Disentangled Dynamic Representations from Unordered Data

Leonhard Helminger [1] Abdelaziz Djelouah Markus Gross [1] Romann M. Weber
[1] ETH Zurich

Motivation

- Better interpretabilty through disentangled representation
- In context of video: static and dynamic information
- Pairwise comparison instead of recurrent architecture

Contribution

We introduce a deep generative model that learns disentangled static and dynamic representations from data without temporal ordering

Model

Generative Model:

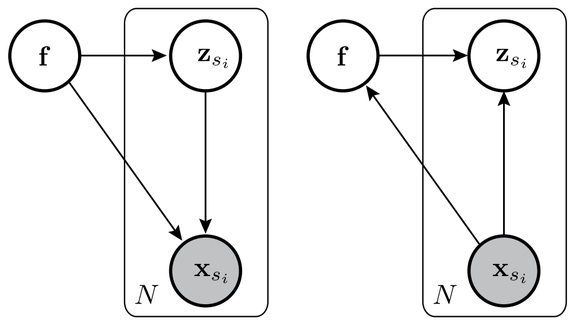
$$p_\theta\left(\mathbf{x}_{1:T}, \mathbf{f}, \mathbf{z}_{1:T}\right) = p_\theta\left(\mathbf{f}\right) \prod_{t=1}^{T} p_\theta\left(\mathbf{x}_t \mid \mathbf{f}, \mathbf{z}_t\right) p_\theta\left(\mathbf{z}_t \mid \mathbf{f}\right)$$

Inference Model:

$$q_\phi\left(\mathbf{f}, \mathbf{z}_{s_{1:N}} \mid \mathbf{x}_S\right) = q_\phi\left(\mathbf{f} \mid \mathbf{x}_S\right) \prod_{i=1}^{N} q_\phi\left(\mathbf{z}_{s_i} \mid \mathbf{f}, \mathbf{x}_{s_i}\right)$$

Learning:

$$\mathcal{L} = \mathbb{E}_{q_\phi(\mathbf{f} \mid \mathbf{x}_S)}\left[\sum_{\mathbf{x} \in \mathbf{x}_S} \mathbb{E}_{q_\phi(\mathbf{z} \mid \mathbf{f}, \mathbf{x})}\left[\log p_\theta\left(\mathbf{x} \mid \mathbf{f}, \mathbf{z}\right) - \mathrm{D}_{KL}\left(q_\phi\left(\mathbf{z} \mid \mathbf{f}, \mathbf{x}\right) \,||\, p_\theta\left(\mathbf{z} \mid \mathbf{f}\right)\right)\right]\right] - \mathrm{D}_{KL}\left(q_\phi\left(\mathbf{f} \mid \mathbf{x}_S\right) \,||\, p_\theta\left(\mathbf{f}\right)\right)$$



Generator

Encoder

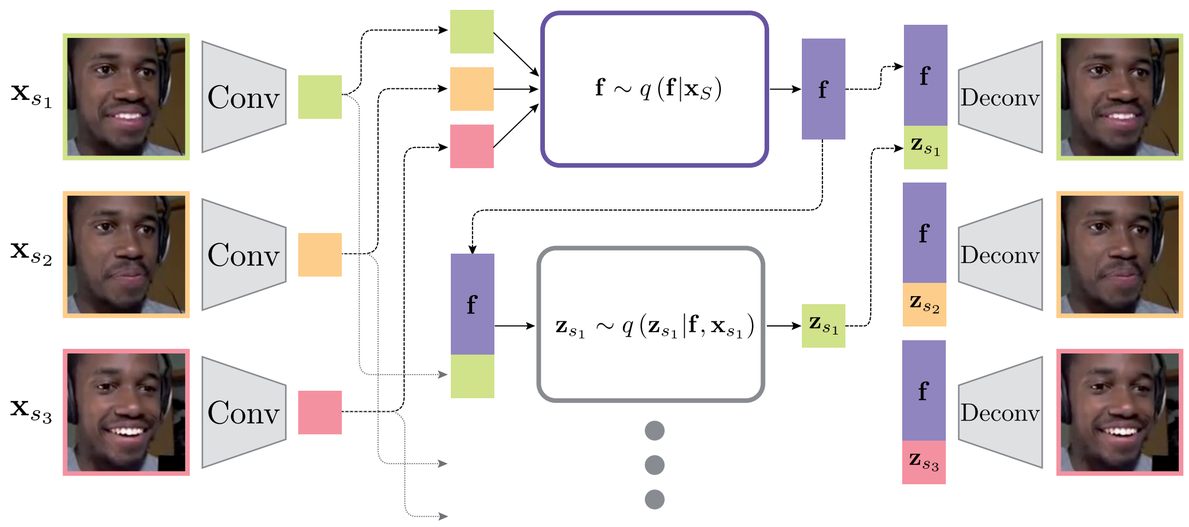
Distributions of the generative model:

