Model Selection for Bayesian Autoencoders

Ba-Hien Tran
EURECOM

Simone Rossi
EURECOM

Dimitrios Milios
EURECOM

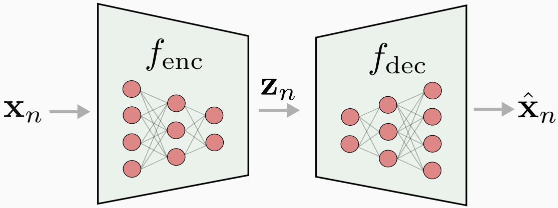
Pietro Michiardi
EURECOM

Edwin V. Bonilla
CSIRO's Data61
The Australian National University
The University of Sydney

Maurizio Filippone
EURECOM

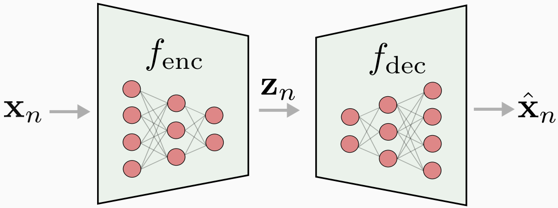
SIAM Conference on Imaging Science 2022

Autoencoders



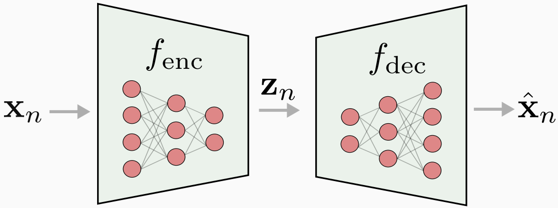
- An autoencoder (AE) is a neural network used for *unsupervised learning*

Autoencoders



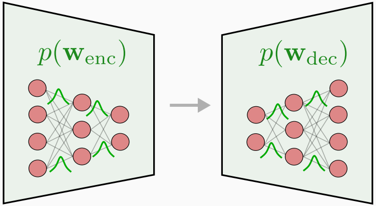
- An autoencoder (AE) is a neural network used for *unsupervised learning*
- *Encoder*: transforms an unlabelled dataset, $\mathbf{x} := \{\mathbf{x}_n\}_n^N$, into latent codes, $\mathbf{z} := \{\mathbf{z}_n\}_n^N$
- *Decoder*: transforms latent codes into reconstructions, $\hat{\mathbf{x}} := \{\hat{\mathbf{x}}_n\}_n^N$

Autoencoders



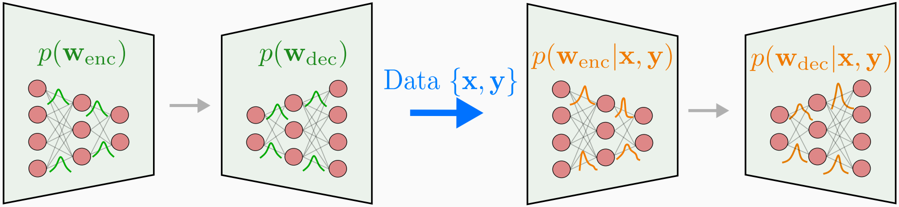
- An autoencoder (AE) is a neural network used for *unsupervised learning*
- *Encoder*: transforms an unlabelled dataset, $\mathbf{x} := \{\mathbf{x}_n\}_n^N$, into latent codes, $\mathbf{z} := \{\mathbf{z}_n\}_n^N$
- *Decoder*: transforms latent codes into reconstructions, $\hat{\mathbf{x}} := \{\hat{\mathbf{x}}_n\}_n^N$
- Typical AE solution: a point estimate of the network's parameters $\mathbf{w} := \{\mathbf{w}_{\text{enc}}, \mathbf{w}_{\text{dec}}\}$

Bayesian Autoencoders



- A Bayesian neural network for unsupervised learning
- Place a prior $p(\mathbf{w})$ over the network's parameters $\mathbf{w} := \{\mathbf{w}_{\mathrm{enc}}, \mathbf{w}_{\mathrm{dec}}\}$

