Disentangled Recurrent Wasserstein AutoEncoder

Jun Han

Dartmouth College

June 17, 2019

Efficient Representation Learning of Sequential Data

Goal

- Disentangle the static (content) latent code and dynamic (motion) latent code of of sequential data
- Unconditionally generate new sequences

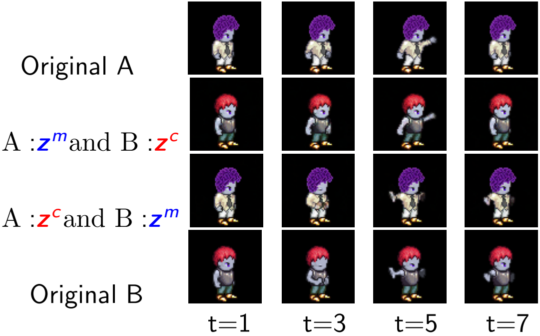
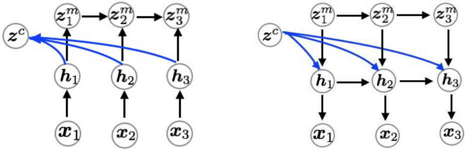


Figure: Illustration of disentangling static and dynamic parts of two sequences (length=8) using **our model** on Sprites. Swap content $\boldsymbol{z}^c$ and motion $\boldsymbol{z}^m$.

Outline

- Structures of Generative Model & Inference Model
- Disentangled Recurrent Wasserstein Autoencoder
- Empirical Results
- Reinforced Model
- Conclusion

Generative Model & Inference Model

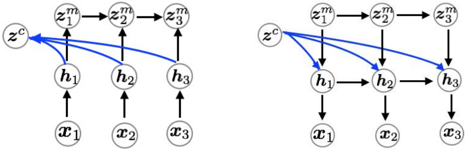


(a) Inference Model (b) Generative Model

Figure: Structures of sequential probabilistic models designed.

Notations: $\boldsymbol{x}_{1:T}$ be sequence. $\boldsymbol{z}_{1:T}$ be latent code. $\boldsymbol{z}_t = (\boldsymbol{z}^c, \boldsymbol{z}_t^m)$, content $\boldsymbol{z}^c$ and motion $\boldsymbol{z}_t^m$. Hidden States $\boldsymbol{h}_t$.

Generative Model & Inference Model



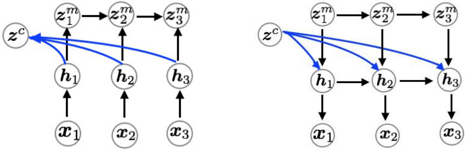
(a) Inference Model (b) Generative Model

Figure: Structures of sequential probabilistic models designed.

Inference Model:

$$q_\phi(\boldsymbol{z}^c, \boldsymbol{z}^m_{1:T} | \boldsymbol{x}_{1:T}) = q_\phi(\boldsymbol{z}^c | \boldsymbol{x}_{1:T}) \prod_{t=1}^{T} q_\phi(\boldsymbol{z}^m_t \mid \boldsymbol{z}^m_{<t}, \boldsymbol{x}_t)$$

Generative Model & Inference Model



(a) Inference Model (b) Generative Model

Figure: Structures of sequential probabilistic models designed.

Inference Model:

$$q_\phi(\boldsymbol{z}^c, \boldsymbol{z}^m_{1:T}|\boldsymbol{x}_{1:T}) = q_\phi(\boldsymbol{z}^c|\boldsymbol{x}_{1:T}) \prod_{t=1}^{T} q_\phi(\boldsymbol{z}^m_t \mid \boldsymbol{z}^m_{<t}, \boldsymbol{x}_t)$$

Generative Model:

$$p(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}) = p(\boldsymbol{z}^c) \prod_{t=1}^{T} p_\psi(\boldsymbol{z}^m_t|\boldsymbol{z}^m_{<t}) p_\theta(\boldsymbol{x}_t|\boldsymbol{z}^m_t, \boldsymbol{z}^c)$$

Disentangled Recurrent Wasserstein AutoEncoder (R-WAE)

- The optimal transport for distributions with $X_{1:T}$ and $Y_{1:T}$ is:

$$W(P_{\mathcal{D}}, P_G) := \inf_{\Gamma \sim \mathcal{P}(X_{1:T} \sim P_{\mathcal{D}}, Y_{1:T} \sim P_G)} \mathbb{E}_{(X_{1:T}, Y_{1:T}) \sim \Gamma}[c(X_{1:T}, Y_{1:T})],$$

$\mathcal{P}(X_{1:T} \sim P_{\mathcal{D}}, Y_{1:T} \sim P_G)$ is a set of all joint distributions with marginals $P_{\mathcal{D}}$ and P_G respectively.

