Linear Disentangled Representations and Unsupervised Action Estimation

Matthew Painter, Jonathon Hare, Adam Prügel-Bennett
Department of Electronics and Computer Science
Southampton University
{mp2u16, jsh2, apb}@ecs.soton.ac.uk

Abstract

Disentangled representation learning has seen a surge in interest over recent times, generally focusing on new models to optimise one of many disparate disentanglement metrics. It was only with Symmetry Based Disentangled Representation Learning that a robust mathematical framework was introduced to define precisely what is meant by a "linear disentangled representation". This framework determines that such representations would depend on a particular decomposition of the symmetry group acting on the data, showing that actions would manifest through irreducible group representations acting on independent representational subspaces. ForwardVAE subsequently proposed the first model to induce and demonstrate a linear disentangled representation in a VAE model. In this work we empirically show that linear disentangled representations are not present in standard VAE models and that they instead require altering the loss landscape to induce them. We proceed to show that such representations are a desirable property with regard to classical disentanglement metrics. Finally we propose a method to induce irreducible representations which forgoes the need for labelled action sequences, as was required by prior work. We explore a number of properties of this method, including the ability to learn from action sequences without knowledge of intermediate states.

1 Introduction

Many learning machines make use of an internal representation [Bengio et al., 2013] to help inform their decisions. For designers of such machines, it is often desirable for these representations to be interpretable in the sense that we can easily understand how individual parts contribute to solving the task at hand. Interpretable representations have slowly become the major goal in the sub-field of deep learning concerned with Variational Auto-Encoders (VAEs) [Kingma and Welling, 2014], seceding from the usual goal of generative models, accurate/realistic sample generation. In this area, representations generally take the form of a multi-dimensional vector space (latent space), and the particular form of interpretability is known as Disentanglement, where each latent dimension (or group of such) is seen to represent an individual (and independent) generative/explanatory factor of the data. Standard examples of such factors include the x position, the y position, an object's rotation, etc. Effectively separating them into separate subspaces, "disentangling them" is the aim of disentanglement research.

Understanding disentangled representations (in the particular case of VAEs) has primarily been based on the application of numerous "disentanglement metrics". These metrics operate on disparate understandings on what constitutes ideal disentanglement, and required large scale work [Locatello et al., 2018] to present correlations between them and solidify the fact that they should be considered jointly and not as individuals. Symmetry based disentangled representation learning (SBDRL) [Higgins et al., 2018] offers a mathematical framework through which a rigorous definition of a linear disentangled

Preprint. Under review.

representation can be formed. In essence, if data has a given symmetry structure (expressed through Group Theory), the linear disentangled representations are exactly those which permit irreducible representations of group elements. Whilst Higgins et al. [2018] do not express methods to determine the symmetry structure for given data, they offers sensible suggestions. Caselles-Dupré et al. [2019] propose modelling and inducing such disentangled representations through observation of action transitions under the framework of reinforcement learning. Their model successfully induces linear disentangled representations with respect to the chosen symmetry structure, demonstrating that the suggestions of Higgins et al. [2018] were indeed sensible and result in interpretable representations. Whilst Caselles-Dupré et al. [2019] assume symmetry structure comprising of purely cyclic groups, Quessard et al. [2020] expand this to the more expressive class of SO(3) matrices.

Current work shows that irreducible representations can be induced in latent spaces, but has yet to determine if they can be found without explicitly structuring the loss landscape. Furthermore, they are not related back to previous disentanglement metrics to demonstrate the utility of such structures being present. Finally, by sticking rigidly to the reinforcement framework, Caselles-Dupré et al. [2019] and Quessard et al. [2020] allow the model direct knowledge of which actions they are observing. This restricts the applicable domain of the models to data where action transition pairs are explicitly labelled.

In this work, we make the following contributions:

- We confirm empirically that irreducible representations are not naturally found in standard VAE models without biasing the loss landscape towards them.
- We determine that inducing such representations in VAE latent spaces garners improved performance on a number of standard disentanglement metrics.
- We introduce a novel disentanglement metric to explicitly measure linear disentangled representations and we modify the mutual information gap metric to be more appropriate for this setting.
- We propose a method to induce irreducible representations without the need for labelled action-transition pairs.
- We demonstrate a number of properties of such a model and show it continues to lead to strong scores on classical disentanglement metrics.