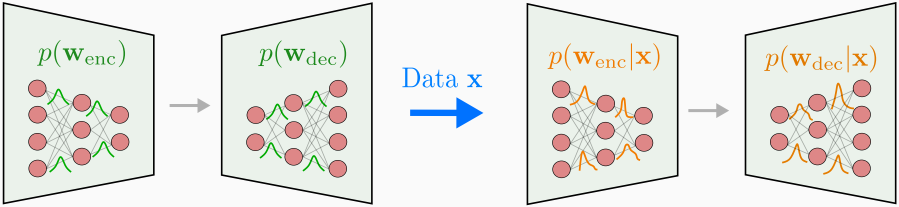
Bayesian Autoencoders



- Place a prior $p(\mathbf{w})$ over the network's parameters $\mathbf{w} := \{\mathbf{w}_{\text{enc}}, \mathbf{w}_{\text{dec}}\}$
- The target is exactly the input, $\mathbf{y}_n = \mathbf{x}_n$
- Compute posterior given a dataset $\{\mathbf{x}, \mathbf{y}\}$:

$$\underbrace{p(\mathbf{w} \mid \mathbf{y})}_{\text{posterior}} \propto \underbrace{p(\mathbf{y} \mid f(\mathbf{x}; \mathbf{w}))}_{\text{likelihood}} \underbrace{p(\mathbf{w})}_{\text{prior}}$$

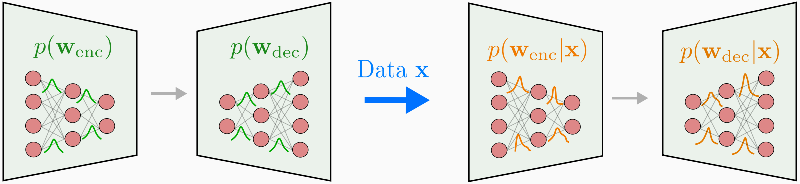
Bayesian Autoencoders



- Place a prior $p(\mathbf{w})$ over the network's parameters $\mathbf{w} := \{\mathbf{w}_{\mathrm{enc}}, \mathbf{w}_{\mathrm{dec}}\}$
- The target is exactly the input, $\mathbf{y}_n = \mathbf{x}_n$
- Compute posterior given a dataset $\{\mathbf{x}\}$:

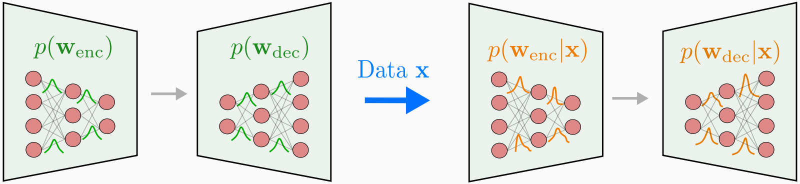
$$\underbrace{p(\mathbf{w} \mid \mathbf{x})}_{\text{posterior}} \propto \underbrace{p(\mathbf{x} \mid f(\mathbf{x}; \mathbf{w}))}_{\text{likelihood}} \underbrace{p(\mathbf{w})}_{\text{prior}}$$

Bayesian Autoencoders



✓ Quantification of uncertainty

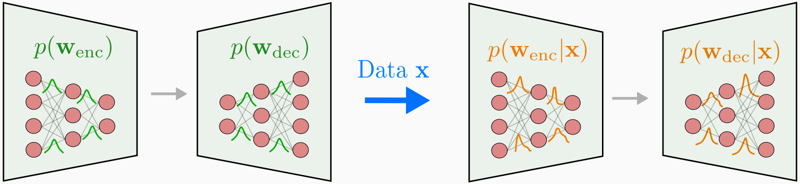
Bayesian Autoencoders



✓ Quantification of uncertainty

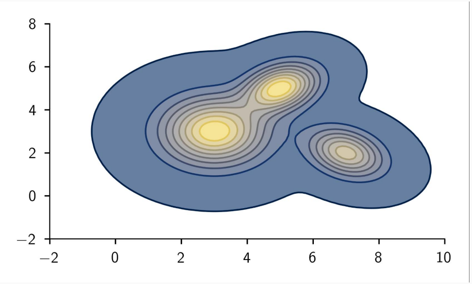
✓ Specifying a prior belief on the network's parameters

