Disentangled Sequential Autoencoder

Y. Li, S. Mandt

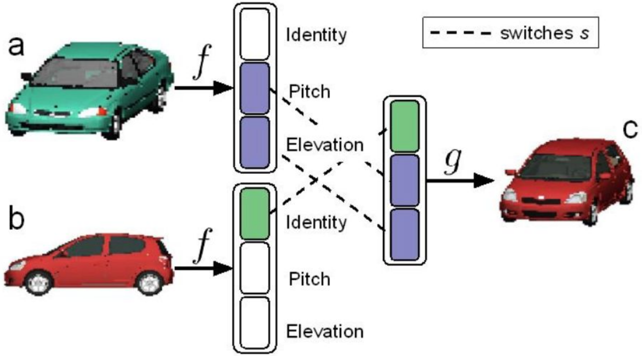
ICML 2018

Shuangshuang Chen

April 2019

Disentangled representation learning

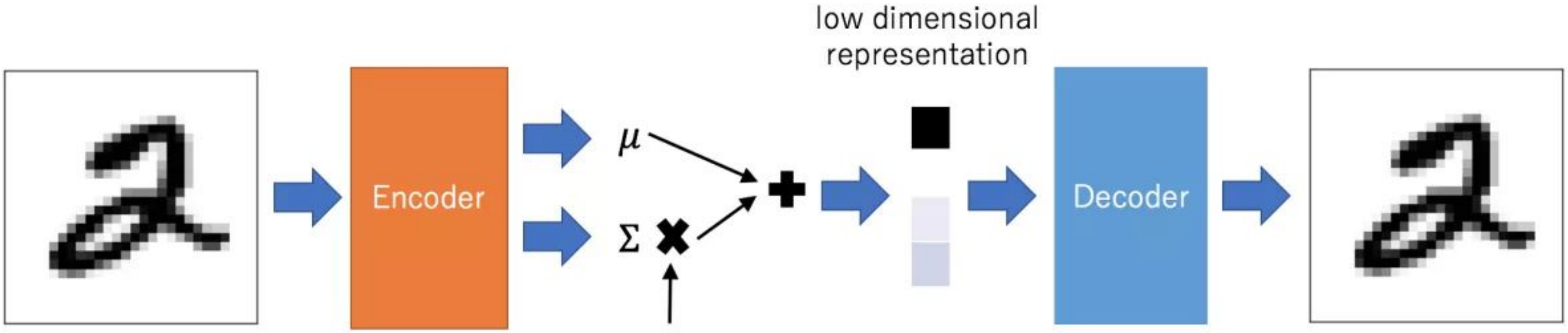
Definition: each learned features refers to a semantically meaningful concept

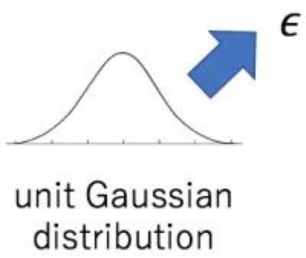


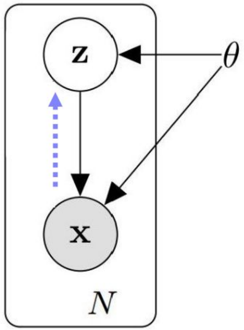
Strategies:

- additional regula lement i.e. beta-VAE
- network structure to enforce factored representations i.e. Siddharth et al. (2017); Bouchacourt et al. (2017)
- mixing both: infoGAN; Mathieu et al. 2016

Recap. Variational Autoencoder







$$\ln p(\mathbf{x}) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p_\theta(\mathbf{x} \mid \mathbf{z})] - \mathrm{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) \,||\, p(\mathbf{z})) =: \mathcal{L}_{\mathrm{SGVB}}(\mathbf{x}, \phi, \theta)$$

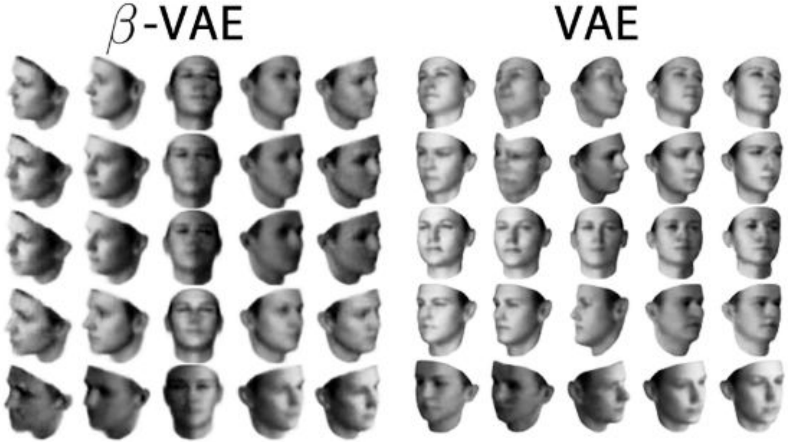
[1] Kingma, D. P. and Welling, M. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

beta-VAE

$= 1 \Rightarrow$ VAE

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

$$\max_{\phi,\theta} \mathbb{E}_{x\sim\mathbf{D}} \left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]\right] \quad \text{subject to} \quad D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) < \epsilon$$

