Tutorial on Variational Autoencoders

David Nagy @dvgnagy
David Szepesvari

EEML 2021, Budapest
Why are VAEs interesting?

• a single model that can both
  • work as a generative model
  • infer latent variables
• both
  • theoretical interpretation in probabilistic modeling framework
  • but also works on naturalistic data
• a building block in many larger models
generative models

Karras et al. 2017
generative models

Silver et al 2016

Besse et al 2019
generative models
generative models

An animation of the gradient descent method predicting a structure for CASP13 target T1008
generative models

latent state $z$

generative model $p(x|z)$

observation $x$
how do we learn a generative model?

- match empirical distribution to predictive distribution

$$\arg\min_{model} KL \left[ p_{data}(x) \parallel p_{model}(x) \right]$$

- equivalent to maximum likelihood

$$\arg\max_{\theta} \log p_{\theta}(x)$$
inference

inference
$p(z|x)$

generative model
$p(x|z)$
inference
inference
inference

• what we see is not the data but an interpretation of the data
• ‘unconscious inference’ over latent variables
how do we do inference?
how do we do inference?
how do we do inference?

lighting \times \text{reflectance} = \text{brightness}
how do we do inference?

lighting \times \text{reflectance} = \text{brightness}
how do we do inference?

lighting \times \text{reflectance} = \text{brightness}

\text{reflection} \rightarrow \text{brightness} → \text{inference}
how do we do inference?

\[ P(x \mid z) \]
how do we do inference?
how do we do inference?

$P(x|z)$
how do we do inference?

- Inference can be performed by inverting the generative model.
- "Vision is inverse graphics."
- Bayesian inference.

\[
p(x \mid z) \propto P(x \mid z) P(z)\]
how do we do inference?

approximation 1: variational bayes

\[ \arg \min_{\phi} KL[q_\phi(z | x) \| p(z | x)] \]

\[ q_\phi(z | x) = \mathcal{N}(z | \mu_\phi, \sigma_\phi) \]
how do we do inference?
approximation 1: variational bayes

\[
\text{arg min}_{\phi} KL[q_\phi(z | x) \mid \mid p(z | x)]
\]

Ferenc notation (inference.vc)

- \( p_\theta(z) \) is very easy 🍦,
- \( p_\theta(x|z) \) is easy 🐒,
- \( p_\theta(x, z) \) is easy 🐒,
- \( p_\theta(x) \) is super-hard 🐒,
- \( p_\theta(z|x) \) is mega-hard 🐒
how do we do inference?

approximation 1: variational bayes

$$q_\phi(z \mid x) = \mathcal{N}(z \mid \mu_\phi, \sigma_\phi)$$

$$p(z \mid x)$$

$$\arg \min_\phi KL[q_\phi(z \mid x) \mid \mid p(z \mid x)]$$
how do we do inference?
approximation 2: amortised inference

\[ f \]

\[
\begin{align*}
x_1 & \mapsto q(z | x_1) \\
x_2 & \mapsto q(z | x_2) \\
& \quad \vdots \\
x_N & \mapsto q(z | x_N)
\end{align*}
\]

- \( f \) mapping between datapoint and posterior
- if we have \( x, q \) pairs for a set of datapoints,
- we can treat learning \( f \) as a supervised regression problem
how do we do inference?

approximation 2: amortised inference

\[ f \approx f_\phi \]

\[ x_1 \mapsto q(z | x_1) \]

\[ x_2 \mapsto q(z | x_2) \]

\[ \vdots \]

\[ x_N \mapsto q(z | x_N) \]

- treat learning \( f \) as a supervised regression problem
- approximate the mapping with a simpler function (NN)
  - inference network
how do we do inference?

reparameterisation trick
how do we do inference?
reparameterisation trick

\[ q_{\phi}(z | x) = \mathcal{N}(z | f_{\phi}^\mu(x), f_{\phi}^\sigma(x)) \]

\[ z \sim q_{\phi}(z | x) \]
how do we do inference?
reparameterisation trick

\[
q_\phi(z | x) = \mathcal{N}(z | f^\mu_\phi(x), f^\sigma_\phi(x))
\]

\[
z \sim q_\phi(z | x)
\]

\[
z = \mu + \sigma \odot \epsilon
\]

\[
\epsilon \sim \mathcal{N}(0, I)
\]
Variational Autoencoder

\[ x \xrightarrow{\text{inference}} z \xrightarrow{\text{generation}} \hat{x} \]
Combined objective

evidence lower bound (ELBO)

\[
\log p_\theta(x) - KL[q_\phi(z \mid x) \mid \mid p(z \mid x)] = \mathcal{L}_{ELBO}(\theta, \phi, x)
\]
Combined objective

evidence lower bound (ELBO)

\[
\log p_\theta(x) - KL[q_\phi(z \mid x) \mid \mid p(z \mid x)] = \mathcal{L}_{ELBO}(\theta, \phi, x)
\]

\[
\mathcal{L}_{ELBO}(\theta, \phi, x) = \mathbb{E}_z \log \mathcal{Z} - \mathbb{E}_z KL[q_\phi(z \mid x) \mid \mid p(z \mid x)]
\]
Combined objective 

evidence lower bound (ELBO)