$$p_\theta\left(\mathbf{f}\right) = \mathcal{N}\left(\mathbf{f} \mid \mathbf{0}, \mathbf{1}\right),$$
$$p_\theta\left(\mathbf{z} \mid \mathbf{f}\right) = \mathcal{N}\left(\mathbf{z} \mid h_{\mu_z}\left(\mathbf{f}\right), \mathrm{diag}\left(h_{\sigma_z^2}\left(\mathbf{f}\right)\right)\right)$$
$$p_\theta\left(\mathbf{x} \mid \mathbf{f}, \mathbf{z}\right) = \mathrm{Ber}_p\left(\mathbf{x} \mid h_x\left(\mathbf{f}, \mathbf{z}\right)\right)$$

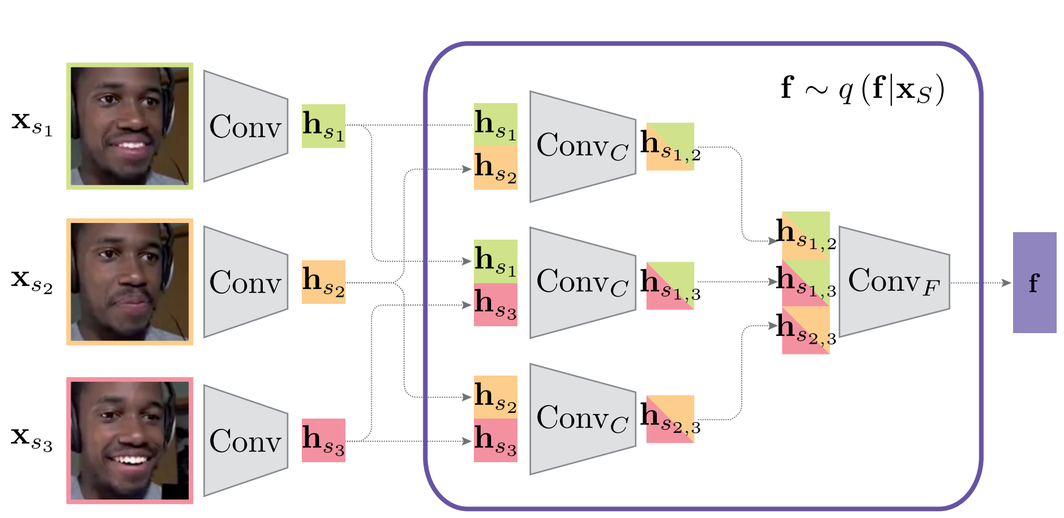
Distributions of the inference model:

$$q_\phi\left(\mathbf{f} \mid \mathbf{x}_S\right) = \mathcal{N}\left(\mathbf{f} \mid g_{\mu_f}\left(\mathbf{x}_S\right), \mathrm{diag}\left(g_{\sigma_f^2}\left(\mathbf{x}_S\right)\right)\right),$$
$$q_\phi\left(\mathbf{z} \mid \mathbf{f}, \mathbf{x}\right) = \mathcal{N}\left(\mathbf{z} \mid g_{\mu_z}\left(\mathbf{f}, \mathbf{x}\right), \mathrm{diag}\left(g_{\sigma_z^2}\left(\mathbf{f}, \mathbf{x}\right)\right)\right)$$
where $\mathbf{x}_S = \left(x_{s_1}, \ldots, x_{s_N}\right),\ N = 3$

