DTU Compute
Department of Applied Mathematics and Computer Science

DTU

Invariant Causal Representation Learning For Out-Of-Distribution Generalization

Frederik Larsen (s174159), Jacob Bendsen (s184328), Marie Pedersen (s174329), Peter Emil Carstensen (s184332), Morten Johnsen (s184334)

1 Introduction

Machine learning models generally assume *the observed variables - both input* **X**, *and output* **Y** *- follow same underlying distribution across the training and test data.* Accordingly, spurious correlations in the training data, are also present in the test data, and model performance on the test data will approximate what is seen during training. When this assumption is violated, the model show poor generalization across datasets. To guarantee performance across environments it is necessary to train the model on **causal relationships**, rather than relying on spurious environmentally dependent correlations. *Lu et al 2022* propose the algorithm, iCaRL, for achieving out-of-distribution generalization by guaranteed identifiability of causal latent variables of **Y**.

iCaRL algorithm phases:

1. Train NF-iVAE and infer the latent variables from the given observations.
2. Use the PC-algorithm to identify causal latent variables.
3. Train the invariant classifier(w) on $Z_{Pa(Y)_{training}}$ from the training set. Infer $Z_{Pa(Y)_{test}}$ from the observed test **X** and predict y_{test}.

2 Analysis

Main questions:

1. Can ≈ 68% accuracy on CMNIST data be achieved by using a VAE and iVAE if phase II and III of iCaRL are performed based on these?
2. Can we achieve a latent space separation of synthetic data, with the simpler iVAE as compared to NF-iVAE in the article?
3. Can we implement iCaRL and NF-iVAE based on the article within the time constraints of the project?

All three different variational autoencoder models VAE, iVAE and NF-iVAE can be inserted in phase I in iCaRL. The models are primarily distinctive through their prior distributions. Both the iVAEs use an exponential family prior and a neural network to condition the prior on the environment and target label through either training the natural parameters or both natural parameters and sufficient statistics for the iVAE and NF-iVAE, respectively. The standard gaussian unconditioned prior of the VAE is shown to be unidentifiable. We use the results from *Lu et all 2022* as benchmark, regarding latent space separation of environments for the synthetic data and label classification accuracy for CMNIST data.

3 Datasets

1. Coloured MNIST dataset with environmentally conditional spurious correlations between *red* and *green* colouring and the target label [*Arjovsky et al 2019*]
2. Synthetic data generated by passing a known causal structure with two parents of Y $Z_{Pa(Y)_1}$, $Z_{Pa(Y)_2}$; **Y**; and one child of Y $Z_{Ch(y)}$ through a randomly initialised neural network [*Lu et al 2022*].

