VAE with a VampPrior

Jakub Tomczak, Max Welling

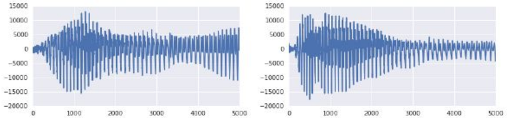
AISTATS 2018

Artificial Intelligence

“ i want to talk to you . ”
“i want to be with you . ”
“i do n’t want to be with you . ”
i do n’t want to be with you .
she did n’t want to be with him .

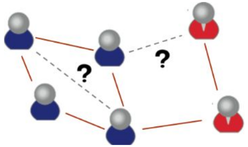
he was silent for a long moment .
he was silent for a moment .
it was quiet for a moment .
it was dark and cold .
there was a pause .
it was my turn .

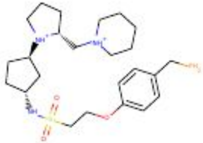
Text analysis



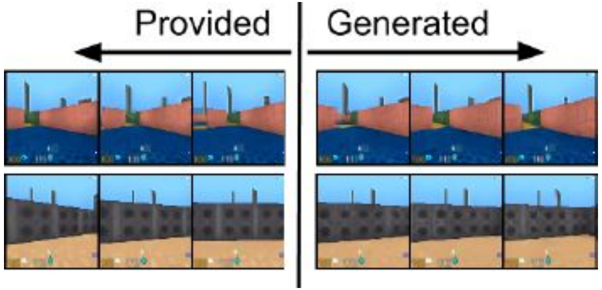
Audio analysis

Image analysis





Graph analysis



Reinforcement Learning

and more...

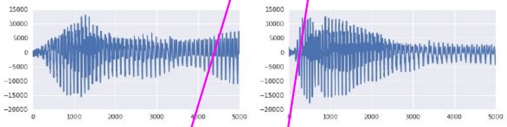
Artificial Intelligence

" i want to talk to you . "
"i want to be with you . "
"i do n't want to be with you . "
i do n't want to be with you .
she did n't want to be with him .

he was silent for a long moment .
he was silent for a moment .
it was quiet for a moment .
it was dark and cold .
there was a pause .
it was my turn .

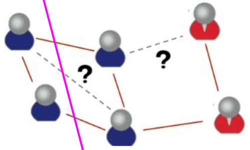
Text analysis

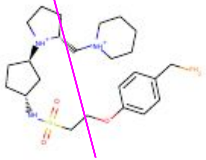
(Deep) Generative Modeling



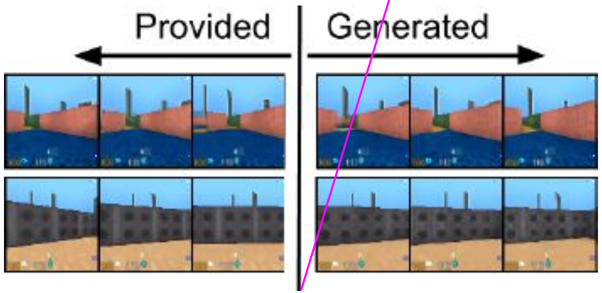
Audio analysis

Image analysis





Graph analysis



Reinforcement Learning

(Deep) generative modeling

Why?

Less labeled data

Hidden structure

Compression

Data simulation

Uncertainty

(Deep) generative modeling

Why?

- **Less labeled data**
- **Hidden structure**
- **Compression**
- **Data simulation**
- **Uncertainty**

How?

- **Fully-observed (*e.g.*, PixelCNN)**
- **Latent variable models**
  - **Implicit models (*e.g.*, GANs)**
  - **Prescribed models (*e.g.*, VAE)**

Variational inference for Latent Variable Models

$$
\begin{aligned}
\log p(\mathbf{x}) &= \log \int p_\theta(\mathbf{x}|\mathbf{z})\ p_\lambda(\mathbf{z})\ \mathrm{d}\mathbf{z} \\
&= \log \int \frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})}\ p_\theta(\mathbf{x}|\mathbf{z})\ p_\lambda(\mathbf{z})\ \mathrm{d}\mathbf{z} \\
&\geq \int q_\phi(\mathbf{z}|\mathbf{x})\ \log \frac{p_\theta(\mathbf{x}|\mathbf{z})\ p_\lambda(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\ \mathrm{d}\mathbf{z} \\
&= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathrm{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p_\lambda(\mathbf{z})]
\end{aligned}
$$

Variational inference for Latent Variable Models

$$\begin{aligned}
\log p(\mathbf{x}) &= \log \int p_\theta(\mathbf{x}|\mathbf{z})\ p_\lambda(\mathbf{z})\ \mathrm{d}\mathbf{z} \\
&= \log \int \frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})}\ p_\theta(\mathbf{x}|\mathbf{z})\ p_\lambda(\mathbf{z})\ \mathrm{d}\mathbf{z} \\
&\geq \int q_\phi(\mathbf{z}|\mathbf{x})\ \log \frac{p_\theta(\mathbf{x}|\mathbf{z})\ p_\lambda(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\ \mathrm{d}\mathbf{z} \\
&= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathrm{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p_\lambda(\mathbf{z})]
\end{aligned}$$

Variational posterior

Variational inference for Latent Variable Models

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

$$= \log \int \frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})}\ p_\theta(\mathbf{x}|\mathbf{z})\ p_\lambda(\mathbf{z})\ \mathrm{d}\mathbf{z}$$

Jensen's inequality

$$\geq \int q_\phi(\mathbf{z}|\mathbf{x}) \log \frac{p_\theta(\mathbf{x}|\mathbf{z})\ p_\lambda(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\ \mathrm{d}\mathbf{z}$$

$$= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathrm{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p_\lambda(\mathbf{z})]$$

Variational inference for Latent Variable Models

$$
\begin{aligned}
\log p(\mathbf{x}) &= \log \int p_\theta(\mathbf{x}|\mathbf{z})\; p_\lambda(\mathbf{z})\; \mathrm{d}\mathbf{z} \\
&= \log \int \frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})}\; p_\theta(\mathbf{x}|\mathbf{z})\; p_\lambda(\mathbf{z})\; \mathrm{d}\mathbf{z} \\
&\geq \int q_\phi(\mathbf{z}|\mathbf{x})\; \log \frac{p_\theta(\mathbf{x}|\mathbf{z})\; p_\lambda(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\; \mathrm{d}\mathbf{z} \\
&= \underbrace{\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]}_{\text{Reconstruction error}} - \underbrace{\mathrm{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p_\lambda(\mathbf{z})]}_{\text{Regularization}}
\end{aligned}
$$

Variational inference for Latent Variable Models

$$
\begin{aligned}
\log p(\mathbf{x}) &= \log \int p_\theta(\mathbf{x}|\mathbf{z})\ p_\lambda(\mathbf{z})\ \mathrm{d}\mathbf{z} \\
&= \log \int \frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})}\ p_\theta(\mathbf{x}|\mathbf{z})\ p_\lambda(\mathbf{z})\ \mathrm{d}\mathbf{z} \\
&\geq \int q_\phi(\mathbf{z}|\mathbf{x})\ \log \frac{p_\theta(\mathbf{x}|\mathbf{z})\ p_\lambda(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\ \mathrm{d}\mathbf{z} \\
&= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathrm{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p_\lambda(\mathbf{z})]
\end{aligned}
$$

decoder

encoder

prior

Variational inference for Latent Variable Models

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

$$= \log \int \frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})}\ p_\theta(\mathbf{x}|\mathbf{z})\ p_\lambda(\mathbf{z})\ \mathrm{d}\mathbf{z}$$

$$\geq \int q_\phi(\mathbf{z}|\mathbf{x})\ \log \frac{p_\theta(\mathbf{x}|\mathbf{z})\ p_\lambda(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\ \mathrm{d}\mathbf{z}$$

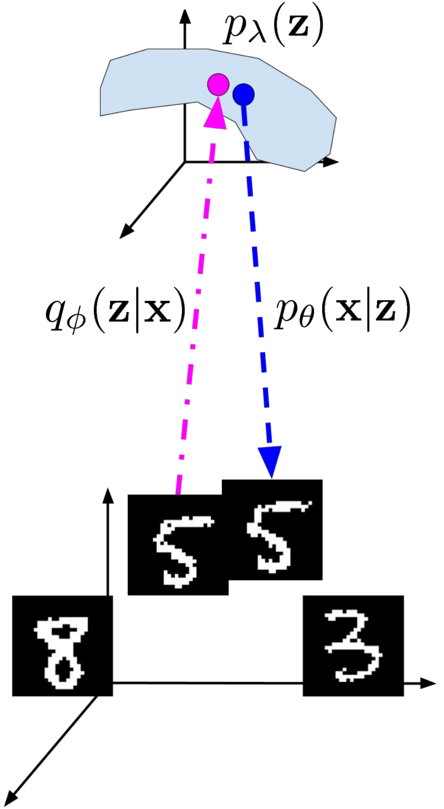
$$= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}[\log \underbrace{p_\theta(\mathbf{x}|\mathbf{z})}_{\text{decoder}}] - \mathrm{KL}[\underbrace{q_\phi(\mathbf{z}|\mathbf{x})}_{\text{encoder}}||\underbrace{p_\lambda(\mathbf{z})}_{\text{prior}}]$$

+ reparameterization trick

= Variational Auto-Encoder

Variational Auto-Encoder

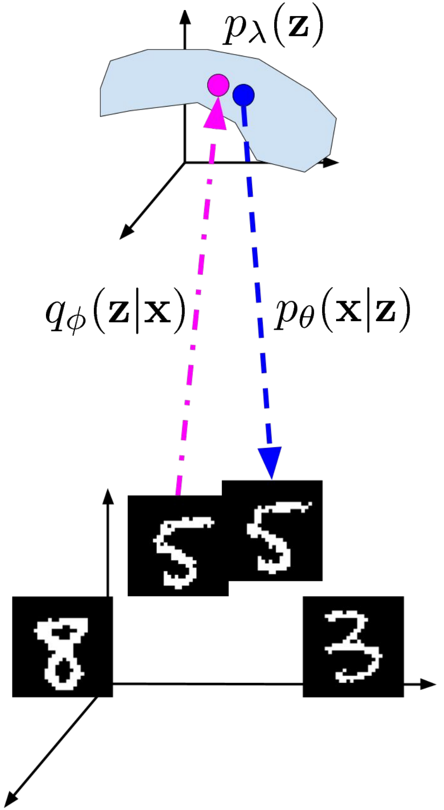
$$q_\phi(\mathbf{z}|\mathbf{x}) \propto p_\theta(\mathbf{x}|\mathbf{z})\ p_\lambda(\mathbf{z})$$



Variational Auto-Encoder

$$q_\phi(\mathbf{z}|\mathbf{x}) \propto p_\theta(\mathbf{x}|\mathbf{z})\, p_\lambda(\mathbf{z})$$

