Taming VAEs

Bayesian Methods in Machine Learning 2018

Anna Kuzina

Denis Prokopenko

Valentina Shumovskaia

β-VAE

Simple VAE if $\beta = 1$.

ELBO optimization with penalization

$$\mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z}, \beta) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta\, D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$$

Helps models to obtain with a high degree of disentanglement in image datasets.

GECO algorithm

ELBO with new constraints.

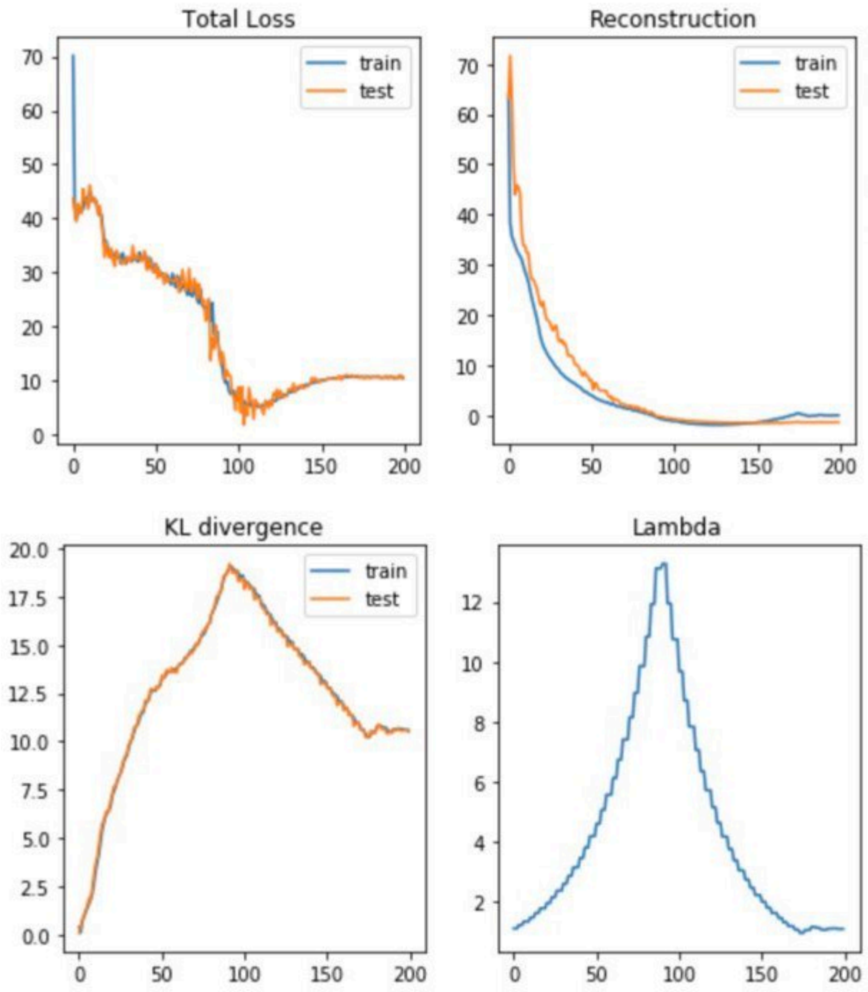
$$\mathcal{L}_{\boldsymbol{\lambda}} = \mathbb{E}_{\rho(\mathbf{x})}\left[\mathrm{KL}\left[q(\mathbf{z}|\mathbf{x}); \pi(\mathbf{z})\right]\right] + \boldsymbol{\lambda}^T \mathbb{E}_{\rho(\mathbf{x})q(\mathbf{z}|\mathbf{x})}\left[\mathcal{C}(\mathbf{x}, g(\mathbf{z}))\right]$$

We use reconstruction error as a reconstruction constraint $\|x - g(x)\|^2 - \kappa^2$

