Learning Hierarchical Priors in VAEs

Alexej Klushyn, Nutan Chen, Richard Kurle, Botond Cseke, Patrick van der Smagt

Machine Learning Research Lab, Volkswagen Group, Munich, Germany

VOLKSWAGEN AKTIENGESELLSCHAFT

ARGMAX.ai VOLKSWAGEN GROUP ML RESEARCH

Abstract

We address the issue of learning informative latent representations of the data in the context of variational autoencoders. To do this we

- propose a hierarchical prior to avoid the over-regularisation resulting from a standard normal prior distribution.
- formulate the learning problem as a constrained optimisation problem.
- introduce a graph-based interpolation method to verify the learned latent representation.

Background: VAEs as a Constrained Optimisation Problem

Rezende & Viola (2018) reformulate the VAE objective as the Lagrangian of a constrained optimisation problem

$$\min_{\theta}\max_{\lambda}\min_{\phi}\ \mathcal{L}(\theta,\phi;\lambda)\quad\text{s.t.}\quad\lambda\geq 0,$$

where

$$\mathcal{L}(\theta,\phi;\lambda)\equiv\mathbb{E}_{p_\mathcal{D}(\mathbf{x})}\Big[\mathbb{KL}\left(q_\phi(\mathbf{z}|\mathbf{x})\|\,p_0(\mathbf{z})\right)+\lambda\big(\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[C_\theta(\mathbf{x},\mathbf{z})\big]-\kappa^2\big)\Big].$$

Optimisation is performed by quasi-gradient ascent/descent

$$\lambda_t=\lambda_{t-1}\cdot\exp\left(\nu\cdot(\hat{C}_t-\kappa^2)\right)\quad\text{and}\quad(\theta_t,\phi_t)=(\theta_{t-1},\phi_{t-1})-\eta_t\,\partial_{(\theta,\phi)}\,\mathcal{L},$$

where $\Delta\lambda_t\cdot\partial_\lambda\mathcal{L}\geq 0$ and ν the update's learning rate. $\mathcal{L}(\theta,\phi;\lambda)$ is an ELBO iff $\lambda=1$, or if $0\leq\lambda<1$, a scaled lower bound on the ELBO; $\max_\lambda\min_\phi\mathcal{L}(\theta,\phi;\lambda)$ corresponds to the E-step of EM.

Hierarchical Priors for Learning Informative Latent Representations

We use a hierarchical prior/two-layer stochastic model

$$p_0(\mathbf{z})\equiv p_\Theta(\mathbf{z})=\int p_\Theta(\mathbf{z}|\zeta)\,p(\zeta)\,\mathrm{d}\zeta.$$

To make computations tractable, we apply an IW upper bound

$$\mathbb{E}_{p_\mathcal{D}(\mathbf{x})}\,\mathbb{KL}\left(q_\phi(\mathbf{z}|\mathbf{x})\|\,p_\Theta(\mathbf{z})\right)\leq\mathcal{F}(\phi,\Theta,\Phi)$$
$$\equiv\mathbb{E}_{p_\mathcal{D}(\mathbf{x})}\,\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log q_\phi(\mathbf{z}|\mathbf{x})-\mathbb{E}_{\zeta_{1:K}\sim q_\Phi(\zeta|\mathbf{z})}\left[\log\frac{1}{K}\sum_{k=1}^{K}\frac{p_\Theta(\mathbf{z}|\zeta_k)\,p(\zeta_k)}{q_\Phi(\zeta_k|\mathbf{z})}\right]\right],$$

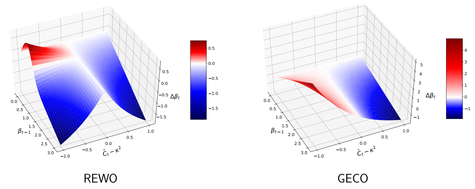
which leads to the objective

$$\mathcal{L}(\theta,\phi;\lambda)\leq\mathcal{L}_{\text{VHP}}(\theta,\phi,\Theta,\Phi;\lambda)$$
$$\equiv\mathcal{F}(\phi,\Theta,\Phi)+\lambda\big(\mathbb{E}_{p_\mathcal{D}(\mathbf{x})}\,\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[C_\theta(\mathbf{x},\mathbf{z})\big]-\kappa^2\big).$$