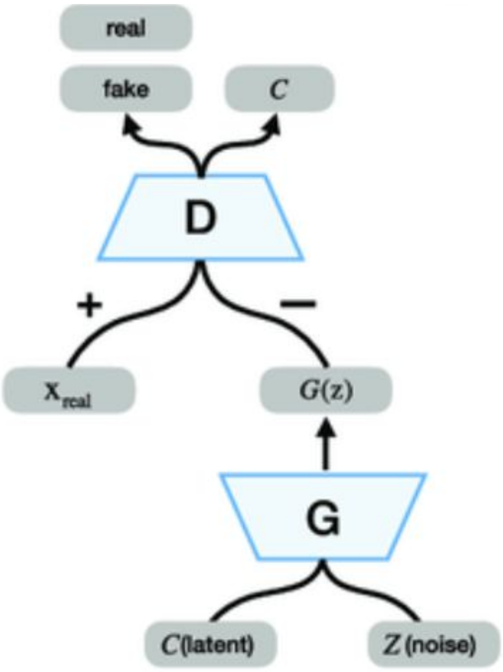


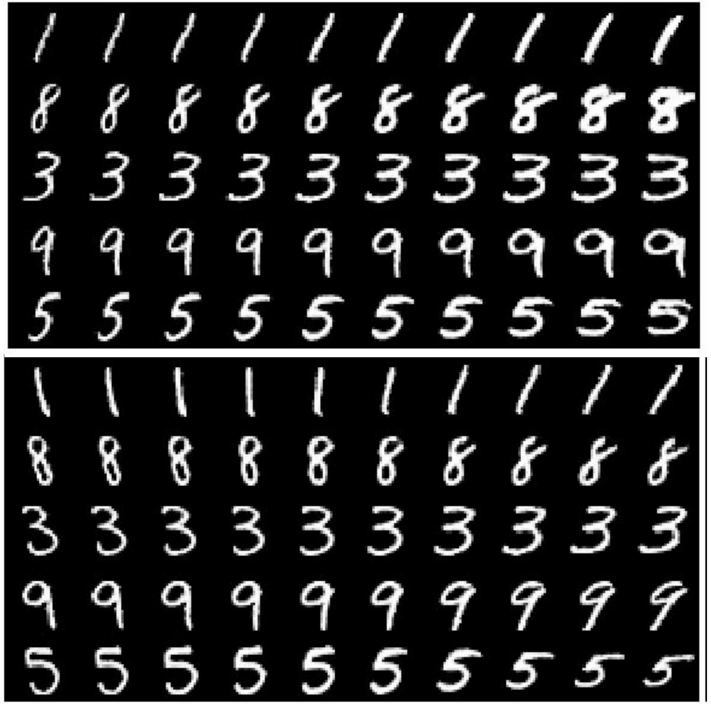
[1] Higgins, I., Matthey, L., Glorot, X., Pal, A., Uria, B., Blundell, C., Mohamed, S., and Lerchner, A. Early visual concept learning with unsupervised deep learning. arXiv preprint arXiv:1606.05579, 2016.

infoGAN

mutual information regularization

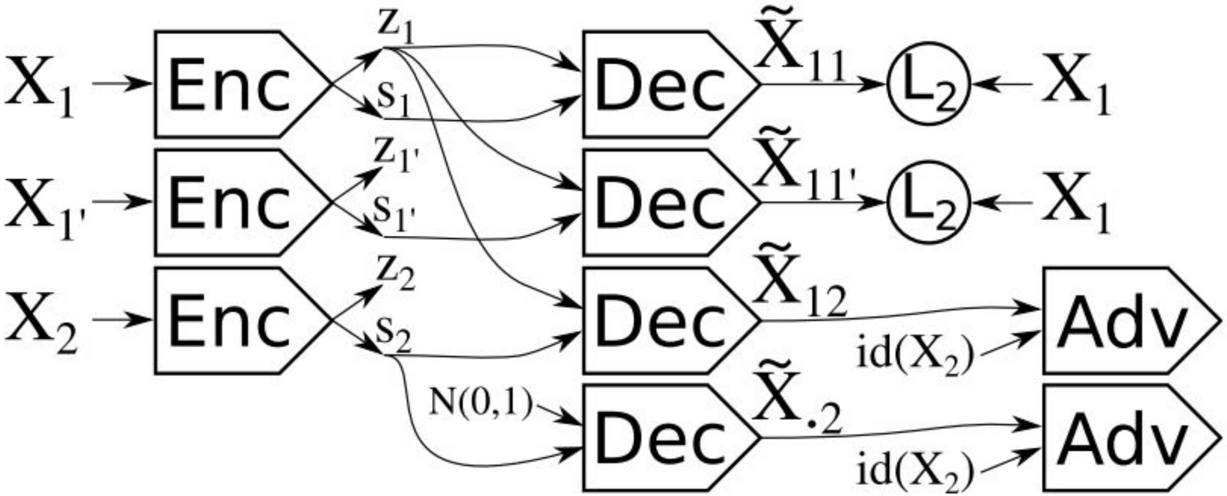
$$\min_G \max_D V_I(D,G) = V(D,G) - \underbrace{\lambda I(c; G(z,c))}_{\text{mutual information regularization}}$$





[1] Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In Advances in Neural Information Processing Systems, pp. 2172–2180, 2016.

Mathieu et al. 2016



$$\mathbb{E}_{q(z \mid x,s)}[-\log p_\theta(x \mid z, s)] + \mathrm{KL}(q(z \mid x, s) \mid\mid p(z)) + \lambda L_{\mathrm{gan}}$$

[1] Mathieu, M. F., Zhao, J. J., Zhao, J., Ramesh, A., Sprech- mann, P., and LeCun, Y. Disentangling factors of variation in deep representation using adversarial training. In Advances in Neural Information Processing Systems, pp. 5040–5048, 2016.

