~~Are~~ Disentangled Representations are Helpful for Abstract Visual Reasoning~~?~~

Sjoerd van Steenkiste
sjoerd@idsia.ch

Francesco Locatello
locatelf@ethz.ch

Jürgen Schmidhuber
juergen@idsia.ch

Olivier Bachem
bachem@google.com

SUPSI IDSIA · ETH zürich · MPI

Summary

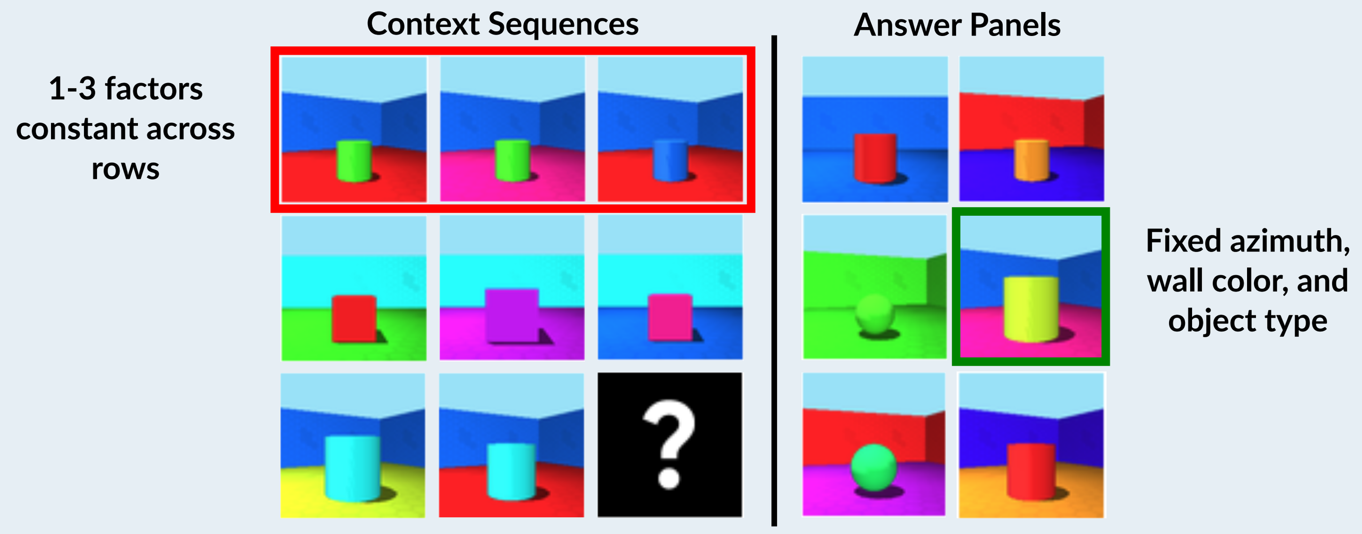
We conduct a large-scale evaluation of disentangled representations on complex abstract visual reasoning tasks to systematically evaluate their benefits

- We create two new abstract reasoning tasks similar to Raven's Progressive Matrices that require reasoning about relatons between objects and background
- We train 360 unsupervised disentanglement models (based on 6 approaches) to acquire disentangled representations
- We train 3600 relational reasoning models that make use of these representations on our abstract reasoning tasks
- We compare the accuracy of these reasoning models to the disentanglement of the representations as determined by five different disentanglement metrics
- We observe compelling evidence that more disentangled representations yield better sample-efficiency in learning to solve the considered abstract visual reasoning tasks