Bayesian Autoencoders

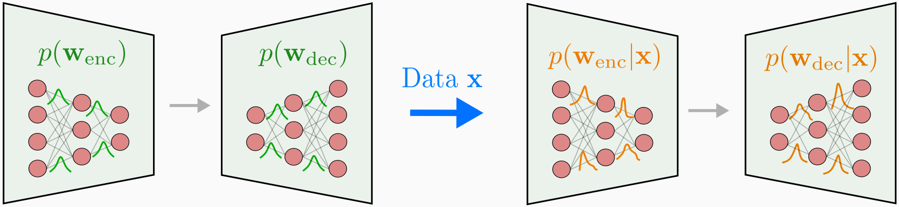


✗ Inference intractability

→ Sampling with stochastic gradient Hamiltonian Monte Carlo (Chen et al., 2014)



Bayesian Autoencoders



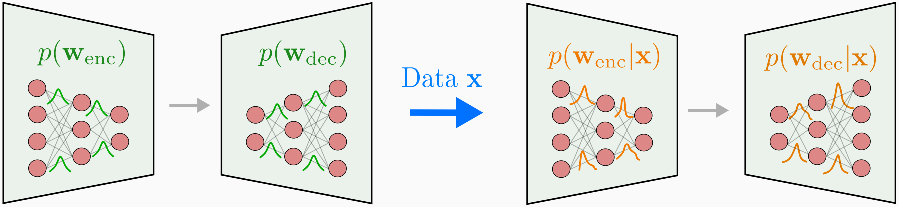
✗ Inference intractability

→ Sampling with stochastic gradient Hamiltonian Monte Carlo (Chen et al., 2014)

✗ Lack of generative modeling

→ Density estimation in *learned* latent space with Dirichlet Process Mixture Model (Blei et al., 2006)

Bayesian Autoencoders



✗ Inference intractability

→ Sampling with stochastic gradient Hamiltonian Monte Carlo (Chen et al., 2014)

✗ Lack of generative modeling

→ Density estimation in *learned* latent space with Dirichlet Process Mixture Model (Blei et al., 2006)

✗ Difficulty of choosing a sensible prior

Functional Priors for Bayesian Autoencoders

✗ Difficulty of choosing a sensible prior

- Assume a prior distribution, $p_{\psi}(\mathbf{w})$, on the parameters
- ψ is the prior hyper-parameters to be chosen

Functional Priors for Bayesian Autoencoders

✗ Difficulty of choosing a sensible prior

- Assume a prior distribution, $p_{\psi}(\mathbf{w})$, on the parameters
- ψ is the prior hyper-parameters to be chosen
- This prior induces a non-trivial effect on the output (functional) prior

$$p_{\psi}(\hat{\mathbf{x}}) = \int f(\mathbf{x}; \mathbf{w}) p_{\psi}(\mathbf{w}) d\mathbf{w},$$

where $\hat{\mathbf{x}} = f(\mathbf{x}; \mathbf{w})$

Functional Priors for Bayesian Autoencoders

✗ Difficulty of choosing a sensible prior

- Assume a prior distribution, $p_{\psi}(\mathbf{w})$, on the parameters
- ψ is the prior hyper-parameters to be chosen
- This prior induces a non-trivial effect on the output (functional) prior

$$p_{\psi}(\hat{\mathbf{x}}) = \int f(\mathbf{x}; \mathbf{w}) p_{\psi}(\mathbf{w}) d\mathbf{w},$$

where $\hat{\mathbf{x}} = f(\mathbf{x}; \mathbf{w})$

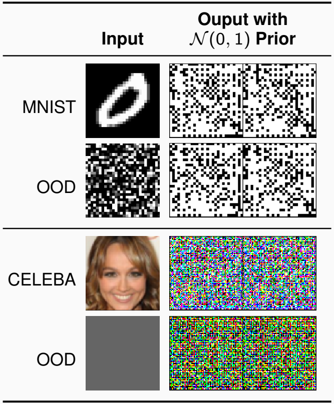


Figure: Realizations sampled from the $\mathcal{N}(0, 1)$ prior given an input image. OOD stands for out-of-distribution.

