The Mutual Autoencoder: Controlling Information in Latent Code Representations

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 $x_i \sim p_{\text{data}}(x)$, 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.](image)

For interesting model classes ($p_{\theta}: \theta \in \Theta$), the log-likelihood is intractable,
$$ \log p(x) = \log \left( \int p_{\theta}(z)p_{\theta}(x|z)dz \right), $$
but can be lower-bounded by
$$ \mathbb{E}_{z \sim q_{\phi}(z)} \log p_{\theta}(z)p_{\theta}(x|z) - KL(q_{\phi}(z)||p(z)) $$
(ELBO) for any $q_{\phi}(z)$.

- VAEs maximise the lower bound jointly in $p_{\theta}$ and $q_{\phi}$.
- The objective can be interpreted as encoding an observation $x_{\text{data}}$ via $q_{\phi}$ into a code $z$, decoding it back into $x_{\text{data}}$, and measuring the reconstruction error.
- The KL term acts as a regulariser.

![Figure 2: VAE objective illustration.](image)

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].](image)

The mutual autoencoder (MAE)

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

Idea:
$$ \max_{\theta, \omega} \mathbb{E}_{x \sim p_{\text{data}}} \left[ p(z)p_{\theta}(x|z)dz \right], $$
subject to
$$ I_{\theta}(z;x) = M, $$
where $M \geq 0$ determines the degree of coupling.

Tractable approximation:
- ELBO to approximate the objective.

Variance infomax bound [1] for the constraint,
$$ I_{\theta}(z;x) = H(z) - H(z|x), $$
subject to
$$ \mathbb{E}_{p(z)} \log p_{\theta}(z) \geq H(z) + \mathbb{E}_{p(z)} \log r_{\omega}(z) $$
for any $r_{\omega}(z)$.

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]).

![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].](image)

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

Splitting the normal

- Data: $x \in \mathbb{R}$, continuous; $p_{\text{true}}(x) = \mathcal{N}(0, 1)$.
- Model: $z \sim \mathcal{N}(0, 1)$.
- $p_{\theta}(x|z), q_{\phi}(z|x), r_{\omega}(z|x)$: 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 $p_{\theta}(x|z)$ for a grid of $z$ (different colours). Different rows correspond to different settings of $I_{\theta}(z;x)$](image)

References