We train λ while in $\beta - \mathrm{VAE}$ it is a hyperparameter

GECO algorithm

Result: Learned parameters θ, η and Lagrange multipliers $\boldsymbol{\lambda}$
Initialize $t = 0$;
Initialize $\boldsymbol{\lambda} = \mathbf{1}$;
while *is training* **do**
    Read current data batch $\mathbf{x}$;
    Sample from variational posterior $\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})$;
    Compute the batch average of the constraint $\hat{C}^t \leftarrow \mathcal{C}(\mathbf{x}^t, g(\mathbf{z}^t))$;
    if $t == 0$ **then**
        Initialize the constraint moving average $C^0_{ma} \leftarrow \hat{C}^0$;
    else
        $C^t_{ma} \leftarrow \alpha C^{t-1}_{ma} + (1-\alpha)\hat{C}^t$;
    end
    $C^t \leftarrow \hat{C}^t + \text{StopGradient}(C^t_{ma} - \hat{C}^t)$;
    Compute gradients $G_\theta \leftarrow \frac{\partial \mathcal{L}_{\boldsymbol{\lambda}}}{\partial \theta}$ and $G_\eta \leftarrow \frac{\partial \mathcal{L}_{\boldsymbol{\lambda}}}{\partial \eta}$;
    Update parameters as $\Delta_{\theta,\eta} \propto -G_{\theta,\eta}$ and Lagrange multiplier(s) $\Delta_{\log(\boldsymbol{\lambda})} \propto C^t$;
    $t \leftarrow t + 1$;
end



Importance Weighted AE

Original VAE:

$$\log p(\mathbf{x}) = \log \mathbb{E}_{q(\mathbf{h}|\mathbf{x})} \left[\frac{p(\mathbf{x}, \mathbf{h})}{q(\mathbf{h}|\mathbf{x})} \right] \geq \mathbb{E}_{q(\mathbf{h}|\mathbf{x})} \left[\log \frac{p(\mathbf{x}, \mathbf{h})}{q(\mathbf{h}|\mathbf{x})} \right] = \mathcal{L}(\mathbf{x}).$$

IWAE improvement:

$$\mathcal{L}_k(\mathbf{x}) = \mathbb{E}_{\mathbf{h}_1, \ldots, \mathbf{h}_k \sim q(\mathbf{h}|\mathbf{x})} \left[\log \frac{1}{k} \sum_{i=1}^{k} \frac{p(\mathbf{x}, \mathbf{h}_i)}{q(\mathbf{h}_i|\mathbf{x})} \right].$$

Dependency on:

$$\log p(\mathbf{x}) \geq \mathcal{L}_{k+1} \geq \mathcal{L}_k.$$

VAE is a particular case of IWAE with $k = 1$

Importance Weighted AE

IWAE improvement:

$$\mathcal{L}_k(\mathbf{x}) = \mathbb{E}_{\mathbf{h}_1,\ldots,\mathbf{h}_k \sim q(\mathbf{h}|\mathbf{x})}\left[\log \frac{1}{k}\sum_{i=1}^{k} \frac{p(\mathbf{x},\mathbf{h}_i)}{q(\mathbf{h}_i|\mathbf{x})}\right].$$

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}\mathcal{L}_k(\mathbf{x}) = \nabla_{\boldsymbol{\theta}}\mathbb{E}_{\mathbf{h}_1,\ldots,\mathbf{h}_k}\left[\log \frac{1}{k}\sum_{i=1}^{k} w_i\right] &= \nabla_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{\epsilon}_1,\ldots,\boldsymbol{\epsilon}_k}\left[\log \frac{1}{k}\sum_{i=1}^{k} w(\mathbf{x},\mathbf{h}(\mathbf{x},\boldsymbol{\epsilon}_i,\boldsymbol{\theta}),\boldsymbol{\theta})\right] \\ &= \mathbb{E}_{\boldsymbol{\epsilon}_1,\ldots,\boldsymbol{\epsilon}_k}\left[\nabla_{\boldsymbol{\theta}} \log \frac{1}{k}\sum_{i=1}^{k} w(\mathbf{x},\mathbf{h}(\mathbf{x},\boldsymbol{\epsilon}_i,\boldsymbol{\theta}),\boldsymbol{\theta})\right] \\ &= \mathbb{E}_{\boldsymbol{\epsilon}_1,\ldots,\boldsymbol{\epsilon}_k}\left[\sum_{i=1}^{k} \widetilde{w_i}\nabla_{\boldsymbol{\theta}} \log w(\mathbf{x},\mathbf{h}(\mathbf{x},\boldsymbol{\epsilon}_i,\boldsymbol{\theta}),\boldsymbol{\theta})\right],\end{aligned}$$

$$\tilde{w}_i = \frac{w_i}{\sum_{j=1}^{K} w_i}$$

IWAE & GECO

Loss function for GECO:

