Towards Nonlinear Disentanglement in Natural Data with Temporal Sparse Coding

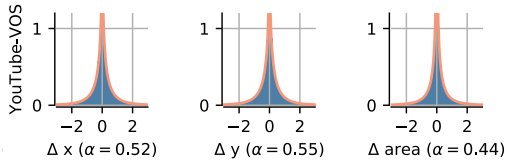
David Klindt*, Lukas Schott*, Yash Sharma*, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge°, Dylan Paiton°

contact: klindt.David@gmail.com

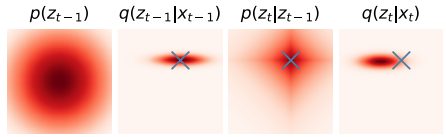
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Abstract

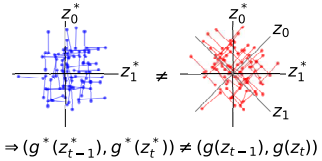
We construct an unsupervised learning model that achieves nonlinear disentanglement of underlying factors of variation in naturalistic videos. Previous work suggests that representations can be disentangled if all but a few factors in the environment stay constant at any point in time. As a result, algorithms proposed for this problem have only been tested on carefully constructed datasets with this exact property, leaving it unclear whether they will transfer to natural scenes. Here we provide evidence that objects in segmented natural movies undergo transitions that are typically small in magnitude with occasional large jumps, which is characteristic of a temporally sparse distribution. We leverage this finding and present SlowVAE, a model for unsupervised representation learning that uses a sparse prior on temporally adjacent observations to disentangle generative factors without any assumptions on the number of changing factors. We provide a proof of identifiability and show that the model reliably learns disentangled representations on several established benchmark datasets, often surpassing the current state-of-the-art. We additionally demonstrate transferability towards video datasets with natural dynamics, Natural Sprites and KITTI Masks, which we contribute as benchmarks for guiding disentanglement research towards more natural data domains.



Natural Transitions in Ground Truth Factors are Sparse. Transition statistics of object masks in natural videos. Red lines indicate fits of a generalized Laplacian distribution (shape parameter alpha).



Model illustration. The left two plots show the posterior and prior for the first image in a pair and they are identical to the standard VAE model. The two plots on the right show the posterior and conditional prior for the second time step, with the mean of the prior centered around the posterior mean of the previous time step.

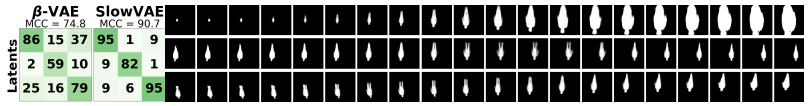


Proof illustration. If both true and fitted model have sparse transitions between pairs of embedded images, but the model representation is rotated with respect to the ground truth latents, then the generated image distributions cannot be matched.

Model (Data)	BetaVAE	FactorVAE	MIG	DCI	Modularity	SAP
β-VAE (*i.i.d.*)	82.3	66.0	10.2	18.6	82.2	4.9
Ada-ML-VAE (LOC)	89.6	70.1	11.5	29.4	89.7	3.6
Ada-GVAE (LOC)	92.3	84.7	26.6	47.9	91.3	7.4
slowVAE (LOC)	90.4 (3.3)	81.35 (7.6)	35.7 (8.3)	52.1 (6.5)	87.6 (2.0)	5.1 (1.4)
slowVAE (LAP)	100.0 (0.0)	98.32 (2.5)	27.8 (7.9)	65.3 (3.1)	97.0 (1.5)	6.1 (2.6)

Model (Data)	SN	Cars3D	Shapes3D	MPI3D
β-VAE (*i.i.d.*)	21.4	8.8	22.0	7.2
Ada-ML-VAE (LOC)	31.1	14.7	50.9	24.1
Ada-GVAE (LOC)	25.6	15.0	56.2	28.4
slowVAE (LOC)	23.2 (1.9)	15.5 (1.4)	66.4 (5.9)	33.1 (1.2)
slowVAE (LAP)	25.4 (0.6)	9.8 (1.4)	62.9 (3.4)	29.6 (1.1)

Performance evaluation on DisLib benchmark datasets. Median and absolute deviation across 10 random seeds. First three rows from [1].