Fully-connected
ConvNets
PixelCNN

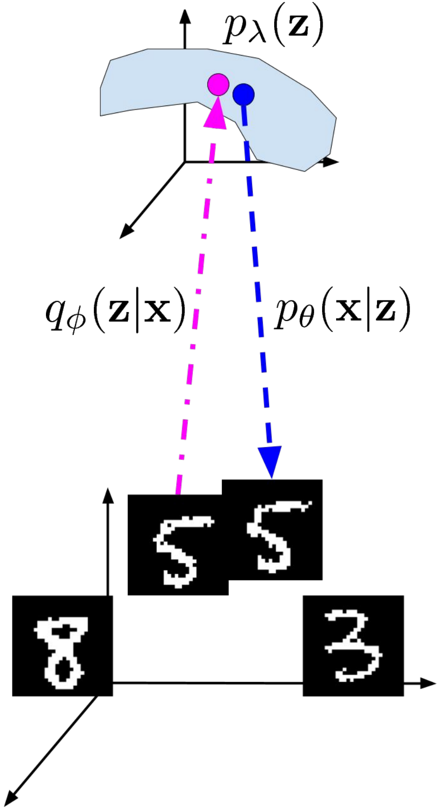


Variational Auto-Encoder

$$q_\phi(\mathbf{z}|\mathbf{x}) \propto p_\theta(\mathbf{x}|\mathbf{z})\, p_\lambda(\mathbf{z})$$

Normalizing flows
Volume-preserving flows

Fully-connected
ConvNets
PixelCNN



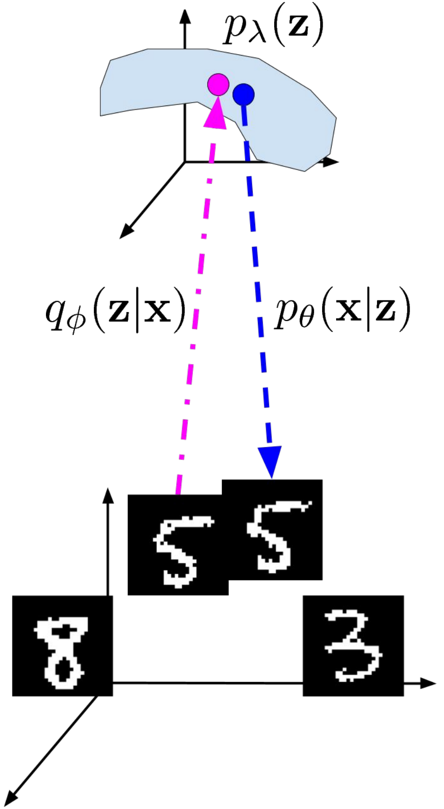
Variational Auto-Encoder

$$q_\phi(\mathbf{z}|\mathbf{x}) \propto p_\theta(\mathbf{x}|\mathbf{z})\, p_\lambda(\mathbf{z})$$

Normalizing flows
Volume-preserving flows

Fully-connected
ConvNets
PixelCNN

Autoregressive Prior
Objective Prior
Stick-Breaking Prior
VampPrior



Variational Auto-Encoder

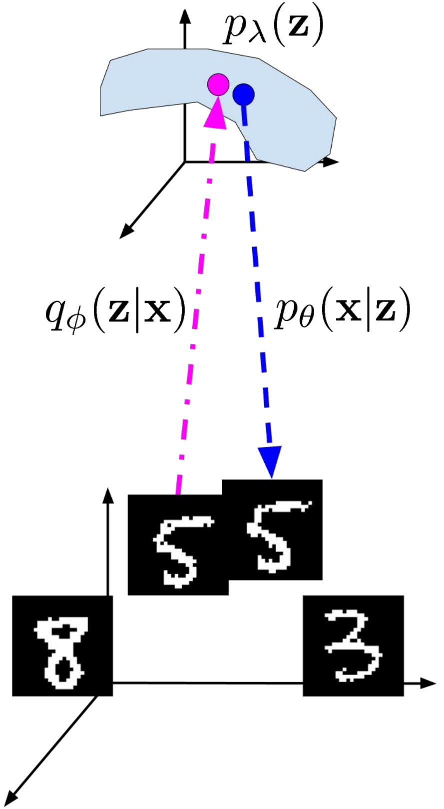
$$q_\phi(\mathbf{z}|\mathbf{x}) \propto p_\theta(\mathbf{x}|\mathbf{z})\, p_\lambda(\mathbf{z})$$

Normalizing flows
Volume-preserving flows

Fully-connected
ConvNets
PixelCNN

Autoregressive Prior
Objective Prior
Stick-Breaking Prior
VampPrior

Importance Weighted AE
Renyi Divergence
Stein Divergence



Variational Auto-Encoder

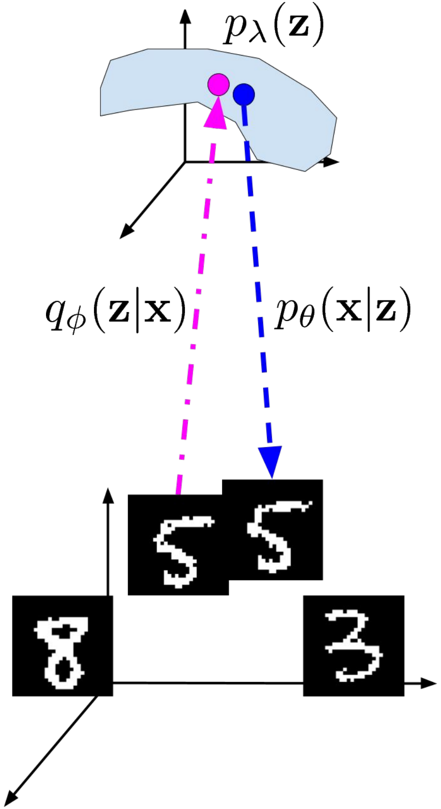
$$q_\phi(\mathbf{z}|\mathbf{x}) \propto p_\theta(\mathbf{x}|\mathbf{z})\, p_\lambda(\mathbf{z})$$

Normalizing flows
Volume-preserving flows

Fully-connected
ConvNets
PixelCNN

Autoregressive Prior
Objective Prior
Stick-Breaking Prior
VampPrior

Importance Weighted AE
Renyi Divergence
Stein Divergence



New Prior

- Let's re-write the ELBO:

$$
\begin{aligned}
\mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}\big[\ln p(\mathbf{x})\big] \geq & \mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}\big[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p_\theta(\mathbf{x}|\mathbf{z})]\big] + \\
& + \mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}\big[\mathbb{H}[q_\phi(\mathbf{z}|\mathbf{x})]\big] + \\
& - \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[-\ln p_\lambda(\mathbf{z})]
\end{aligned}
$$

New Prior

- Let's re-write the ELBO:

$$\mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}\left[\ln p(\mathbf{x})\right] \geq \mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}\left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p_\theta(\mathbf{x}|\mathbf{z})]\right] + \\ + \mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}\left[\mathbb{H}[q_\phi(\mathbf{z}|\mathbf{x})]\right] + \\ - \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[-\ln p_\lambda(\mathbf{z})]$$

Empirical distribution

New Prior

- Let's re-write the ELBO:

Reconstruction error

$$
\begin{aligned}
\mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}\big[\ln p(\mathbf{x})\big] \geq\ & \mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}\big[\mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})}[\ln p_{\theta}(\mathbf{x}|\mathbf{z})]\big] + \\
& + \mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}\big[\mathbb{H}[q_{\phi}(\mathbf{z}|\mathbf{x})]\big] + \\
& - \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[-\ln p_{\lambda}(\mathbf{z})]
\end{aligned}
$$

New Prior

- Let's re-write the ELBO:

$$\begin{aligned}\mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}\left[\ln p(\mathbf{x})\right] \geq & \mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}\left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p_\theta(\mathbf{x}|\mathbf{z})]\right] + \\ & + \mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}\left[\mathbb{H}[q_\phi(\mathbf{z}|\mathbf{x})]\right] + \\ & - \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[-\ln p_\lambda(\mathbf{z})]\end{aligned}$$

Encoder's entropy

New Prior

- Let's re-write the ELBO:

$$
\begin{aligned}
\mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}\left[\ln p(\mathbf{x})\right] \geq & \mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}\left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p_\theta(\mathbf{x}|\mathbf{z})]\right] + \\
& + \mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}\left[\mathbb{H}[q_\phi(\mathbf{z}|\mathbf{x})]\right] + \\
& - \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[-\ln p_\lambda(\mathbf{z})]
\end{aligned}
$$

Cross Entropy

New Prior

- Let's re-write the ELBO:

$$
\begin{aligned}
\mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}\big[\ln p(\mathbf{x})\big] \geq & \,\mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}\Big[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p_\theta(\mathbf{x}|\mathbf{z})]\Big] + \\
& + \mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}\Big[\mathbb{H}[q_\phi(\mathbf{z}|\mathbf{x})]\Big] + \\
& - \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}\big[-\ln p_\lambda(\mathbf{z})\big]
\end{aligned}
$$

Aggregated posterior

$$
\begin{aligned}
q(\mathbf{z}) &= \mathbb{E}_{q(\mathbf{x})}[q_\phi(\mathbf{z}|\mathbf{x})] \\
&= \frac{1}{N}\sum_{n=1}^{N} q_\phi(\mathbf{z}|\mathbf{x}_n)
\end{aligned}
$$

New Prior

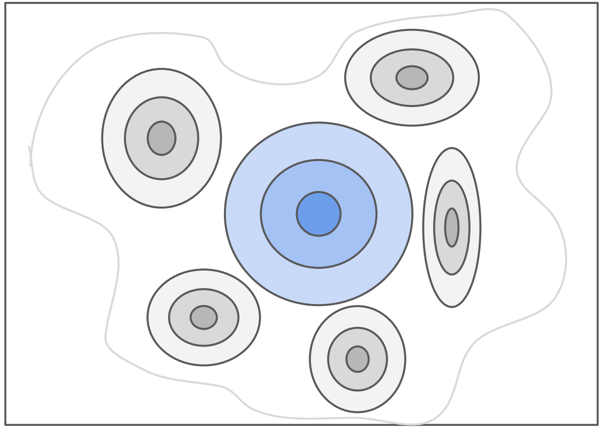
- Let's re-write the ELBO:

$$
\textbf{max.}\ \mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}\left[\ln p(\mathbf{x})\right] \geq \mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}\left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p_\theta(\mathbf{x}|\mathbf{z})]\right] + \\
+ \mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}\left[\mathbb{H}[q_\phi(\mathbf{z}|\mathbf{x})]\right] + \\
- \boxed{\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[-\ln p_\lambda(\mathbf{z})]}
$$

$$
\textbf{min.}\ \boxed{\mathbb{H}[q(\mathbf{z})] + KL[q(\mathbf{z})||p_\lambda(\mathbf{z})]}
$$

New Prior

$$\text{min. } \mathbb{H}[q(\mathbf{z})] + KL[q(\mathbf{z})||p_{\lambda}(\mathbf{z})]$$