2 Symmetry Based Disentangled Representation Learning and Prior Works

This section provides a high level overview of the SBDRL framework without the mathematical grounding in Group and Representation theory on which it is based. The work and appendices of Higgins et al. [2018] offer a concise overview of the necessary definitions and theorems. We encourage the reader to first study this work since they provide intuition and examples which we cannot present here given space constraints.

SBDRL VAE representation learning is concerned with the mapping from an observation space (generally images) $\mathcal{O} \subset \mathbb{R}^{n_x \times n_y}$ to a vector space forming the latent space $\mathcal{Z} \subset \mathbb{R}^l$. SBDRL includes the additional construct of a world space $\mathcal{W} \subset \mathbb{R}^d$ containing the possible states of the world which observations represent. There exists a generative process $b : \mathcal{W} \to \mathcal{O}$ and a inference process $h : \mathcal{O} \to \mathcal{Z}$, the latter being accessible and parametrised by the VAE encoder. SBDRL assumes for convenience that both h and b are injective. For this work, it will be convenient to further assume that $|\mathcal{W}| = |\mathcal{O}| = |\mathcal{Z}| = C \in \mathbb{N}$, i.e. there are a finite number of world states and there is no occlusion in observations. We should note however that neither of these are strictly required.

SBDRL proposes to disentangle *symmetries* of the world space, transformations that preserve some (mathematical or physical) feature, often object identity. Specific transformations may be translation of an object or its rotation, both independent of any other motions (note the similarity to generative data factors). Such symmetries are described by symmetry groups and the symmetry structure of the world space is represented by a group with decomposition $G = G_1 \times \cdots \times G_N$ acting on $\mathcal{W}$ through action $\cdot_{\mathcal{W}} : G \times \mathcal{W} \to \mathcal{W}$. The component groups G_i reflect the individual symmetries and the particular decomposition need not be unique.

SBDRL calls $\mathcal{Z}$ disentangled with respect to decomposition $G = G_1 \times \cdots \times G_N$ if:

1. There is a group action $\cdot_{\mathcal{Z}} : G \times \mathcal{Z} \to \mathcal{Z}$
2. The composition $f = h \circ b : \mathcal{W} \to \mathcal{Z}$ is equivariant with respect to the group actions on $\mathcal{W}$ and $\mathcal{Z}$. i.e. $g \cdot_{\mathcal{Z}} f(w) = f(g \cdot_{\mathcal{W}} w)\ \forall w \in \mathcal{W}, g \in G$.
3. There is a decomposition $\mathcal{Z} = \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_N$ such that $\mathcal{Z}_i$ is fixed by the action of all $G_j, j \neq i$ and affected only by G_i

Since we assume $\mathcal{Z}$ is a vector space, SBDRL refines this through group representations, constructs which preserve the linear structure. Define a group representation $\rho : G \to GL(V)$ as disentangled with respect to $G = G_1 \times \cdots \times G_N$ if there exists a decomposition $V = V_1 \oplus \cdots \oplus V_N$ and representations $\rho_i : G_i \to GL(V_i)$ such that $\rho = \rho_1 \oplus \cdots \oplus \rho_N$, i.e. $\rho(g_1, \ldots, g_N)(v_1, \ldots, v_N) = (\rho_1(g_1)v(1), \ldots, \rho_N(g_N)v_N)$. A consequence of this definition is that ρ is disentangled if each factor of ρ is irreducible. Note that we can associate an action with a group representation through $g \cdot w = \rho(g)(w)$.

Given this, a linear disentangled representation is defined in SBDRL to be any $f : \mathcal{W} \to \mathcal{Z}$ that admits a disentangled group representation with respect to the decomposition $G = G_1 \times \cdots \times G_N$. As such we can look for mappings to $\mathcal{Z}$ where actions by G_i are equivalent to irreducible representations.

It will be useful in later sections to know that the (real) irreducible representations of the cyclic group C_n are the rotation matrices with angle $\frac{2\pi}{n}$.