Disentangled Recurrent Wasserstein AutoEncoder (R-WAE)

- With $P(X_t|Z_t)$ and $Y_t = G(Z_t)$, we derive

$$W(P_{\mathcal{D}}, P_G) = \inf_{\underbrace{Q_{Z^c} = P_{Z^c};\, Q_{Z^m_{1:T}} = P_{Z^m_{1:T}}}_{\text{Constraints}}} \sum_t \mathbb{E}_{P_{\mathcal{D}}} \mathbb{E}_Q [\, \underbrace{c(X_t, G(Z_t))}_{\text{Reconstructed Cost}} \,],$$

Disentangled Recurrent Wasserstein AutoEncoder (R-WAE)

- With $P(X_t|Z_t)$ and $Y_t = G(Z_t)$, we derive

$$W(P_{\mathcal{D}}, P_G) = \inf_{\underbrace{Q_{Z^c} = P_{Z^c};\, Q_{Z^m_{1:T}} = P_{Z^m_{1:T}}}_{\text{Constraints}}} \sum_t \mathbb{E}_{P_{\mathcal{D}}} \mathbb{E}_Q [\, \underbrace{c(X_t, G(Z_t))}_{\text{Reconstructed Cost}} \,],$$

- We have an upper bound,

$$W(P_{\mathcal{D}}, P_G) \leq \inf_{Q \in \mathcal{S}} \sum_t \mathbb{E}_{P_{\mathcal{D}}} \mathbb{E}_Q [c(X_t, G(Z_t))],$$

where the subset $\mathcal{S} = \{Q : Q_{Z^c} = P_{Z^c}, Q_{Z^m_1} = P_{Z^m_1}, Q_{Z^m_t|Z^m_{<t}} = P_{Z^m_t|Z^m_{<t}}\}$.

Loss Function in R-WAE

- With divergence $\mathbb{D}$ between two distributions,

$$\sum_{t=1}^{T} \mathbb{E}_{q(\mathbf{z}_t|\mathbf{z}_{<t},\mathbf{x}_t)}[c(\mathbf{x}_t, G(\mathbf{z}_t))] + \beta_1 \, \mathbb{D}(q_{\mathbf{z}^c}, p_{\mathbf{z}^c}) + \beta_2 \sum_{t=1}^{T} \mathbb{D}(q_{\mathbf{z}_t^m|\mathbf{z}_{<t}^m}, p_{\mathbf{z}_t^m|\mathbf{z}_{<t}^m}),$$

where G is a decoder function.

Loss Function in R-WAE

- With divergence $\mathbb{D}$ between two distributions,

$$\sum_{t=1}^{T} \mathbb{E}_{q(\boldsymbol{z}_t|\boldsymbol{z}_{<t},\boldsymbol{x}_t)}[c(\boldsymbol{x}_t, G(\boldsymbol{z}_t))] + \beta_1\, \mathbb{D}(q_{\boldsymbol{z}^c}, p_{\boldsymbol{z}^c}) + \beta_2 \sum_{t=1}^{T} \mathbb{D}(q_{\boldsymbol{z}_t^m|\boldsymbol{z}_{<t}^m}, p_{\boldsymbol{z}_t^m|\boldsymbol{z}_{<t}^m}),$$

where G is a decoder function.

- R-WAE(GAN): $\mathbb{D}_{\mathrm{JS}}(q_{\boldsymbol{z}^c}, p_{\boldsymbol{z}^c})$; $\mathrm{MMD}_k(q_{\boldsymbol{z}_t^m|\boldsymbol{z}_{<t}^m}, p_{\boldsymbol{z}_t^m|\boldsymbol{z}_{<t}^m})$.

- R-WAE(MMD): Scaled $\mathrm{MMD}_{k_\gamma}(q_{\boldsymbol{z}^c}, p_{\boldsymbol{z}^c})$; $\mathrm{MMD}_k(q_{\boldsymbol{z}_t^m|\boldsymbol{z}_{<t}^m}, p_{\boldsymbol{z}_t^m|\boldsymbol{z}_{<t}^m})$.

