forms of deep generative models

• Training of DGMs via an EM style framework

  • Sampling / data augmentation
    \[ z = \{z_1, z_2\} \]
    \[ z_1^{\text{new}} \sim p(z_1 | z_2, x) \]
    \[ z_2^{\text{new}} \sim p(z_2 | z_1^{\text{new}}, x) \]

  • Variational inference
    \[ \log p(x) \geq \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x, z)] - \text{KL}(q_\phi(z|x) \| p(z)) := \mathcal{L}(\theta, \phi; x) \]
    \[ \max_{\theta, \phi} \mathcal{L}(\theta, \phi; x) \]

  • Wake sleep
    \[ \text{Wake: } \min_{\theta} \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] \]
    \[ \text{Sleep: } \min_{\phi} \mathbb{E}_{p_\theta(x|z)}[\log q_\phi(z|x)] \]
Resurgence of deep generative models

- Variational autoencoders (VAEs) [Kingma & Welling, 2014]
  / Neural Variational Inference and Learning (NVIL) [Mnih & Gregor, 2014]

Figure courtesy: Kingma & Welling, 2014
Resurgence of deep generative models

- Variational autoencoders (VAEs) [Kingma & Welling, 2014] / Neural Variational Inference and Learning (NVIL) [Mnih & Gregor, 2014]
- Generative adversarial networks (GANs)

\[ G_\theta : \text{generative model} \]
\[ D_\phi : \text{discriminator} \]
Outline

• Theoretical Basis of deep generative models
  • Wake sleep algorithm
  • Variational autoencoders
  • Generative adversarial networks

• A unified view of deep generative models
  • New formulations of deep generative models
  • Symmetric modeling of latent and visible variables
Synonyms in the literature

• Posterior Distribution -> Inference model
  • Variational approximation
  • Recognition model
  • Inference network (if parameterized as neural networks)
  • Recognition network (if parameterized as neural networks)
  • (Probabilistic) encoder

• "The Model" (prior + conditional, or joint) -> Generative model
  • The (data) likelihood model
  • Generative network (if parameterized as neural networks)
  • Generator
  • (Probabilistic) decoder
Recap: Variational Inference

• Consider a generative model $p_\theta(x|z)$, and prior $p(z)$
  • Joint distribution: $p_\theta(x,z) = p_\theta(x|z)p(z)$
• Assume variational distribution $q_\phi(z|x)$
• Objective: Maximize lower bound for log likelihood

\[
\log p(x) = KL \left( q_\phi(z|x) \| p_\theta(z|x) \right) + \int_z q_\phi(z|x) \log \frac{p_\theta(x,z)}{q_\phi(z|x)} \\
\geq \int_z q_\phi(z|x) \log \frac{p_\theta(x,z)}{q_\phi(z|x)} \\
:= \mathcal{L}(\theta, \phi; x)
\]

• Equivalently, minimize free energy

\[
F(\theta, \phi; x) = -\log p(x) + KL(q_\phi(z|x) \| p_\theta(z|x))
\]
Recap: Variational Inference

Maximize the variational lower bound $\mathcal{L}(\theta, \phi; x)$

- **E-step**: maximize $\mathcal{L}$ wrt. $\phi$ with $\theta$ fixed

  $$\max_{\phi} \mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_{\phi}(z|x)}[\log p_{\theta}(x|z)] + KL(q_{\phi}(z|x)||p(z))$$

- If with closed form solutions

  $$q^*_\phi(z|x) \propto \exp[\log p_{\theta}(x, z)]$$

- **M-step**: maximize $\mathcal{L}$ wrt. $\theta$ with $\phi$ fixed

  $$\max_{\theta} \mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_{\phi}(z|x)}[\log p_{\theta}(x|z)] + KL(q_{\phi}(z|x)||p(z))$$
Wake Sleep Algorithm

• [Hinton et al., Science 1995]

• Train a separate inference model along with the generative model
  • Generally applicable to a wide range of generative models, e.g., Helmholtz machines

• Consider a generative model $p_{\theta}(x|z)$ and prior $p(z)$
  • Joint distribution $p_{\theta}(x, z) = p_{\theta}(x|z)p(z)$
  • E.g., multi-layer brief nets

• Inference model $q_{\phi}(z|x)$

• Maximize data log-likelihood with two steps of loss relaxation:
  • Maximize the lower bound of log-likelihood, or equivalently, minimize the free energy
    \[
    F(\theta, \phi; x) = -\log p(x) + KL(q_{\phi}(z|x) \| p_{\theta}(z|x))
    \]

  • Minimize a different objective (reversed KLD) wrt $\phi$ to ease the optimization
    • Disconnect to the original variational lower bound loss
    \[
    F'(\theta, \phi; x) = -\log p(x) + KL(p_{\theta}(z|x) \| q_{\phi}(z|x))
    \]
Wake Sleep Algorithm

• Free energy:

\[ F(\theta, \phi; x) = -\log p(x) + KL(q_\phi(z|x) \| p_\theta(z|x)) \]

• Minimize the free energy wrt. \( \theta \) of \( p_\theta \) \( \rightarrow \) wake phase

\[ \max_\theta E_{q_\phi(z|x)} \left[ \log p_\theta(x, z) \right] \]