ForwardVAE Higgins et al. [2018] provides the framework for linear disentanglement however intentionally restrained from empirical findings. Caselles-Dupré et al. [2019] presented the first results inducing such representations in VAE latent spaces. Through observing transitions induced by actions in a grid world with $G = C_x \times C_y$ structure, their model successfully learns the rotation matrix representations corresponding to the known irreducible representations of C_n.

In implementation, the model stores a learnable parameter matrix (a representation) for each possible action, which is applied when that action is observed. Under the reinforcement learning setting of environment-state-action sequences, they have labelled actions for each observation pair, allowing the action selection at each step. This is suitable for reinforcement problems, however cannot be applied for problems which lack action labelling.

Caselles-Dupré et al. [2019] subsequently show that linear disentangled representations are advantageous to the downstream task of predicting which action is applied between two sequential representations. This is similar to the β-VAE metric [Higgins et al., 2017] which comparatively attempts to predict a fixed action between two representations. We shall explicitly evaluate the β-VAE metric alongside other disentanglement metrics in Section 4 and the downstream task in the appendix.

Finally, they offer two theoretical proofs centred around linear disentangled representations. We briefly outline the theorems here: 1) Without interaction with the environment (observing action transitions), you are not guaranteed to learn linear disentangled representations with respect to any given symmetry structure. 2) It is impossible to learn linear disentangled representation spaces $\mathcal{Z}$ of order 2 for the Flatland problem, i.e. learn 1D representations for each component cyclic group.

Further symmetry structures ForwardVAE only explored cyclic symmetry structures $G = C_n \times C_n$, albeit with continuous representations that are expressive enough for $SO(2)$. Subsequent work by Quessard et al. [2020] explored (outside of the VAE framework) linear disentangled representations of $SO(2)$ and non-abelian (non-commutative) $SO(3)$. Similar to ForwardVAE, they required knowledge of the action at each step.

Subspace Diffusion Pfau et al. [2020] explored extracting linear disentangled planes in the data manifold under a metric. Assuming this metric is sufficiently strong, they find their method can extract up to 5 independent planes (i.e. symmetries) in a completely unsupervised way. However, unlike ForwardVAE and our work, this method fails when applied directly to the data or to a representation learnt by a VAE, since they do not provide sufficient metric structure. They find that the ideal metric is the labelled generative factors.

Learning Observed Actions Section 5 revolves around predicting observed actions, a concept which has some prior work with two particularly relevant works. Rybkin et al. [2018] learn a (composable) mapping from pre to post-action latents, conditioned on observing the action. Edwards

Figure 1: Sequential observations from the Flatland toy problem. If the agent would contact the boundary, it is instead warped to the opposing side (e.g. between the third and fourth observations).

et al. [2019] learn both a forward dynamics model to predict post action states (given state and action) and a distribution over actions given the initial state. Contrary to our work, both methods allow arbitrarily non-linear actions (parametrised by neural networks) which makes them unsuitable for our task. Furthermore they differ significantly in implementation.

3 Which Spaces Admit Linear Disentangled Representations

This section empirically explores admission of irreducible representations in latent spaces. In particular we look at standard VAE baselines and search for known true symmetry structure.

Problem Setting We shall use the Flatland toy problem [Caselles-Dupré et al., 2018] for consistency with ForwardVAE, a grid world with symmetry structure $G = C_x \times C_y$, manifesting as translation (in 5px steps) of a circle/agent (radius 15px) around a canvas of size $64px \times 64px$. Under the SBDRL framework, we have world space $\mathcal{W} = \{(x_i, y_i)\}$, the set of all possible locations of the agent. Note that the agent is warped to the opposite boundary if within a distance less than r. The observation space $\mathcal{O}$, renderings of this space as binary images (see Fig 1), is generated with the PyGame framework [Shinners] which represents the generative process b. The inference process h is parametrised by candidate VAE models, specifically the encoder parameters θ. The order of the cyclic groups is given by, $(64 - 2 * 15)/5 \approx 7$ which leads to the approximate phase angle of $\alpha \approx 0.924$. All candidate representation spaces shall be restricted to 4 dimensional for simplicity and consistency. All experiments report errors as one standard deviation over 3 runs (if applicable), use a randomly selected validation splits of 10% and run on a single GTX1070Ti consumer GPU. We shall evaluate the following baseline models, VAE [Kingma and Welling, 2014], β-VAE [Higgins et al., 2017], CC-VAE [Burgess et al., 2018], FactorVAE [Kim and Mnih, 2018] and DIP-VAE-I/II [Kumar et al., 2017a].