Results on Stochastic Moving MNIST Dataset

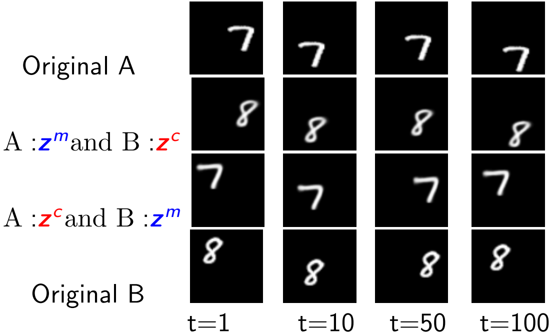


Figure: Illustration of disentangling motions $\{z_t^m\}$ and contents $\{z^c\}$ of two sequences with $T = 100$.

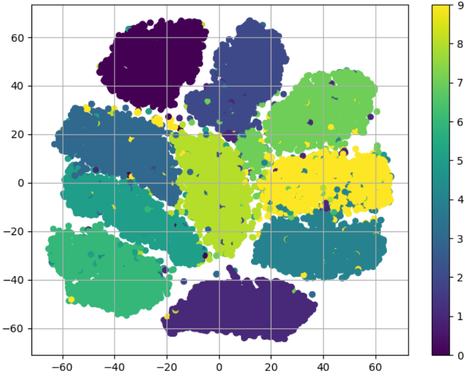


Figure: 2D manifold of content $\{z^c\}$ encoded from R-WAE(MMD) by t-SNE.

Quantitative Results on Benchmark Datasets

Datasets / Methods	Sprites				MNIST
	actions	skin color	pants	tops	digits
DS-VAE(S)	8.11%	3.98%	1.82%	0.34%	3.31%
R-WAE(S)	**5.83%**	**2.45%**	1.73%	0.32%	**1.78%**
DS-VAE(C)	10.37%	4.86%	3.62%	0.53%	4.26%
R-WAE(C)	**7.72%**	**3.31%**	3.20%	0.40%	**2.84%**

Figure: Comparison of averaged classification error when fixing one encoded attribute and randomly sampling others for simple/complex structures on Sprites and SM-MNIST. Lower values in the table indicate better disentanglement.

Unconditional Sequence Generation

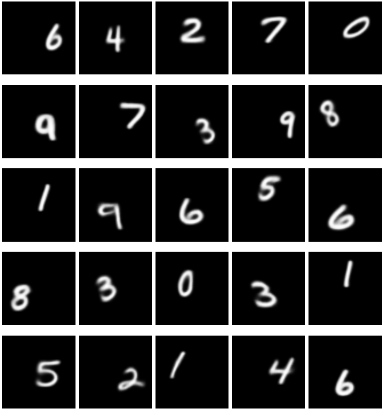


Figure: Generated video (length=20) in R-WAE(MMD) are randomly taken from generated samples with $T = 100$. See video with Adobe.

Baseline

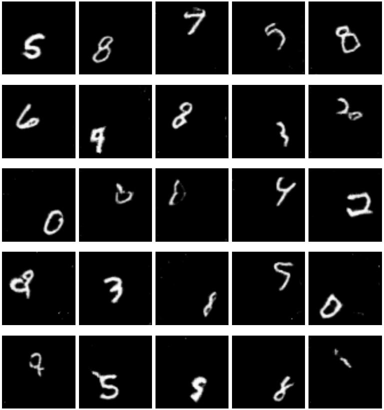


Figure: MoGoGAN(length=20). See video with Adobe.

Baseline

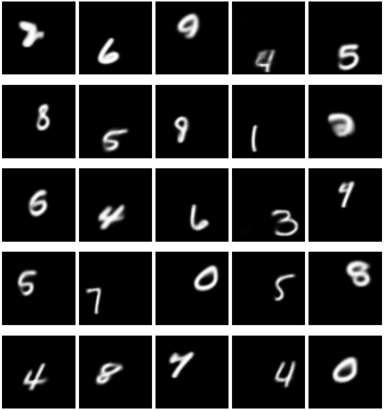
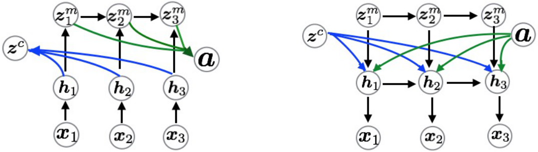


Figure: DS-VAE(length=20). See video with Adobe.

Reinforced Model



(a) Reinforced Inference (b) Reinforced Generative

Figure: Structures of sequential probabilistic models.