• Get samples from \( q_\phi(z|x) \) through inference on hidden variables
• Use the samples as targets for updating the generative model \( p_\theta(z|x) \)
• Correspond to the variational M step

[Figure courtesy: Maei’s slides]
Wake Sleep Algorithm

• Free energy:
  \[ F(\theta, \phi; x) = -\log p(x) + KL(q_\phi(z|x) \| p_\theta(z|x)) \]

• Minimize the free energy wrt. \( \phi \) of \( q_\phi(z|x) \)
  • Correspond to the variational E step
  • Difficulties:
    • Optimal \( q_\phi^*(z|x) = \frac{p_\theta(z, x)}{\int p_\theta(z, x) \, dz} \) intractable
    • High variance of direct gradient estimate \( \nabla_\phi F(\theta, \phi; x) = \cdots + \nabla_\phi \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(z, x)] + \cdots \)
      • Gradient estimate with the log-derivative trick:
        \[ \nabla_\phi \mathbb{E}_{q_\phi} [\log p_\theta] = \int \nabla_\phi q_\phi \log p_\theta = \int q_\phi \log p_\theta \nabla_\phi \log q_\phi = \mathbb{E}_{q_\phi} [\log p_\theta \nabla_\phi \log q_\phi] \]
    • Monte Carlo estimation:
      \[ \nabla_\phi \mathbb{E}_{q_\phi} [\log p_\theta] \approx \mathbb{E}_{z_i\sim q_\phi} [\log p_\theta(x, z_i) \nabla_\phi q_\phi(z_i|x)] \]
  • The scale factor \( \log p_\theta \) of the derivative \( \nabla_\phi \log q_\phi \) can have arbitrary large magnitude
Wake Sleep Algorithm

• Free energy:
\[ F(\theta, \phi; x) = -\log p(x) + KL(q_\phi(z|x) || p_\theta(z|x)) \]

• WS works around the difficulties with the sleep phase approximation

• Minimize the following objective \( \rightarrow \) sleep phase
\[ F'(\theta, \phi; x) = -\log p(x) + KL(p_\theta(z|x) || q_\phi(z|x)) \]
\[ \max_\phi E_{p_\theta(z|x)} [\log q_\phi(z|x)] \]

• “Dreaming” up samples from \( p_\theta(x|z) \) through top-down pass
• Use the samples as targets for updating the inference model

• (Recent approaches other than sleep phase is to reduce the variance of gradient estimate: slides later)

[Figure courtesy: Maei’s slides]
**Wake Sleep Algorithm**

Wake sleep
- Parametrized inference model \( q_{\phi}(z|x) \)
- Wake phase:
  - minimize \( KL(q_{\phi}(z|x) \| p_{\theta}(z|x)) \) wrt. \( \theta \)
  - \( E_{q_{\phi}(z|x)} [\nabla_{\theta} \log p_{\theta}(x|z)] \)
- Sleep phase:
  - minimize \( KL(p_{\theta}(z|x) \| q_{\phi}(z|x)) \) wrt. \( \phi \)
  - \( E_{p_{\theta}(z,x)} [\nabla_{\phi} \log q_{\phi}(z,x)] \)
  - low variance
  - Learning with generated samples of \( x \)
- Two objective, not guaranteed to converge

**Variational EM**
- Variational distribution \( q_{\phi}(z|x) \)
- Variational M step:
  - minimize \( KL(q_{\phi}(z|x) \| p_{\theta}(z|x)) \) wrt. \( \theta \)
  - \( E_{q_{\phi}(z|x)} [\nabla_{\theta} \log p_{\theta}(x|z)] \)
- Variational E step:
  - minimize \( KL(q_{\phi}(z|x) \| p_{\theta}(z|x)) \) wrt. \( \phi \)
  - \( q_{\phi} \propto \exp[\log p_{\theta}] \) if with closed-form
  - \( \nabla_{\phi} E_{q_{\phi}} [\log p_{\theta}(z,x)] \)
    - need variance-reduce in practice
  - Learning with real data \( x \)
- Single objective, guaranteed to converge
Variational Autoencoders (VAEs)

- [Kingma & Welling, 2014]
- Use variational inference with an inference model
  - Enjoy similar applicability with wake-sleep algorithm
- Generative model $p_\theta(x|z)$, and prior $p(z)$
  - Joint distribution $p_\theta(x, z) = p_\theta(x|z)p(z)$
- Inference model $q_\phi(z|x)$

Figure courtesy: Kingma & Welling, 2014
Variational Autoencoders (VAEs)

• Variational lower bound

\[ \mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x, z)] - \text{KL}(q_\phi(z|x) \| p(z)) \]

• Optimize \( \mathcal{L}(\theta, \phi; x) \) wrt. \( \theta \) of \( p_\theta(x|z) \)
  • The same with the wake phase

• Optimize \( \mathcal{L}(\theta, \phi; x) \) wrt. \( \phi \) of \( q_\phi(z|x) \)

\[ \nabla_\phi \mathcal{L}(\theta, \phi; x) = \cdots + \nabla_\phi \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] + \cdots \]