Evaluation Method Once we have a candidate representation (latent) space $\mathcal{Z}$, we need to locate potential irreducible representations $\rho(g)$ of action a by group element g. For the defined symmetry structure G, we know that the irreducible representations take the form of 2D rotation matrices. We further restrict to observations of actions by either of the cyclic generators g_x, g_y or their inverses g_x^{-1}, g_y^{-1}. Consequently, we store 4 learnable matrices to represent each of these possible actions. Since there is no requirement for representations to be admissible on axis aligned planes, we also learn a change of basis matrix for each representation. To locate group representations, we encode the pre-action observation and apply the corresponding matrix, optimising to reconstruct the latent representation of the post-action observation. As we iterate through the data and optimise for best latent reconstruction, the matrices should converge towards irreducible representations if admissible.

If cyclic representations are admissible, then our matrices should learn to estimate the post action latent with low error. However, since not all models will encode to the same space, we also compare to the expected distance between latent codes. If the reconstruction error is significantly lower than the expected distance, then we can say that the space is admissible to (in this case) cyclic irreducible representations. We also introduce a metric to explicitly measure the extent to which actions operate on independent subspaces as required by the definition of a linear disentangled representation. We call this metric the independence score and define it as a function of the latent code z and the latent code after applying action a, denoted z_a for actions of $G = G_0 \times \cdots \times G_N$,

$$\text{Independence Score} = \frac{1}{N!} \sum_{i,\, j \neq i} \max_{a \in G_i, b \in G_j} \left(\left| \frac{\mathbf{z} - \mathbf{z}_a}{||\mathbf{z} - \mathbf{z}_a||_2} \cdot \frac{\mathbf{z} - \mathbf{z}_b}{||\mathbf{z} - \mathbf{z}_b||_2} \right| \right) \quad . \tag{1}$$

Results For comparison, we first discuss results with ForwardVAE, where cyclic representations are known to reside in axis aligned planes. We provide in section B of the appendix explicit errors for post action estimation whilst restricted to each possible axis aligned plane. Comparing ForwardVAE with

Table 1: Mean reconstruction values through learnt estimation of a cyclic representation.

	Forward	VAE	β-VAE	CC-VAE	FactorVAE	DIP-I	DIP-II
$\|\|\hat{\mathbf{x}_a} - \mathbf{x}_a\|\|_1$	$\mathbf{0.011}_{\pm.004}$	$0.096_{\pm.004}$	$0.108_{\pm.020}$	$0.093_{\pm.015}$	$0.060_{\pm.027}$	$0.054_{\pm.011}$	$0.045_{\pm.013}$
$\|\|\hat{\mathbf{z}}_a - \mathbf{z}_a\|\|_1$	$\mathbf{0.034}_{\pm.023}$	$0.328_{\pm.029}$	$0.255_{\pm.037}$	$0.151_{\pm.028}$	$1.286_{\pm.748}$	$0.431_{\pm.140}$	$0.270_{\pm.106}$
$\|\|\hat{\alpha} - \alpha\|\|_1$	$\mathbf{0.009}_{\pm.013}$	$0.054_{\pm.366}$	$0.07_{\pm.316}$	$0.100_{\pm.297}$	$0.260_{\pm.190}$	$0.139_{\pm.147}$	$0.051_{\pm.062}$
Indep	$\mathbf{0.926}_{\pm.063}$	$0.791_{\pm.109}$	$0.581_{\pm.475}$	$0.289_{\pm.465}$	$0.547_{\pm.362}$	$0.655_{\pm.216}$	$0.814_{\pm.119}$

baseline VAEs, Table 1 reports mean reconstruction errors for the post action observations x_a and latent codes z_a, the estimated phase angle for cyclic representations (against known true value α) and the independence score. We can see that none of the standard baselines achieve reconstruction errors close to that of ForwardVAE. Neither do we see good approximations of the true cyclic representation angles (α) in any model but ForwardVAE. Whilst the mean α learnt for the VAE and β-VAE are close to the ground truth, the deviation is very large. Dip-II achieves lower error however it is still an order of magnitude worse than ForwardVAE. This is further reflected in the independence scores of each model, with ForwardVAE performing strongly and consistently whilst the baselines perform worse and with high variance. These results suggests that none of the baseline models have effectively learnt a linear disentangled representation.

