Mutual Information in Variational Autoencoders

Felipe N. Ducau, Sony Trénous

NYU

Motivation

State-of-the-Art deep generative models are able to learn expressive latent representations of images. It has been shown that within the latent representations, there are directions corresponding to semantic properties of the images, such as the thickness of stroke in hand-written digits. It would be a desirable property to *disentangle* these directions, such that individual dimensions of the latent code correspond to independent semantic features of the images.

The *infoGAN* model achieved this by enforcing high Mutual Information (MI) between the latent code and its output.

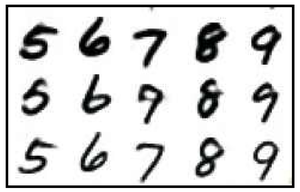


Figure 1. Example of disentanglement in images generated by InfoGAN.

Motivated by this work, we investigate the role of MI in variational autoencoders (VAEs).

Mutual Information

$$\mathbf{I}(X, Z) = \mathbf{H}(Z) - \mathbf{H}(X \mid Z)$$

A symmetric measure of how much information the outcome of random variable Z gives you about X, and vice versa.

VAE Structure

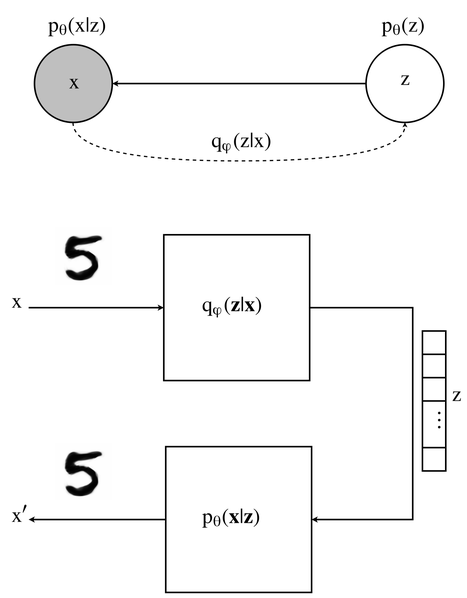


Figure 2. Top: Graphical model of a variational autoencoder. Bottom: Structure of a variational autoencoder seen as a neural network.

Variational Lower Bounds

Intractable Posterior $p_\theta(x|z) \rightarrow$ Variational approximations to $\mathcal{L}(x), \mathbf{I}(x, z)$

$$\mathcal{L}(x) \geq \underbrace{\mathop{\mathbf{E}}_{q_\phi(\mathbf{z}|\mathbf{x}^{(i)})}\left[\log p_\theta(\mathbf{x}^{(i)}|\mathbf{z})\right]}_{\text{Reconstruction Term}} - \underbrace{\mathbf{D}_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x}^{(i)}) \,\middle\|\, p_\theta(\mathbf{z})\right)}_{\text{Regularizer}}$$

$$\mathbf{I}(x, z) \geq \mathop{\mathbf{E}}_{\mathbf{z}\sim p_\theta(\mathbf{z}), \mathbf{x}\sim p_\theta(\mathbf{x}|\mathbf{z})}\left[\log q_\phi(\mathbf{z} \mid \mathbf{x})\right] + \mathbf{H}(\mathbf{z})$$

Implementation & Results

New loss function for our VAE implementation:

$$\mathbf{L} = \mathbf{D}_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x}^{(i)}) \,\middle\|\, p_\theta(\mathbf{z})\right) + \mathop{\mathbf{E}}_{q_\phi(\mathbf{z}|\mathbf{x}^{(i)})}\left[\log p_\theta(\mathbf{x}^{(i)}|\mathbf{z})\right] - \lambda \mathbf{I}(z, x')$$

