VISUAL REPRESENTATION LEARNING DOES NOT GENERALIZE STRONGLY WITHIN THE SAME DOMAIN

Lukas Schott[1], Julius von Kügelgen[2, 3, 4], Frederik Träuble[2, 4],
Peter Gehler[4], Chris Russell[4], Matthias Bethge[1, 4], Bernhard Schölkopf[2, 4],
Francesco Locatello[4, †], Wieland Brendel[1, †]

[1]University of Tübingen
[2]Max Planck Institute for Intelligent Systems
[3]University of Cambridge
[4]Amazon Web Services
[†]Joint senior authors
lukas.schott@bethgelab.org

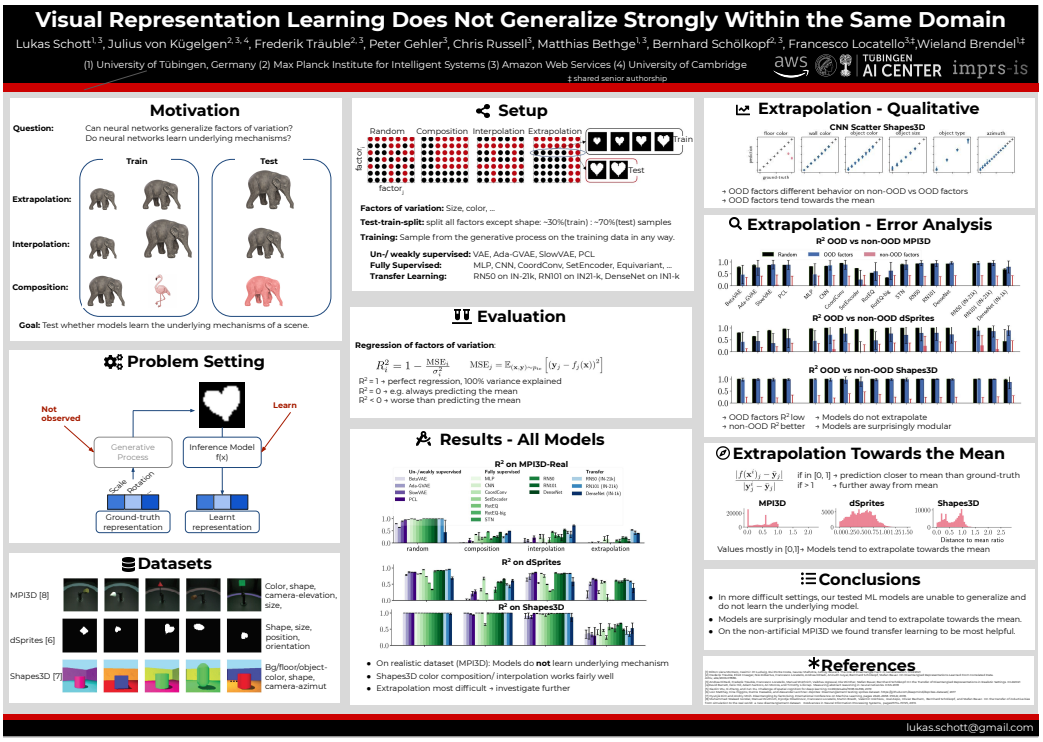
ABSTRACT

An important goal in machine learning is to uncover the underlying latent factors of variation as well as the true mechanism through which each factor acts in the world.

In this paper, we test whether 17 representation learning approaches correctly infer the generative factors of variation in simple datasets (DSprites, Shapes3D, MPI3D). In contrast to prior work that introduces novel factors of variation during test time, such as blur or other (un)structured noise, we here recompose, interpolate, or extrapolate only existing factors of variation from the training data set (e.g. small and medium sizes squares during training and large squares during testing). Models that learn the correct mechanism should be to generalise to this benchmark.