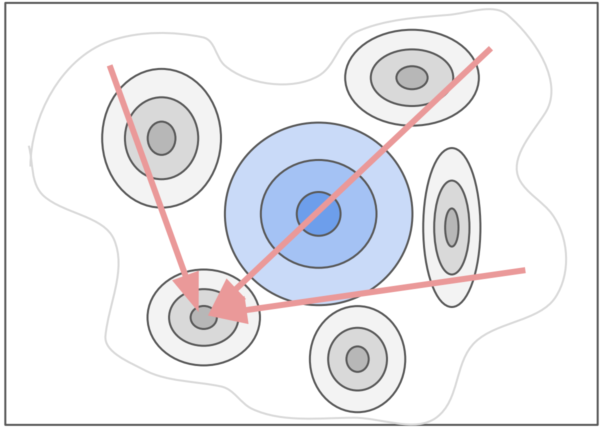


Prior

Aggregated posterior

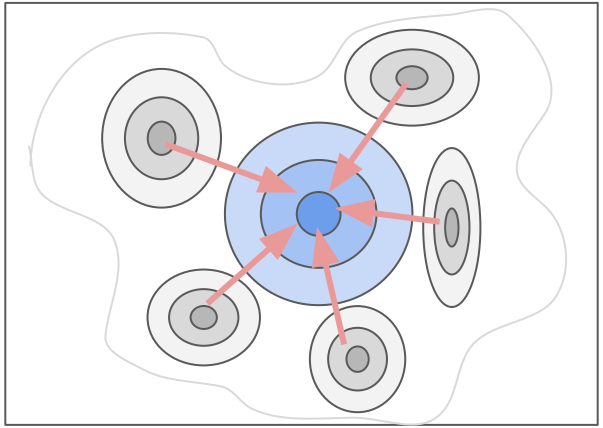
New Prior

$$\text{min. } \mathbb{H}[q(\mathbf{z})] + KL[q(\mathbf{z})||p_\lambda(\mathbf{z})]$$



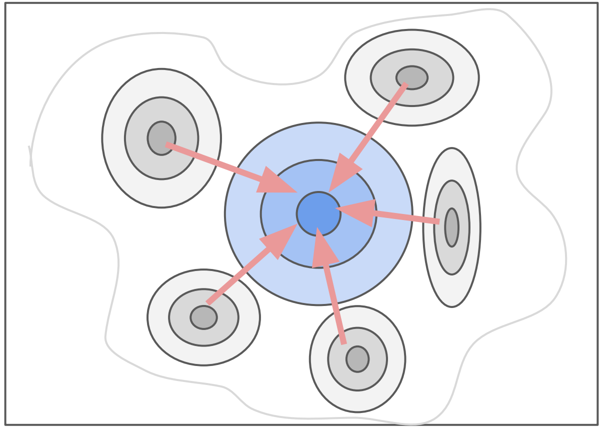
New Prior

min. $\mathbb{H}[q(\mathbf{z})] + KL[q(\mathbf{z})||p_\lambda(\mathbf{z})]$



New Prior

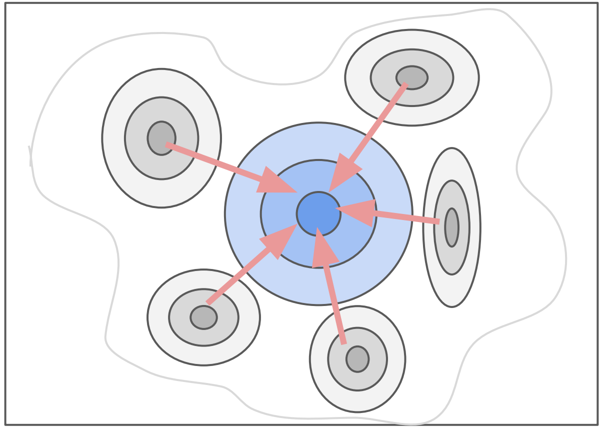
$$\text{min. } \mathbb{H}[q(\mathbf{z})] + KL[q(\mathbf{z})||p_{\lambda}(\mathbf{z})]$$



Standard prior is too strong and overregularizes the encoder.

New Prior

$$\text{min. } \mathbb{H}[q(\mathbf{z})] + KL[q(\mathbf{z})||p_\lambda(\mathbf{z})]$$



Standard prior is too strong and overregularizes the encoder.

What is the “optimal” prior?

New Prior (**Va**riational **M**ixture of **P**osteriors **Prior**)

- We look for **the optimal prior** using the Lagrange function:

$$\max_{p_\lambda(\mathbf{z})} -\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[-\ln p_\lambda(\mathbf{z})] + \beta\Big(\int p_\lambda(\mathbf{z})\mathrm{d}\mathbf{z} - 1\Big)$$

- The solution is simply **the aggregated posterior**.

- We approximate it using *K* **pseudo-inputs** instead of *N* observations:

$$p_\lambda(\mathbf{z}) = \frac{1}{K}\sum_{k=1}^{K} q_\phi(\mathbf{z}|\mathbf{u}_k)$$

New Prior (**Va**riational **M**ixture of **P**osteriors **Prior**)

- We look for **the optimal prior** using the Lagrange function:

$$\max_{p_\lambda(\mathbf{z})} -\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[-\ln p_\lambda(\mathbf{z})] + \beta\left(\int p_\lambda(\mathbf{z})\mathrm{d}\mathbf{z} - 1\right)$$

- The solution is simply **the aggregated posterior**. $\boxed{p^*_\lambda(\mathbf{z}) = \frac{1}{N}\sum_{n=1}^{N} q_\phi(\mathbf{z}|\mathbf{x}_n)}$

- We approximate it using K **pseudo-inputs** instead of N observations:

$$p_\lambda(\mathbf{z}) = \frac{1}{K}\sum_{k=1}^{K} q_\phi(\mathbf{z}|\mathbf{u}_k)$$

New Prior (Variational Mixture of Posteriors Prior)

- We look for **the optimal prior** using the Lagrange function:

$$\max_{p_\lambda(\mathbf{z})} -\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[-\ln p_\lambda(\mathbf{z})] + \beta\left(\int p_\lambda(\mathbf{z})\mathrm{d}\mathbf{z} - 1\right)$$

- The solution is simply **the aggregated posterior**.

$$\boxed{p^*_\lambda(\mathbf{z}) = \frac{1}{N}\sum_{n=1}^{N} q_\phi(\mathbf{z}|\mathbf{x}_n)}$$

- We approximate it using K **pseudo-inputs** instead of N observations:

infeasible

$$p_\lambda(\mathbf{z}) = \frac{1}{K}\sum_{k=1}^{K} q_\phi(\mathbf{z}|\mathbf{u}_k)$$

New Prior (**Va**riational **M**ixture of **P**osteriors **Prior**)

- We look for **the optimal prior** using the Lagrange function:

$$\max_{p_\lambda(\mathbf{z})} -\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[-\ln p_\lambda(\mathbf{z})] + \beta\Big(\int p_\lambda(\mathbf{z})\mathrm{d}\mathbf{z} - 1\Big)$$

- The solution is simply **the aggregated posterior**.

- We approximate it using *K* **pseudo-inputs** instead of *N* observations:

$$p_\lambda(\mathbf{z}) = \frac{1}{K}\sum_{k=1}^{K} q_\phi(\mathbf{z}|\mathbf{u}_k)$$

New Prior (**Va**riational **M**ixture of **P**osteriors **Prior**)

- We look for **the optimal prior** using the Lagrange function:

$$\max_{p_\lambda(\mathbf{z})} -\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[-\ln p_\lambda(\mathbf{z})] + \beta\Big(\int p_\lambda(\mathbf{z})\mathrm{d}\mathbf{z} - 1\Big)$$

- The solution is simply **the aggregated posterior**.

- We approximate it using K **pseudo-inputs** instead of N observations:

$$p_\lambda(\mathbf{z}) = \frac{1}{K}\sum_{k=1}^{K} q_\phi(\mathbf{z}|\mathbf{u}_k)$$

they are trained from scratch by SGD

New Prior (**Va**riational **M**ixture of **P**osteriors **Prior**)

- Is the VampPrior different than the Mixture of Gaussians? $p_{\lambda}(\mathbf{z}) = \frac{1}{K}\sum_{k=1}^{K}\mathcal{N}(\mu_k, \mathrm{diag}(\sigma_k^2))$
- VampPrior: the prior and the posterior must "**cooperate**" during training.

VampPrior

$$\frac{1}{K}\sum_{k=1}^{K}\Big\{\Big(\frac{q_{\phi}(\mathbf{z}_{\phi}^{(l)}|\mathbf{x})\ \frac{\partial}{\partial\phi_i}q_{\phi}(\mathbf{z}_{\phi}^{(l)}|\mathbf{u}_k) - q_{\phi}(\mathbf{z}_{\phi}^{(l)}|\mathbf{u}_k)\ \frac{\partial}{\partial\phi_i}q_{\phi}(\mathbf{z}_{\phi}^{(l)}|\mathbf{x})}{\frac{1}{K}\sum_{k=1}^{K}q_{\phi}(\mathbf{z}_{\phi}^{(l)}|\mathbf{u}_k)\ q_{\phi}(\mathbf{z}_{\phi}^{(l)}|\mathbf{x})}\Big)+$$
$$+\Big(\frac{\big(q_{\phi}(\mathbf{z}_{\phi}^{(l)}|\mathbf{x})\ \frac{\partial}{\partial\mathbf{z}_{\phi}}q_{\phi}(\mathbf{z}_{\phi}^{(l)}|\mathbf{u}_k) - q_{\phi}(\mathbf{z}_{\phi}^{(l)}|\mathbf{u}_k)\ \frac{\partial}{\partial\mathbf{z}_{\phi}}q_{\phi}(\mathbf{z}_{\phi}^{(l)}|\mathbf{x})\big)\ \frac{\partial}{\partial\phi_i}\mathbf{z}_{\phi}^{(l)}}{\frac{1}{K}\sum_{k=1}^{K}q_{\phi}(\mathbf{z}_{\phi}^{(l)}|\mathbf{u}_k)\ q_{\phi}(\mathbf{z}_{\phi}^{(l)}|\mathbf{x})}\Big)\Big\}$$

standard/ MoG

$$\frac{1}{p_{\lambda}(\mathbf{z}_{\phi}^{(l)})\ q_{\phi}(\mathbf{z}_{\phi}^{(l)}|\mathbf{x})}\Big(q_{\phi}(\mathbf{z}_{\phi}^{(l)}|\mathbf{x})\frac{\partial}{\partial\mathbf{z}_{\phi}}p_{\lambda}(\mathbf{z}_{\phi}^{(l)}) - p_{\lambda}(\mathbf{z}_{\phi}^{(l)})\frac{\partial}{\partial\mathbf{z}_{\phi}}q_{\phi}(\mathbf{z}_{\phi}^{(l)}|\mathbf{x})\Big)\frac{\partial}{\partial\phi_i}\mathbf{z}_{\phi}^{(l)}$$

New Prior (**Va**riational **M**ixture of **P**osteriors **Prior**)

- Is the VampPrior different than the Mixture of Gaussians? $p_\lambda(\mathbf{z}) = \frac{1}{K}\sum_{k=1}^{K}\mathcal{N}(\mu_k, \mathrm{diag}(\sigma_k^2))$
- VampPrior: the prior and the posterior must “**cooperate**” during training.