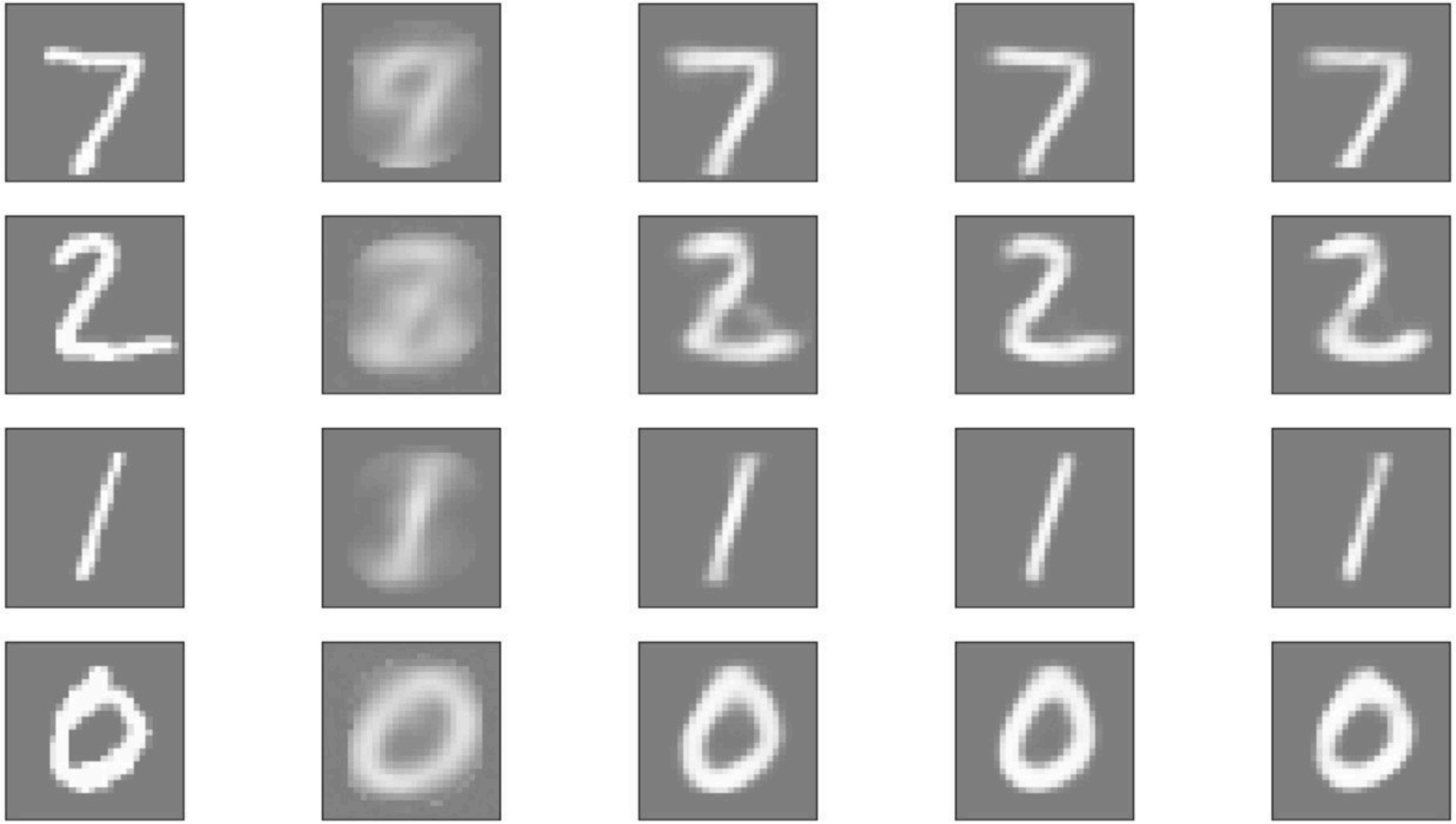
$$\mathcal{L} = \lambda \mathbb{E}_z \log p(x|z) + KL(q(z|x)||p(z))$$

$$\mathcal{L}_{IWAE} = \mathbb{E}_z \log \frac{p(x,z)}{q_{IW}(z|x)} \qquad q_{IW}(z|x) = \frac{p(x,z)}{\sum_{i=1}^{K} \frac{p(x,z_i)}{q(z_i,x)}}$$

$$\mathcal{L}_{GECO} = \lambda \log p(x|z) + KL(q_{IW}(z|x)||p(z)) = (\lambda - 1) \log p(x|z) + \mathcal{L}_{IWAE}$$