Disentangled sequential representation learning

Time-independent representation (i.e. for video sequence modeling: identity of the object in scene); time-dependent representations (time-varying position & orientation)

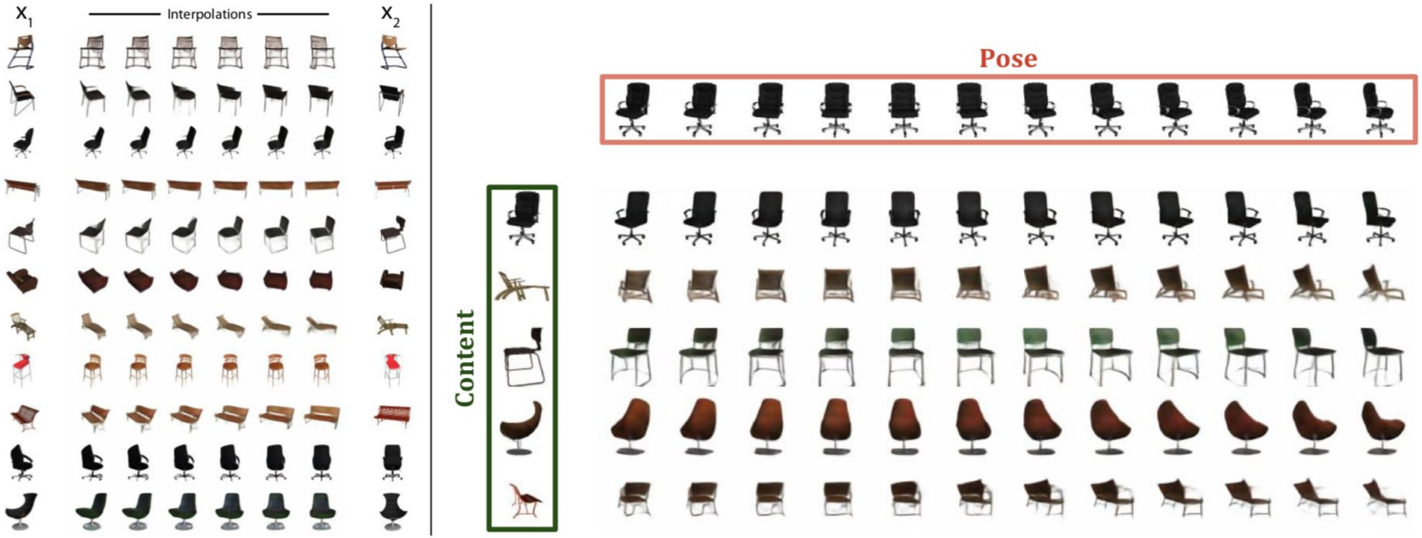


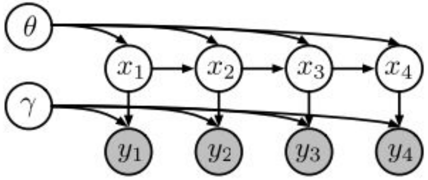
Fig. Left: example of linear interpolation in pose space; right: generated sequences according to extracted pose and content [1]

[1] Denton, Emily L. "Unsupervised learning of disentangled representations from video." *Advances in neural information processing systems*. 2017.

Sequential disentangled representation learning

- Structured VAEs, Johnson et al. (2016)
- Factorised VAEs, Deng et al. (2017)
- Factorised Hierarchical VAE, Hsu et al. (2017)
- Villegas et al. (2017)
- Denton & Birodkar (2017)

Structured VAE

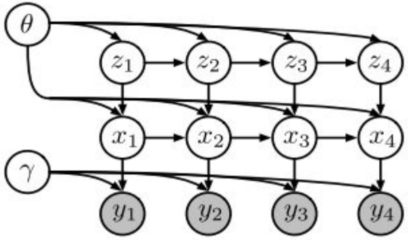


(c) Latent LDS

Latent state follows Gaussian linear dynamical system

NOTE: x is latent variable in the graphic model

$$x_n = Ax_{n-1} + Bu_n, \qquad u_n \overset{\text{iid}}{\sim} \mathcal{N}(0, I), \qquad A, B \in \mathbb{R}^{m \times m}$$



(d) Latent SLDS

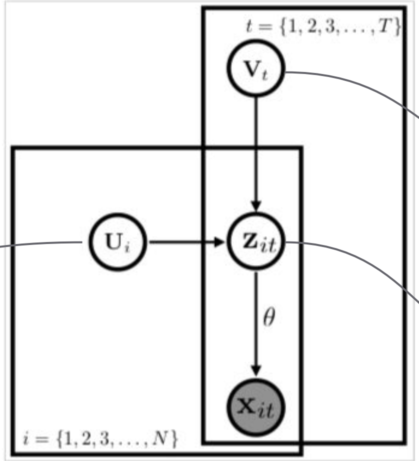
Latent state follow the hidden Markov model

NOTE: x is latent variable in the graphic model

$$z_n \mid z_{n-1}, \pi \sim \pi_{z_{n-1}}, \qquad x_n = A_{z_n} x_{n-1} + B_{z_n} u_n, \qquad u_n \overset{\text{iid}}{\sim} \mathcal{N}(0, I),$$

[1] Johnson, M., Duvenaud, D. K., Wiltschko, A., Adams, R. P., and Datta, S. R. Composing graphical models with neural networks for structured representations and fast inference. In Advances in neural information processing systems, pp. 2946–2954, 2016.