VampPrior

$$\frac{1}{K}\sum_{k=1}^{K}\Bigg\{ \Bigg(\frac{q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{x})\ \frac{\partial}{\partial\phi_i}q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{u}_k) - q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{u}_k)\ \frac{\partial}{\partial\phi_i}q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{x})}{\frac{1}{K}\sum_{k=1}^{K}q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{u}_k)\ q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{x})}\Bigg)+$$

$$+\Bigg(\frac{\big(q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{x})\ \frac{\partial}{\partial\mathbf{z}_\phi}q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{u}_k) - q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{u}_k)\ \frac{\partial}{\partial\mathbf{z}_\phi}q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{x})\big)\ \frac{\partial}{\partial\phi_i}\mathbf{z}_\phi^{(l)}}{\frac{1}{K}\sum_{k=1}^{K}q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{u}_k)\ q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{x})}\Bigg)\Bigg\}$$

standard/ MoG

$$\frac{1}{p_\lambda(\mathbf{z}_\phi^{(l)})\ q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{x})}\Big(q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{x})\frac{\partial}{\partial\mathbf{z}_\phi}p_\lambda(\mathbf{z}_\phi^{(l)}) - p_\lambda(\mathbf{z}_\phi^{(l)})\frac{\partial}{\partial\mathbf{z}_\phi}q_\phi(\mathbf{z}_\phi^{(l)}|\mathbf{x})\Big)\frac{\partial}{\partial\phi_i}\mathbf{z}_\phi^{(l)}$$

New Prior (**Va**riational **M**ixture of **P**osteriors **Prior**)

- VampPrior is closely related to the **Empirical Bayes**.
  - We propose a new approach that learns parameters of the prior and combines the variational inference with the EB approach.

- VampPrior is closely related to the **Information Bottleneck**.
  - The aggregated posterior naturally plays the role of the prior.
  - The VampPrior brings the VAE and the IB formulations together.

New Prior (Variational Mixture of Posteriors Prior)

- Is it advantageous to take *K*=*N*?
  - Not necessarily...
  - Let's re-write (one more time) the ELBO:

$$\text{maximize } \mathbb{E}_{q(\mathbf{x})}[\ln p(\mathbf{x}|\mathbf{z})] - \mathrm{I}(\mathbf{x};\mathbf{z}) - \mathrm{KL}[q(\mathbf{z})||p(\mathbf{z})]$$

  - We will see this effect also during experiments.

New Prior (**Va**riational **M**ixture of **P**osteriors **Prior**)

- Is it advantageous to take *K*=*N*?
  - Not necessarily...
  - Let's re-write (one more time) the ELBO:

$$\text{maximize } \mathbb{E}_{q(\mathbf{x})}[\ln p(\mathbf{x}|\mathbf{z})] - \mathrm{I}(\mathbf{x};\mathbf{z}) - \underbrace{\mathrm{KL}[q(\mathbf{z})||p(\mathbf{z})]}_{\approx 0 \text{ for } p(\mathbf{z}) \approx q(\mathbf{z})}$$

  - We will see this effect also during experiments.

New Prior (**Va**riational **M**ixture of **P**osteriors **Prior**)

- Is it advantageous to take *K*=*N*?
  - Not necessarily…
  - Let's re-write (one more time) the ELBO:

$$\text{maximize } \mathbb{E}_{q(\mathbf{x})}[\ln p(\mathbf{x}|\mathbf{z})] - \mathrm{I}(\mathbf{x};\mathbf{z}) - \underbrace{\mathrm{KL}[q(\mathbf{z})||p(\mathbf{z})]}_{\approx 0 \text{ for } p(\mathbf{z}) \approx q(\mathbf{z})}$$

⇒ x independent of z !

  - We will see this effect also during experiments.

Hierarchical VampPrior VAE

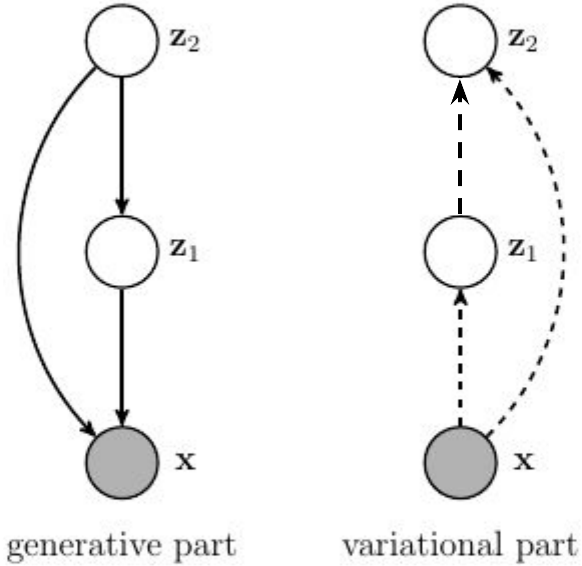
Typical issue in hierarchical VAE: **inactive stochastic units**

$$p(\mathbf{z}_2) = \frac{1}{K}\sum_{k=1}^{K} q_\psi(\mathbf{z}_2|\mathbf{u}_k),$$

$$p_\lambda(\mathbf{z}_1|\mathbf{z}_2) = \mathcal{N}\big(\mathbf{z}_1|\mu_\lambda(\mathbf{z}_2), \mathrm{diag}(\sigma^2_\lambda(\mathbf{z}_2))\big),$$

$$q_\phi(\mathbf{z}_1|\mathbf{x},\mathbf{z}_2) = \mathcal{N}\big(\mathbf{z}_1|\mu_\phi(\mathbf{x},\mathbf{z}_2), \mathrm{diag}(\sigma^2_\phi(\mathbf{x},\mathbf{z}_2))\big),$$

$$q_\psi(\mathbf{z}_2|\mathbf{x}) = \mathcal{N}\big(\mathbf{z}_2|\mu_\psi(\mathbf{x}), \mathrm{diag}(\sigma^2_\psi(\mathbf{x}))\big)$$



Hierarchical VampPrior VAE

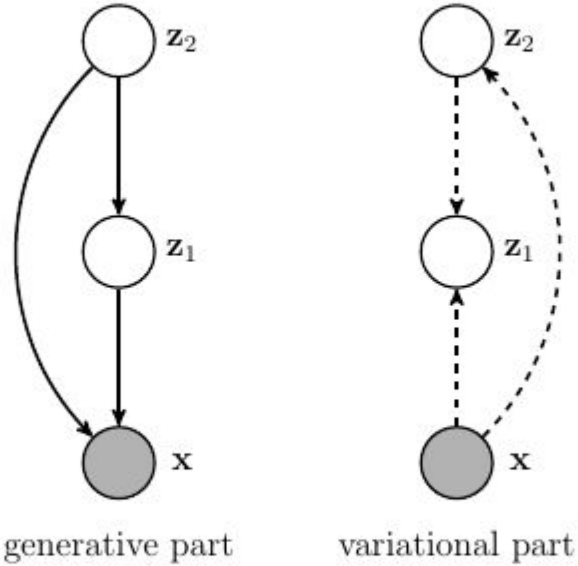
Typical issue in hierarchical VAE: **inactive stochastic units**

$$p(\mathbf{z}_2) = \frac{1}{K} \sum_{k=1}^{K} q_\psi(\mathbf{z}_2 | \mathbf{u}_k),$$

$$p_\lambda(\mathbf{z}_1|\mathbf{z}_2) = \mathcal{N}\big(\mathbf{z}_1|\mu_\lambda(\mathbf{z}_2), \mathrm{diag}(\sigma^2_\lambda(\mathbf{z}_2))\big),$$

$$q_\phi(\mathbf{z}_1|\mathbf{x}, \mathbf{z}_2) = \mathcal{N}\big(\mathbf{z}_1|\mu_\phi(\mathbf{x}, \mathbf{z}_2), \mathrm{diag}(\sigma^2_\phi(\mathbf{x}, \mathbf{z}_2))\big),$$

$$q_\psi(\mathbf{z}_2|\mathbf{x}) = \mathcal{N}\big(\mathbf{z}_2|\mu_\psi(\mathbf{x}), \mathrm{diag}(\sigma^2_\psi(\mathbf{x}))\big)$$



Hierarchical VampPrior VAE

Typical issue in hierarchical VAE: **inactive stochastic units**

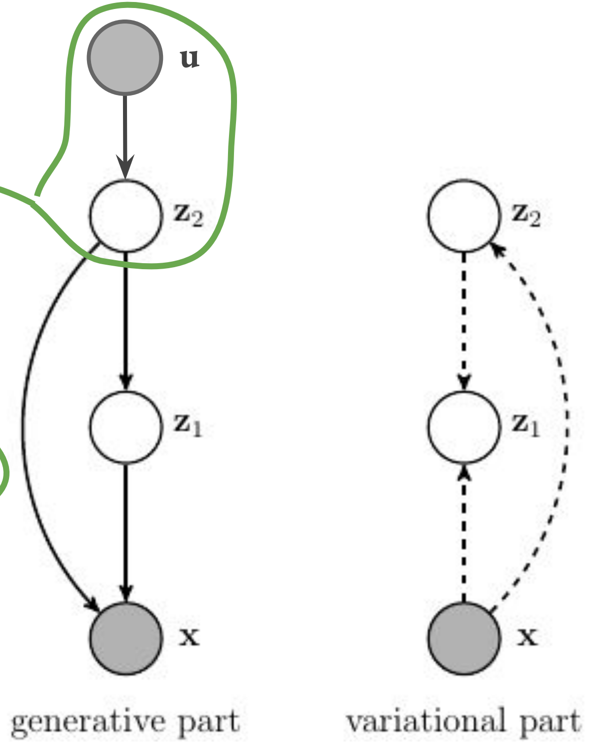
$$p(\mathbf{z}_2) = \frac{1}{K}\sum_{k=1}^{K} q_\psi(\mathbf{z}_2|\mathbf{u}_k),$$

$$p_\lambda(\mathbf{z}_1|\mathbf{z}_2) = \mathcal{N}\big(\mathbf{z}_1|\mu_\lambda(\mathbf{z}_2), \mathrm{diag}(\sigma^2_\lambda(\mathbf{z}_2))\big),$$

$$q_\phi(\mathbf{z}_1|\mathbf{x},\mathbf{z}_2) = \mathcal{N}\big(\mathbf{z}_1|\mu_\phi(\mathbf{x},\mathbf{z}_2), \mathrm{diag}(\sigma^2_\phi(\mathbf{x},\mathbf{z}_2))\big),$$

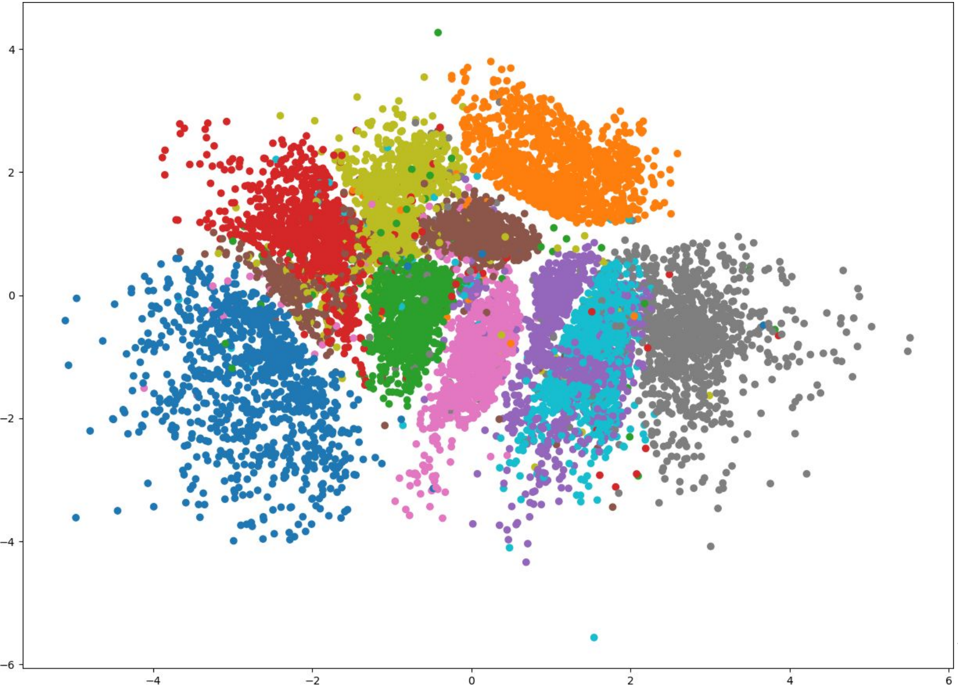
$$q_\psi(\mathbf{z}_2|\mathbf{x}) = \mathcal{N}\big(\mathbf{z}_2|\mu_\psi(\mathbf{x}), \mathrm{diag}(\sigma^2_\psi(\mathbf{x}))\big)$$

It counteracts inactive stochastic hidden units problem!