Visual Reasoning Tasks

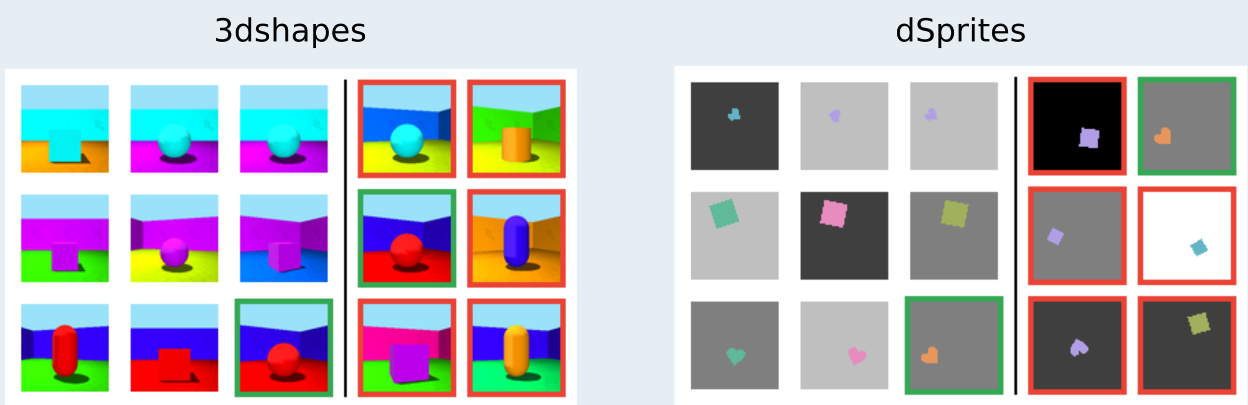
We adapt dSprites and 3dshapes to obtain two new abstract visual reasoning tasks similar to Raven's Progessive Matrices

Task: Complete the final sequence by choosing from answer panels



Requires inferring relationships between context panels, and applying this knowledge to the partial sequence in relation to the anwer panels

- Answer panels are generated to include difficult alternatives
- Difficult task for neural networks that can not easily be solved by correlating image statistics
- Need to reason about image content, which makes this a good benchmark for evaluating disentangled representations



Setup

We train 360 unsupervised disentangled representation learning models on the panels of the reasoning tasks to obtain (disentangled) representations

We consider recent approaches that use a regularized variational auto-encoding objective

$$\mathbb{E}_{p(x)}[\mathbb{E}_{q_\phi(z|x)}[-\log p_\theta(x|z)]] + \lambda_1 \mathbb{E}_{p(x)}[R_1(q_\phi(z|x))] + \lambda_2 R_2(q_\phi(z))$$

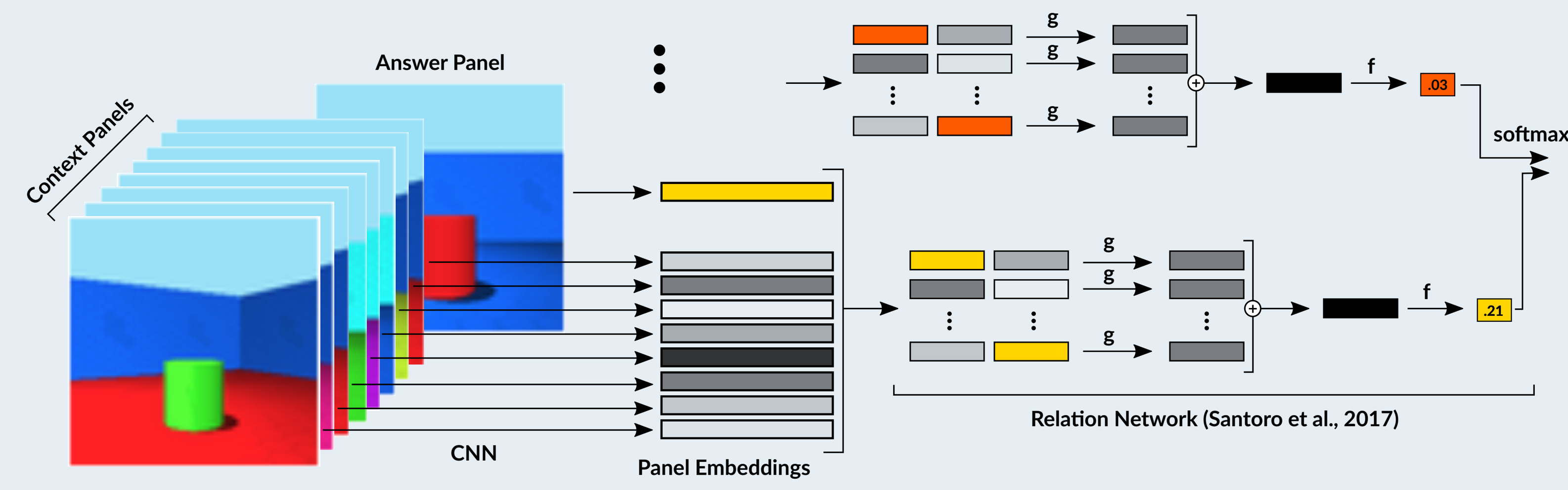
- β-VAE: $R_1 := D_{KL}[q_\phi(z|x)||p(z)]$
- β-TCVAE: $R_2 := TC(q_\phi(z))$ Monte Carlo estimator
- FactorVAE: $R_2 := TC(q_\phi(z))$ density ratio estimator
- DIP-VAE: $R_2 := ||\mathrm{Cov}_{q_\phi(z)} - I||_F^2$