Factorised VAEs



"global" variable
individual encoding

"global" variable
dynamic encoding

"local" variable

$\mathbf{z}_{it} \sim \mathcal{N}(\mathbf{U}_i \circ \mathbf{V}_t, \mathbf{I})$

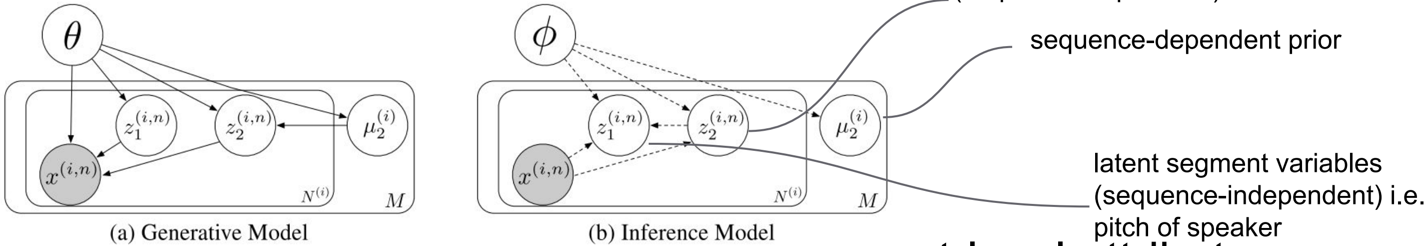
hadamard product

$$\mathcal{L}(\theta, \lambda, \mathbf{U}, \mathbf{V}) = \mathbb{E}_q[\log p_\theta(\mathbf{x}|\mathbf{z})] - KL(q_\lambda(\mathbf{z}|\mathbf{x})||\mathcal{N}(\mathbf{U} \circ \mathbf{V}, \mathbf{I})) + \log p(\mathbf{U}) + \log p(\mathbf{V})$$

[1] Deng, Z., Navarathna, R., Carr, P., Mandt, S., Yue, Y., Matthews, I., and Mori, G. Factorized variational autoencoders for modeling audience reactions to movies. In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, pp. 6014–6023. IEEE, 2017.

Factorised Hierarchy VAEs

- sequence-level attributes + segment-level attributes



(a) Generative Model (b) Inference Model

- discriminative objective to impose z2 to encode segment-level attributes

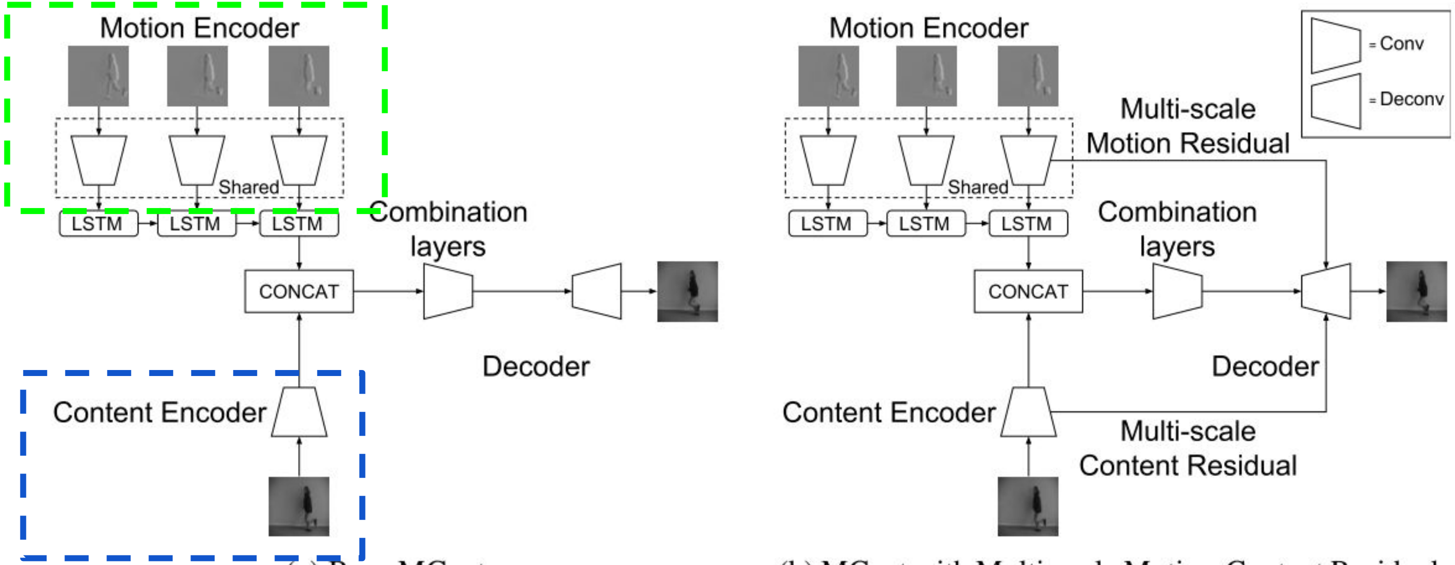
$$\log p(i|\boldsymbol{z}_2^{(i,n)}) = \log p(\boldsymbol{z}_2^{(i,n)}|i) - \log \sum_{j=1}^{M} p(\boldsymbol{z}_2^{(i,n)}|j) \quad (p(i) \text{ is assumed uniform})$$

$$:= \log p_\theta(\boldsymbol{z}_2^{(i,n)}|\tilde{\boldsymbol{\mu}}_2^{(i)}) - \log\Big(\sum_{j=1}^{M} p_\theta(\boldsymbol{z}_2^{(i,n)}|\tilde{\boldsymbol{\mu}}_2^{(j)})\Big),$$

[1] Hsu, W.-N., Zhang, Y., and Glass, J. Unsupervised learning of disentangled and interpretable representations from sequential data. In Advances in neural information processing systems, pp. 1876–1887, 2017

Villegas et al. (2017)

generate future prediction $\hat{\mathbf{x}}_{t+1}$ given $\mathbf{x}_{1:t}$.