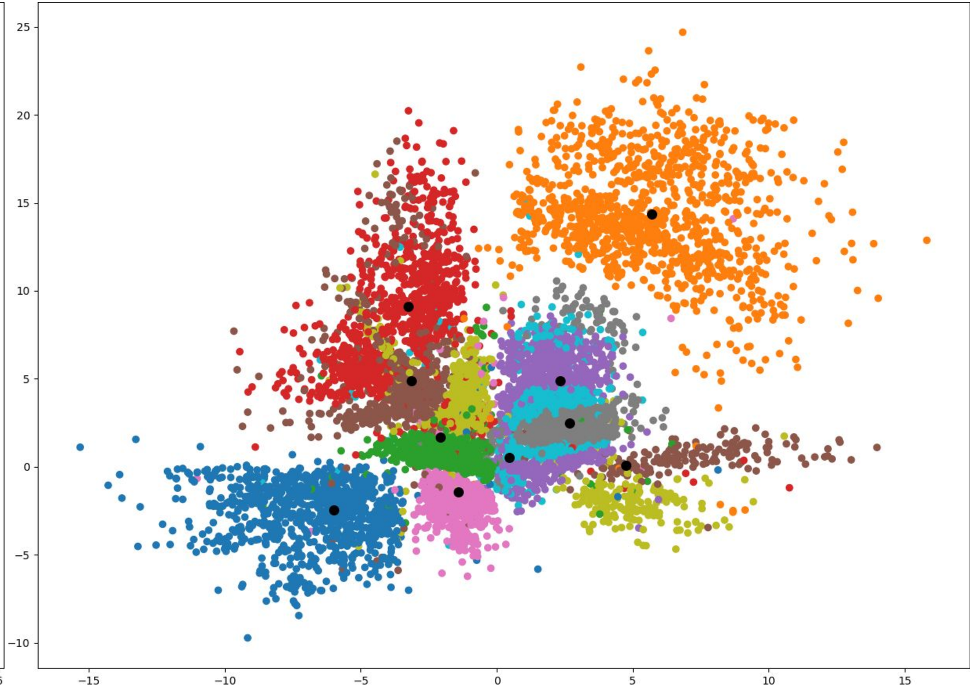


Toy problem (MNIST): VAE with dim(z)=2

Latent space representation + psedoinputs (black dots)



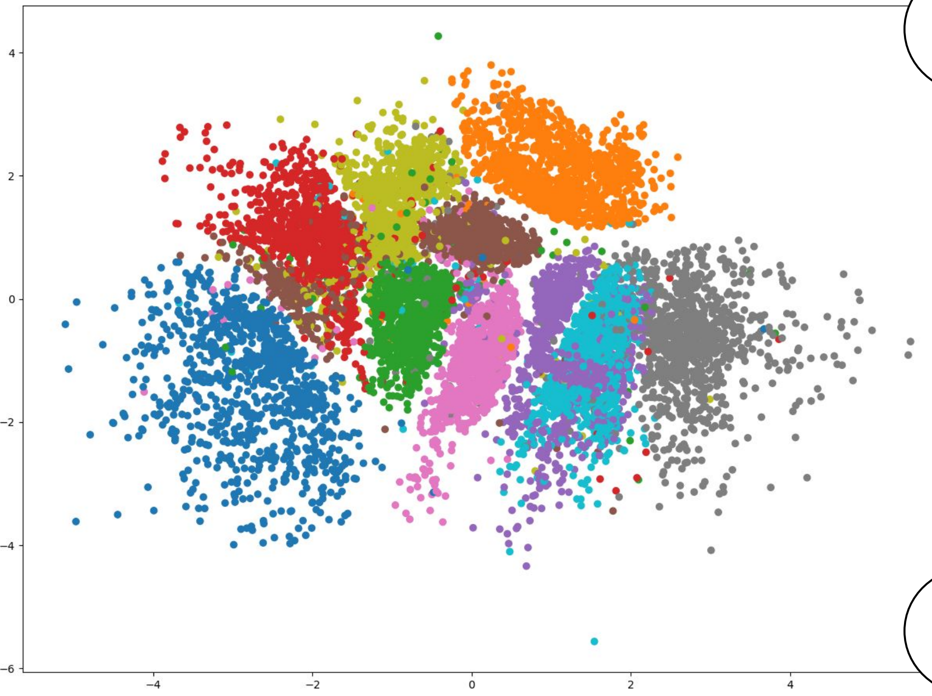
standard



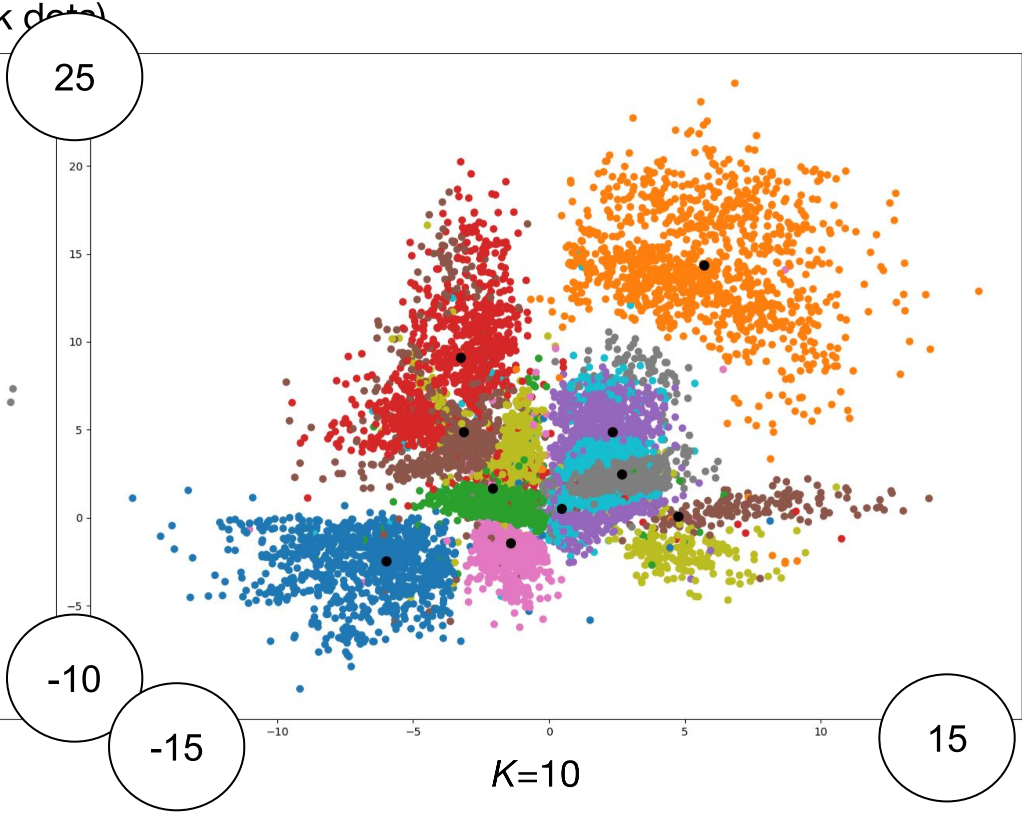
K=10

Toy problem (MNIST): VAE with dim(z)=2

Latent space representation + psedoinputs (black dots)



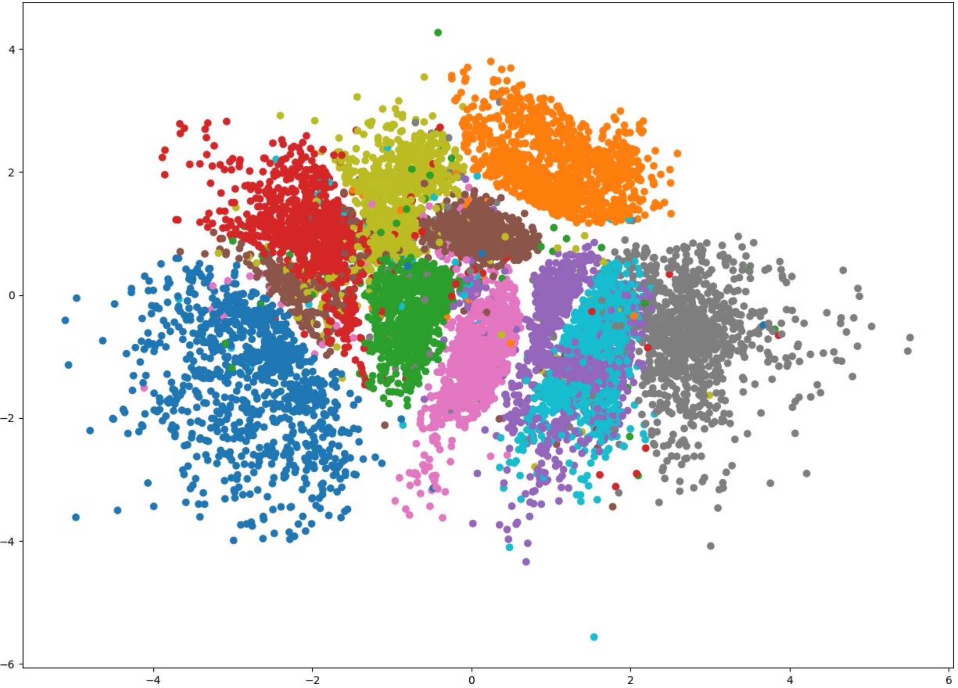
standard



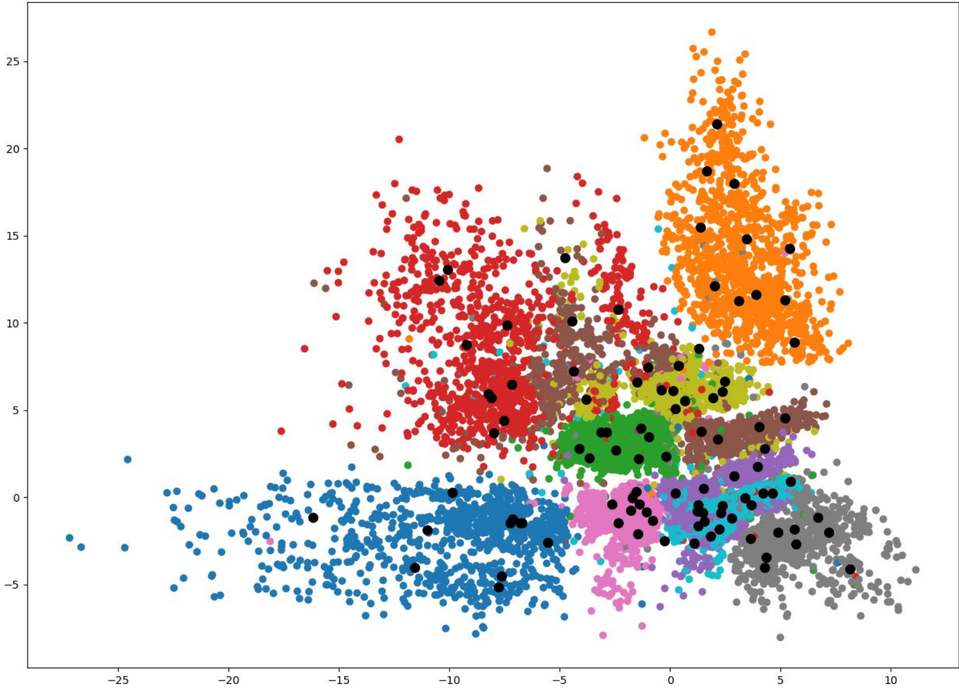
K=10

Toy problem (MNIST): VAE with dim(z)=2

Latent space representation + psedoinputs (black dots)



