VAE 4: Challenging Assumptions behind Disentanglement

Shengyu Feng

Sep 28, 2021

Outline

Disentanglement review

Disentanglement is impossible without inductive bias

Experimental results

Future directions

Outline

Disentanglement review

Disentanglement is impossible without inductive bias

Experimental results

Future directions

Disentanglement

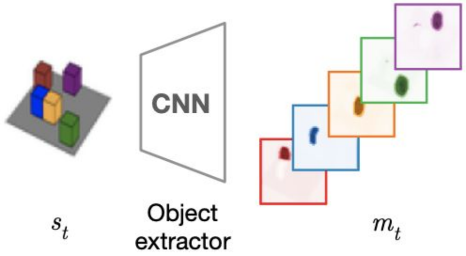
No formal definition, briefly, separating the distinct, informative factors of variations in the data

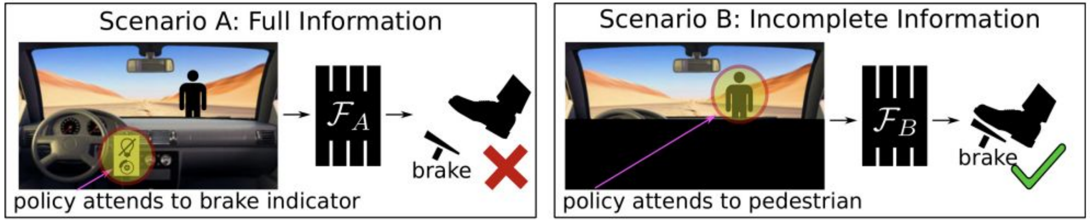
Entanglement: a random variable X = [a+b, a-b, a+2b, b]

Disentanglement: Z = [a,b]

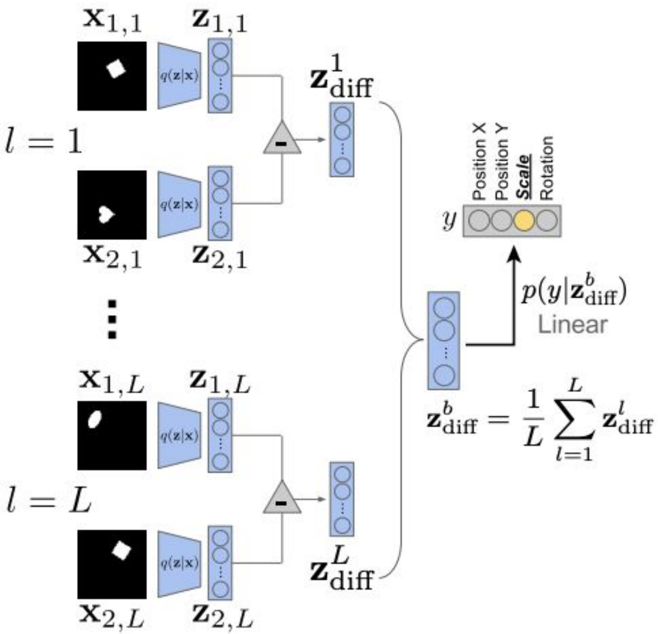
Why is it useful?

- Customized generation
- Causality: $p(x)\underline{p(y|x)}$, $p(y)p(x|y)$
- Robot tasks





β-VAE



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

Outline

Disentanglement review

Disentanglement is impossible without inductive bias

Experimental results

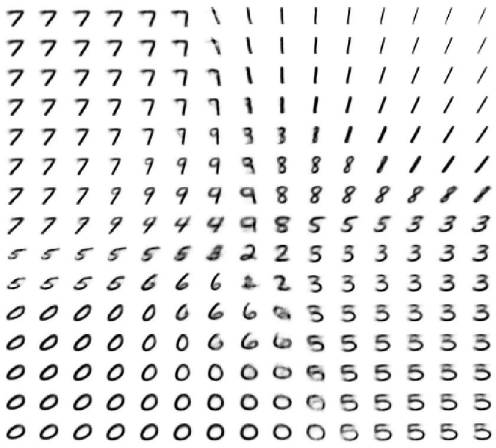
Future directions

Disentanglement representations is impossible without inductive bias

For any disentanglement representation, we can find infinite many equivalent representations. It's impossible to get the disentanglement representation without inductive bias.