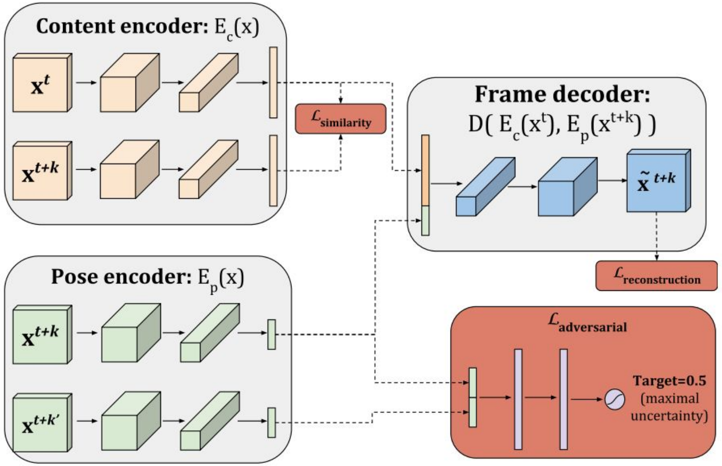
(a) Base MCnet (b) MCnet with Multi-scale Motion-Content Residuals

[1] Villegas, R., Yang, J., Hong, S., Lin, X., and Lee, H. Decomposing motion and content for natural video sequence prediction. In ICLR, 2017.

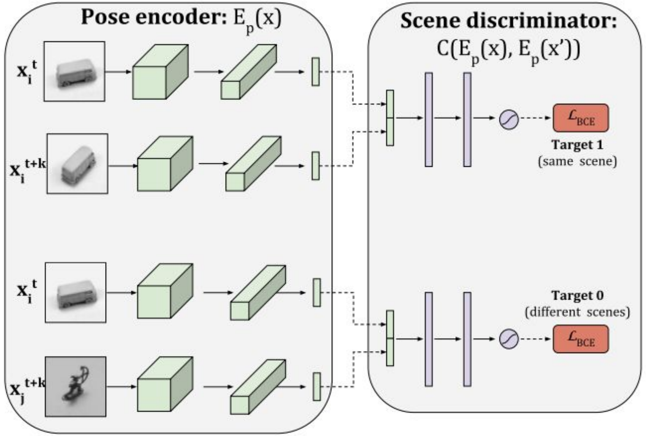
Denton et al. (2017)

DRNET

2 encoders - pose encoder Ep + content encoder Ec
Frame Decoder D - map content encoding + pose encoding to prediction



Scene Discriminator C to predict pose vectors come from the same scenes



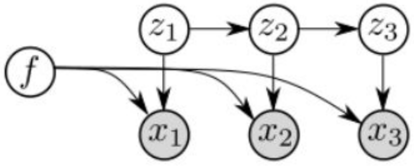
[1] Denton, Emily L. "Unsupervised learning of disentangled representations from video." *Advances in neural information processing systems*. 2017.

Disentangled Sequential Autoencoder

- ★ Disentanglement is achieved by the design of graphic model
  - invariant latent variables represents ***content***
  - variant latent variables represents ***dynamical information***
- ★ New metric to verify disentanglement
  - KL similarity measure
- ★ Efficient encoding
  - smaller dimensionality of variant latent variables
  - data efficient
- ★ Controlled sequence generation
  - manipulate sequence with random dynamics + fixed content or fixed dynamics + random content

Disentangled Sequential Autoencoder

Generative model



(a) generator

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}, \boldsymbol{f}) = \underbrace{p_{\boldsymbol{\theta}}(\boldsymbol{f})}_{\text{Time-invariant}} \underbrace{\prod_{t=1}^{T} \overbrace{p_{\boldsymbol{\theta}}(\boldsymbol{z}_t | \boldsymbol{z}_{<t})}^{\text{Transition}} \overbrace{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t | \boldsymbol{z}_t, \boldsymbol{f})}^{\text{Emission}}}_{\text{Time-variant}}$$

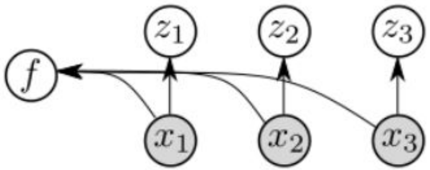
ELBO

$$\mathbb{E}_{p_{\mathcal{D}}(\boldsymbol{x}_{1:T})}\left[\mathbb{E}_{q_\phi}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}, \boldsymbol{f})}{q_\phi(\boldsymbol{z}_{1:T}, \boldsymbol{f} | \boldsymbol{x}_{1:T})}\right]\right]$$

Disentangled Sequential Autoencoder

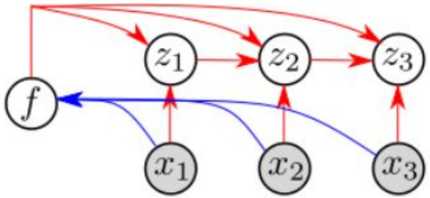
Variational Inference model (recognition model)

partially factorized q



$$q_\phi(\boldsymbol{z}_{1:T}, \boldsymbol{f} | \boldsymbol{x}_{1:T}) = q_\phi(\boldsymbol{f} | \boldsymbol{x}_{1:T}) \prod_{t=1}^{T} q_\phi(\boldsymbol{z}_t | \boldsymbol{x}_t)$$