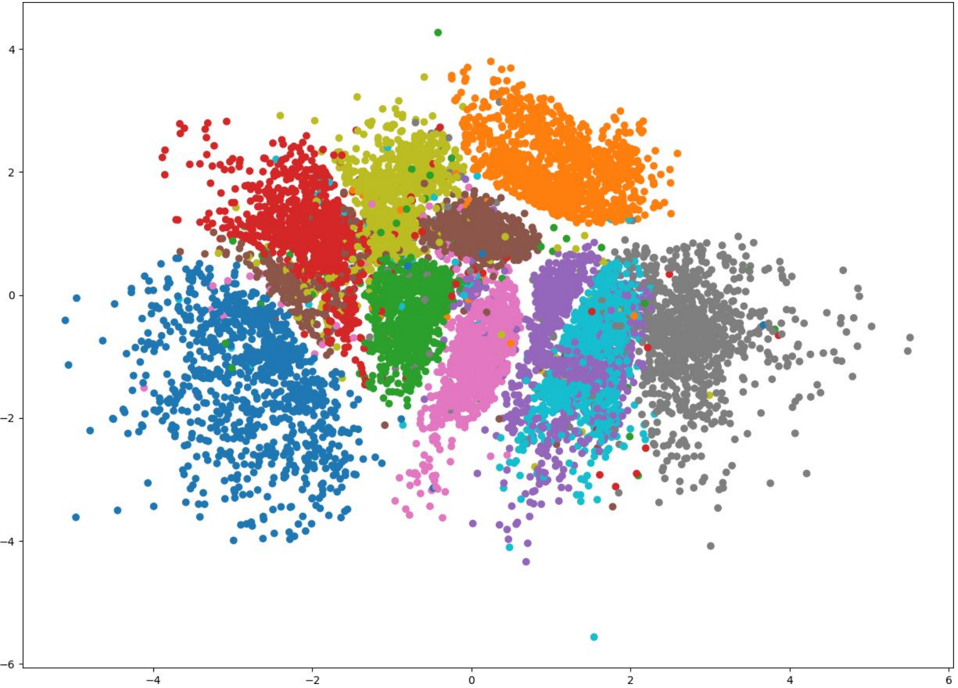
standard



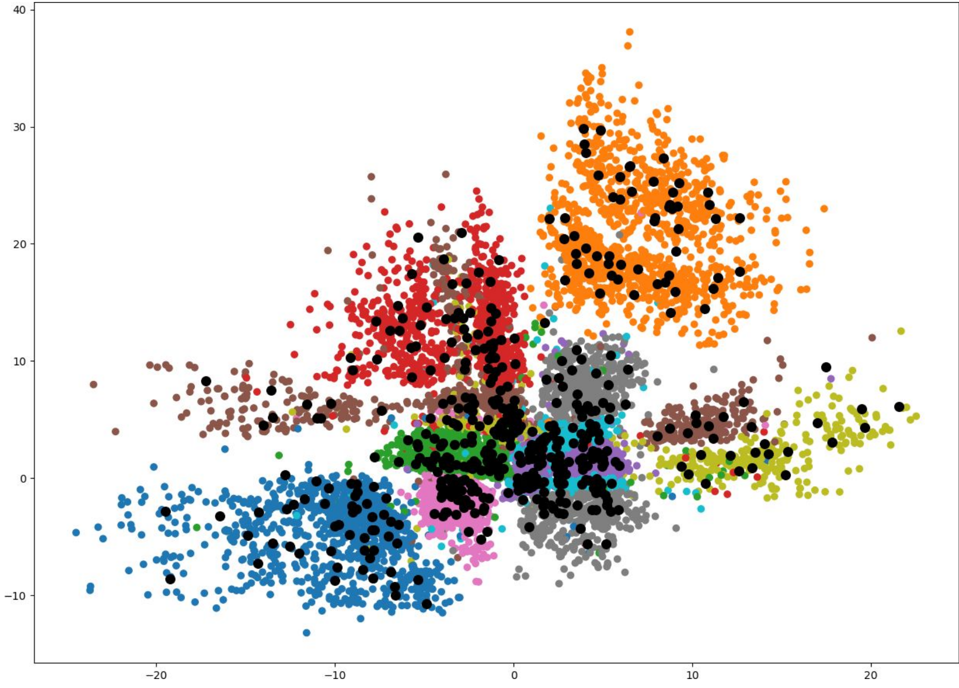
K=100

Toy problem (MNIST): VAE with dim(z)=2

Latent space representation + psedoinputs (black dots)



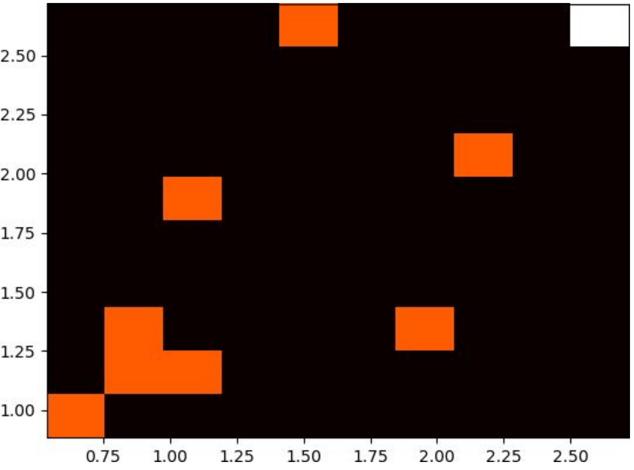
standard



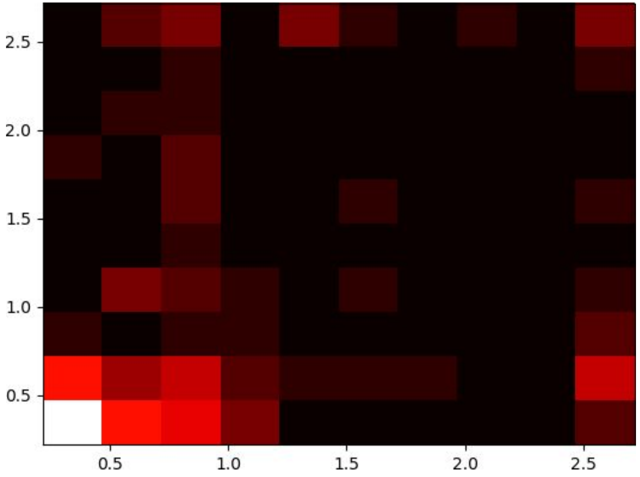
K=500

Toy problem (MNIST): VAE with dim(z)=2

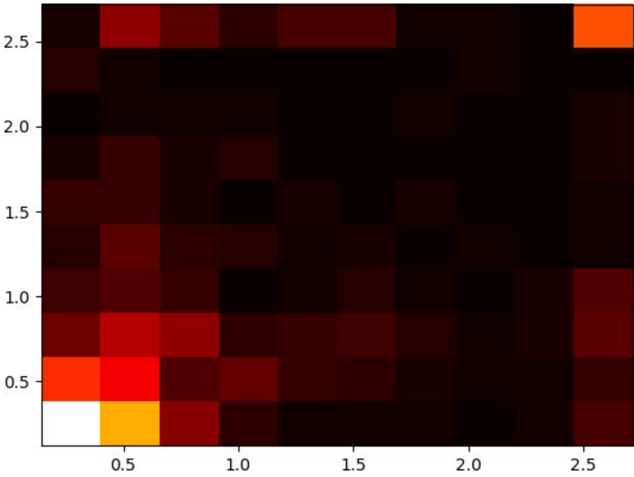
Standard deviations of the encoder for given pseudoinputs:



K=10



K=100



K=500

Experiments

DATASET	VAE ($L=1$) standard	VAE ($L=1$) VampPrior	HVAE ($L=2$) standard	HVAE ($L=2$) VampPrior	CONVHVAE ($L=2$) standard	CONVHVAE ($L=2$) VampPrior	PIXELHVAE ($L=2$) standard	PIXELHVAE ($L=2$) VampPrior
staticMNIST	-88.56	$\mathbf{-85.57}$	-86.05	$\mathbf{-83.19}$	-82.41	$\mathbf{-81.09}$	-80.58	$\mathbf{-79.78}$
dynamicMNIST	-84.50	$\mathbf{-82.38}$	-82.42	$\mathbf{-81.24}$	-80.40	$\mathbf{-79.75}$	-79.70	$\mathbf{-78.45}$
Omniglot	-108.50	$\mathbf{-104.75}$	-103.52	$\mathbf{-101.18}$	-97.65	$\mathbf{-97.56}$	-90.11	$\mathbf{-89.76}$
Caltech 101	-123.43	$\mathbf{-114.55}$	-112.08	$\mathbf{-108.28}$	-106.35	$\mathbf{-104.22}$	$\mathbf{-85.51}$	-86.22
Frey Faces	4.63	**4.57**	4.61	**4.51**	4.49	**4.45**	4.43	**4.38**
Histopathology	6.07	**6.04**	5.82	**5.75**	5.59	**5.58**	4.84	**4.82**

Experiments

Table 2: Test LL for static MNIST.

Model	LL
VAE ($L = 1$) + NF [32]	-85.10
VAE ($L = 2$) [6]	-87.86
IWAE ($L = 2$) [6]	-85.32
HVAE ($L = 2$) + SG	-85.89
HVAE ($L = 2$) + MoG	-85.07
HVAE ($L = 2$) + VampPrior *data*	-85.71
HVAE ($L = 2$) + VampPrior	$\mathbf{-83.19}$
AVB + AC ($L = 1$) [28]	-80.20
VLAE [7]	$\mathbf{-79.03}$
VAE + IAF [18]	-79.88
convHVAE ($L = 2$) + VampPrior	-81.09
PixelHVAE ($L = 2$) + VampPrior	-79.78

Table 3: Test LL for dynamic MNIST.