We consider β-VAE with and without annealing, and two different estimators for DIP-VAE

Using 6 different regularization strengths and 5 different seeds for 6 methods we obtain 180 models per dataset from which we can obtain (disentangled) representations

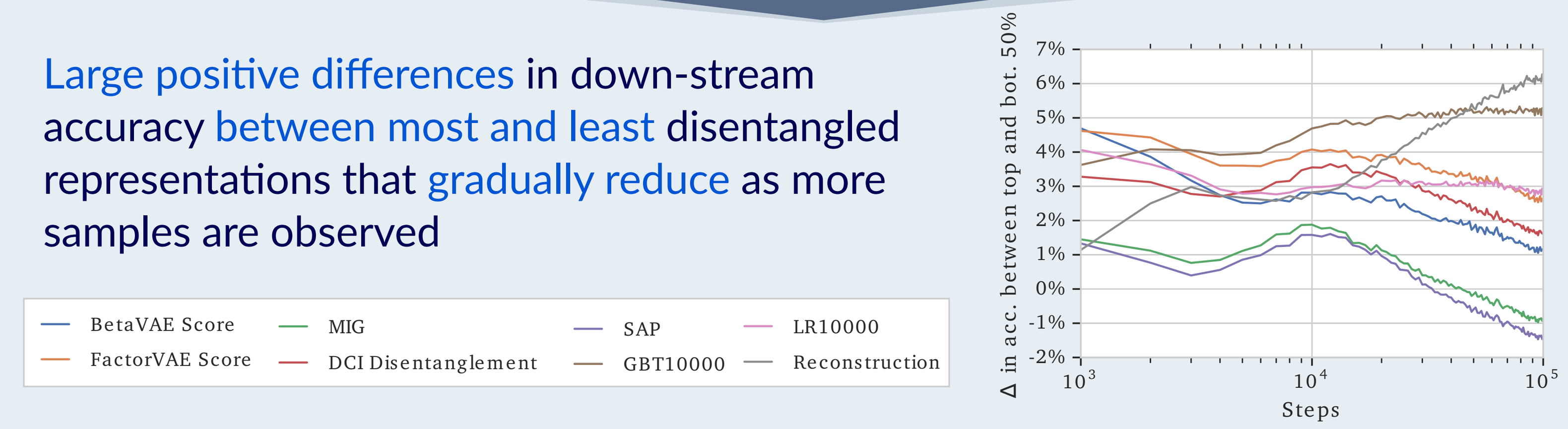
We train 3600 abstract visual reasoning models using the representations as panel embeddings

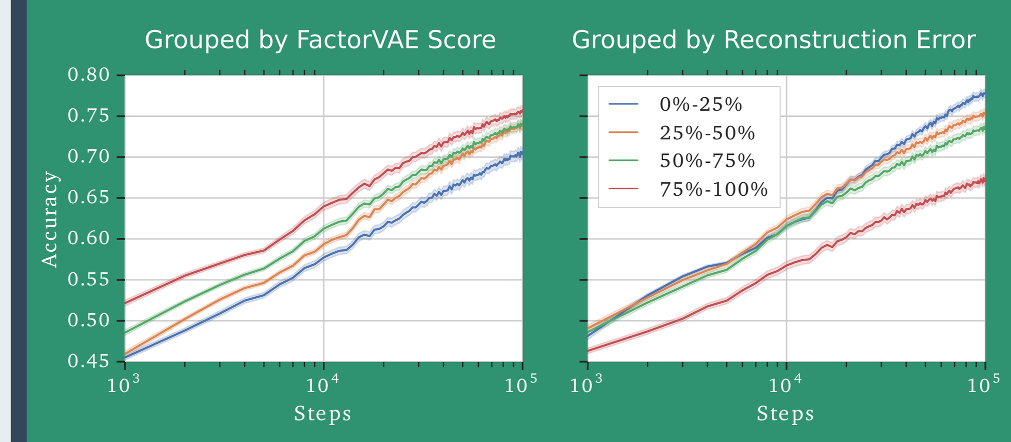
We make use of the Wild Relational Network (Barret et al., 2018), which incorporates a relational inductive bias, to perform the reasoning task



