The Mutual Autoencoder: Controlling Information in Latent Code Representations

Bui Thi Mai Phuong[1]
Nate Kushman[2]
Sebastian Nowozin[2]
Ryota Tomioka[2]
Max Welling[3]

[2] Microsoft Research
[1] IST AUSTRIA Institute of Science and Technology
[3] Universiteit van Amsterdam

Summary

- Variational autoencoders fail to learn a representation when an expressive model class is used.
- We propose to explicitly constrain the mutual information between data and the representation.
- On small problems, our method learns useful representations even if a trivial solution exists.

Variational autoencoders (VAEs)

- VAEs: popular approach to generative modelling, i.e. given samples $x_i \sim p_{\text{true}}(x)$, we want to approximate $p_{\text{true}}(x)$.
- Consider the model $p_\theta(x) = p(z)p_\theta(x|z)$, where z is unobserved (latent) and $p(z) = \mathcal{N}(z|0, I)$.

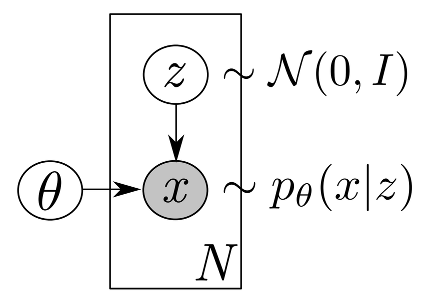

Figure 1: The VAE model.

- For interesting model classes $\{p_\theta : \theta \in \Theta\}$, the log-likelihood is intractable,

$$\log p(x) = \log \int p(z)p_\theta(x|z)\mathrm{d}z,$$

but can be lower-bounded by

$$\mathop{\mathbb{E}}_{z\sim q_\theta(.|x)} \log p_\theta(x|z) - \mathrm{KL}[q_\theta(z|x)\|p(z)] \qquad \text{(ELBO)}$$

for any $q_\theta(z|x)$.
- VAEs maximise the lower bound jointly in p_θ and q_θ.
- The objective can be interpreted as encoding an observation x_{data} via q_θ into a code z, decoding it back into x_{gen}, and measuring the reconstruction error.
- The KL term acts as a regulariser.

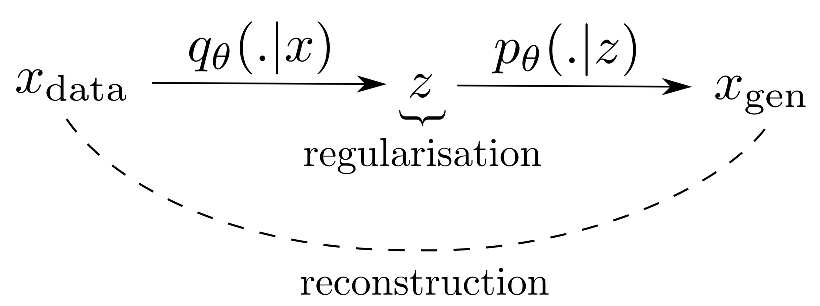

Figure 2: VAE objective illustration.

VAEs for representation learning

- VAEs can learn meaningful representations (latent codes).

Figure 3: Example of a VAE successfully learning a representation (here angle and emotion of a face). Shown are samples from $p_\theta(x|z)$ for a grid of z. Adapted from [6].

VAEs can fail to learn a representation

- Consider setting $p_\theta(x|z) = p_\theta(x)$.
- The ELBO and the log-likelihood attain a global maximum for

$$p_\theta(x|z) = p_{\text{true}}(x) \quad \text{and} \quad q_\theta(z|x) = p(z),$$

but z, x are independent.
- $\Rightarrow$ Useless representation!
- The representation must come from the limited capacity of the decoder family $\{p_\theta : \theta \in \Theta\}$.