Model	LL
VAE ($L = 2$) + VGP [40]	-81.32
CaGeM-0 ($L = 2$) [25]	-81.60
LVAE ($L = 5$) [36]	-81.74
HVAE ($L = 2$) + VampPrior *data*	-81.71
HVAE ($L = 2$) + VampPrior	$\mathbf{-81.24}$
VLAE [7]	-78.53
VAE + IAF [18]	-79.10
PixelVAE [15]	-78.96
convHVAE ($L = 2$) + VampPrior	-79.78
PixelHVAE ($L = 2$) + VampPrior	$\mathbf{-78.45}$

Experiments

Table 4: Test LL for OMNIGLOT.

Model	LL
VR-max ($L = 2$) [24]	-103.72
IWAE ($L = 2$) [6]	-103.38
LVAE ($L = 5$) [36]	-102.11
HVAE ($L = 2$) + VampPrior	$\mathbf{-101.18}$
VLAE [7]	-89.83
convHVAE ($L = 2$) + VampPrior	-97.56
PixelHVAE ($L = 2$) + VampPrior	$\mathbf{-89.76}$

Table 5: Test LL for Caltech 101 Silhouettes.

Model	LL
IWAE ($L = 1$) [24]	-117.21
VR-max ($L = 1$) [24]	-117.10
HVAE ($L = 2$) + VampPrior	$\mathbf{-108.28}$
VLAE [7]	$\mathbf{-78.53}$
convHVAE ($L = 2$) + VampPrior	-104.22
PixelHVAE ($L = 2$) + VampPrior	-86.22

Experiments

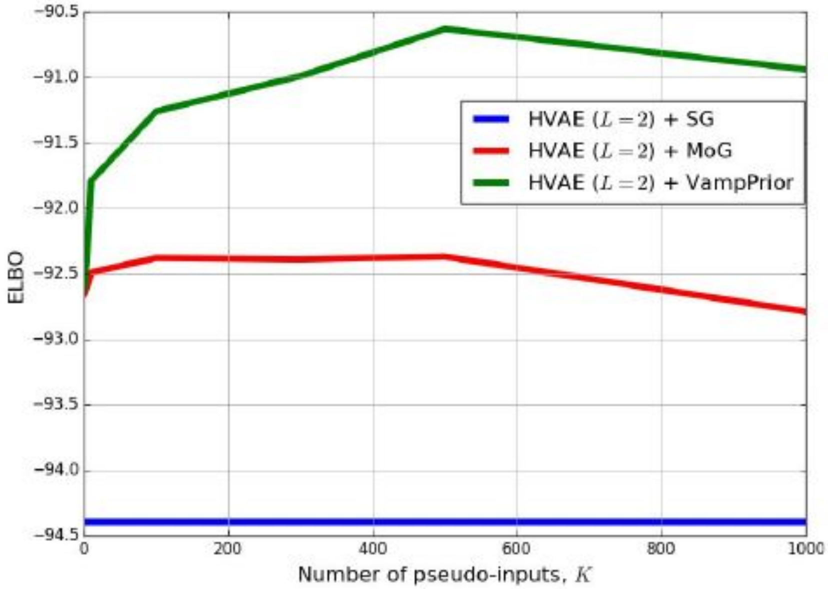


Figure 2: A comparison between the HVAE ($L = 2$) with SG prior, MoG prior and VampPrior in terms of ELBO and varying number of pseudo-inputs/components on static MNIST.

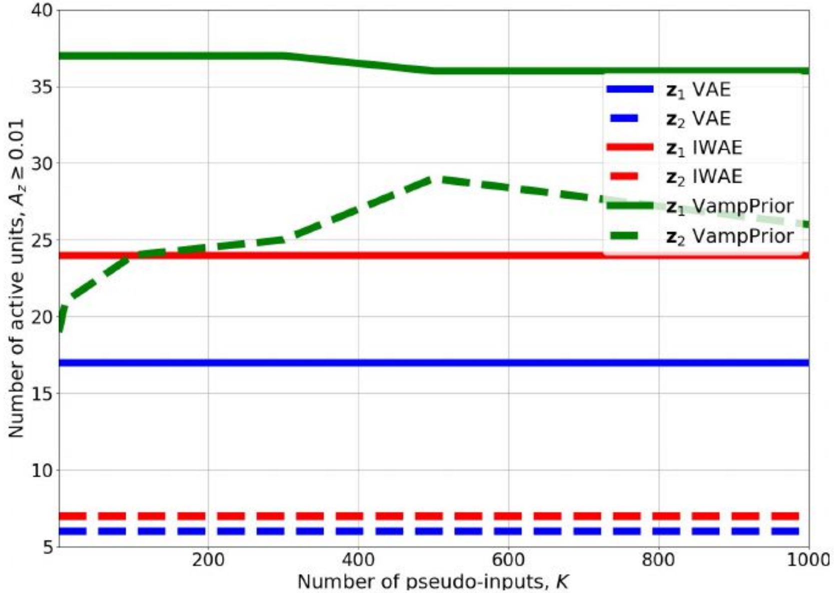


Figure 3: A comparison between two-level VAE and IWAE with the standard normal prior and theirs VampPrior counterpart in terms of number of active units for varying number of pseudo-inputs on static MNIST.

Experiments

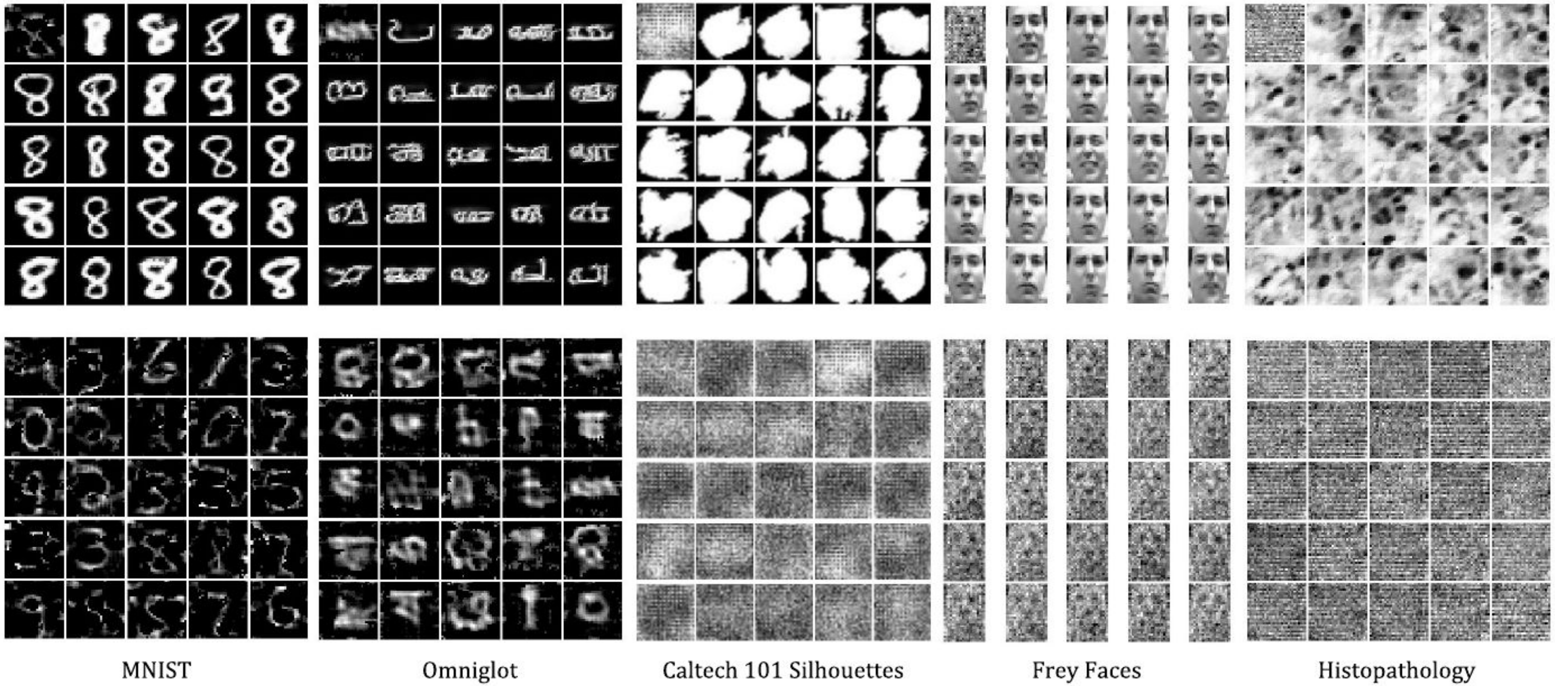


Figure 4: (*top row*) Images generated by PixelHVAE + VampPrior for chosen pseudo-input in the left top corner. (*bottom row*) Images represent a subset of trained pseudo-inputs for different datasets.

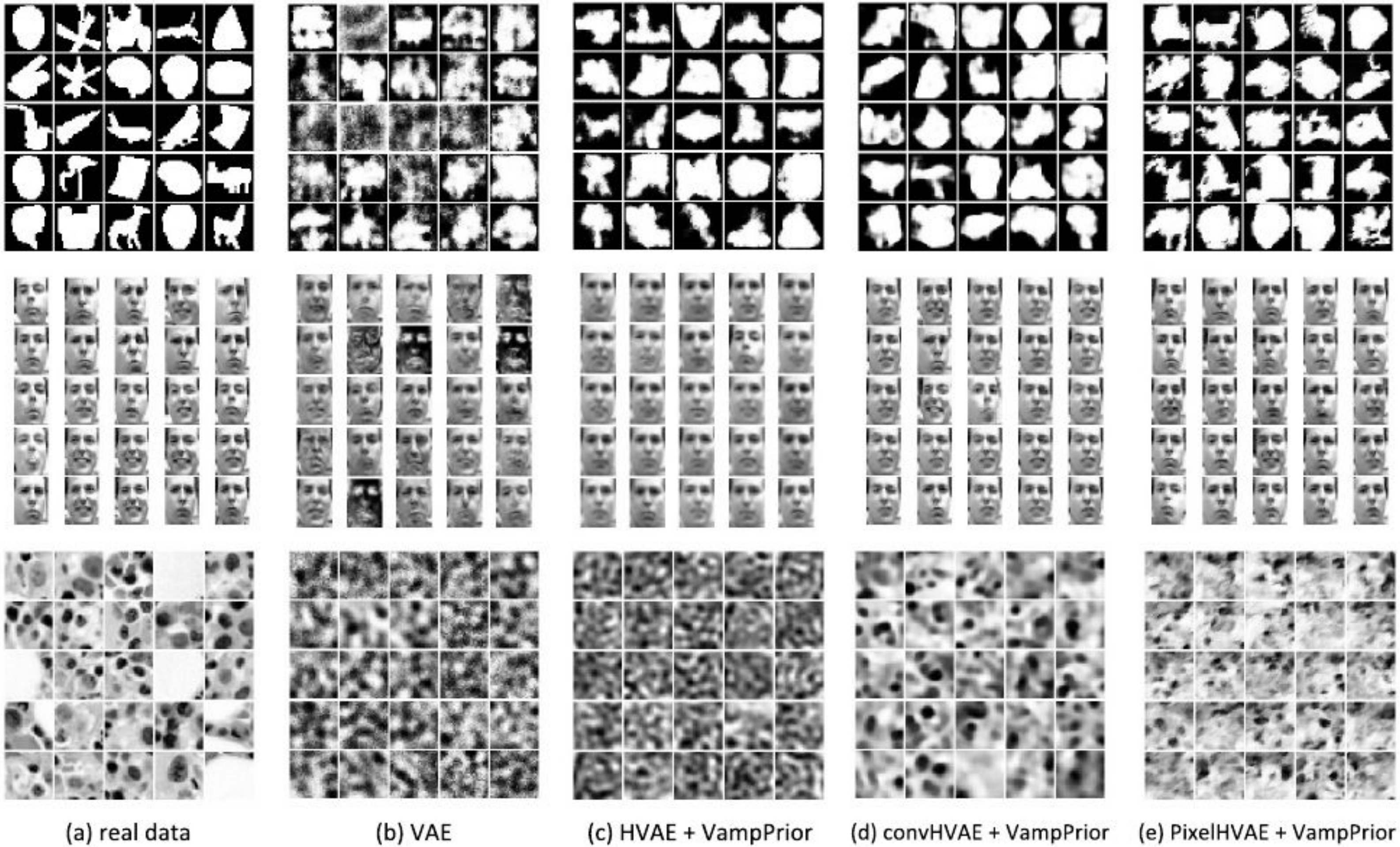


Figure 5: (a) Real images from test sets and images generated by (b) the vanilla VAE, (c) the HVAE ($L = 2$) + VampPrior, (d) the convHVAE ($L = 2$) + VampPrior and (e) the PixelHVAE ($L = 2$) + VampPrior.