• Use reparameterization trick to reduce variance
• Alternatives: use control variates as in reinforcement learning [Mnih & Gregor, 2014; Paisley et al., 2012]
Reparametrized gradient

• Optimize $\mathcal{L}(\theta, \phi; x)$ wrt. $\phi$ of $q_\phi(z|x)$
  • Recap: gradient estimate with log-derivative trick:
    \[ \nabla_\phi \mathbb{E}_{q_\phi}[\log p_\theta(x, z)] = \mathbb{E}_{q_\phi}[\log p_\theta(x, z) \nabla_\phi \log q_\phi] \]
  • High variance: \[ \nabla_\phi \mathbb{E}_{q_\phi}[\log p_\theta] \approx \mathbb{E}_{z_i \sim q_\phi}[\log p_\theta(x, z_i) \nabla_\phi q_\phi(z_i|x)] \]
    • The scale factor $\log p_\theta(x, z_i)$ of the derivative $\nabla_\phi \log q_\phi$ can have arbitrary large magnitude

• gradient estimate with reparameterization trick
  \[ z \sim q_\phi(z|x) \iff z = g_\phi(\epsilon, x), \quad \epsilon \sim p(\epsilon) \]
  \[ \nabla_\phi \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x, z)] = \mathbb{E}_{\epsilon \sim p(\epsilon)}[\nabla_\phi \log p_\theta(x, z_\phi(\epsilon))] \]
  • (Empirically) lower variance of the gradient estimate
  • E.g., $z \sim N(\mu(x), L(x)L(x)^T) \iff \epsilon \sim N(0,1), \quad z = \mu(x) + L(x)\epsilon$
VAEs: example results

- VAEs tend to generate blurred images due to the mode covering behavior (more later)

![Celebrity faces](Radford 2015)

- Latent code interpolation and sentences generation from VAEs [Bowman et al., 2015].

```
"i want to talk to you."
"i want to be with you."
"i do n’t want to be with you."
i do n’t want to be with you.
she did n’t want to be with him.
```
Generative Adversarial Nets (GANs)

• [Goodfellow et al., 2014]
• Generative model $x = G_\theta(z), \ z \sim p(z)$
  • Map noise variable $z$ to data space $x$
  • Define an \textit{implicit distribution} over $x$: $p_{g_\theta}(x)$
    • a stochastic process to simulate data $x$
    • Intractable to evaluate likelihood
• Discriminator $D_\phi(x)$
  • Output the probability that $x$ came from the data rather than the generator
• No explicit inference model
• No obvious connection to previous models with inference networks like VAEs
  • We will build formal connections between GANs and VAEs later
Generative Adversarial Nets (GANs)

• Learning
  • A minimax game between the generator and the discriminator
  • Train $D$ to maximize the probability of assigning the correct label to both training examples and generated samples
  • Train $G$ to fool the discriminator

\[
\begin{align*}
\max_D L_D &= \mathbb{E}_{x \sim p_{data}} \left[ \log D(x) \right] + \mathbb{E}_{x \sim G(z), z \sim p(z)} \left[ \log (1 - D(x)) \right] \\
\min_G L_G &= \mathbb{E}_{x \sim G(z), z \sim p(z)} \left[ \log (1 - D(x)) \right].
\end{align*}
\]
GANs: example results

Generated bedrooms [Radford et al., 2016]
The Zoo of DGMs

• Variational autoencoders (VAEs) [Kingma & Welling, 2014]
  • Adversarial autoencoder [Makhzani et al., 2015]
  • Importance weighted autoencoder [Burda et al., 2015]
  • Implicit variational autoencoder [Mescheder., 2017]

• Generative adversarial networks (GANs) [Goodfellos et al., 2014]
  • InfoGAN [Chen et al., 2016]
  • CycleGAN [Zhu et al., 2017]
  • Wasserstein GAN [Arjovsky et al., 2017]

• Autoregressive neural networks
  • PixelRNN / PixelCNN [Oord et al., 2016]
  • RNN (e.g., for language modeling)

• Generative moment matching networks (GMMNs) [Li et al., 2015; Dziugaite et al., 2015]

• Restricted Boltzmann Machines (RBMs) [Smolensky, 1986]
Alchemy Vs Chemistry
Outline

• Theoretical backgrounds of deep generative models
  • Wake sleep algorithm
  • Variational autoencoders
  • Generative adversarial networks

• A unified view of deep generative models
  • New formulations of deep generative models
  • Symmetric modeling of latent and visible variables

Z Hu, Z YANG, R Salakhutdinov, E Xing,
“On Unifying Deep Generative Models”, arxiv 1706.00550
Generative Adversarial Nets (GANs):

• Implicit distribution over \( x \sim p_\theta(x|y) \)

\[
p_\theta(x|y) = \begin{cases} 
  p_{g_\theta}(x) & y = 0 \\
  p_{data}(x) & y = 1.
\end{cases}
\]

• \( x \sim p_{g_\theta}(x) \iff x = G_\theta(z), \ z \sim p(z|y = 0) \)