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

reconstruction

regularisation
$$-KL[q_\phi(z|x) \mid \mid p(z)]$$

regularisation

$$\mathbb{E}_{z \sim q_\phi(z|x)}[\log p_\theta(x|z)] = |x - \hat{x}|^2$$ for gaussian noise model
\[-KL[q_\phi(z \mid x) \mid \mid p(z)]\] 
regularisation 

\[q_\phi(z \mid x) = \mathcal{N}(z \mid f_\phi^\mu(x), f_\phi^\sigma(x))\]

\[p(z) = \mathcal{N}(z \mid 0, I)\]

analytical result for KL if Gaussian prior and posterior:

\[\frac{1}{2}(1 + \log \sigma^2 - \mu^2 - \sigma^2)\]

\[\mathbb{E}_{z \sim q_\phi(z \mid x)}[\log p_\theta(x \mid z)]\] 
reconstruction
Reconstruction

\[-KL[q_\phi(z \mid x) \mid \mid p(z)]\]

Regularisation

\[\mathbb{E}_{z \sim q_\phi(z \mid x)}[\log p_\theta(x \mid z)]\]

Low KL

\[q_\phi(z \mid x)\]
\[-KL[q_\phi(z \mid x) \mid p(z)]\text{ regularisation}\]

\[\mathbb{E}_{z \sim q_\phi(z \mid x)}[\log p_\theta(x \mid z)]\text{ reconstruction}\]

\[p(z) \quad q_\phi(z \mid x) \quad \text{high KL}\]

\[z_1 \quad z_2 \quad x \quad \hat{x}\]
\[-KL[q_\phi(z | x) || p(z)]\]
regularisation

\[\mathbb{E}_{z \sim q_\phi(z | x)}[\log p_\theta(x | z)]\]
reconstruction

zero KL
Demo

https://github.com/eemlcommunity/PracticalSessions2021/tree/main/generative
Part II
\[ \mathcal{L}(\theta, \phi, x) = \mathbb{E}_{z \sim q_{\phi}(z|\mathbf{x})}[\log p_{\theta}(\mathbf{x}|z)] - \beta \text{KL}[q_{\phi}(z|\mathbf{x}) || p(z)] \]
Compression view
Optimal compression

Each encoding is a point on the RD plane.
Optimal compression

memory resource

realisable codes

optimal codes

expected distortion

$R$

$D$
Optimal compression

memory resource

lossless compression

realisable codes

optimal codes

compression without retaining information

expected distortion
Compression view of ELBO

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

- **reconstruction**
- **regularisation**
Compression view of ELBO
Compression view of ELBO

- low expected KL
- low mutual information

- high expected KL
- high mutual information
RD plane

no constraints on model family

- encoder
- decoder
- prior

(Alemi et al. 2017)
Auto-encoding limit

- lossless compression
- store and return data points
- each region of $Z$ corresponds to a datapoint

(Alemi et al. 2017)
Auto-decoding limit

- compression without retaining information
- $z$ independent of $x$
- density estimation without representation learning

(Alemi et al. 2017)
Intermediate points

- intermediate trade-offs between rate and distortion
- these points can’t be targeted through β-VAE objective since they all belong to β=1
- point is selected implicitly through model architecture and initial conditions

(Alemi et al. 2017)
Constrained model family

no constraints on model family

• varying $\beta$ within single architecture can interpolate between auto-encoding and auto-decoding behaviour

(Alemi et al. 2017)
Reconstruction

**autoencoder**
\[- \rightarrow + v \]
\[\beta = 0.1\]
\[(R, D) = (156.0, 4.8)\]
\[\text{ELBO} = -162.0\]

<table>
<thead>
<tr>
<th>data</th>
<th>sample</th>
<th>average</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

(Alemi et al. 2017)
Reconstruction

<table>
<thead>
<tr>
<th>autoencoder</th>
<th>syntactic encoder</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta = 0.1$</td>
<td>$\beta = 0.9$</td>
</tr>
<tr>
<td>$(R, D) = (156.0, 4.8)$</td>
<td>$(R, D) = (120.3, 8.1)$</td>
</tr>
<tr>
<td>ELBO = $-162.0$</td>
<td>ELBO = $-128.0$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>data</th>
<th>sample average</th>
<th>sample average</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

(Alemi et al. 2017)
Reconstruction

<table>
<thead>
<tr>
<th>autoencoder</th>
<th>syntactic encoder</th>
<th>semantic encoder</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta = 0.1 ) &amp; ( \beta = 0.9 ) &amp; ( \beta = 1.0 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>((R, D) = (156.0, 4.8)) &amp; ((R, D) = (120.3, 8.1)) &amp; ((R, D) = (6.2, 74.1))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \text{ELBO} = -162.0 ) &amp; ( \text{ELBO} = -128.0 ) &amp; ( \text{ELBO} = -80.3 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(data) sample average sample average sample average

\(2\) \(2\) \(2\) \(2\) \(2\) \(2\) \(2\) \(2\)