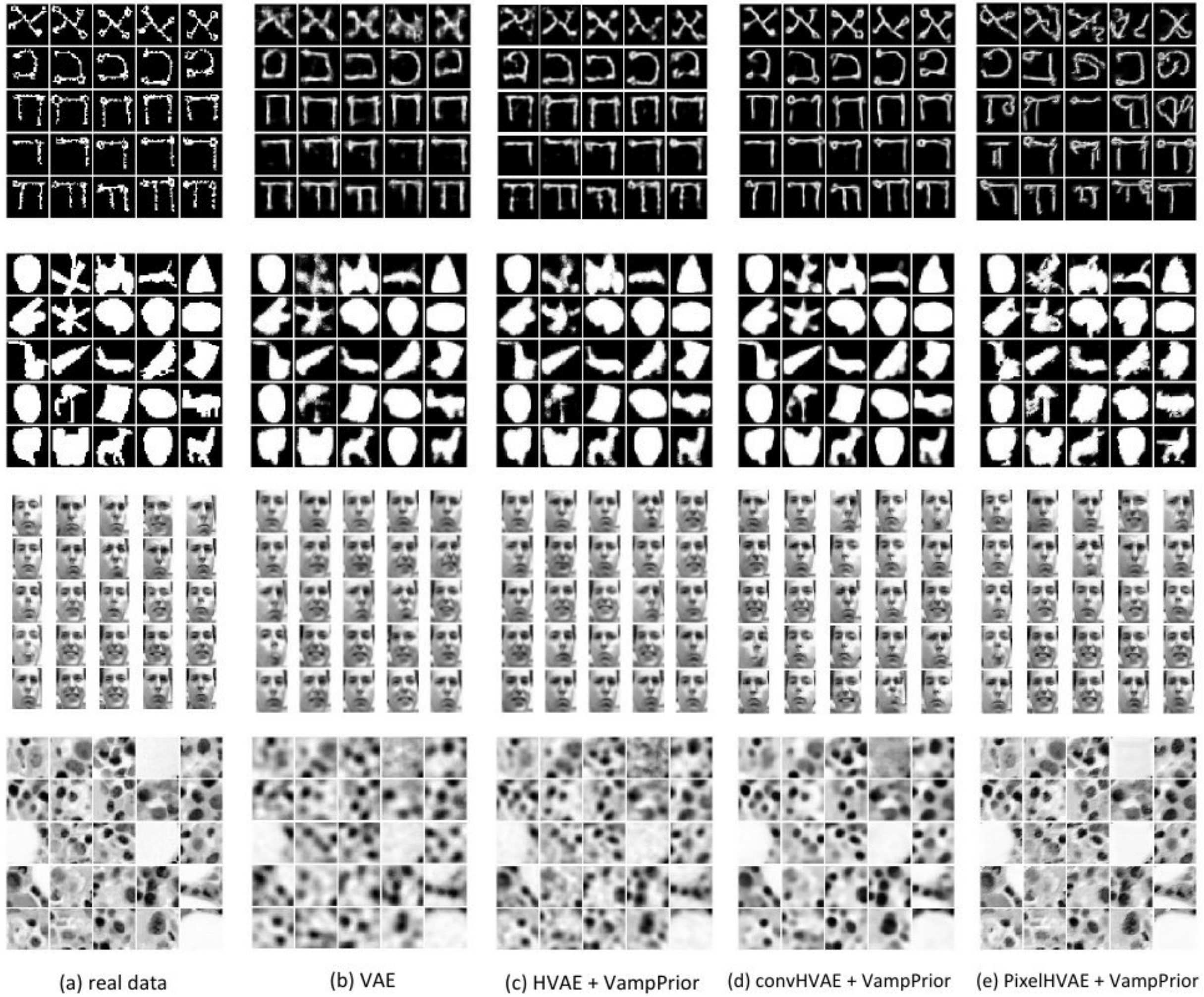


Figure 6: (a) Real images from test sets, (b) reconstructions given by the vanilla VAE, (c) the HVAE ($L = 2$) + VampPrior, (d) the convHVAE ($L = 2$) + VampPrior and (e) the PixelHVAE ($L = 2$) + VampPrior.

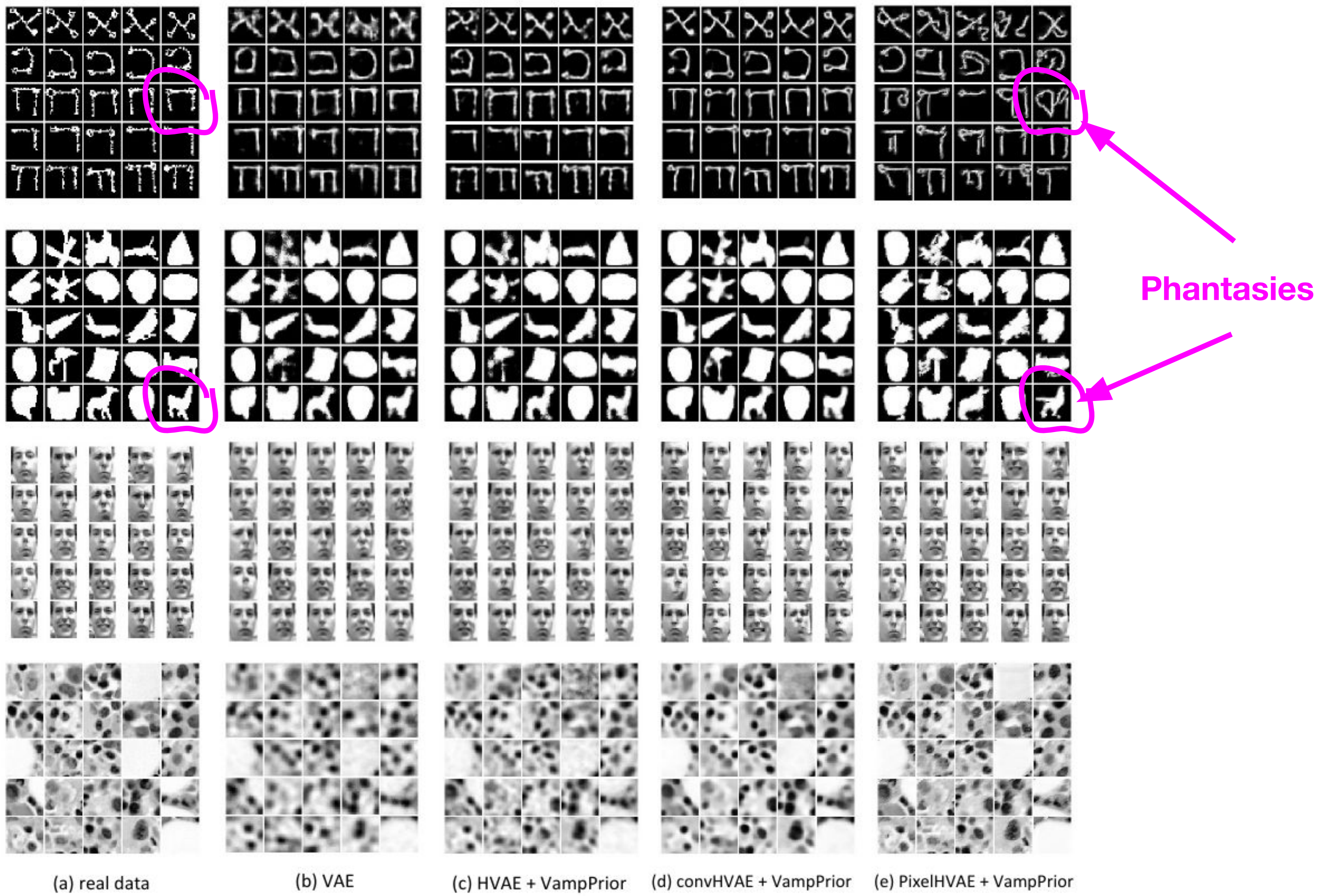


Figure 6: (a) Real images from test sets, (b) reconstructions given by the vanilla VAE, (c) the HVAE ($L = 2$) + VampPrior, (d) the convHVAE ($L = 2$) + VampPrior and (e) the PixelHVAE ($L = 2$) + VampPrior.

Conclusion

The **prior** in VAE is extremely important.

VampPrior = approximated aggregated posterior as the optimal prior

Hierarchical VampPrior VAE → **less** inactive stochastic units.

Multimodal prior → **better** generative process

Conclusion

The **prior** in VAE is extremely important.

VampPrior = approximated aggregated posterior as the optimal prior

Hierarchical VampPrior VAE → **less** inactive stochastic units.

Multimodal prior → **better** generative process

Conclusion

The prior in VAE is extremely important.

VampPrior = approximated aggregated posterior as the optimal prior

Hierarchical VampPrior VAE → **less** inactive stochastic units.

Multimodal prior → **better** generative process

Conclusion

The prior in VAE is extremely important.

VampPrior = **approximated aggregated posterior** **as the** **optimal prior**

Hierarchical VampPrior VAE → **less** inactive stochastic units.

Multimodal prior → **better** generative process

Conclusion

The prior in VAE is extremely important.

VampPrior = approximated aggregated posterior as the optimal prior

Hierarchical VampPrior VAE → **less** inactive stochastic units.

Multimodal prior → **better** generative process

Webpage:

https://jmtomczak.github.io/

Code on github:

https://github.com/jmtomczak/

Contact:

jakubmkt@gmail.com

Marie Skłodowska-Curie Actions

The research conducted by Jakub M. Tomczak was funded by the European Commission within the Marie Skłodowska-Curie Individual Fellowship (Grant No. 702666, ”Deep learning and Bayesian inference for medical imaging”).