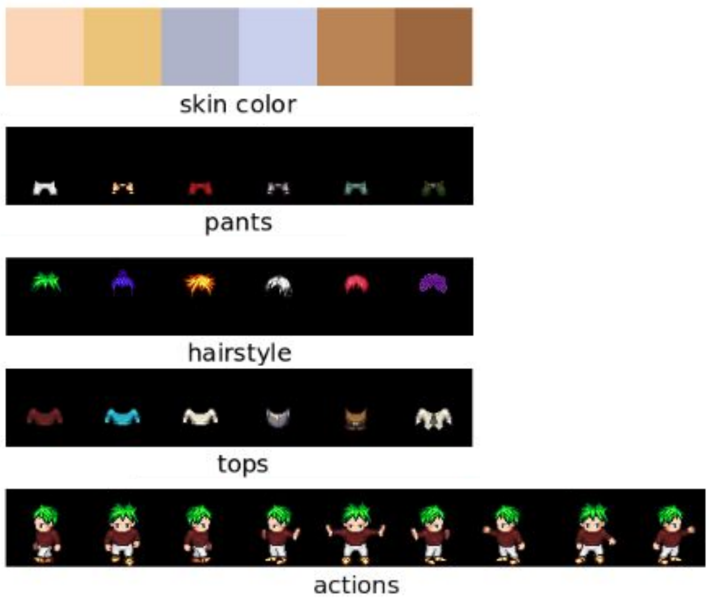
full factorized q



$$q_\phi(\boldsymbol{z}_{1:T}, \boldsymbol{f} | \boldsymbol{x}_{1:T}) = q_\phi(\boldsymbol{f} | \boldsymbol{x}_{1:T}) q_\phi(\boldsymbol{z}_{1:T} | \boldsymbol{f}, \boldsymbol{x}_{1:T})$$

Experiments: Sprites video sequences

- Controllable attribute variants
- 1296 time-invariant characters (1000 for training/validation; rest for testing)
- T = 8 sequences; no label provided for training



Qualitative analysis

Unconditional generation

- synthesize sequence by sampling latent variables from prior and decoding them
- fixing dynamics or f to generate controlled sequence

(a) random test data sequences (b) reconstruction (c) reconstruction with randomly sampled $\boldsymbol{f}$ (d) reconstruction with randomly sampled $\boldsymbol{z}_{1:T}$

Qualitative analysis

Conditional generation

- generating sequence given $\boldsymbol{x}_{1:T}$ sampling $\boldsymbol{f} \sim q(\boldsymbol{f}|\boldsymbol{x}_{1:T})$ and $\boldsymbol{z}_{1:T} \sim p(\boldsymbol{z}_{1:T})$

(a) random test data sequences (e) reconstruction with swapped encoding $\boldsymbol{f}$ (f) reconstruction with swapped encoding $\boldsymbol{z}_{1:T}$

Feature swapping

- given two sequences $\boldsymbol{x}^a_{1:T}$ and $\boldsymbol{x}^b_{1:T}$
- sampling $\boldsymbol{f}^a \sim q(\boldsymbol{f}|\boldsymbol{x}^a_{1:T})$ sampling $\boldsymbol{z}^b_{1:T} \sim q(\boldsymbol{z}_{1:T}|\boldsymbol{x}^b_{1:T})$

(g) generated sequences with fixed $\boldsymbol{f}$ (h) generated sequences with fixed $\boldsymbol{z}_{1:T}$

Quantitative analysis

- Supervised-learning classifier of each attributes trained on labelled frame on the generated sequences to provide probability of frame in original sequence and reconstructed one respectively
- Quantitative measures:
  - disagreement: predicted max probability $\max_i[\boldsymbol{p}_{recon}(i)] \neq \max_i[\boldsymbol{p}_{data}(i)]$
  - KL-recon: $\mathrm{KL}[\boldsymbol{p}_{recon}||\boldsymbol{p}_{data}]$
  - KL-random: $\mathrm{KL}[\boldsymbol{p}_{random}||\boldsymbol{p}_{data}]$

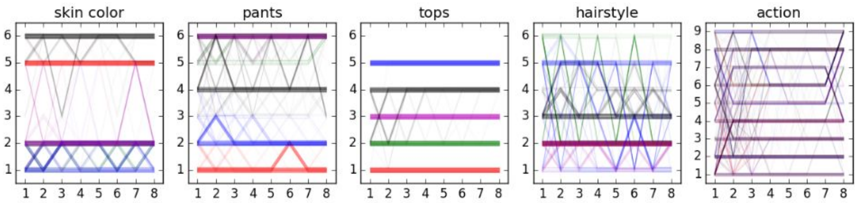
$$\boldsymbol{p}_{random} = (1/N_{\text{class}}, ..., 1/N_{\text{class}})$$

attributes	disagreement	KL-recon	KL-random
skin colour	3.98%	0.7847	8.8859
pants	1.82%	0.3565	8.9293
tops	0.34%	0.0647	8.9173
hairstyle	0.06%	0.0126	8.9566
action	8.11%	0.9027	13.7510

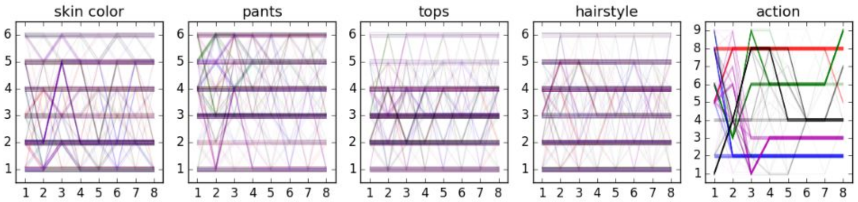
Quantitative analysis

Evaluate the static attributes of generated sequences

- sample 200 sequences with ***same f*** but ***different latent dynamics*** from generator
    - most attributes are preserved over time
    - some trajectory for attributes drift away from majority class i.e. hairstyle
- sample sequences with ***same dynamics***
    - trajectory diverse on static attributes
    - “almost” constant in action
    - “multi-modality” in action domain



(a) Trajectory plots on the generated sequences with shared $\boldsymbol{f}$.