4 Results & Discussion

Synthetic data - 2 Parents, 1 child

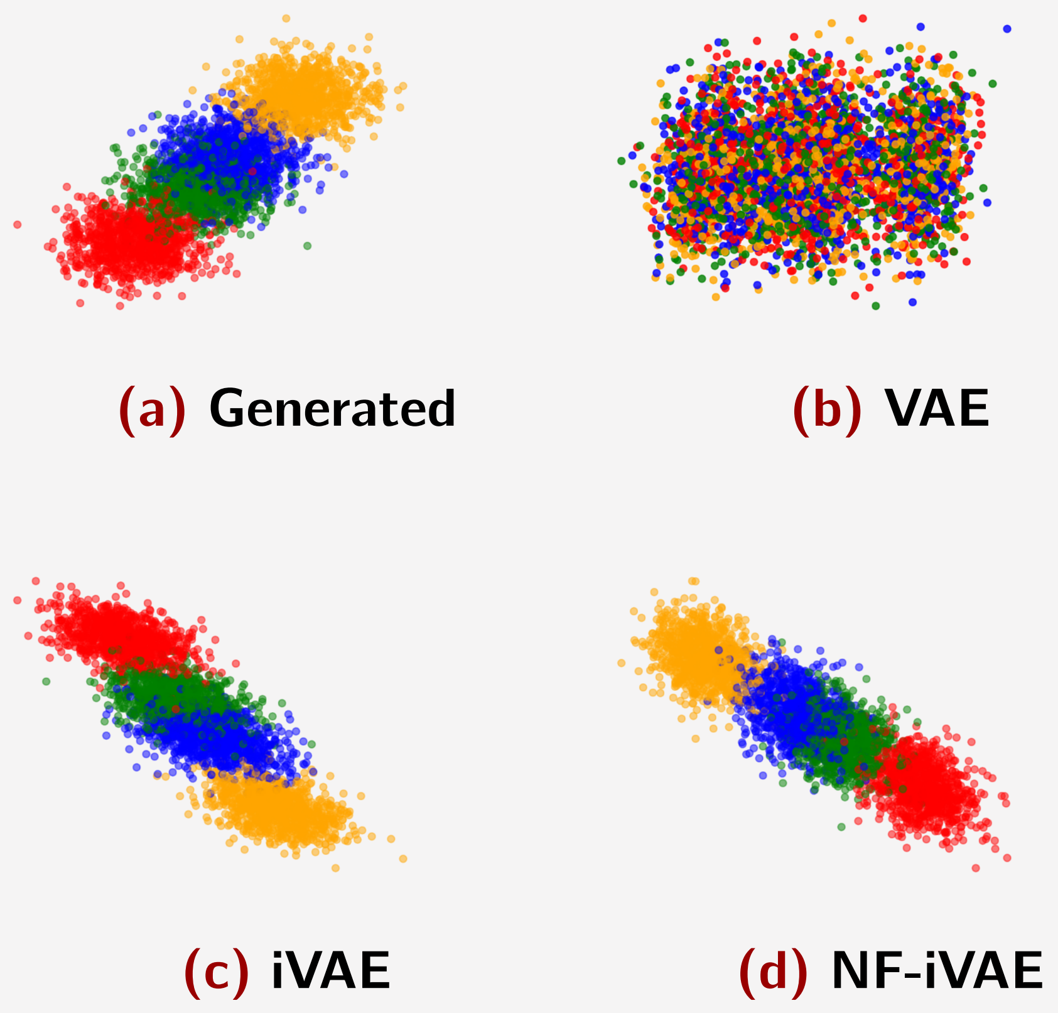
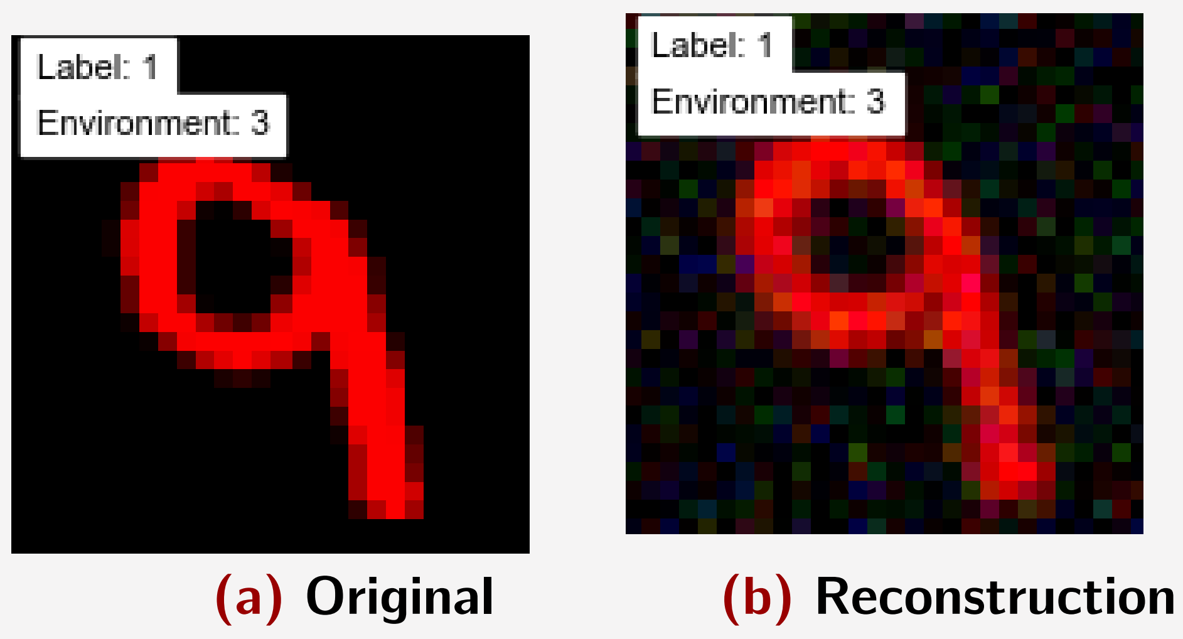