- Additional inferece model: $q_\phi(\boldsymbol{a}|\boldsymbol{x}_{1:T}, \boldsymbol{z}^m_{1:T})$, $\boldsymbol{a}$ is one-hot vector inferred by an LSTM through motions $\boldsymbol{z}^m_{1:T}$.
- Optimize with regularization $\mathbb{D}_{\mathrm{KL}}(q_\phi(\boldsymbol{a}|\boldsymbol{x}_{1:T}, \boldsymbol{z}^m_{1:T}), p(\boldsymbol{a}))$ (prior $p(\boldsymbol{a})$).

Results on MUG facial dataset

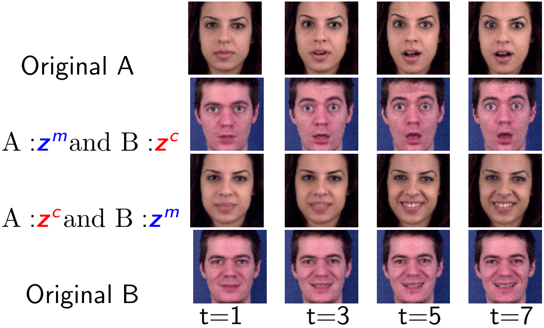


Figure: Illustration of disentangling motions $\{z_t^m\}$ and contents $\{z^c\}$ of two video sequences with $T = 10$.

Results on MUG facial dataset

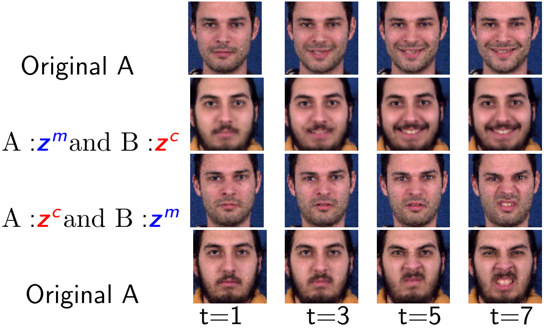


Figure: Illustration of disentangling motions $\{z_t^m\}$ and contents $\{z^c\}$ of two video sequences with $T = 10$.

Quantitative Results on MUG Facial Dataset

Methods / Metrics	MocoGAN	DS-VAE(NA)	DS-VAE(W)	R-WAE(MMD)	R-WAE(GAN)
Accuracy ↑	75.50%	66.73%	82.84 %	88.62%	**90.15%**
Intra-entropy ↓	0.26	0.28	0.23	0.17	**0.15**
Inter-entropy ↑	1.78	1.77	1.78	**1.79**	**1.79**
IS ↑	4.60	4.44	4.71	5.05	**5.16**

Figure: Quantitative results on generated samples. R-WAE(MMD) and R-WAE(GAN) are ours.

Unconditional Sequence Generation

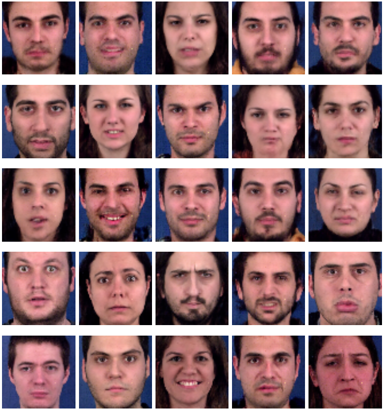


Figure: Generated video (length=10) in R-WAE(GAN). See video with Adobe.

Baseline

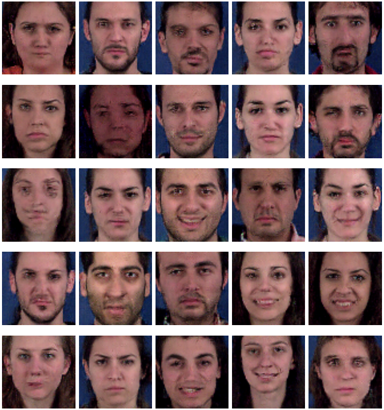


Figure: MoGoGAN(length=10). See video with Adobe.

Baseline

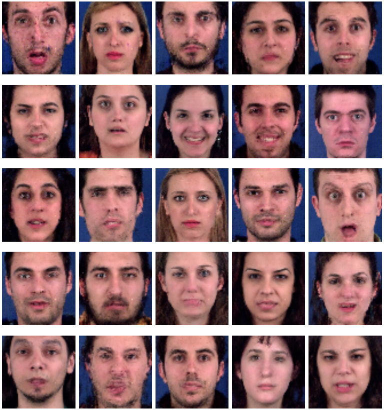


Figure: DS-VAE(length=10). See video with Adobe.

Conclusion

1. First sequential Wasserstein AutoEncoder framework

2. Two efficient methods by effective regularizers

3. A reinforced method to improve disentanglement