• \( x \sim p_{data}(x) \)
  • the code space of \( z \) is degenerated
  • sample directly from data

\[
z_{gen} \xrightarrow{G_\theta} x_{gen} \xrightarrow{D_\phi} y
\]

\( \text{code} \quad \text{data/gen} \)
A new formulation

- Rewrite GAN objectives in the ”variational-EM” format
  \[
  \max_\phi \mathcal{L}_\phi = \mathbb{E}_{x=G_\theta(z), z \sim p(z|y=0)} [\log (1 - D_\phi(x))] + \mathbb{E}_{x \sim p_{data}(x)} [\log D_\phi(x)] \\
  \max_\theta \mathcal{L}_\theta = \mathbb{E}_{x=G_\theta(z), z \sim p(z|y=0)} [\log D_\phi(x)] + \mathbb{E}_{x \sim p_{data}(x)} [\log (1 - D_\phi(x))]
  \]

- Recap: conventional formulation:
  - Implicit distribution over \( x \sim p_\theta(x|y) \)
    \[ x = G_\theta(z), \; z \sim p(z|y) \]
  - Discriminator distribution \( q_\phi(y|x) \)
    \[ q_\phi^r(y|x) = q_\phi(1 - y|x) \]

- Rewrite in the new form
  - \( \max_\phi \mathcal{L}_\phi = \mathbb{E}_{p_\theta(x|y)p(y)} [\log q_\phi(y|x)] \)
  - \( \max_\theta \mathcal{L}_\theta = \mathbb{E}_{p_\theta(x|y)p(y)} [\log q_\phi^r(y|x)] \)
**GANs vs. Variational EM**

### Variational EM

- **Objectives**
  \[
  \begin{align*}
  \max_\phi & \mathcal{L}_{\phi, \theta} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] + KL(q_\phi(z|x) || p(z)) \\
  \max_\theta & \mathcal{L}_{\phi, \theta} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] + KL(q_\phi(z|x) || p(z))
  \end{align*}
  \]

- Single objective for both $\theta$ and $\phi$
- Extra prior regularization by $p(z)$
- **The reconstruction term**: maximize the conditional log-likelihood of $x$ with the generative distribution $p_\theta(x|z)$ conditioning on the latent code $z$ inferred by $q_\phi(z|x)$

  - $p_\theta(x|z)$ is the generative model
  - $q_\phi(z|x)$ is the inference model

### GAN

- **Objectives**
  \[
  \begin{align*}
  \max_\phi & \mathcal{L}_{\phi} = \mathbb{E}_{p_\theta(x|y)p(y)}[\log q_\phi(y|x)] \\
  \max_\theta & \mathcal{L}_{\theta} = \mathbb{E}_{p_\theta(x|y)p(y)}[\log q_\theta^r(y|x)]
  \end{align*}
  \]

- Two objectives
- Have global optimal state in the game theoretic view
- **The objectives**: maximize the conditional log-likelihood of $y$ (or $1 - y$) with the distribution $q_\phi(y|x)$ conditioning on data/generation $x$ inferred by $p_\theta(x|y)$

  - Interpret $q_\phi(y|x)$ as the generative model
  - Interpret $p_\theta(x|y)$ as the inference model

---

### Table 1: Correspondence between different approaches in the proposed formulation.

<table>
<thead>
<tr>
<th>Method</th>
<th>Feature Distribution</th>
<th>Generation Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAN</td>
<td>$p(x)$</td>
<td>$q_\phi(x)$</td>
</tr>
<tr>
<td>ADA</td>
<td>$p(x)$</td>
<td>$q_\phi(x)$</td>
</tr>
<tr>
<td>InfoGAN</td>
<td>$p(x)$</td>
<td>$q_\phi(x)$</td>
</tr>
<tr>
<td>ADA with a degenerated source domain</td>
<td>$p(x)$</td>
<td>$q_\phi(x)$</td>
</tr>
<tr>
<td>VAE</td>
<td>$p(x)$</td>
<td>$q_\phi(x)$</td>
</tr>
</tbody>
</table>

---

**Adversarial Domain Adaptation**

- ADA aims to transfer prediction knowledge learned from a source domain with labeled data to a target domain.
- ADA with a degenerated source domain, and reveal close relations to VAEs and wake-sleep algorithm.
- This section formally explores these connections.

---

**Materials**

- For ease of presentation and to establish a systematic notation for the paper, we start with a new interpretation of the objectives. Table 1 lists the correspondence of each component in these approaches.
GANs vs. Variational EM

Variational EM

- Objectives
  \[ \max_{\phi} \mathcal{L}_{\phi, \theta} = \mathbb{E}_{q_{\phi}(z|x)}[\log p_{\theta}(x|z)] + KL(q_{\phi}(z|x) || p(z)) \]
  \[ \max_{\theta} \mathcal{L}_{\phi, \theta} = \mathbb{E}_{q_{\phi}(z|x)}[\log p_{\theta}(x|z)] + KL(q_{\phi}(z|x) || p(z)) \]
  - Single objective for both \( \theta \) and \( \phi \)
  - Extra prior regularization by \( p(z) \)
  - The reconstruction term: maximize the conditional log-likelihood of \( x \) with the generative distribution \( p_{\theta}(x|z) \) conditioning on the latent code \( z \) inferred by \( q_{\phi}(z|x) \)

