An Information-Theoretic Analysis of Deep Latent-Variable Models

Alexander A. Alemi, Ben Poole, Ian Fischer, Joshua V. Dillon, Rif A. Saurous, Kevin Murphy

(under review for ICLR 2018)
Is Maximum Likelihood Useful for Representation Learning?

A few weeks ago at the DALI Theory of GANs workshop we had a great discussion about what GANs are even useful for. Pretty much everybody agreed that generating random images from a model is not really our goal. We either want to use GANs to train conditional probabilistic models (like we do for image super-resolution or speech synthesis, or something along those lines), or as a means of unsupervised representation learning. Indeed, many papers examine the latent space representations that GANs learn.

But the elephant in the room is that nobody really agrees on what unsupervised representation learning really means, and why any GAN variant should be any better or worse at it than others, whether GANs or VAEs are better for that. So I thought I'd write a post to address this, focusing now on maximum likelihood learning and variational autoencoders, but many of these things holds true for variants of GANs as well.
Maximum likelihood over all LVMs

usefulness of $p_{\text{model}}(z|x)$

negative log likelihood $KL[p_D(x) \parallel p_{\text{model}}(x)]$
Maximum likelihood within model class $Q$

usefulness of $p_{\theta}(z|x)$

negative log likelihood $KL[p_D(x) \| p_{\theta}(x)]$

$\theta_{ML}$
Max. likelihood with overly flexible $p_\theta(x|z)$

usefulness of $p_\theta(z|x)$

negative log likelihood $KL[p_D(x)||p_\theta(x)]$
framework

(whiteboard)
true data dist: \( p^*(x) \)

encoder: \( e(z \mid x) \)

joint: \( p_e(x, z) = p^*(x) \cdot e(z \mid x) \)

aggregated pot: \( p_e(z) = \int dx \ p_e(x, z) \)

decoder: \( p_e(x \mid z) = \frac{p_e(x, z)}{p_e(z)} \)

representation usefuless: \( I_{\text{rep}}(x, z) \)

\[ I_{\text{rep}}(x, z) = \langle KL[p_e(x, z) \mid p^*(x), p_e(z)] \rangle_{p^*} \]

\[ \text{Var. approx} \]

\[ \frac{d(e(x \mid z))}{\sim p_e(x \mid z)} \]

\[ \frac{m(z)}{\sim p_e(z)} \]

reconstruction loss:

\[ -H[p^*] - \langle e(z \mid x) \log d(x \mid z) \rangle_{p^*} \leq I_{\text{rep}} \leq \langle e(z \mid x) \cdot \log \frac{e(z \mid x)}{m(z)} \rangle_{p^*} \]

D discrimination

\[ L_{\text{ELBO}} = -D - R \]

sandwich of: \( H - D - R \leq 0 \)
For discrete data:

\[
H \geq 0, 0 \leq P \leq 1
\]

\[
P \geq 0
\]

\[
R \geq 0, (R \leq \text{div})
\]

---

**auto-decoding limit**

\[
e(\hat{x} | x) = \mu(x)
\]

- lowest possible distortion due to \(H\)
- density estimation w/o learned representation

\[\text{under decoder worse approximation}\]

**autoencoding limit**

- lowest possible rate: \(H\)
- \(d(x | \hat{x}) = \text{Pe}(x | \hat{x})\)

\[\rightarrow \text{higher rate with no change in } D\]

by \(\mu(t)\) worse approx. to \(\text{Pe}(t)\)

- \(D\) not a func. of \(\mu(t)\)
- lossless compressor

for parametric families

\[
D
\]

\[\text{Worse } \mu(t)\]

\[\text{worse decoder}\]
min D at fixed rate

or

min D + β R for fixed β
3. Optimisation

\[ \min D \text{ at fixed } R \]  
\begin{align*}
&\text{(Constrained optimisation)}
\end{align*}

\[ \text{OR} \]

\[ \text{Logistic transform.} \]

\[ \text{Find optimal } R \text{ and } D \text{ for fixed } \beta = \frac{\partial D}{\partial R} \]

\[ \min \mathbb{E}_{q(z|\theta)}[D + \beta R] \]

* with \( \beta = 1 \) this is the ELBO (VAE)

* \( \beta \)-VAE

4. Toy Model

**Target**: \( \text{target } R = 1(z, x) = 0.5 \) and minimise \( D \)

**Model's Gen. dist.**
\[ q(x) = \sum_{z} w(z) \pi(x | z) \]

**Empirical data reconstruction distribution**
\[ d(x) = \sum_{z} \sum_{z'} \frac{\pi(z') e(\pm 1 x')}{N} \pi(x | z) \]

\[ \tilde{p}(x) \approx q(x) \approx d(x) \]

**VAE (\( \beta = 1 \))** → \( K = 0.0002 \) units
toy model

\[ z = \text{flip}(0.7) \]
\[ h = \text{normal}(\mu(z), \sigma(z)) \]
\[ x = \text{discretize}(h, \text{bins}) \]

\[ g(x) = \sum_z m(z)d(x|z) \]
\[ d(x) = \sum_{x'} \sum_z \hat{p}(x')e(z|x')d(x|z) \]
\[ p^*(x) \approx g(x) \approx d(x) \]
model form
<table>
<thead>
<tr>
<th>encoder</th>
<th>decoder</th>
<th>marginal</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN w 64dim Gaussian</td>
<td>Deconvolutional net</td>
<td>Gaussian</td>
</tr>
<tr>
<td>CNN then 4 step mean-only IAF</td>
<td>PixelCNN++</td>
<td>CNN then 4 step mean-only IAF</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>VampPrior</td>
</tr>
</tbody>
</table>
MNIST

(a) Distortion vs Rate

(b) ELBO ($R + D$) vs Rate
reconstruction generation

(a) MNIST Reconstructions: $z \sim e(z|x), \hat{x} \sim d(x|z)$

(b) MNIST Generations: $z \sim m(z), \hat{x} \sim d(x|z)$

smoothly interpolate between pure autoencoding and autodecoding behaviour by varying $\beta$
compression model
$\lambda = \lambda_1$

compression model
$\lambda = \lambda_2$

generative models
$\lambda \rightarrow \infty$

$D$

$R + \lambda_1 D = \text{const}$

$R + \lambda_2 D = \text{const}$

(Ballé et al, 2017)
Omniglot

(a) Distortion vs Rate

(b) ELBO ($R + D$) vs Rate
<table>
<thead>
<tr>
<th></th>
<th>data</th>
<th>sample average</th>
<th>sample average</th>
<th>sample average</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta = 1.10$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta = 0.80$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta = 0.40$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**reconstruction**

(a) Omniglot Reconstructions: $z \sim e(z|x)$, $\hat{x} \sim d(x|z)$

(b) Omniglot Generations: $z \sim m(z)$, $\hat{x} \sim d(x|z)$
β-VAE

(Higgins et al, 2017)
\(\beta\text{-VAE}\)

- DC-IGN
- InfoGAN
- \(\beta\text{-VAE}\)
- VAE

(a) Azimuth (rotation)

(b) Lighting

(c) Elevation

(Higgins et al, 2017)
sketch-rnn

$w_{KL} = 1.00$

$w_{KL} = 0.50$

$w_{KL} = 0.25$

(Ha, Eck, 2017)
Figure 13: Reconstructions of sketch images using models with various $w_{KL}$ settings.

(Ha, Eck, 2017)