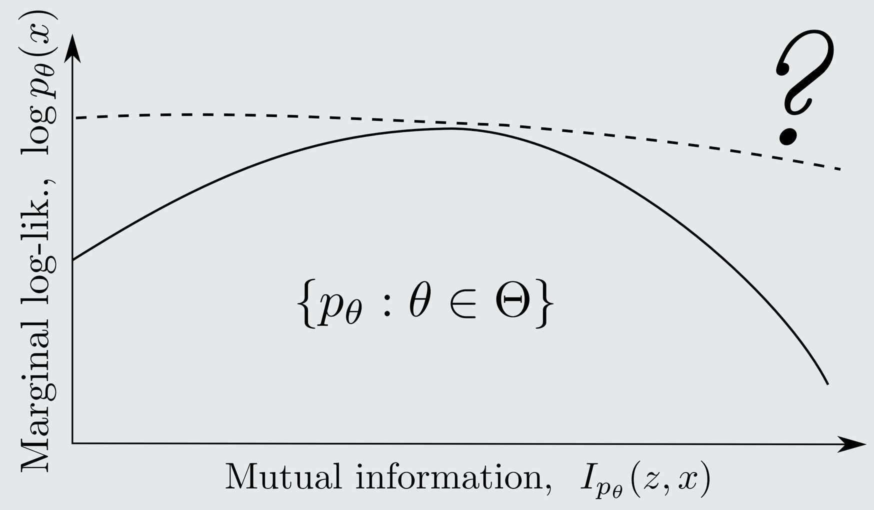

Figure 4: Maximising the log-likelihood (y-axis) enforces mutual information between x and z for appropriately restricted model classes (solid), but not for expressive ones (dashed). Also see [4].

The mutual autoencoder (MAE)

Aims:

- Explicit control of information between x and z.
- Representation learning with powerful decoders.

Idea:

$$\max_\theta \mathop{\mathbb{E}}_{x\sim p_{\text{data}}} \log \int p(z)p_\theta(x|z)\mathrm{d}z,$$
$$\text{subject to} \quad I_{p_\theta}(z, x) = M,$$

where $M \geq 0$ determines the degree of coupling.

Tractable approximation:

- ELBO to approximate the objective.
- Variational infomax bound [1] for the constraint,

$$\begin{aligned} I_{p_\theta}(z, x) &= H(z) - H(z|x) \\ &= H(z) + \mathop{\mathbb{E}}_{z,x\sim p_\theta} \log p_\theta(z|x) \\ &\geq H(z) + \mathop{\mathbb{E}}_{z,x\sim p_\theta} \log r_\omega(z|x) \end{aligned}$$

for any $r_\omega(z|x)$.

Related literature

- In [2], the LSTM decoder learns trivial latent codes, unless weakened via word drop-out.
- In [3], the authors show how to encode specific information in z by deliberate construction of the decoder family.
- For powerful decoders, the KL term in ELBO is commonly annealed from 0 to 1 during training (used e.g. in [2], [5]).

MAE, categorical example

- Data: $x \in \{0, \dots, 9\}$, discrete; $p_{\text{true}} = \mathrm{Uniform}(\{0, \dots, 9\})$.
- Model: $z \sim \mathcal{N}(0, 1)$.
- $p_\theta(x|z)$: 2-layer FC net with softmax output.
- $q_\theta(z|x), r_\omega(z|x)$: normal with means and log-variances modelled by 2-layer FC nets.