(b) Trajectory plots on the generated sequences with shared $\boldsymbol{z}_{1:T}$.

Speech data: TIMIT

- 6300 utterances with 10 sententces from 630 speakers (70% male + 30% female)
- split to 200ms subsequences; pre-processing to 200 dimensional log-magnitude spectrum of sub-sequences of every 10ms
- $T = 20$
- speaker identity (static representations) + content of speech (dynamic representations)

Voice conversion on TIMIT

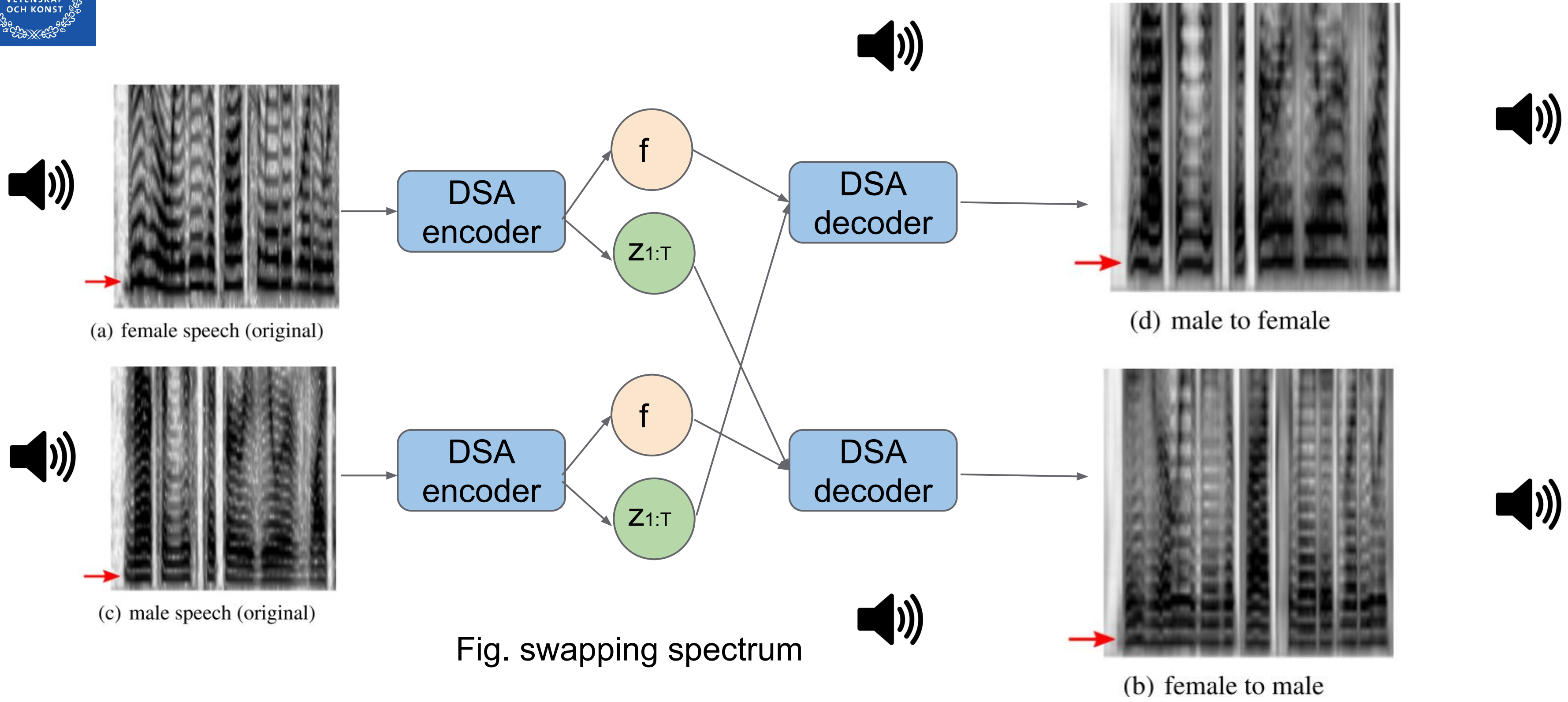


Fig. swapping spectrum

*Sound reconstructed by Griffin-Lim algorithm from spectrogram

Speech data: TIMIT

Evaluation - speaker verification

- identity confirmed by cosine similarity of “features”
- equal error rate EER (where false rejection = false acceptance rate)
- MC estimator to approximate mean of “features”

$$\boldsymbol{\mu}_{\boldsymbol{f}} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\mu}_{\boldsymbol{f}^n}, \quad \boldsymbol{\mu}_{\boldsymbol{f}^n} = \mathbb{E}_{q(\boldsymbol{f}^n|\boldsymbol{x}^n_{1:T})}[\boldsymbol{f}^n],$$

$$\boldsymbol{\mu}_{\boldsymbol{z}} = \frac{1}{TN}\sum_{t=1}^{T}\sum_{n=1}^{N}\boldsymbol{\mu}_{\boldsymbol{z}^n_t}, \quad \boldsymbol{\mu}_{\boldsymbol{z}^n_t} = \mathbb{E}_{q(\boldsymbol{z}^n_t|\boldsymbol{x}^n_{1:T})}[\boldsymbol{z}^n_t]$$