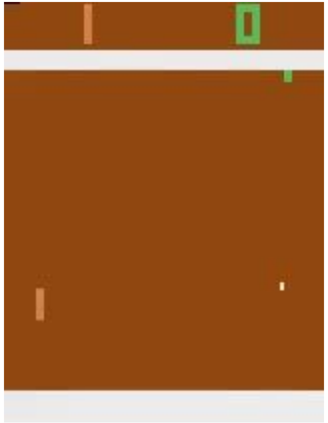
Prior knowledge

Theorem 1. *For* $d > 1$, *let* $\mathbf{z} \sim P$ *denote any distribution which admits a density* $p(\mathbf{z}) = \prod_{i=1}^{d} p(\mathbf{z}_i)$. *Then, there exists an infinite family of bijective functions* $f : \mathrm{supp}(\mathbf{z}) \to \mathrm{supp}(\mathbf{z})$ *such that* $\frac{\partial f_i(\mathbf{u})}{\partial u_j} \neq 0$ *almost everywhere for all* i *and* j *(i.e.,* $\mathbf{z}$ *and* $f(\mathbf{z})$ *are completely entangled) and* $P(\mathbf{z} \leq \mathbf{u}) = P(f(\mathbf{z}) \leq \mathbf{u})$ *for all* $\mathbf{u} \in \mathrm{supp}(\mathbf{z})$ *(i.e., they have the same marginal distribution).*

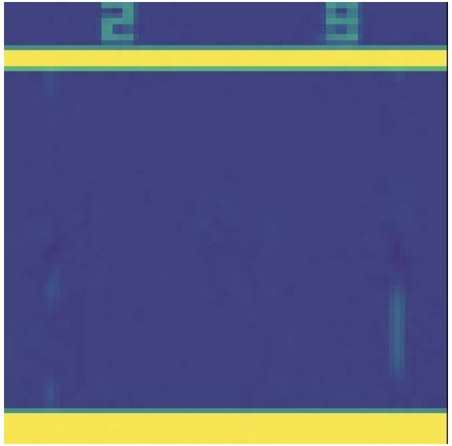


β-VAE generated results (on Atari Pong)

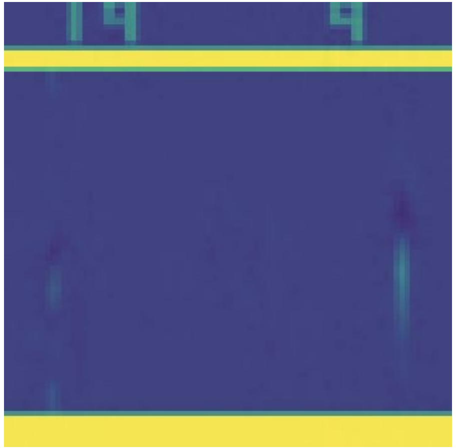
Data



Generated: $z = [0, 0, \dots, 0]$



Generated: $z = [0, 0, 0, 1, \dots, 0]$



Outline

Disentanglement review

Disentanglement is impossible without inductive bias

Experimental results

Future directions

Experiments

Methods:
- β-VAE
- AnnealedVAE
- FactorVAE
- β-TCVAE
- DIP-VAE-I
- DIP-VAE-II

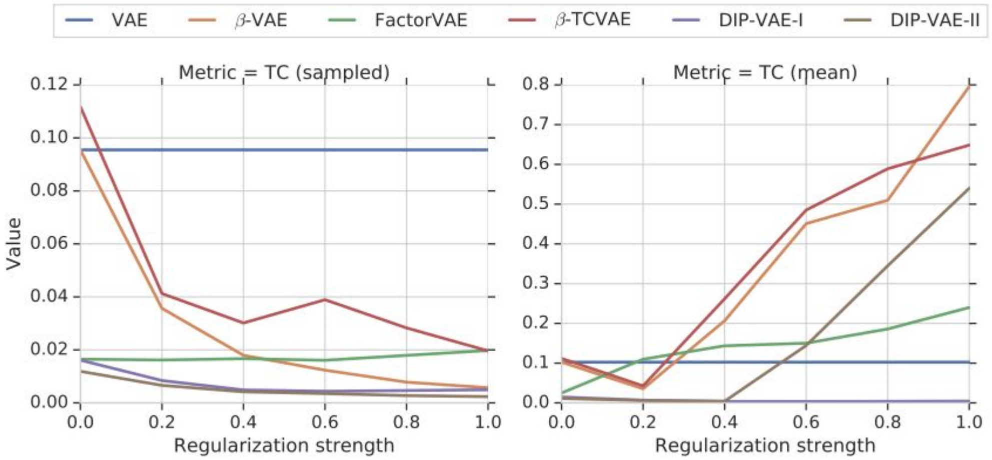
Metrics:
- β-VAE score
- FactorVAE score
- Mutual Information Gap (MIG)
- DCI Disentanglement
- Modularity
- SAP score