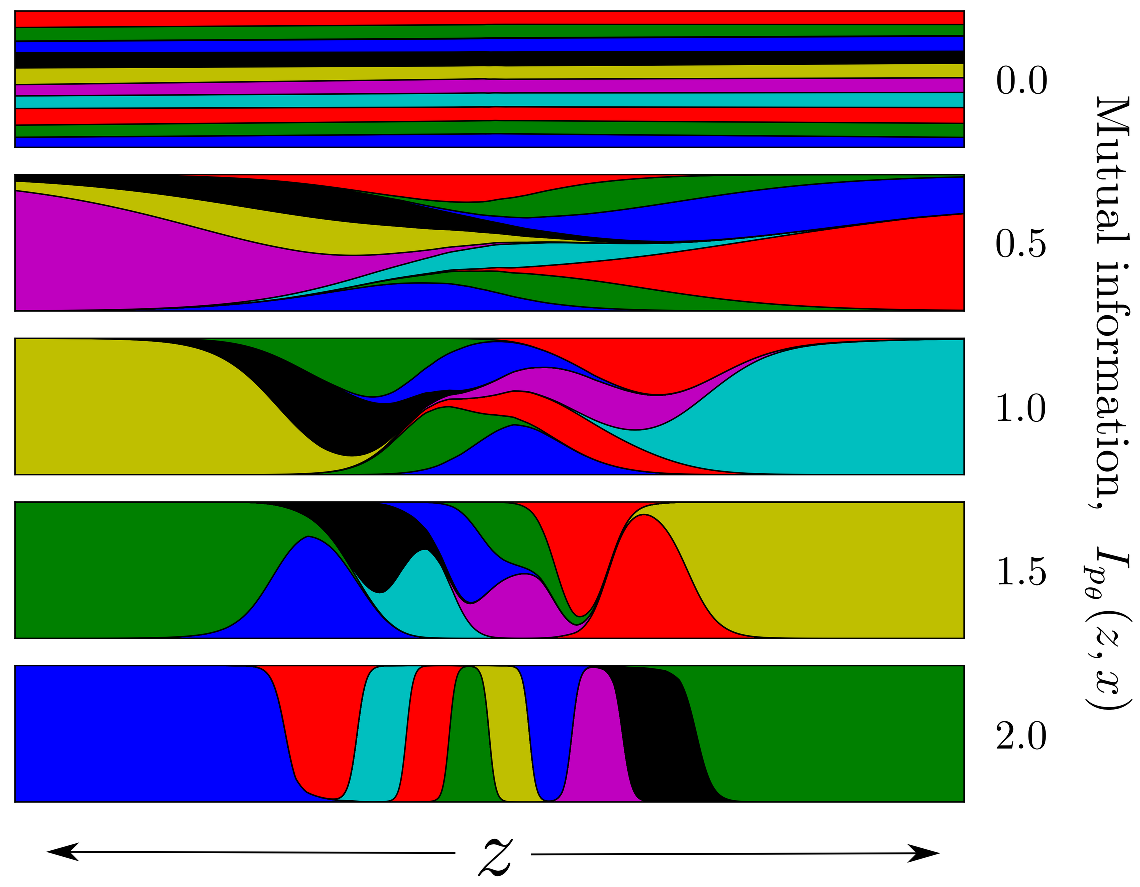

Figure 5: Each row shows the learnt $p_\theta(x|z)$ as a function of z. Different rows correspond to different settings of $I_{p_\theta}(z, x)$.

Splitting the normal

- Data: $x \in \mathbb{R}$, continuous; $p_{\text{true}} = \mathcal{N}(0, 1)$.
- Model: $z \sim \mathcal{N}(0, 1)$.
- $p_\theta(x|z), q_\theta(z|x), r_\omega(z|x)$: normal with means and log-variances modelled by 2-layer FC nets.
- The model has to learn to represent a normal as an infinite mixture of normals.
- A trivial solution ignoring z exists and is recovered by VAEs. Can MAEs obtain an informative representation?