The resulting optimsation problem is formulated as

$$\min_{\Theta,\Phi}\min_{\theta}\max_{\lambda}\min_{\phi}\ \mathcal{L}_{\text{VHP}}(\theta,\phi,\Theta,\Phi;\lambda)\quad\text{s.t.}\quad\lambda\geq 0$$

and leads to the following double-loop method: (i) update the upper bound via (Θ,Ψ); (ii) solve the resulting constrained optimisation problem w.r.t. (θ,λ,ψ). We experienced that the double-loop method behaves as a layer-wise pre-training: first, we optimise w.r.t. (Θ,Ψ) until $\hat{C}_t<\kappa^2$; then, we train all parameters jointly.



β-update scheme: $\Delta\beta_t=\beta_t-\beta_{t-1}$ as a function of β_{t-1} and $\hat{C}_t-\kappa^2$ for $\nu=1$ and $\tau=3$.

To obtain a tight lower bound on the ELBO, we use $\lambda\geq 1$ instead of $\lambda\geq 0$. In terms of $\beta=1/\lambda$, as in previous literature, this results in $0<\beta\leq 1$. We achieve $0<\beta\leq 1$ by applying the following update:

$$\beta_t=\beta_{t-1}\cdot\exp\left[\nu\cdot f_\beta(\beta_{t-1},\hat{C}_t-\kappa^2;\tau)\cdot(\hat{C}_t-\kappa^2)\right],$$

where we define

$$f_\beta(\beta,\delta;\tau)=\big(1-H(\delta)\big)\cdot\tanh\big(\tau\cdot(\beta-1)\big)-H(\delta).$$

H is the Heaviside function and τ a slope parameter.

We implement the resulting pre-training as follows:

- Initial phase: we start with $\beta\ll 1$ to enforce a reconstruction optimisation and keep β,Θ,Φ constant until $\hat{C}_t<\kappa^2$.
- Main phase: after $\hat{C}_t<\kappa^2$ is fulfilled, we also optimise Θ,Φ, and update β.

Reconstruction-error-based weighting of the objective function (REWO)

Initialise $t=1$
Initialise $\beta\ll 1$
Initialise InitialPhase = True
while training **do**
  Read current data batch $\mathbf{x}_{\text{ba}}$
  Sample from variational posterior $\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x}_{\text{ba}})$
  Compute $\hat{C}_{\text{ba}}$ (batch average)
  $\hat{C}_t=(1-\alpha)\cdot\hat{C}_{\text{ba}}+\alpha\cdot\hat{C}_{t-1},\quad(\hat{C}_0=\hat{C}_{\text{ba}})$
  if $\hat{C}_t<\kappa^2$ **then**
    InitialPhase = False
  end if
  if InitialPhase **then**
    Optimise $\mathcal{L}_{\text{VHP}}(\theta,\phi,\Theta,\Phi;\beta)$ w.r.t θ,ϕ
  else
    $\beta\leftarrow\beta\cdot\exp\left[\nu\cdot f_\beta(\beta_{t-1},\hat{C}_t-\kappa^2;\tau)\cdot(\hat{C}_t-\kappa^2)\right]$
    Optimise $\mathcal{L}_{\text{VHP}}(\theta,\phi,\Theta,\Phi;\beta)$ w.r.t θ,ϕ,Θ,Φ
  end if
  $t\leftarrow t+1$
end while

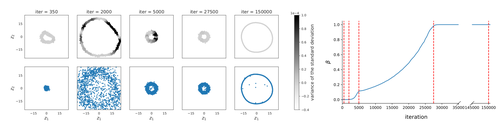
Graph-Based Interpolation

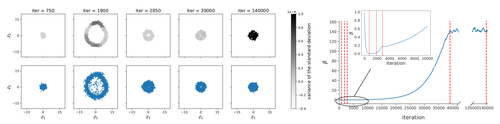
The nodes of the graph $\mathbf{Z}=\{\mathbf{z}_1,\ldots,\mathbf{z}_N\}$ are obtained by randomly sampling N samples from the respective prior distribution:

$$\mathbf{z}_n,\zeta_n\sim p_\Theta(\mathbf{z}|\zeta)\,p(\zeta),\quad n=1,\ldots,N.$$

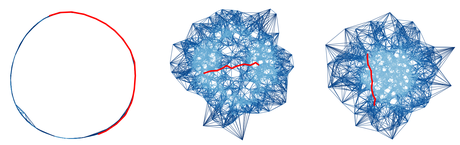
The graph is constructed by connecting each node by undirected edges to its k-nearest neighbours. The edge weights are Euclidean distances in the latent space between the related node pairs.