- \( p_{\theta}(x|z) \) is the generative model
- \( q_{\phi}(z|x) \) is the inference model

GAN

- Objectives
  \[ \max_{\phi} \mathcal{L}_{\phi} = \mathbb{E}_{p_{\theta}(x|y)p(y)}[\log q_{\phi}(y|x)] \]
  \[ \max_{\theta} \mathcal{L}_{\theta} = \mathbb{E}_{p_{\theta}(x|y)p(y)}[\log q_{\phi}^{\tau}(y|x)] \]
  - Two objectives
  - Have global optimal state in the game theoretic view
  - The objectives: maximize the conditional log-likelihood of \( y \) (or \( 1 - y \)) with the distribution \( q_{\phi}(y|x) \) conditioning on data/generation \( x \) inferred by \( p_{\theta}(x|y) \)

- Interpret \( q_{\phi}(y|x) \) as the generative model
- Interpret \( p_{\theta}(x|y) \) as the inference model

- Interpret \( x \) as latent variables
- Interpret generation of \( x \) as performing inference over latent variables
GANs: minimizing KLD

• As in Variational EM, we can further rewrite in the form of minimizing KLD to reveal more insights into the optimization problem

• For each optimization step of \( p_\theta(x|y) \) at point \( (\theta = \theta_0, \phi = \phi_0) \), let
  - \( p(y) \): uniform prior distribution
  - \( p_{\theta=\theta_0}(x) = \mathbb{E}_{p(y)}[p_{\theta=\theta_0}(x|y)] \)
  - \( q^r(x|y) \propto q_{\phi=\phi_0}^r(y|x)p_{\theta=\theta_0}(x) \)

• **Lemma 1:** The updates of \( \theta \) at \( \theta_0 \) have
  \[
  \nabla_\theta \left[ - \mathbb{E}_{p_\theta(x|y)p(y)} \left[ \log q_{\phi=\phi_0}^r(y|x) \right] \right] \bigg|_{\theta=\theta_0} = \\
  \nabla_\theta \left[ \mathbb{E}_{p(y)} \left[ KL(p_\theta(x|y)||q^r(x|y)) \right] - JSD(p_\theta(x|y=0)||p_\theta(x|y=1)) \right] \bigg|_{\theta=\theta_0},
  \]

  - KL: KL divergence
  - JSD: Jensen-shannon divergence
GANs: minimizing KLD

- **Lemma 1**: The updates of \( \theta \) at \( \theta_0 \) have
  \[
  \nabla_\theta \left[ - \mathbb{E}_{p(y)} \left[ \log q^r_{\phi=0}(y|x) \right] \right]_{\theta=\theta_0} = \\
  \nabla_\theta \left[ \mathbb{E}_{p(y)} \left[ \text{KL}(p_\theta(x|y)||q^r(x|y)) - \text{JSD}(p_\theta(x|y=0)||p_\theta(x|y=1)) \right] \right]_{\theta=\theta_0}
  \]

- **Connection to variational inference**
  - See \( x \) as latent variables, \( y \) as visible
  - \( p_{\theta=\theta_0}(x) \): prior distribution
  - \( q^r(x|y) \propto q^r_{\phi=0}(y|x)p_{\theta=\theta_0}(x) \): posterior distribution
  - \( p_{\theta}(x|y) \): variational distribution
    - Amortized inference: updates model parameter \( \theta \)

- **Suggests relations to VAEs, as we will explore shortly**
We now take a closer look at the form of Eq.(4) which is essentially reconstructing the real/fake

**Lemma 1**

Let \( y \) be a point of \( \mathbb{R}^d \).

**Objective:**

For instance, the objective of the generative parameters is to

- Recover the classical form by unfolding over the parameter space, we have
  
  \[ p_{\theta=\theta_0}(x|y=1) = p_{data}(x) \]
  \[ p_{\theta=\theta_0}(x|y=0) = p_{g_{\theta=\theta_0}}(x) \]

**Figure 2:** One optimization step of the parameter \( \theta \)

- KL divergence, \( KL(p_{\theta}(x|y=1)||q^r(x|y=1)) \) : constant, no free parameters
- KL divergence, \( KL(p_{\theta}(x|y=0)||q^r(x|y=0)) \) : parameter \( \theta \) to optimize
  
  - \( q^r(x|y=0) \propto q^{r}_{\phi=\theta_0}(y=0|x)p_{\theta=\theta_0}(x) \)
  - seen as a mixture of \( p_{g_{\theta=\theta_0}}(x) \) and \( p_{data}(x) \)
  - mixing weights induced from \( q^{r}_{\phi=\theta_0}(y=0|x) \)

\[ \Rightarrow \text{Drives } p_{g_{\theta}}(x|y) \text{ to mixture of } p_{g_{\theta=\theta_0}}(x) \text{ and } p_{data}(x) \]
GANs: minimizing KLD

$p_{\theta=\hat{o}}(x|y=1) = p_{\text{data}}(x)$ \hspace{1cm} $p_{\theta=\hat{o}}(x|y=0) = p_{g_{\theta=\hat{o}}}(x)$