Model Selection for Bayesian Autoencoders

✗ Difficulty of choosing a sensible prior

$\rightarrow$ Estimating prior hyper-parameters, $\boldsymbol{\psi}$, based on the *empirical Bayes* approach

Model Selection for Bayesian Autoencoders

✗ Difficulty of choosing a sensible prior

→ Estimating prior hyper-parameters, ψ, based on the *empirical Bayes* approach

- Marginal likelihood

$$p_{\psi}(\mathbf{x}) = \int p(\mathbf{x} \mid \hat{\mathbf{x}}) p_{\psi}(\hat{\mathbf{x}}) d\hat{\mathbf{x}},$$

where $p(\mathbf{x} \mid \hat{\mathbf{x}})$ is the likelihood, and $\hat{\mathbf{x}} = f(\mathbf{x}; \mathbf{w})$

Model Selection for Bayesian Autoencoders

✗ Difficulty of choosing a sensible prior

→ Estimating prior hyper-parameters, ψ, based on the *empirical Bayes* approach

- Marginal likelihood

$$p_{\psi}(\mathbf{x}) = \int p(\mathbf{x} \,|\, \hat{\mathbf{x}}) p_{\psi}(\hat{\mathbf{x}}) d\hat{\mathbf{x}},$$

  where $p(\mathbf{x} \,|\, \hat{\mathbf{x}})$ is the likelihood, and $\hat{\mathbf{x}} = f(\mathbf{x}; \mathbf{w})$

- Equivalence between maximum likelihood estimation and KL-divergence minimization

$$\arg\max_{\psi} \int \pi(\mathbf{x}) \log p_{\psi}(\mathbf{x}) d\mathbf{x} = \arg\min_{\psi} \mathrm{KL}[\pi(\mathbf{x}) || p_{\psi}(\mathbf{x})],$$

  where $\pi(\mathbf{x})$ is the data-generating distribution

Model Selection for Bayesian Autoencoders

✗ Difficulty of choosing a sensible prior

→ Estimating prior hyper-parameters, ψ, based on the *empirical Bayes* approach

- Marginal likelihood

$$p_{\psi}(\mathbf{x}) = \int p(\mathbf{x} \mid \hat{\mathbf{x}}) p_{\psi}(\hat{\mathbf{x}}) d\hat{\mathbf{x}},$$

  where $p(\mathbf{x} \mid \hat{\mathbf{x}})$ is the likelihood, and $\hat{\mathbf{x}} = f(\mathbf{x}; \mathbf{w})$

- Equivalence between maximum likelihood estimation and KL-divergence minimization

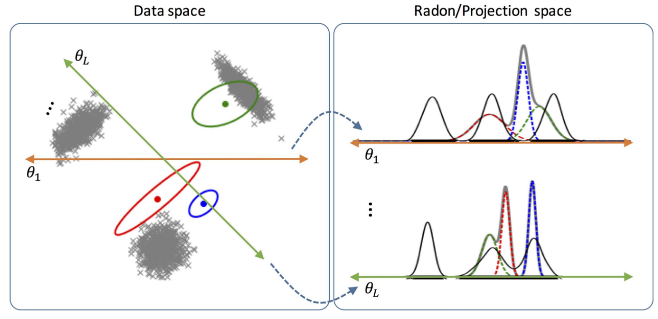
$$\arg\max_{\psi} \int \pi(\mathbf{x}) \log p_{\psi}(\mathbf{x}) d\mathbf{x} = \arg\min_{\psi} \mathrm{KL}[\pi(\mathbf{x}) || p_{\psi}(\mathbf{x})],$$

  where $\pi(\mathbf{x})$ is the data-generating distribution

- Matching these two distributions is non-trivial!

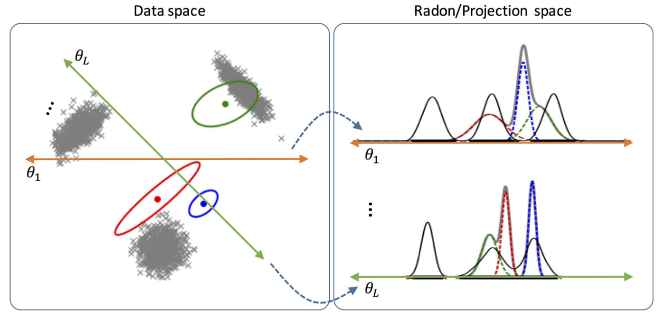
Model Selection for Bayesian Autoencoders

