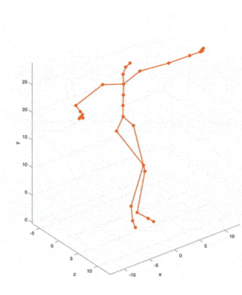
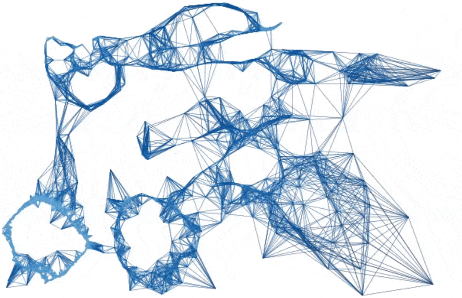
Learning Hierarchical Priors in VAEs

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

Machine Learning Research Lab, Volkswagen Group





VAEs as a Constrained Optimisation Problem

- In the context of VAEs, it is a desired ability to being able to control the reconstruction quality

- Therefore, Rezende & Viola (2018) proposed to formulate the learning problem as

$$\min_{\phi} \underbrace{\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\, \mathrm{KL}\big(q_\phi(\mathbf{z}|\mathbf{x}) \| \, p(\mathbf{z})\big) \,\right]}_{\text{optimisation objective}} \quad \text{s.t.} \quad \underbrace{\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\, \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\mathcal{C}_\theta(\mathbf{x}, \mathbf{z})\right] \leq \kappa^2}_{\text{inequality constraint}}$$

- $\mathcal{C}_\theta(\mathbf{x}, \mathbf{z})$ is defined as the reconstruction-error-related term in $-\log p_\theta(\mathbf{x}|\mathbf{z})$

Hierarchical Priors for Learning Informative Latent Representations

- The optimal empirical Bayes prior is the aggregated posterior distribution

$$p^*(\mathbf{z}) = \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[q_\phi(\mathbf{z}|\mathbf{x})\right]$$

- In order to express $p^*(\mathbf{z})$, we use a hierarchical model

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

- The parameters are learned by applying an importance-weighted lower bound

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

Lagrangian & Optimisation Problem

- The corresponding Lagrangian is

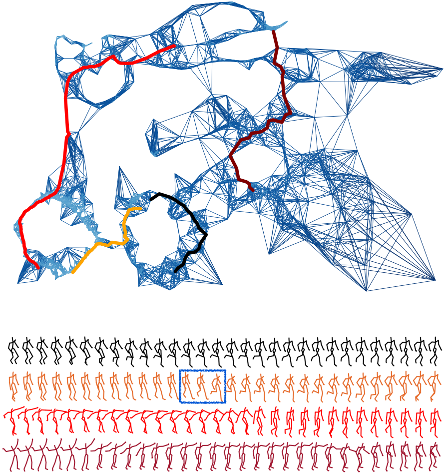
$$\mathcal{L}(\theta, \phi, \Theta, \Phi; \lambda) = \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \Big[\log q_\phi(\mathbf{z}|\mathbf{x}) - \mathcal{F}(\Theta, \Phi; \mathbf{z}) + \lambda \big(\mathcal{C}_\theta(\mathbf{x}, \mathbf{z}) - \kappa^2 \big) \Big]$$

- As a result, we arrive to the optimisation problem

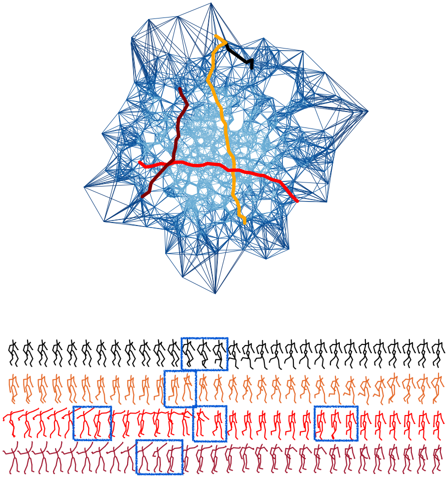
$$\underbrace{\min_{\Theta,\Phi}}_{\substack{\text{empirical}\\\text{Bayes}}} \overbrace{\min_{\theta}}^{\text{M-step}} \underbrace{\max_{\lambda} \min_{\phi}}_{\text{E-step}} \mathcal{L}(\theta, \phi, \Theta, \Phi; \lambda) \quad \text{s.t.} \quad \lambda \geq 0$$

- $\min_\theta \mathcal{L}$ and $\max_\lambda \min_\phi \mathcal{L}$ can be interpreted as the corresponding steps of the original EM algorithm for training VAEs

CMU Human Motion Data



VHP-VAE



IWAE

3D Faces

Poster #153

Today, 05:30–07:30 PM