Analysis

Large positive differences in down-stream accuracy between most and least disentangled representations that gradually reduce as more samples are observed





Representations that are more disentangled give rise to better relative performance throughout all phases of training

Ordering is less pronounced for reconstruction error

dSprites (best 50%)

	1K	2K	5K	10K	20K	50K	100K
BetaVAE Score	78	77	76	59	53	39	28
FactorVAE Score	68	73	74	59	54	40	25
MIG	35	31	48	63	52	16	-2
DCI Disentanglement	61	59	68	66	53	28	11
SAP	37	34	49	63	54	21	3
Reconstruction Error	35	19	26	49	40	1	-16

dSprites (worst 50%)

	1K	2K	5K	10K	20K	50K	100K
BetaVAE Score	60	48	34	28	31	18	6
FactorVAE Score	63	58	49	42	43	30	18
MIG	24	28	25	27	26	12	6
DCI Disentanglement	40	40	31	30	30	18	11
SAP	12	14	11	13	12	2	-4
Reconstruction Error	-38	-49	-48	-48	-49	-56	-61

Filtering out the best / worst performing models reveals a sharp contrast between disentanglement and reconstruction error in terms of correlation

Main Result

In the few sample regime (modularity-based) disentanglement is postively correlated with down-stream reasoning accuracy

dSprites

	1K	2K	5K	10K	20K	50K	100K
BetaVAE Score	67	59	52	45	41	27	20
FactorVAE Score	69	67	63	56	53	39	32
MIG	22	19	27	33	24	-0	-8
DCI Disentanglement	47	42	43	42	34	15	7
SAP	16	11	19	26	17	-5	-12
GBT10000	60	67	71	69	67	64	60
LR10000	66	62	54	41	43	39	35
Reconstruction Error	-26	-43	-42	-34	-42	-62	-67

3dshapes

	1K	2K	5K	10K	20K	50K	100K
BetaVAE Score	40	56	59	42	39	18	18
FactorVAE Score	59	71	72	65	37	-8	-13
MIG	37	26	12	15	-8	-40	-43
DCI Disentanglement	31	35	24	19	17	4	6
SAP	40	42	34	29	10	-21	-26
GBT10000	32	38	29	20	26	19	22
LR10000	1	10	21	7	34	62	63
Reconstruction Error	-1	-16	-30	-17	-38	-55	-52

We make the following observations:

- In the few-sample regime disentanglement is postively correlated with down-stream accuracy
- When enough training data is provided the benefit of disentanglement disappears
- Large differences between various notions of disentanglement. Intervention-based metrics that test for modularity correlate best
- Reconstruction error is only strongly negatively correlated with down-stream performance when many samples are given
- Simple single-class classification metrics correlate well with down-stream accuracy on this task

Disentanglement Metrics

Test mostly whether latents are associated with only a single factor (modularity)

- BetaVAE Score: accuracy of linear classifier that predicts the index of fixed factor
- FactorVAE Score: accuracy of majority vote classifier that predicts the index of a fixed factor
- DCI Disentanglement Score: Entropy of the latent / factor predictive importance over factors

Test mostly whether factors are associated with only a single latent (compactness)

- Mutual Information Gap: normalized gap in latent / factor MI between top two latents
- Separated Attribute Predictability: avg. difference in latent / factor prediction error between top two latents

GitHub/google-research/disentanglement_lib#abstract-reasoning-experiments