4 Are Linear Disentangled Representations Advantageous for Classical Disentanglement

This section expands on the single downstream performance task evaluated by Caselles-Dupré et al. [2019]. We evaluate models with a suite of disentanglement metrics from the literature. By Locatello et al. [2018], it is known that not all such metrics correlate, and as such it is useful to contrast linear disentanglement with previous understandings.

Problem Setting Flatland has two generative factors, the x and y coordinates of the agent. With these we can evaluate standard disentanglement metrics in an effort to discern commonalities between previous understandings. In this work we adapt the open source code of Locatello et al. [2018] to PyTorch and evaluate using the following metrics: Higgins metric (Beta) [Higgins et al., 2017], mutual information gap (MIG) [Chen et al., 2018], DCI Disentanglement metric [Eastwood and Williams, 2018], Modularity metric [Ridgeway and Mozer, 2018], SAP metric [Kumar et al., 2017b]. In addition, we evaluate two further metrics that expand on previous works: Factor Leakage (FL) and the previously introduced Independence score.

Factor Leakage Our Factor Leakage metric is descended from the MIG metric which measures, for each generative factor, the difference in information content between the most informative latent dimension and the second most informative dimension. Considering linear disentangled group representations are frequently of dimension 2 or higher, this would result in low MIG scores, which would imply entanglement despite being obviously disentangled. We extend it by measuring the information of all latent dimensions for each generative factor (i.e. each group action). The FL metric then reports the mean of the areas under this curve for each action/factor. Intuitively, entangled representations encode actions over all dimensions; the fewer dimensions the action is encoded in, the more disentangled it is considered with respect to this metric.

Results Table 2 reports classical disentanglement metrics alongside our FL and independence metrics with comparisons of ForwardVAE to baseline VAEs. ForwardVAE clearly results in stronger scores on metrics which do not assume each generative factor is encoded in a single latent dimension. Naturally, the 2D cyclic representations of FlatLand will not perform well in these cases. Indeed, by Caselles-Dupré et al. [2019], no action can be learnt in a single dimension on this problem. We can see however that it performs extremely strongly and consistently on the other metrics, including Factor Leakage which shows the information gap between factors is large. We also report the correlation of the independence scores against all other metrics. We see quite strong correlations with all metrics other than SAP and MIG, which both depend on single latent dimensions per factor. From the performance of ForwardVAE and the correlation of independence and other metrics, we can see linear disentangled representations are beneficial for the relevant classical metrics.

Table 2: Disentanglement metrics for baseline models on Flatland and the spearman correlation of independece score with each (italics indicate low confidence).

	Beta	MIG	DCI	Mod	SAP	FL	Indep
Forward	$\mathbf{1.000}_{\pm.001}$	$0.025_{\pm.007}$	$0.935_{\pm.009}$	$\mathbf{0.974}_{\pm.000}$	$0.293_{\pm.002}$	$\mathbf{0.329}_{\pm.005}$	$\mathbf{0.961}_{\pm.013}$
VAE	$0.998_{\pm.003}$	$\mathbf{0.303}_{\pm.164}$	$0.209_{\pm.165}$	$0.607_{\pm.069}$	$\mathbf{0.733}_{\pm.160}$	$0.359_{\pm.106}$	$0.740_{\pm.026}$
β-VAE	$0.857_{\pm.128}$	$0.194_{\pm.080}$	$0.056_{\pm.065}$	$0.348_{\pm.214}$	$0.400_{\pm.204}$	$0.485_{\pm.118}$	$0.317_{\pm.432}$
cc-VAE	$0.721_{\pm.123}$	$0.076_{\pm.047}$	$0.013_{\pm.021}$	$0.228_{\pm.065}$	$0.166_{\pm.075}$	$0.653_{\pm.193}$	$0.066_{\pm.030}$
Factor	$0.999_{\pm.002}$	$0.058_{\pm.061}$	$0.249_{\pm.126}$	$0.865_{\pm.054}$	$0.519_{\pm.160}$	$0.444_{\pm.074}$	$0.864_{\pm.027}$
DIP-I	$0.858_{\pm.130}$	$0.071_{\pm.121}$	$0.205_{\pm.332}$	$0.528_{\pm.343}$	$0.487_{\pm.317}$	$0.633_{\pm.273}$	$0.764_{\pm.154}$
DIP-II	$0.898_{\pm.171}$	$0.040_{\pm.046}$	$0.151_{\pm.205}$	$0.575_{\pm.345}$	$0.573_{\pm.313}$	$0.636_{\pm.244}$	$0.747_{\pm.222}$
Corr	0.800	-0.467	0.933	0.833	0.233	-0.600	1.000