Reconstruction comparison

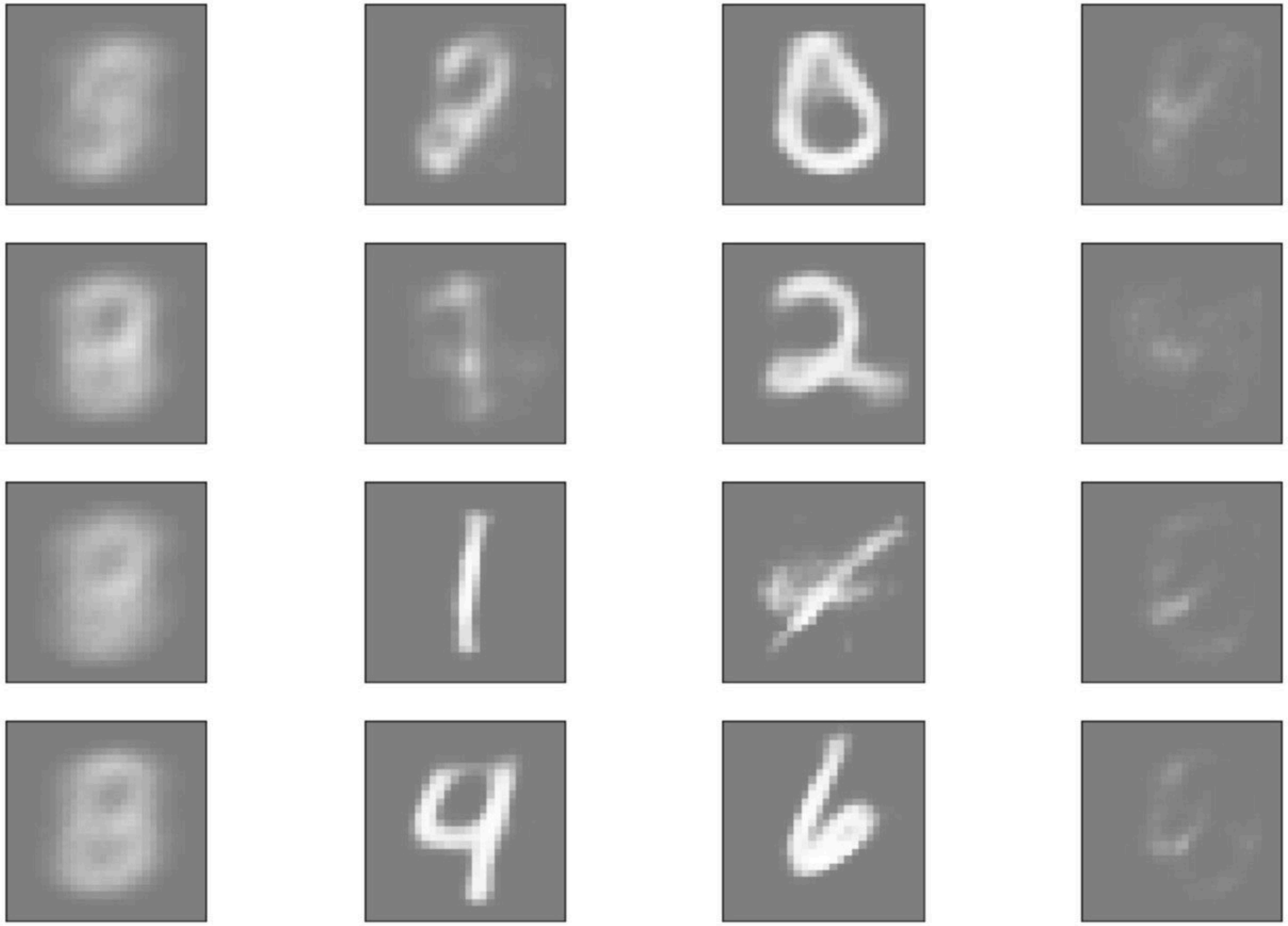
MNIST



Init | VAE | β = 0.5 VAE | GECO | GECO + IWAE

Generation comparison

MNIST



VAE

β = 0.5 VAE

GECO

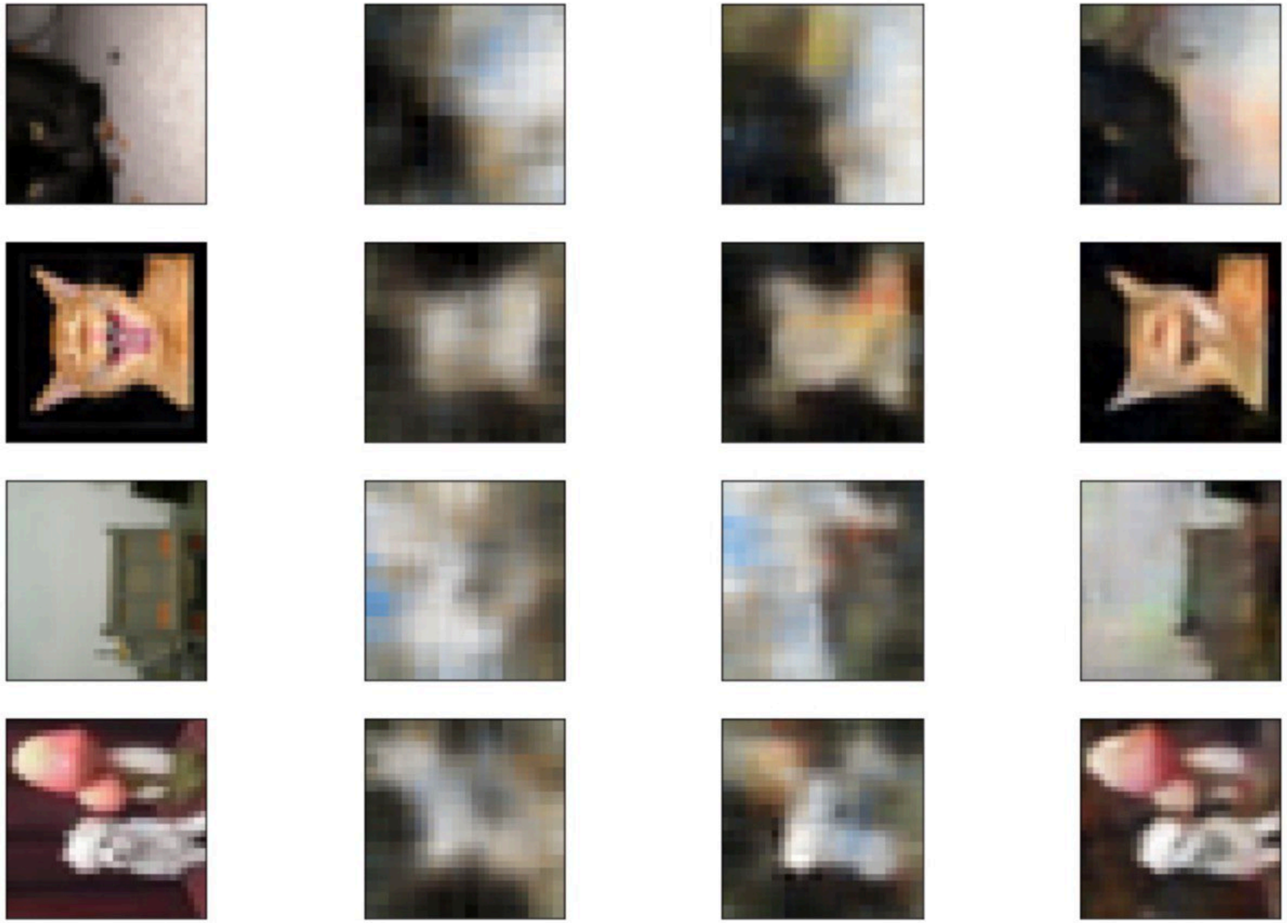
GECO + IWAE

Numerical comparison

MNIST

Model	Marginal KL	Reconstruction loss
VAE	1.5813	55.8484
β-VAE	10.5938	15.1434
GECO	10.4975	14.6269
GECO + IWAE	1.3640	12.9214

Reconstruction comparison
CIFAR 10



Init | VAE | $\beta = 0.5$ VAE | GECO

Numerical comparison

CIFAR 10

Model	Marginal KL	Reconstruction loss
VAE	31.9662	77.6570
β-VAE	59.4998	57.9653
GECO	383.0849	11.4413

Conclusions

- GECO is a good algorithm for VAEs training, showed itself better than simple VAE and β-VAE:
  - Good reconstructions,
  - Good generations
- GECO + IWAE algorithm:
  - Better reconstructions,
  - Worse generations;