We propose to use the distributional sliced 2-Wasserstein distance (Nguyen et al., 2020)



Model Selection for Bayesian Autoencoders

We propose to use the distributional sliced 2-Wasserstein distance (Nguyen et al., 2020)



✓ DSW distance addresses two major constraints

- Computational scalability thanks to using random projection
- Curse of dimensionality

Model Selection for Bayesian Autoencoders

We propose to use the distributional sliced 2-Wasserstein distance (Nguyen et al., 2020)

$$\psi^\star = \arg\min_{\psi} \left[DSW_2(p_{\psi}(\mathbf{x}), \pi(\mathbf{x})) \right]$$

✓ The objective is *fully sampled-based* and can be optimized with gradient descent algorithms

$\longrightarrow$ Not necessary to know the closed-form of either $p_{\psi}(\mathbf{x})$ or $\pi(\mathbf{x})$

$\longrightarrow$ Only requirement is that we can draw samples from these two distributions

Model Selection for Bayesian Autoencoders

We propose to use the distributional sliced 2-Wasserstein distance (Nguyen et al., 2020)

$$\psi^\star = \arg\min_{\psi} \Big[DSW_2(p_\psi(\mathbf{x}), \pi(\mathbf{x})) \Big]$$

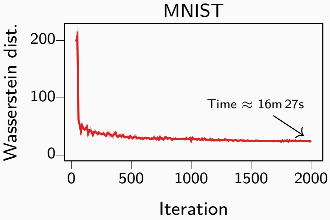
✓ The objective is *fully sampled-based* and can be optimized with gradient descent algorithms

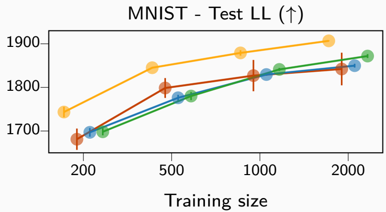
$\longrightarrow$ Not necessary to know the closed-form of either $p_\psi(\mathbf{x})$ or $\pi(\mathbf{x})$

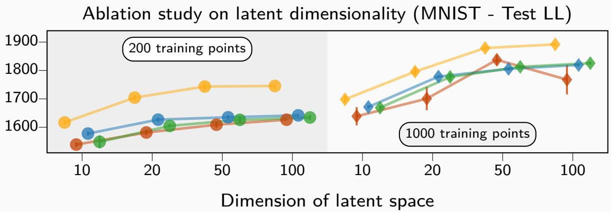
$\longrightarrow$ Only requirement is that we can draw samples from these two distributions

To sample from $p_\psi(\mathbf{x})$
$\rightarrow$ Sample $\mathbf{w}$ from prior $p_\psi(\mathbf{w})$
$\rightarrow$ Compute the output $\hat{\mathbf{x}} = f(\mathbf{x}; \mathbf{w})$
$\rightarrow$ Sample from likelihood $p(\mathbf{x} \mid \hat{\mathbf{x}})$

Experiments on MNIST







VAE ★ β-VAE ★ BAE + $\mathcal{N}(0, 1)$ Prior ★ BAE + Optim. Prior

Inductive Bias of the Optimized Priors

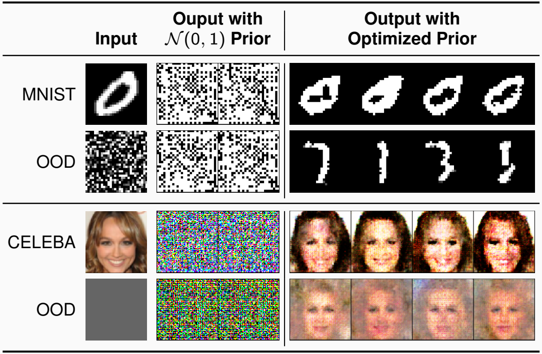


Figure: Realizations sampled from different priors given an input image. OOD stands for out-of-distribution.