In total, we train and test 2000+ unsupervised, weakly-supervised and fully-supervised models and observe that all of them struggle to learn the underlying mechanism regardless of supervision signal and architectural bias. Moreover, the generalization capabilities of all tested models drop significantly as we move from artificial data sets towards more realistic real world data sets. Despite their inability to identify the correct mechanism, the models are surprisingly modular as their ability to infer other in-distribution factors remains fairly stable providing only a single factor is OOD.

These results point to an important yet understudied problem of learning mechanistic models of observations that can facilitate generalization.


Visual Representation Learning Does Not Generalize Strongly Within the Same Domain
Lukas Schott[1,3], Julius von Kügelgen[2,3,4], Frederik Träuble[2,3], Peter Gehler[3], Chris Russell[3], Matthias Bethge[1,3], Bernhard Schölkopf[2,3], Francesco Locatello[3‡], Wieland Brendel[1‡]
(1) University of Tübingen, Germany (2) Max Planck Institute for Intelligent Systems (3) Amazon Web Services (4) University of Cambridge
‡ shared senior authorship
Motivation
Question:
Can neural networks generalize factors of variation?
Do neural networks learn underlying mechanisms?
Train
Test
Extrapolation:
Interpolation:
Composition:
Goal: Test whether models learn the underlying mechanisms of a scene.
Problem Setting
Not observed
Learn
Generative Process
Inference Model f(x)
Scale
Rotation
Ground-truth representation
Learnt representation
Datasets
MPI3D [8]
Color, shape, camera-elevation, size,
dSprites [6]
Shape, size, position, orientation
Shapes3D [7]
Bg/floor/object-color, shape, camera-azimut
Setup
Random
Composition
Interpolation
Extrapolation
Train
Test
factor_i
factor_j
Factors of variation: Size, color, ...
Test-train-split: split all factors except shape: ~30%(train) : ~70%(test) samples
Training: Sample from the generative process on the training data in any way.
Un-/ weakly supervised: VAE, Ada-GVAE, SlowVAE, PCL
Fully Supervised: MLP, CNN, CoordConv, SetEncoder, Equivariant, ...
Transfer Learning: RN50 on IN-21k, RN101 on IN21-k, DenseNet on IN1-k
Evaluation
Regression of factors of variation:
$R_i^2 = 1 - \frac{\mathrm{MSE}_i}{\sigma_i^2}$ $\quad \mathrm{MSE}_j = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_{\text{test}}}\left[(y_j - f_j(\mathbf{x}))^2\right]$
R² = 1 → perfect regression, 100% variance explained
R² = 0 → e.g. always predicting the mean
R² < 0 → worse than predicting the mean
Results - All Models
R² on MPI3D-Real
R² on dSprites
R² on Shapes3D
random
composition
interpolation
extrapolation
On realistic dataset (MPI3D): Models do not learn underlying mechanism
Shapes3D color composition/ interpolation works fairly well
Extrapolation - Qualitative
CNN Scatter Shapes3D
floor color
wall color
object color
object size
object type
azimuth
→ OOD factors different behavior on non-OOD vs OOD factors
→ OOD factors tend towards the mean
Extrapolation - Error Analysis
R² OOD vs non-OOD MPI3D
R² OOD vs non-OOD dSprites
R² OOD vs non-OOD Shapes3D
→ OOD factors R² low
→ non-OOD R² better
→ Models do not extrapolate
→ Models are surprisingly modular
Extrapolation Towards the Mean
if in [0, 1] → prediction closer to mean than ground-truth
if > 1 → further away from mean
MPI3D
dSprites
Shapes3D
Distance to mean ratio
Values mostly in [0,1] → Models tend to extrapolate towards the mean
Conclusions
In more difficult settings, our tested ML models are unable to generalize and do not learn the underlying model.
Models are surprisingly modular and tend to extrapolate towards the mean.
On the non-artificial MPI3D we found transfer learning to be most helpful.
References
lukas.schott@gmail.com