Datasets
- dSprites
- Cars3D
- SmallNORB
- Shapes3D
- Color-dSprites
- Noisy-dSprites
- Scream-dSprites

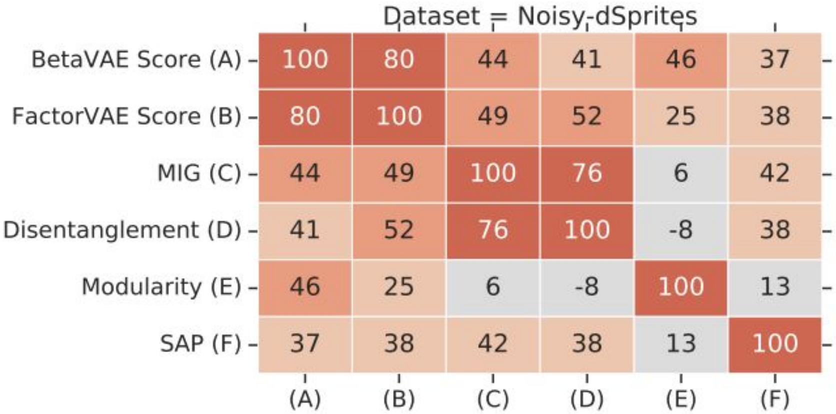
Mean representation of the latent variables are correlated

Common practice in latent representation:

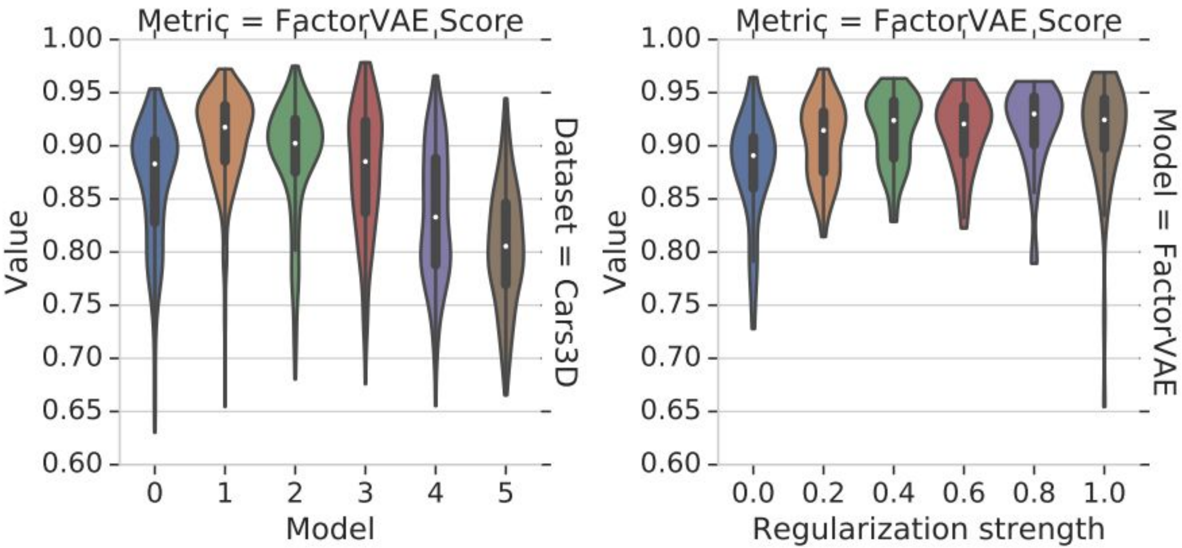
- Training: $z \sim N(\mu(z), \boldsymbol{\sigma}(z))$
- Testig: $\mu(z)$