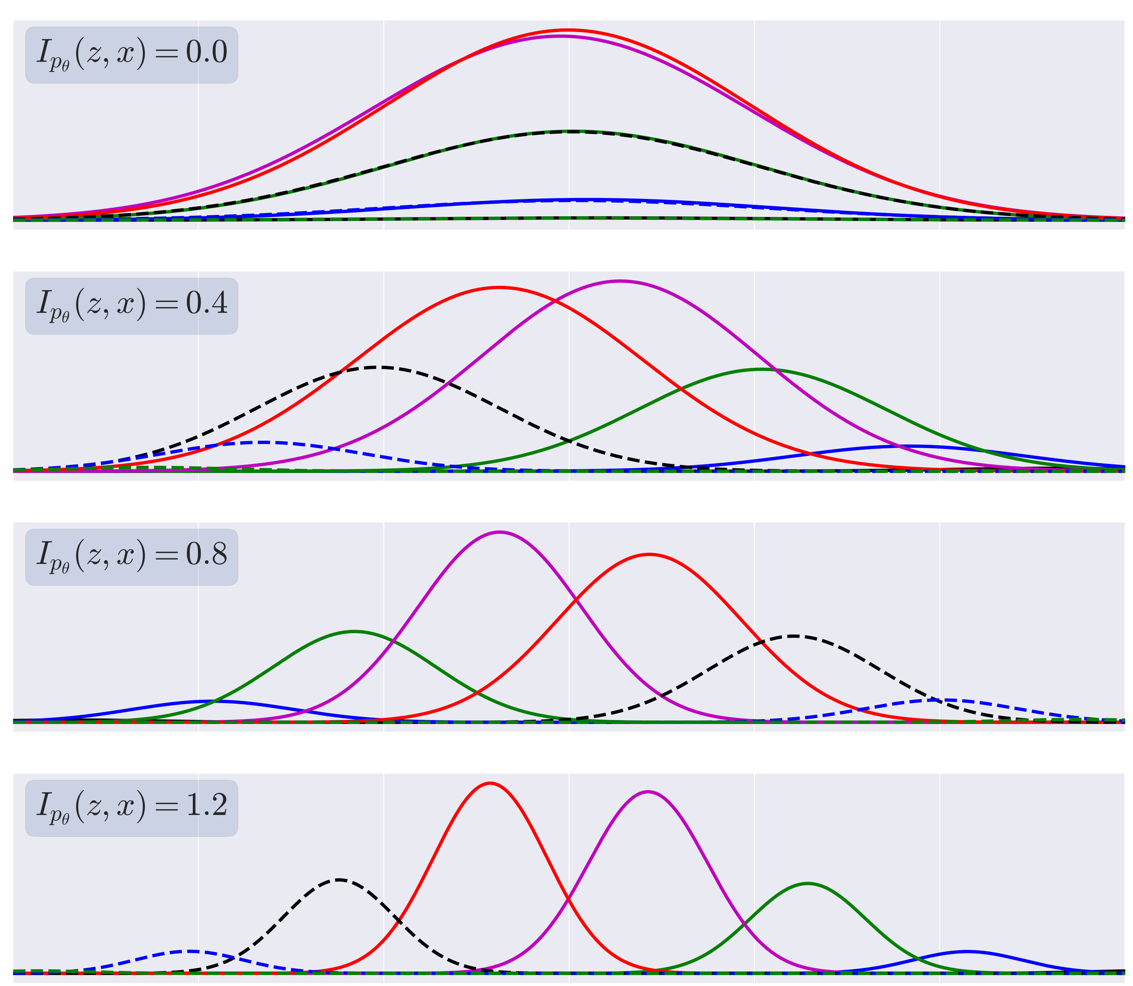

Figure 6: Each row shows the learnt $p_\theta(x|z)p(z)$ (a Gaussian curve) for a grid of z (different colours). Different rows correspond to different settings of $I_{p_\theta}(x, z)$.

References

[1] David Barber and Felix Agakov.
The IM algorithm: a variational approach to information maximization.
In *NIPS*, 2003.

[2] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz, and Samy Bengio.
Generating sentences from a continuous space.
arXiv:1511.06349, 2015.

[3] Xi Chen, Diederik P. Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel.
Variational lossy autoencoder.
arXiv:1611.02731, 2016.

[4] Ferenc Huszár.
Is maximum likelihood useful for representation learning?
http://www.inference.vc/maximum-likelihood-for-representation-learning-2/, 2017.

[5] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther.
How to train deep variational autoencoders and probabilistic ladder networks.
arXiv:1602.02282, 2016.

[6] Diederik P. Kingma and Max Welling.
Auto-encoding variational Bayes.
arXiv:1312.6114, 2013.