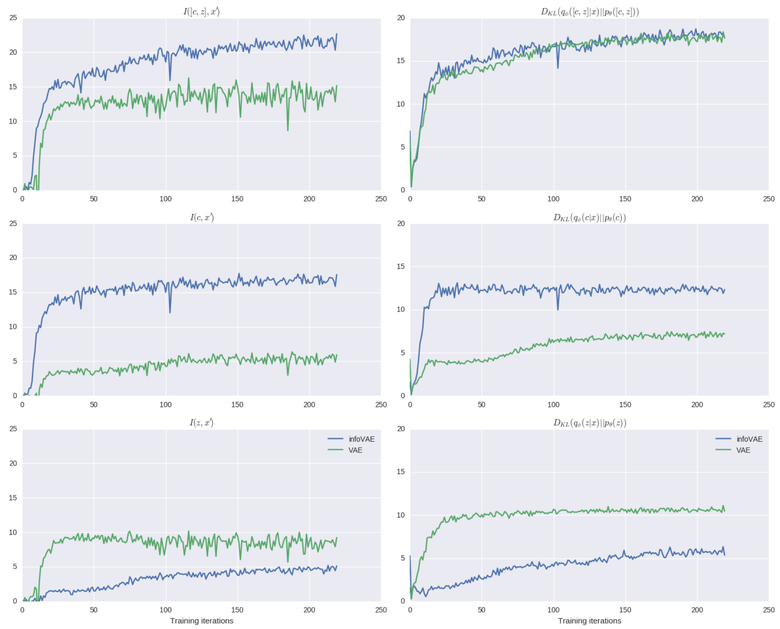


Figure 3. Experimental results enforcing mutual information in half of the dimensions of the latent representation (blue trace) and not enforcing mutual information (green trace).

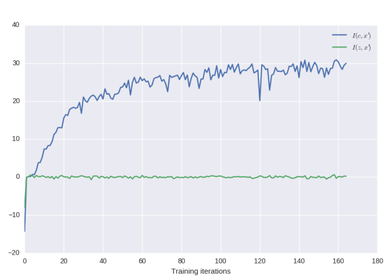


Figure 4. Behavior when enforcing MI in a large code.

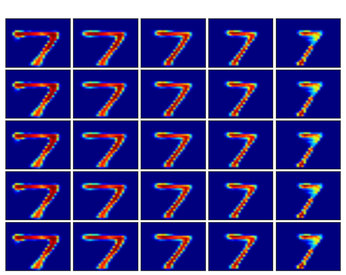
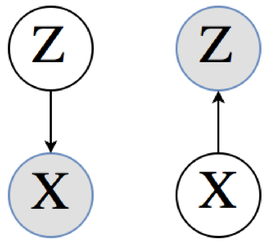


Figure 5. Reconstruction when enforcing MI in a large code.

Observations

- Trade Off between MI and Regularizer is the same as tradeof between Reconstruction Term and Regularizer.
- When Model Capability large enough, the non-regularized part of the code is ignored.
- MI identifies dimensions which are meaningful in the reconstruction.

Duality MI ↔ Reconstruction



MI lower bound is the Reconstruction Term of a *dual* VAE where we observe $\mathbf{z} \sim p_\theta(\mathbf{z})$.

In the same way, the reconstruction term is the MI lower bound in this dual VAE.

Reconstruction Term thus maximizes MI between $\mathbf{x} \sim p_{\text{data}}(\mathbf{x}), \mathbf{z} \sim p_\theta(\mathbf{x})$ – MI lower bound between $\mathbf{z} \sim p_\theta(z), \mathbf{x} \sim p_\theta(\mathbf{x}) = \int_z p_\theta(z) p_\theta(\mathbf{x} \mid \mathbf{z})\, \mathrm{d}z$.

Open Questions

We have shown that maximizing mutual information does not neccessarily help in disentangling latent representations. The question is then:

> *Which properties of the InfoGAN model lead to a disentangled representation?*

Intuitively, we believe the choice of priors which are attuned to independent features in the data is crucial. Thickness of stroke in MNIST digits might follow a uniform distribution, so that the most mutual information between a uniformly distributed variable and the MNIST data is precisely this feature of the data.

References

- Xi Chen et al. "InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets".
- D. P Kingma and M. Welling. "Auto-Encoding Variational Bayes".