Artificial Pendulum

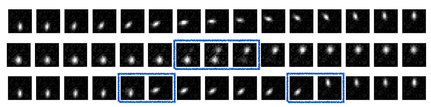




Latent representation of the pendulum data at different iteration steps when optimising $\mathcal{L}_{\text{VHP}}(\theta,\phi,\Theta,\Phi;\beta)$ with REWO and GECO, respectively. The top row shows the approximate posterior; the greyscale encodes the variance of its standard deviation. The bottom row shows the hierarchical prior.



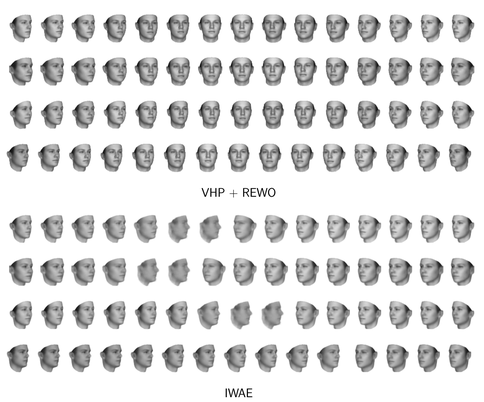
Graph-based interpolation of the pendulum movement. The graph is based on the respective prior. The red curves depict the interpolations, the bluescale indicates the edge weight.



top: VHP + REWO, middle: VHP + GECO, bottom: IWAE

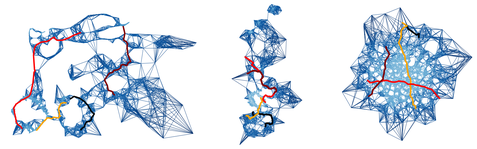
Pendulum reconstructions of the graph-based interpolations (red curve). Discontinuities are marked by blue boxes.

3D Faces

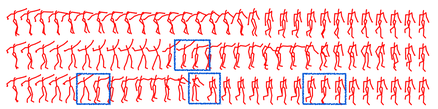


Graph-based interpolations along the learned 32-dimensional latent manifold. The graph is based on samples from the respective prior distribution.

Human Motion

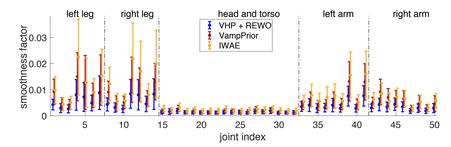


Graph-based interpolation of human motions. The graphs are based on the (learned) prior distributions. The bluescale indicates the edge weight. The coloured lines represent four interpolated movements.



top: VHP + REWO, middle: VampPrior, bottom: IWAE

Human-movement reconstructions of the graph-based interpolations (red curve). Discontinuities are marked by blue boxes.



Smoothness measure of the human-movement interpolations. For each joint, the mean and standard deviation of the smoothness factor are displayed. Smaller values correspond to smoother movements.

MNIST, Fashion-MNIST, & OMNIGLOT

Negative test log-likelihood estimated with 5,000 importance samples

	DYNAMIC MNIST	STATIC MNIST	FASHION-MNIST	OMNIGLOT
VHP + REWO	78.88	82.74	225.37	101.78
VHP + GECO	95.01	96.32	234.73	108.97
VAMPPRIOR	80.42	84.02	232.78	101.97
IWAE (L=1)	81.36	84.46	226.83	101.57
IWAE (L=2)	80.66	82.83	225.39	101.83

References

Burda, Y. et al. Importance weighted autoencoders. *ICLR*, 2016.
Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. *ICML*, 2014.
Rezende, D. J. and Viola, F. Taming VAEs. *CoRR*, 2018.
Tomczak, J. and Welling, M. VAE with a VampPrior. *AISTATS*, 2018.

Contact

alexej.klushyn@argmax.ai
botond.cseke@argmax.ai