Details



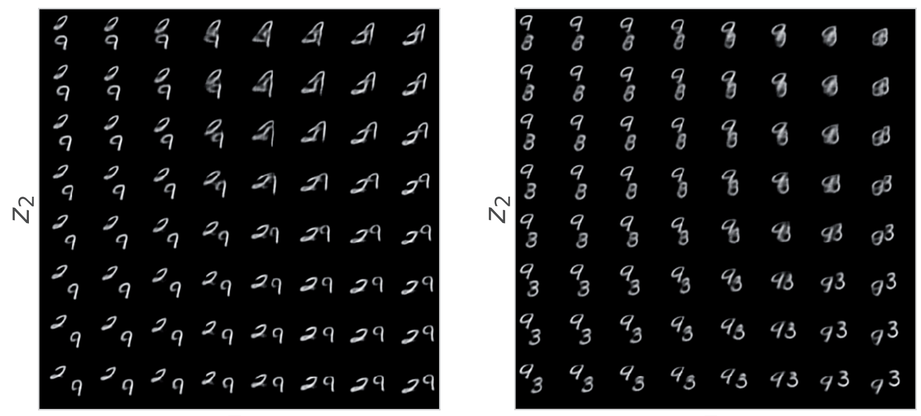
Visualization of the full model



Visualization of the encoder for the static latent variable model

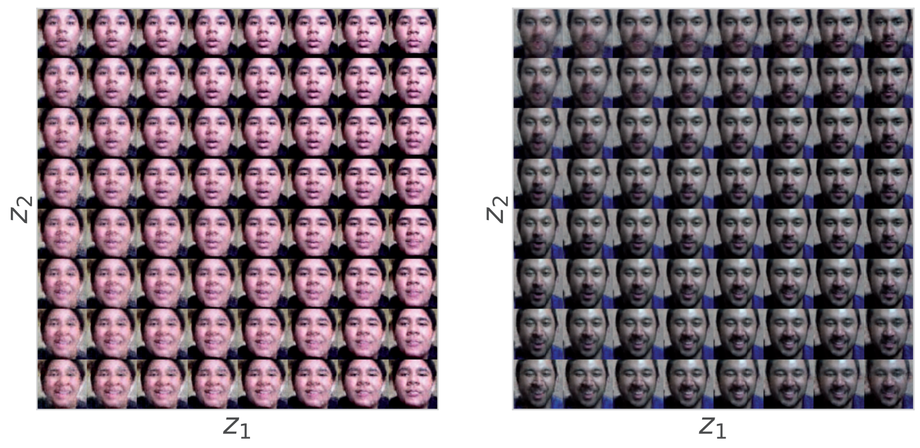
Results

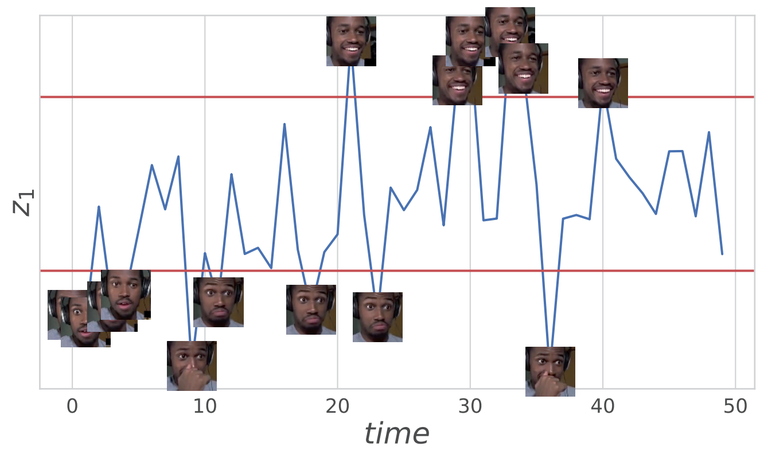
Learned dynamic data manifold of the 2d latent space



- Digits and style of the handwritten numbers are consistent over the spanned latent space
- Coherence of the dynamic space between different identities

Naive emotion detection in a 2d latent space





- Two-dimensional latent space captures dynamics of faces
- Naive analysis of the dynamic plot can already extract some meaningful facial expressions (smiling / astonished)