$q^r(x|y=0)$

missed mode

$\text{KL}(p_{g_{\theta}}(x)||q^r(x|y=0))$

$= \int p_{g_{\theta}}(x) \log \frac{p_{g_{\theta}}(x)}{q^r(x|y=0)} dx$

• Missing mode phenomena of GANs
  • Asymmetry of KLD
    • Concentrates $p_{\theta}(x|y=0)$ to large modes of $q^r(x|y)$
      ⇒ $p_{g_{\theta}}(x)$ misses modes of $p_{\text{data}}(x)$
  • Symmetry of JSD
    • Does not affect the behavior of mode missing

• Large positive contribution to the KLD in the regions of $x$ space where $q^r(x|y=0)$ is small, unless $p_{g_{\theta}}(x)$ is also small
  ⇒ $p_{g_{\theta}}(x)$ tends to avoid regions where $q^r(x|y=0)$ is small
GANs: minimizing KLD

- **Lemma 1**: The updates of $\theta$ at $\theta_0$ have

  $\nabla_\theta \left[ -\mathbb{E}_{p(y)p(y)\mathbb{E}_{p(y)}} \left[ \log q_{\phi_0}^r (y|x) \right] \right]_{\theta=\theta_0} = $

  $\nabla_\theta \left[ \mathbb{E}_{p(y)} \left[ KL (p_\theta(x|y) || q^r(x|y)) \right] - JSD (p_\theta(x|y = 0) || p_\theta(x|y = 1)) \right]_{\theta=\theta_0}$

- No assumption on optimal discriminator $q^r_{\phi_0} (y|x)$

  - Previous results usually rely on (near) optimal discriminator
    - $q^*(y = 1|x) = p_{data}(x)/(p_{data}(x) + p_g(x))$
  - Optimality assumption is impractical: limited expressiveness of $D_\phi$ [Arora et al 2017]
  - Our result is a generalization of the previous theorem [Arjovský & Bottou 2017]
    - Plug the optimal discriminator into the above equation, we recover the theorem
      $\nabla_\theta \left[ -\mathbb{E}_{p_\theta(x|y)p(y)} \left[ \log q_{\phi_0}^r (y|x) \right] \right]_{\theta=\theta_0} = \nabla_\theta \left[ \frac{1}{2} KL (p_{g_\theta} || p_{data}) - JSD (p_{g_\theta} || p_{data}) \right]_{\theta=\theta_0}$
  - Give insights on the generator training when discriminator is optimal
GANs: minimizing KLD

In summary:

• Reveal connection to variational inference
  • Build connections to VAEs (slides soon)
  • Inspire new model variants based on the connections

• Offer insights into the generator training
  • Formal explanation of the missing mode behavior of GANs
  • Still hold when the discriminator does not achieve its optimum at each iteration
GANs vs InfoGAN

We frame our new interpretation of ADA, and review conventional formulations in the supplementary which is intractable to evaluate likelihood but easy to sample from:

more in the next section. Note that the only (but critical) difference between the objective of conditioned on feature entropy with respect to distribution over domains. The objectives of ADA are therefore given as:

The conventional definition of V AEs is written as:

Specifically, we again introduce the real/fake variable

This is where the adversarial mechanism comes about.

The conventional definition of V AEs is written as:

Arjovsky and Bottou[1] derive a similar result of minimizing the KL divergence between the generative model by reconstructing observed real examples, sharing similarity to the wake phase.

The conventional definition of V AEs is written as:

\[ \text{max}_\phi L_\phi = \mathbb{E}_{p_\theta(x|y)p(y)} [\log q_\phi(y|x)] \]
\[ \text{max}_\theta L_\theta = \mathbb{E}_{p_\theta(x|y)p(y)} [\log q_\phi^r(y|x)] \]
\[ \text{max}_\theta L_\theta = \mathbb{E}_{p_\theta(x|y)p(y)} [\log q_\eta(z|x,y)q_\phi(y|x)] \]
\[ \text{max}_\theta,\eta L_{\theta,\eta} = \mathbb{E}_{p_\theta(x|y)p(y)} [\log q_\eta(z|x,y)q_\phi^r(y|x)] \]
Relates VAEs with GANs

- Resemblance of GAN generator learning to variational inference
  - Suggest strong relations between VAEs and GANs

- Indeed, VAEs are basically minimizing KLD with an opposite direction, and with a degenerated adversarial discriminator
GANs vs VAEs side by side

<table>
<thead>
<tr>
<th></th>
<th>GANs (InfoGAN)</th>
<th>VAEs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generative distribution</td>
<td>( p_\theta(x</td>
<td>y) = \begin{cases} p_{g_\theta}(x) &amp; y = 0 \ p_{data}(x) &amp; y = 1 \end{cases} )</td>
</tr>
<tr>
<td>Discriminator distribution</td>
<td>( q_\phi(y</td>
<td>x) )</td>
</tr>
<tr>
<td>( z )-inference model</td>
<td>( q_\eta(z</td>
<td>x, y) ) of InfoGAN</td>
</tr>
<tr>
<td>KLD to minimize</td>
<td>( \min_\theta \text{KL} \left( p_\theta(x</td>
<td>y) \parallel q^r(x</td>
</tr>
<tr>
<td></td>
<td>( \sim \min_\theta \text{KL}(P_\theta \parallel Q) )</td>
<td>( \sim \min_\theta \text{KL}(Q \parallel p_\theta) )</td>
</tr>
</tbody>
</table>
Asymmetry of KLDs inspires combination of GANs and VAEs