Inductive Bias of the Optimized Priors

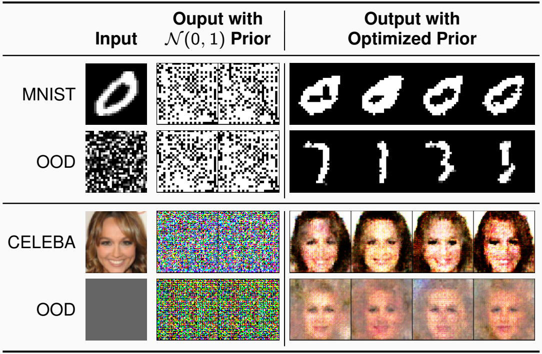


Figure: Realizations sampled from different priors given an input image. OOD stands for out-of-distribution.

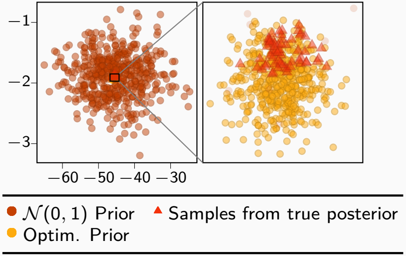


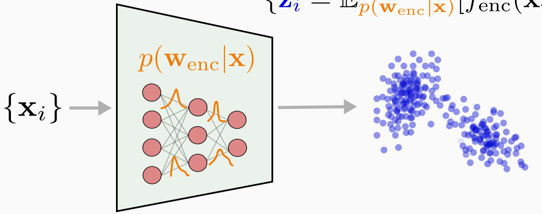
Figure: Visualization in 2D of samples from priors and posteriors of BAE parameters.

The hypothesis space of the optimized prior is reduced to regions close to the true posterior

Generative Modeling for Bayesian Autoencoders

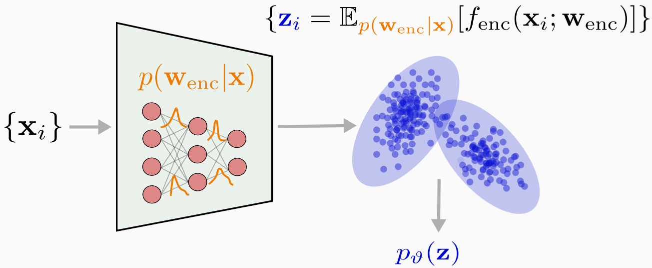
Use a Dirichlet process mixture model (Blei and Jordan, 2006) for density estimation in latent space

$$\{\mathbf{z}_i = \mathbb{E}_{p(\mathbf{w}_{\text{enc}}|\mathbf{x})}[f_{\text{enc}}(\mathbf{x}_i; \mathbf{w}_{\text{enc}})]\}$$



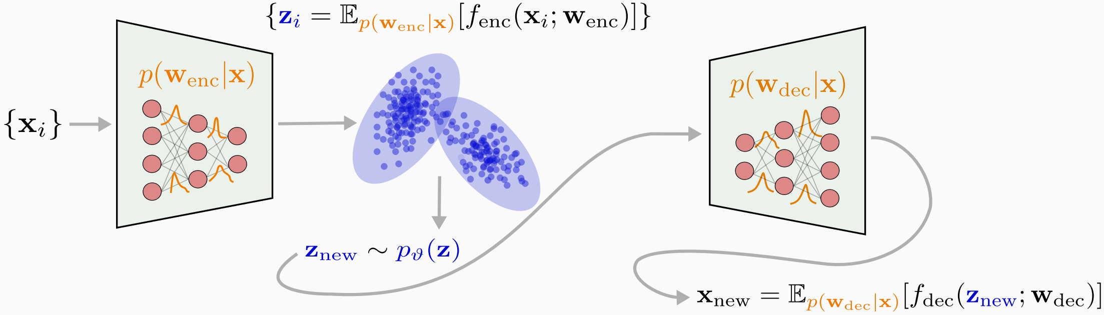
Generative Modeling for Bayesian Autoencoders

Use a Dirichlet process mixture model (Blei and Jordan, 2006) for density estimation in latent space