Most disentanglement metrics are correlated

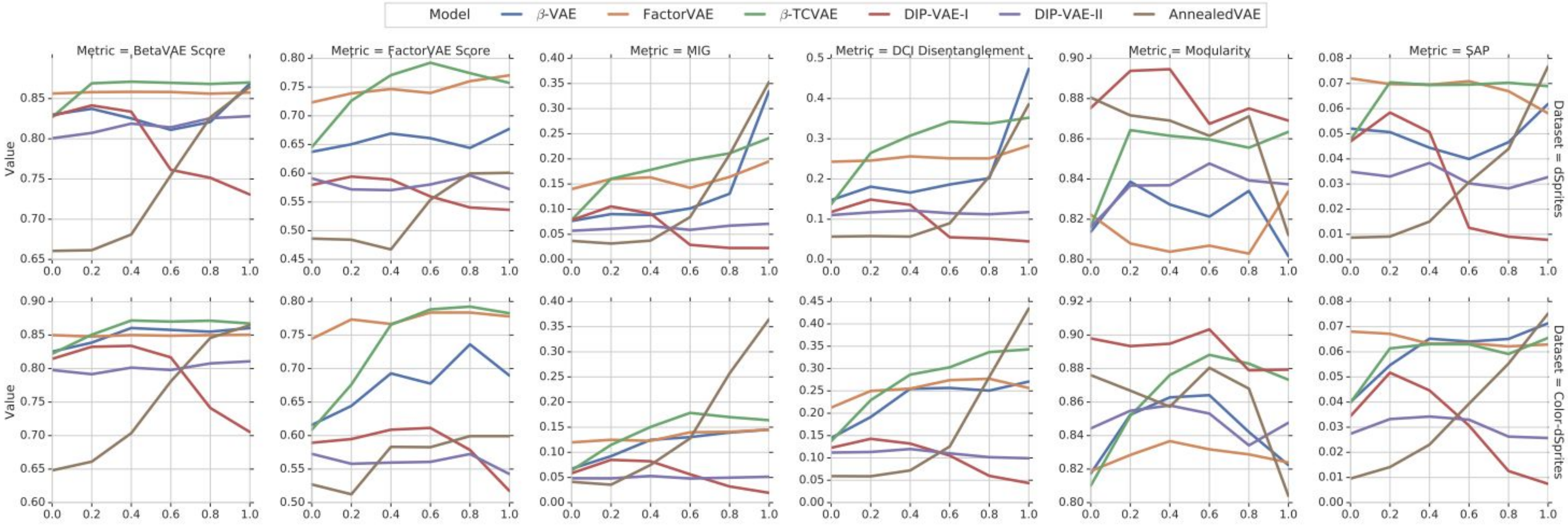


Hyperparameters and random seeds are important than models



Recipes for hyperparameter selection

1. Strategy should not depend on the score, which needs labels and the full generative model
2. No hyperparameters and models work the best every time



Unsupervised loss vs. disentanglement scores

Dataset = dSprites

	BetaVAE Score	FactorVAE Score	MIG	DCI Disentanglement	Modularity	SAP
Reconstruction	2	3	37	28	-5	11
TC (sampled)	-1	22	5	11	-12	10
KL	-57	-51	-73	-78	31	-56
ELBO	-8	-5	28	17	2	4

Dataset = Color-dSprites

	BetaVAE Score	FactorVAE Score	MIG	DCI Disentanglement	Modularity	SAP
Reconstruction	12	18	46	34	-6	20
TC (sampled)	11	35	7	18	-8	13
KL	-68	-53	-73	-80	21	-61
ELBO	5	13	40	26	-3	15

Dataset = Noisy-dSprites

	BetaVAE Score	FactorVAE Score	MIG	DCI Disentanglement	Modularity	SAP
Reconstruction	-4	-12	37	27	4	6
TC (sampled)	11	22	-2	6	-11	11
KL	28	22	-43	-42	32	-6
ELBO	13	-0	31	20	19	9

Dataset = Scream-dSprites

	BetaVAE Score	FactorVAE Score	MIG	DCI Disentanglement	Modularity	SAP
Reconstruction	-9	-18	-15	-10	-35	-7
TC (sampled)	-21	-25	-25	-25	-17	-20
KL	-53	-46	-47	-48	-9	-55
ELBO	-14	-24	-20	-14	-34	-13

Dataset = SmallNORB

	BetaVAE Score	FactorVAE Score	MIG	DCI Disentanglement	Modularity	SAP
Reconstruction	-83	-73	-59	-54	69	-38
TC (sampled)	-12	41	-21	-2	-12	-46
KL	20	49	-19	-18	-22	-39
ELBO	-79	-57	-72	-65	61	-55

Dataset = Cars3D

	BetaVAE Score	FactorVAE Score	MIG	DCI Disentanglement	Modularity	SAP
Reconstruction	-19	9	42	20	38	12
TC (sampled)	-18	-6	-3	-14	2	0
KL	-14	-10	-46	-45	-40	-5
ELBO	-24	5	35	12	29	11

