VISUAL REPRESENTATION LEARNING DOES NOT GENERALIZE STRONGLY WITHIN THE SAME DOMAIN

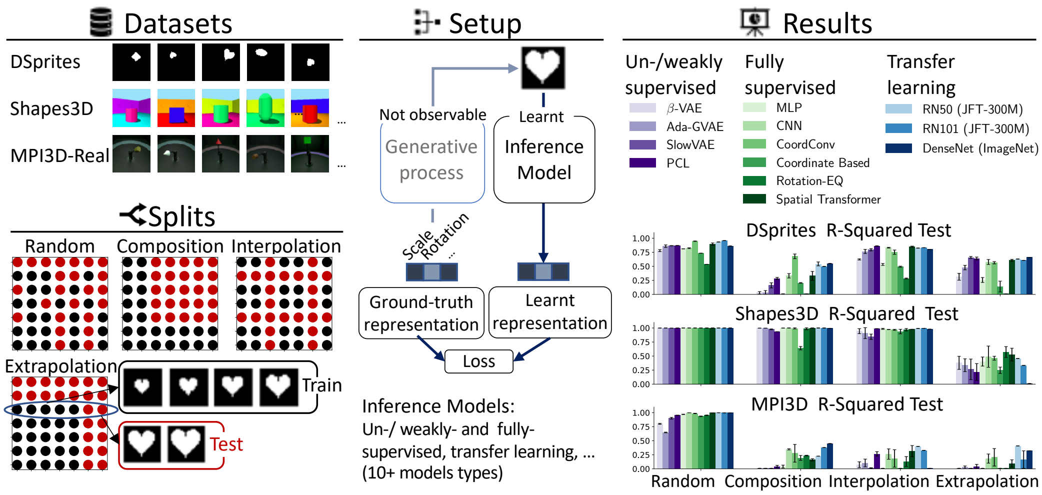
Anonymous authors
Paper under double-blind review

ABSTRACT

A common goal of many fields connected to machine learning such as inverse graphics, causality, disentanglement, or generalization is to learn the underlying latent factors of variation as well as the true mechanism through which each factor acts in the world (Fig.1, *center*).

In this paper, we test 13 representation learning approaches (Fig.1, *top right*) to infer the generative factors of variation, i.e., to invert the underlying generative process, of simple datasets (DSprites, Shapes3D, MPI3D; Fig.1, *top left*). Opposed to prior work that introduce novel factors of variation at test time such as blur or other (un)structured noise, we here only extrapolate, interpolate, or recompose (Fig.1, *bottom left*) the existing factors of variation from the training dataset (e.g., recognizing small squares at test time, despite having only seen small hearts and large squares during training). A model that learns the correct mechanism should be able to infer the factors of variation *and* be able to generalise.

We train and test a wide variety of 2000+ unsupervised, weakly-supervised and fully-supervised models and observe that architectural bias seems to matter more than disentanglement, while extensive pre-training seems to be most helpful. However, as soon as we move towards more realistic datasets, the generalization capabilities of all the tested architectures drop significantly. This can for instance be seen on the MPI3-real extrapolation test split where no approach achieves an r-squared above 50% (Fig.1 *bottom right*), compared to even 100% r-squared in distribution. Even when only a single factor is out-of-distribution the performance of inferring other in-distribution factors likewise degrades. This suggests that the considered models lack modularity. We observed that this holds for disentangled representations as well and regardless of the supervision signal. Our results thus point to an important yet understudied problem of learning mechanistic models of observations that can facilitate generalization.


Datasets
DSprites
Shapes3D
MPI3D-Real
Splits
Random
Composition
Interpolation
Extrapolation
Train
Test
Setup
Not observable
Generative process
Learnt
Inference Model
Scale
Rotation
Ground-truth representation
Learnt representation
Loss
Inference Models:
Un-/ weakly- and fully-supervised, transfer learning, ...
(10+ models types)
Results
Un-/weakly supervised
β-VAE
Ada-GVAE
SlowVAE
PCL
Fully supervised
MLP
CNN
CoordConv
Coordinate Based
Rotation-EQ
Spatial Transformer
Transfer learning
RN50 (JFT-300M)
RN101 (JFT-300M)
DenseNet (ImageNet)
DSprites R-Squared Test
Shapes3D R-Squared Test
MPI3D R-Squared Test
1.00
0.75
0.50
0.25
0.00
Random
Composition
Interpolation
Extrapolation