Latent traversals on Kitti Masks. Left, MCC scores. Right, corresponding latent traversals.

Model (Data)	MCC
β-VAE (C)	42.6 (4.7)
slowVAE (C)	**49.1** (4.0)

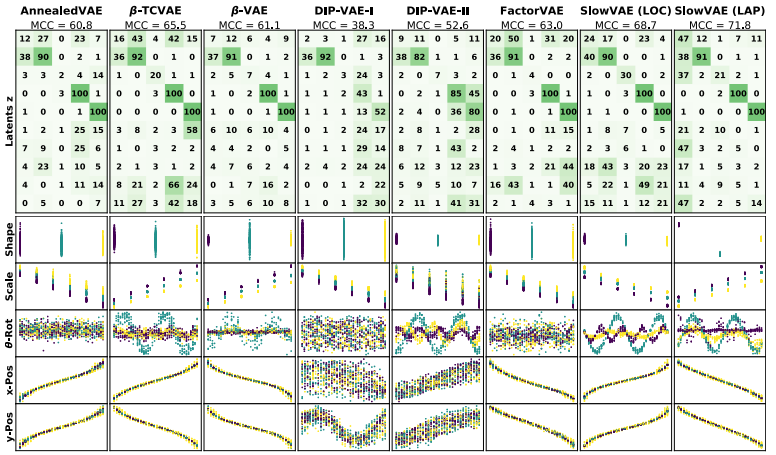
Table 4: **Continuous Natural Sprites**. Mean (s.d.) over 10 random seeds.

Model (frame separation)	MCC
β-VAE	62.7 (7.1)
slowVAE (Δ t=1)	66.1 (4.5)
slowVAE (Δ t=5)	**79.6** (5.8)

Table 5: **KITTI-Masks**. Mean (s.d.) over 10 random seeds.

Model (Data)	BetaVAE	FactorVAE	MIG	DCI	Modularity	SAP	MCC
β-VAE	78.1 (3.0)	60.6 (6.0)	4.6 (1.9)	10.3 (1.8)	87.8 (2.3)	2.1 (1.0)	41.7 (3.4)
slowVAE	**82.6** (2.2)	**76.2** (4.8)	**11.7** (5.0)	**18.9** (5.5)	88.1 (3.6)	**4.4** (2.3)	**52.6** (4.1)

Table 3: **Discrete Natural Sprites.** Mean (s.d.) performance levels over 10 random seeds. Best models with statistical significance ($p = 0.05$) are indicated in bold.



Dsprites latent visualizations. Top half are MCC correlation matrices (greedily matched latents) for different models from DisLib. Bottom, latent over ground truth for different factors, coloured by shape.

Conclusions

- We present a set of video datasets that are increasingly more natural using measurements from natural scenes.
- We provide evidence that natural generative factors undergo sparse changes across time, which we exploit by extending the VAE framework with a sparse temporal prior.
- We provide theoretical justification for our model by proving that it is identifiable assuming temporal sparsity.
- We demonstrate improved disentanglement over previous models using quantitative metric evaluation across several datasets as well as visualizations of the learned manifolds.

References

[1] Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. *arXiv preprint arXiv:2002.02886*, 2020.

Funding

We thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for support. This work was supported by the German Federal Ministry of Education and Research (BMBF) through the Tübingen AI Center (FKZ: 01IS18039A), by the Deutsche Forschungsgemein- schaft (DFG) in the priority program 1835 under grant BR2321/5-2 and by SFB 1233, Robust Vision: Inference Principles and Neural Mechanisms (TP3), project number: 276693517.