Dataset = Shapes3D

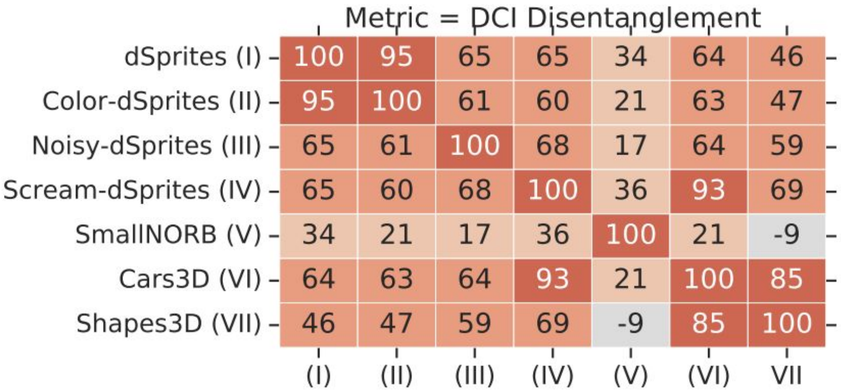
	BetaVAE Score	FactorVAE Score	MIG	DCI Disentanglement	Modularity	SAP
Reconstruction	-30	-4	59	22	-21	27
TC (sampled)	1	5	-11	-8	-11	-2
KL	-14	-1	-38	-31	-11	-29
ELBO	-38	-9	48	9	-25	15

Bad transfer performance across datasets and metrics

The same dataset and metric: 80.7%

Different datasets and the same metric: 59.3%

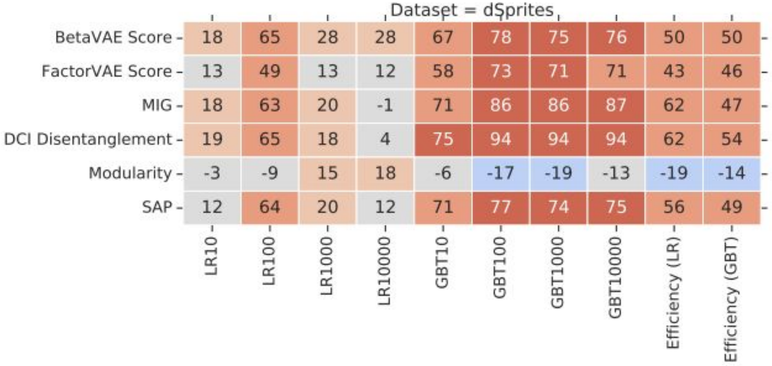
Different datasets and metrics: 54.9%

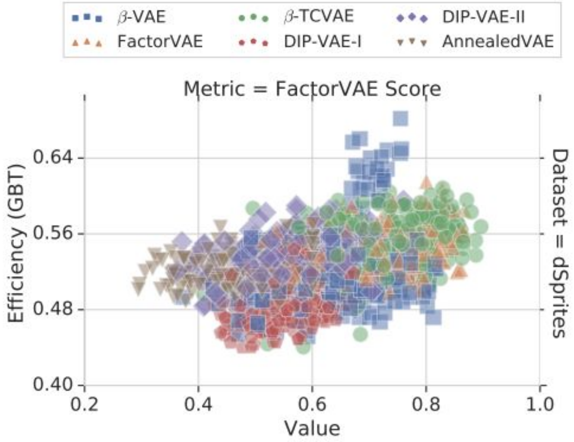


Downstream tasks grounded on the disentangled representation

Basically, better disentanglement score leads to better performance in downstream tasks

But no clear evidence it leads to better sample complexity





Outline

Disentanglement review

Disentanglement is impossible without inductive bias

Experimental results

Future directions

Future directions

Make inductive bias explicit, figure out how to select the hyperparameters without labels

Concrete the benefits of the disentanglement

Evaluate disentanglement methods over diverse datasets

References

- 2017 (ICLR): I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, A. Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. ICLR, 2017.
- 2019 (ICML): F. Locatello, S. Bauer, M. Lucic, G. RÃ¤tsch, S. Gelly, B. Scholkopf, O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. ICML, 2019.
- 2020 (ICLR): T. Kipf, E. Pol, M. Welling. Contrastive Learning of Structured World Models. ICLR, 2020.
- 2019 (NeurIPS): P. Haan, D. Jayaraman, S. Levine. Causal Confusion in Imitation Learning. NeurIPS, 2019.