- **GANs**: \( \min \theta \text{KL}(p_\theta(x|y) \parallel q^r(x|z, y)) \)
  - \( \approx \min \theta \text{KL}(P_\theta \parallel Q) \)

- **VAEs**: \( \min \theta \text{KL}(q_\eta(z|x, y)q_*^r(y|x) \parallel p_\theta(z, y|x)) \)
  - \( \approx \min \theta \text{KL}(Q \parallel P_\theta) \)

- Asymmetry of KLDs inspires combination of GANs and VAEs
  - GANs: \( \min \theta \text{KL}(P_\theta \parallel Q) \) tends to missing mode
  - VAEs: \( \min \theta \text{KL}(Q \parallel P_\theta) \) tends to cover regions with small values of \( p_{data} \)
## Mutual exchanges of ideas: augment the loss

<table>
<thead>
<tr>
<th></th>
<th>GANs (InfoGAN)</th>
<th>VAEs</th>
</tr>
</thead>
<tbody>
<tr>
<td>KLD to minimize</td>
<td>( \min_{\theta} \text{KL}(p_{\theta}(x</td>
<td>y) \</td>
</tr>
</tbody>
</table>

- Asymmetry of KLDs inspires combination of GANs and VAEs
  - GANs: \( \min_{\theta} \text{KL}(P_{\theta} \ || \ Q) \) tends to missing mode
  - VAEs: \( \min_{\theta} \text{KL}(Q \ || \ P_{\theta}) \) tends to cover regions with small values of \( p_{\text{data}} \)
- Augment VAEs with GAN loss [Larsen et al., 2016]
  - Alleviate the mode covering issue of VAEs
  - Improve the sharpness of VAE generated images
- Augment GANs with VAE loss [Che et al., 2017]
  - Alleviate the mode missing issue of GAN
### Mutual exchanges of ideas: augment the model

<table>
<thead>
<tr>
<th>Discriminator distribution</th>
<th>GANs (InfoGAN)</th>
<th>VAEs</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q_\phi(y</td>
<td>x)$</td>
<td>$q_*(y</td>
</tr>
</tbody>
</table>

- Activate the adversarial mechanism in VAEs
  - Enable adaptive incorporation of fake samples for learning
  - Straightforward derivation by making symbolic analog to GANs

---

Vanilla VAEs

![Vanilla VAEs](image1)

Adversary Activated VAEs

![Adversary Activated VAEs](image2)
Adversary Activated VAEs (AAVAE)

• Vanilla VAEs:

\[
\max_{\theta, \eta} L_{\theta, \eta}^{\text{vae}} = \mathbb{E}_{p_{\theta_0}(x)} \left[ \mathbb{E}_{q_{\eta}(z|x,y)} q_{\star}(y|x) \log p_{\theta}(x|z, y) - \text{KL}(q_{\eta}(z|x,y)q_{\star}(y|x) || p(z|y)p(y)) \right]
\]

• Replace \( q_{\star}(y|x) \) with learnable one \( q_{\phi}(y|x) \) with parameters \( \phi \)
  • As usual, denote reversed distribution \( q_{\phi}^r(y|x) = q_{\phi}(y|x) \)

\[
\max_{\theta, \eta} L_{\theta, \eta}^{\text{aavae}} = \mathbb{E}_{p_{\theta_0}(x)} \left[ \mathbb{E}_{q_{\eta}(z|x,y)} q_{\phi}^r(y|x) \log p_{\theta}(x|z, y) - \text{KL}(q_{\eta}(z|x,y)q_{\phi}^r(y|x) || p(z|y)p(y)) \right]
\]
AAVAE: empirical results

- Applied the adversary activating method on
  - vanilla VAEs
  - class-conditional VAEs (CVAE)
  - semi-supervised VAEs (SVAE)

- Evaluated test-set variational lower bound on MNIST
  - The higher the better

<table>
<thead>
<tr>
<th>Train Data Size</th>
<th>VAE</th>
<th>AA-VAE</th>
<th>CVAE</th>
<th>AA-CVAE</th>
<th>SVAE</th>
<th>AA-SVAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>-122.89</td>
<td>-122.15</td>
<td>-125.44</td>
<td>-122.88</td>
<td>-108.22</td>
<td>-107.61</td>
</tr>
<tr>
<td>10%</td>
<td>-104.49</td>
<td>-103.05</td>
<td>-102.63</td>
<td>-101.63</td>
<td>-99.44</td>
<td>-98.81</td>
</tr>
<tr>
<td>100%</td>
<td>-92.53</td>
<td>-92.42</td>
<td>-93.16</td>
<td>-92.75</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

- X-axis: the ratio of training data for learning: (1%, 10%, 100%)
- Y-axis: value of test-set lower bound
AAVAE: empirical results