(a) Generated
(b) VAE
(c) iVAE
(d) NF-iVAE

Figure 1: Note: A poorly performing model will have data that seem unorganised and mixed, whilst a good performing model should indicate a clear separation in data.

For the synthetic data we included a child of Y, to ensure that **X** is jointly influenced by both causal and non-causal variables, and thus - in opposition to the same calculations by *Lu et al 2022* - we test the identifiability of the models in a situation where not all latent variables are causal. From figure (1) it can be seen that both the iVAE and NF-iVAE models are - as predicted by *Khemakhem et al 2020* and *Lu et at 2022* - capable of identifying the causal latent variables up to a simple componentwise transformation. The VAE can't identify the latent environmetally dependent latent variables.



(a) Original
(b) Reconstruction

Figure 2: Original CMNIST data and NF-iVAE reconstruction

The CMNIST results are precarious and contingent on the fact that the PC-algorithm for step 2 in iCaRL currently yield inconsistent results, and the results given here, are consequently based on iterations through semi-random choices of "parent" latent variables.

Table 1: Accuracies obtained from classification on CMNIST data

CMNIST	Train acc. [%]	Validation acc. [%]
All Z	≈ 85	≈ 10
VAE_{Z_p}	]53;79[	]19;49[
$iVAE_{Z_p}$	]70;85[	]23;33[
$NF-iVAE_{Z_p}$	]28;82[	]28;73[
Paper $iCaRL$	70.56	68.75
Random guess	50	50
Optimal	75	75

Depending on the inclusion of different combinations of latent variables in the CMNIST classification algorithm, we found a wide array of accuracies of the classification algorithm (table 1). This is expected when shuffling causal and non-causal input variables. Random inclusion/exclusion of children of Y and inclusion/exclusion of parents, should have a drastic impact on accuracy. Interestingly, these preliminary results seem show a pattern where the NF-iVAE has a higher potential for high accuracy on the test data. When including all the latent variables all models show a test accuracy of ≈ 10% as the classifier is learning the spurious color correlations, instead of whether the number on the image is less than 5. Only including two latent variables, did in most cases lead to 50%-50% accuracies (random guess), as they did not contain enough information for classification training.

5 Conclusion

Data synthesization and Phase I and III of the iCaRL algorithm which spans, a VAE, iVAE, NF-iVAE, the invariant classifier and latent space inference based on the trained decoder, have been implemented. But, in order to properly assess the performance of VAE and iVAE against NF-iVAE on the CMNIST data, it is necessary to successfully implement phase II. The used PC-algorithm package currently yield inconsistent results. Our preliminary results when excluding phase II have shown that accuracies towards the ≈ 68% might be possible for the NF-iVAE, with no indication of the iVAE and VAE scoring that high. Additionally, in contrast to the synthetic data results by *Lu et al* 2022 we found that the iVAE can achieve approximately the same recovery of the original synthetic data, as the NF-iVAE.