\(0\) \(0\) \(0\) \(0\) \(0\) \(0\) \(0\) \(0\)

\(/\) \(/\) \(/\) \(/\) \(/\) \(/\) \(/\) \(/\)

\(8\) \(8\) \(8\) \(8\) \(8\) \(8\) \(8\) \(8\)
### Reconstruction

<table>
<thead>
<tr>
<th>autoencoder</th>
<th>syntactic encoder</th>
<th>semantic encoder</th>
<th>autodecoder</th>
</tr>
</thead>
<tbody>
<tr>
<td>$+v$</td>
<td>$+v$</td>
<td>$+v$</td>
<td>$+v$</td>
</tr>
<tr>
<td>$\beta = 0.1$</td>
<td>$\beta = 0.9$</td>
<td>$\beta = 1.0$</td>
<td>$\beta = 1.1$</td>
</tr>
<tr>
<td>$(R, D) = (156.0, 4.8)$</td>
<td>$(R, D) = (120.3, 8.1)$</td>
<td>$(R, D) = (6.2, 74.1)$</td>
<td>$(R, D) = (0.0, 80.6)$</td>
</tr>
<tr>
<td>ELBO = $-162.0$</td>
<td>ELBO = $-128.0$</td>
<td>ELBO = $-80.3$</td>
<td>ELBO = $-80.6$</td>
</tr>
</tbody>
</table>

- **Data**
  - 2
  - 0
  - /
  - 8

- **Sample average**
  - 2
  - 0
  - /
  - 8

- **Sample average**
  - 2
  - 0
  - /
  - 8

- **Sample average**
  - 2
  - 4
  - 4
  - 3

(Alemi et al. 2017)
## Generation

<table>
<thead>
<tr>
<th>autoencoder</th>
<th>syntactic encoder</th>
<th>semantic encoder</th>
<th>autodecoder</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta = 0.1$</td>
<td>$\beta = 0.9$</td>
<td>$\beta = 1.0$</td>
<td>$\beta = 1.1$</td>
</tr>
<tr>
<td>$(R, D) = (156.0, 4.8)$</td>
<td>$(R, D) = (120.3, 8.1)$</td>
<td>$(R, D) = (6.2, 74.1)$</td>
<td>$(R, D) = (0.0, 80.6)$</td>
</tr>
<tr>
<td>ELBO = −162.0</td>
<td>ELBO = −128.0</td>
<td>ELBO = −80.3</td>
<td>ELBO = −80.6</td>
</tr>
</tbody>
</table>

| sample average | sample average | sample average | sample average |
|               |               |               |               |

(Alemi et al. 2017)
Disentangled representations

• we were able to control the information content of the latent representation with $\beta$
• we would also prefer representations that are disentangled
  • disentangled = factorised + interpretable (Bengio, 2013)
  • one generative factor per latent dimension
  • capture symmetry transformations
<table>
<thead>
<tr>
<th>Colour</th>
<th>Brightness</th>
<th>Object ID</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wall</td>
<td>Wall</td>
<td>Camera Rotation</td>
<td>Spawn</td>
</tr>
<tr>
<td>Floor</td>
<td>Floor</td>
<td></td>
<td>-3</td>
</tr>
<tr>
<td>Object</td>
<td>Object</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
dsprites dataset
β-VAE

\[ \beta = 1 \]

\[ \beta = 150 \]

(Burgess et al. 2017)
β-VAE

(a) azimuth
DC-IGN

(b) width
InfoGAN

(c) leg style
β-VAE

Factor not learnt

Factor not learnt

Factor not learnt

VAE
Why does $\beta$-VAE disentangle?

- mainly because of the diagonal normal approximate posterior
  - different generative factors give different average contributions to reconstruction
  - these factors should have different capacities allocated to them
  - with a diagonal posterior, this necessitates using different dimensions for each factor
Why does $\beta$-VAE disentangle?

- mainly because of the diagonal normal approximate posterior
  - different generative factors give different average contributions to reconstruction
  - these factors should have different capacities allocated to them
  - with a diagonal posterior, this necessitates using different dimensions for each factor

- Rolinek et al., 2019. “Variational Autoencoders Pursue Pca Directions (by Accident).” https://doi.org/10.1109/CVPR.2019.01269
$$L(\theta, \phi; x(f), z, C) = \mathbb{E}_{q_\phi(z|f)}[\log p_\theta(x|z)] - \gamma D_{KL}(q_\phi(z|f) \parallel p(z)) - C$$

(Burgess et al. 2017)
Modified objective

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

\[ \mathcal{L}'(\theta, \phi, x, C) = \mathbb{E}_{z \sim q_{\phi}(z \mid x)}[\log p_{\theta}(x \mid z)] - \gamma \text{KL}[q_{\phi}(z \mid x) \mid \mid p(z)] - C \]
Modified objective

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

\[ \mathcal{L}'(\theta, \phi, x, C) = \mathbb{E}_{z \sim q_\phi(z|x)}[\log p_\theta(x | z)] - \gamma | KL[q_\phi(z | x) || p(z)] - C | \]
Demo

https://github.com/eemlcommunity/PracticalSessions2021/tree/main/generative