- Evaluated classification accuracy of SVAE and AA-SVAE

<table>
<thead>
<tr>
<th></th>
<th>1%</th>
<th>10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVAE</td>
<td>0.9412±.0039</td>
<td>0.9768±.0009</td>
</tr>
<tr>
<td>AASVAE</td>
<td><strong>0.9425±.0045</strong></td>
<td><strong>0.9797±.0010</strong></td>
</tr>
</tbody>
</table>

- Used 1% and 10% data labels in MNIST
Importance weighted GANs (IWGAN)

• Generator learning in vanilla GANs

$$\max_{\theta} \mathbb{E}_{x \sim p_{\theta}(x|y)p(y)} \left[ \log q^r_{\phi_0}(y|x) \right]$$

• Generator learning in IWGAN

$$\max_{\theta} \mathbb{E}_{x_1,...,x_k \sim p_{\theta}(x|y)p(y)} \left[ \sum_{i=1}^{k} \frac{q^r_{\phi_0}(y|x_i)}{q_{\phi_0}(y|x_i)} \log q^r_{\phi_0}(y|x_i) \right]$$

• Assigns higher weights to samples that are more realistic and fool the discriminator better.
IWGAN: empirical results

- Evaluated on MNIST and SVHN
- Used pre-trained NN to evaluate:
  - Inception scores of samples from GANs and IW-GAN
    - Confidence of a pre-trained classifier on generated samples + diversity of generated samples

<table>
<thead>
<tr>
<th></th>
<th>MNIST</th>
<th>SVHN</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAN</td>
<td>8.34±.03</td>
<td>5.18±.03</td>
</tr>
<tr>
<td>IWGAN</td>
<td>8.45±.04</td>
<td>5.34±.03</td>
</tr>
</tbody>
</table>

- Classification accuracy of samples from CGAN and IW-CGAN

<table>
<thead>
<tr>
<th></th>
<th>MNIST</th>
<th>SVHN</th>
</tr>
</thead>
<tbody>
<tr>
<td>CGAN</td>
<td>0.985±.002</td>
<td>0.797±.005</td>
</tr>
<tr>
<td>IWCGAN</td>
<td>0.987±.002</td>
<td>0.798±.006</td>
</tr>
</tbody>
</table>
Symmetric modeling of latent & visible variables

Empirical distributions over visible variables

• Impossible to be explicit distribution
  • The only information we have is the observe data examples
  • Do not know the true parametric form of data distribution

• Naturally an implicit distribution
  • Easy to sample from, hard to evaluate likelihood

Prior distributions over latent variables

• Traditionally defined as explicit distributions, e.g., Gaussian prior distribution
  • Amiable for likelihood evaluation
  • We can assume the parametric form according to our prior knowledge

• New tools to allow implicit priors and models
  • GANs, density ratio estimation, approximate Bayesian computations
  • E.g., adversarial autoencoder [Makhzani et al., 2015] replaces the Gaussian prior of vanilla VAEs with implicit priors
Symmetric modeling of latent & visible variables

• No difference in terms of formulations
  • with implicit distributions and black-box NN models
  • just swap the symbols $x$ and $z$

$$z \sim p_{\text{prior}}(z)$$
$$x \sim f_{\text{black-box}}(z)$$

$$x \sim p_{\text{data}}(x)$$
$$z \sim f'_{\text{black-box}}(x)$$
Symmetric modeling of latent & visible variables

• No difference in terms of formulations
  • with implicit distributions and black-box NN models

• Difference in terms of space complexity
  • depend on the problem at hand
  • choose appropriate tools:
    • implicit/explicit distribution, adversarial/maximum-likelihood optimization, …

6 Discussions

Our new interpretations of GANs and VAEs have revealed strong connections between them, and linked the emerging new approaches to the classic wake-sleep algorithm. The generality of the proposed formulation offers a unified statistical insight of the broad landscape of deep generative modeling, and encourages mutual exchange of improvement ideas across research lines. It is interesting to further generalize the framework to connect to other learning paradigms such as reinforcement learning as previous works have started exploration [14, 44]. GANs simultaneously learn a metric (defined by the discriminator) to guide the generator learning, which resembles the iterative teacher-student distillation framework [23, 24] where a teacher network is simultaneously learned from structured knowledge (e.g., logic rules) and provides knowledge-informed learning signals for student networks of interest. It will be intriguing to build formal connections between these approaches and enable incorporation of structured knowledge in deep generative modeling.
Conclusions

• Deep generative models research have a long history
  • Deep belief nets / Helmholtz machines / Predictability Minimization / …

• Unification of deep generative models
  • GANs and VAEs are essentially minimizing KLD in opposite directions
    • Extends two phases of classic wake sleep algorithm, respectively
  • A general formulation framework useful for
    • Analyzing broad class of existing DGM and variants: ADA/InfoGAN/Joint-models/…
    • Inspiring new models and algorithms by borrowing ideas across research fields

• Symmetric view of latent/visible variables
  • No difference in formulation with implicit prior distributions and black-box NN transformations
  • Difference in space complexity: choose appropriate tools
Thank You