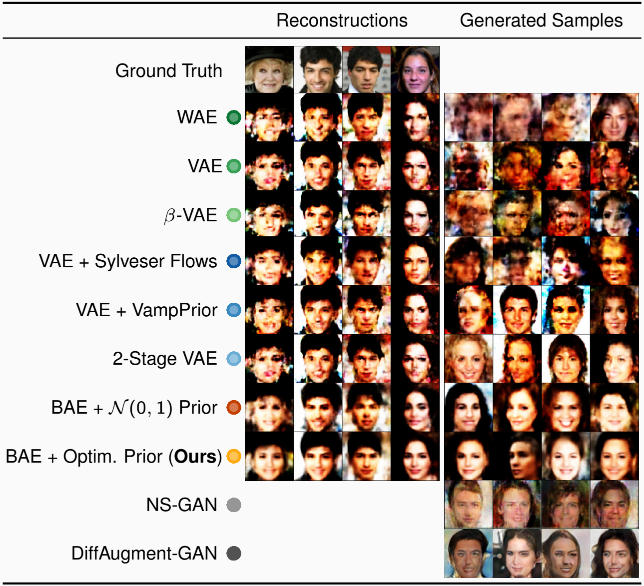


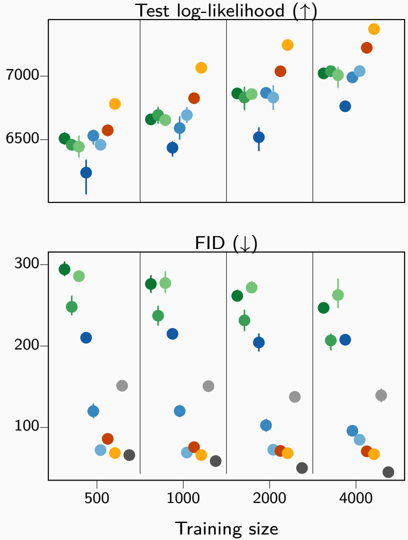
Generative Modeling for Bayesian Autoencoders

Use a Dirichlet process mixture model (Blei and Jordan, 2006) for density estimation in latent space



Experiments on CelebA Dataset





Experiments on CelebA Dataset

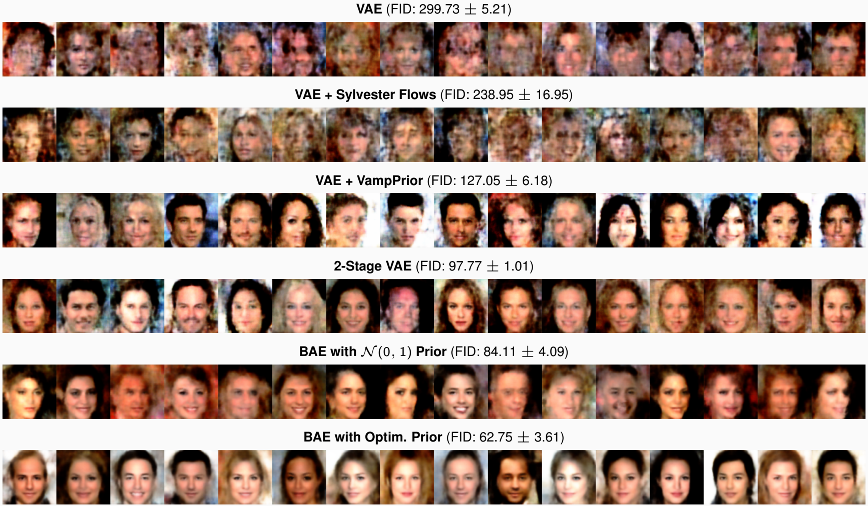


Figure: Qualitative and quantitative evaluation of generated samples with the *truncated Gaussian likelihood*. Here, we use 500 CelebA samples for inference.

Conclusions

- Revisited the Bayesian treatment of autoencoders

Conclusions

- Revisited the Bayesian treatment of autoencoders
- Proposed a novel approach of choosing priors for Bayesian autoencoders
  - Inspired by the empirical Bayes approach

Conclusions

- Revisited the Bayesian treatment of autoencoders
- Proposed a novel approach of choosing priors for Bayesian autoencoders
  - Inspired by the empirical Bayes approach
  - Showed state-of-the-art results, outperforming multiple competitive baselines

Conclusions

- Revisited the Bayesian treatment of autoencoders
- Proposed a novel approach of choosing priors for Bayesian autoencoders
  - Inspired by the empirical Bayes approach
  - Showed state-of-the-art results, outperforming multiple competitive baselines
- Ongoing work: extend to other types of data such as text, graph and heterogeneous data

Check the full paper at bit.ly/bae_prior