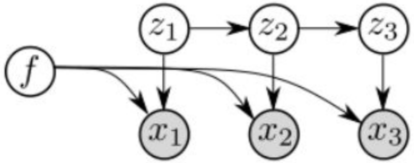
Speech data: TIMIT

model	feature	dim	EER
-	i-vector	200	9.82%
FHVAE ($\alpha = 0$)	μ_2	16	5.06%
FHVAE ($\alpha = 10$)	μ_2	32	2.38%
	μ_1	32	22.47%
factorised q	μ_f	16	4.78%
	μ_z	16	17.84%
factorised q	μ_f	64	4.94%
	μ_z	64	17.49%
full q	μ_f	16	5.64%
	μ_z	16	19.20%
full q	μ_f	64	4.82%
	μ_z	64	18.89%

- lower EER → more similar
- FHVAE sensitive to “tuning” disriminative objective trade-off
- μ_f performs better than baseline
- μ_z does not contain much information about identity
- structured inference network improve disentanglement

Stochastic VS deterministic dynamics

Comparing to deteriministic dynamics generative model

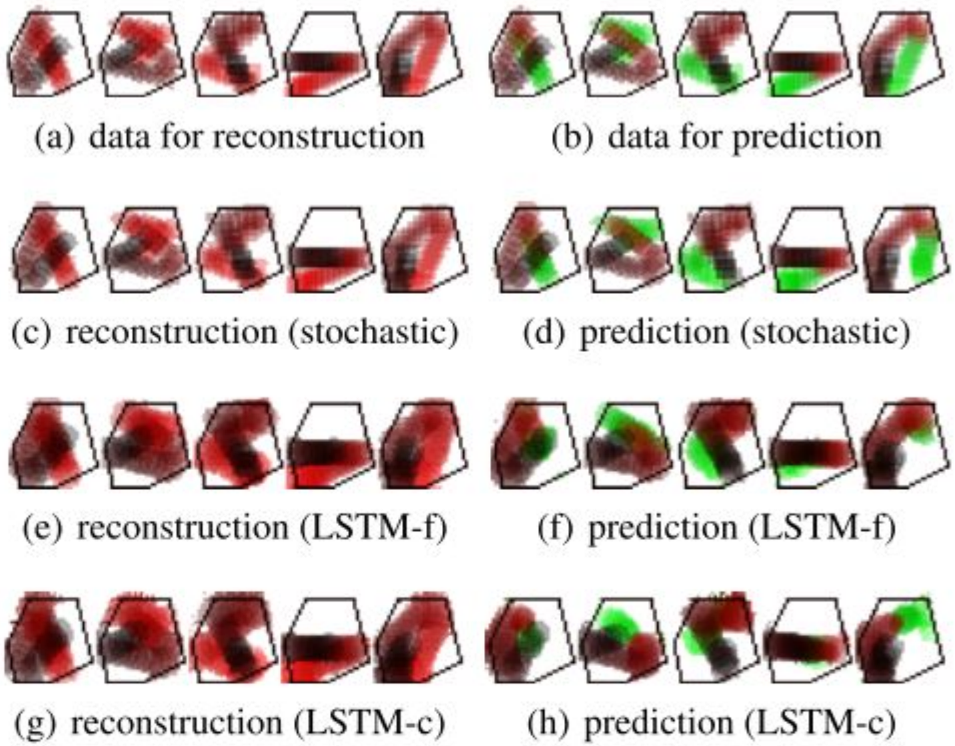


(a) generator

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}, \boldsymbol{f}) = \underbrace{p_{\boldsymbol{\theta}}(\boldsymbol{f})}_{\text{Time-invariant}} \underbrace{\prod_{t=1}^{T} \overbrace{p_{\boldsymbol{\theta}}(\boldsymbol{z}_t|\boldsymbol{z}_{<t})}^{\text{Transition}} \overbrace{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{z}_t, \boldsymbol{f})}^{\text{Emission}}}_{\text{Time-variant}}$$

$$p(\boldsymbol{x}_{1:T}, \boldsymbol{z}, \boldsymbol{f}) = p(\boldsymbol{f})p(\boldsymbol{z}) \underbrace{\prod_{t=1}^{T} p(\boldsymbol{x}_t|\boldsymbol{z}, \boldsymbol{f})}_{\text{model by } \textbf{\textit{deteriministic}} \text{ RNN + NN}(h_t,f)}$$

Stochastic VS deterministic dynamics



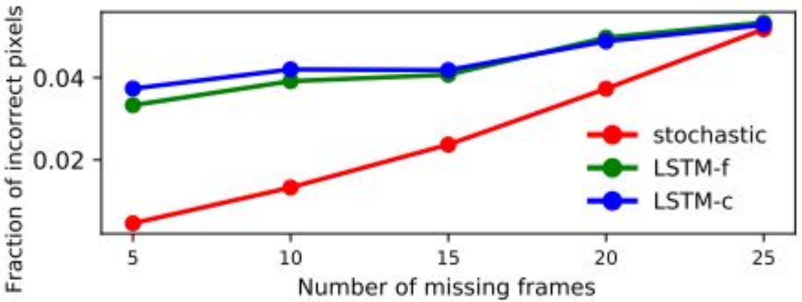
LSTM-f

$$h_0 = z, h_t = \mathrm{LSTM}(h_{t-1})$$

LSTM-c, similar to FHVAE

$$h_0 = 0, h_t = \mathrm{LSTM}(h_{t-1}, z))$$

stochastic transition model → realistic dynamics



Symmary

- proposed simple generative model disentangles “local” time-dependent features from “global” time-independent features
- empirically show applicable in speech synthesis and videa generation with controlled latent features
- stochastic RNN is more efficient than deterministic one