5 Unsupervised Action Estimation

ForwardVAE explicitly requires the action at each step to be able to induce linear disentangled structures in latent spaces. We now show that policy gradients allow jointly learning to estimate the observed action alongside learning the latent representation mapping. We will then examine properties of the model such as learning over longer term action sequences and temporal consistency. Experimental details are reported in appendix section A and code will be made public upon publication. Before introducing the model, we briefly outline the RL methods that it will utilise.

Policy Gradients Policy gradient methods will allow us to optimise through a Categorical sampling of η-parametrised distribution $p(A|\psi, s)$ over possible choices $\{A_1, \dots, A_N\}$ and conditioned on state s. The policy gradient loss in the REINFORCE [Williams, 1992] setting where action A_i receives reward $R(A_i, s)$ is given by,

$$\mathcal{L}_{policy} = \begin{cases} -\log(p(A_i|\psi, s)) \cdot R(A_i, s) & \text{if } R(A_i, s) > 0 \\ -\log(1 - p(A_i|\psi, s)) \cdot |R(A_i)| & \text{if } R(A_i, s) < 0 \end{cases} \quad . \tag{2}$$

We find that minimising the regret $R(A_i, s) - \max_j R(A_j, s)$ instead of maximising reward provides more stable learning. Furthermore, to encourage exploration we subtract the weighted entropy $0.01H(p(A_i|\psi, s))$, a technique used by methods such as Soft Actor-Critic [Haarnoja et al., 2018].

Reinforced GroupVAE Reinforced GroupVAE (RGrVAE) is our proposed method for learning linear disentangled representation in a VAE without the constraint of action supervision. A schematic diagram is given in Figure 2. Alongside a standard VAE we use a small CNN (parameter ψ) which infers from each image pair a distribution over a set of possible (learnable) representation matrices. Unlike ForwardVAE, we restrict these to being irreducible cyclic representations by simply learning a rotational phase angle (we explore generic matrices in appendix B). Further we introduce a decay loss on each representation towards the identity representation, since we prefer to use the minimum number of representations possible. The policy selection distribution is sampled categorically and this representation matrix is applied to the latent code of the pre-action image with the aim of reconstructing the latent code of the post-action image. The policy selection network is optimised through regret minimising policy gradients of the previous paragraph with entropy regularisation to encourage exploration. Rewards are provided as a function of the pre-action latent code $\mathbf{z}$, the post action latent code $\mathbf{z}_a$ and the predicted post action code $\hat{\mathbf{z}}_a$,

$$R(a, \mathbf{x}, \mathbf{x}_a) = ||\mathbf{z} - \mathbf{z}_a||_2^2 - ||\hat{\mathbf{z}}_a - \mathbf{z}_a||_2^2 \quad . \tag{3}$$

Similar to ForwardVAE, we then train a standard VAE with the additional policy gradient loss and a weighted prediction loss given by

$$\mathcal{L}_{pred} = ||\hat{\mathbf{z}}_a - \mathbf{z}_a||_2^2 \quad , \quad \mathcal{L}_{\text{total}} = \mathcal{L}_{\text{VAE}} + \mathcal{L}_{\text{policy}} + \gamma\mathcal{L}_{\text{pred}} \quad . \tag{4}$$

Flatland To demonstrate the effectiveness of the policy gradient network, we evaluate RGrVAE on Flatland. We allow 4 latent dimensions and initialise 2 cyclic representations per latent pair with random angles but alternating signs to speed up convergence. Examples of the learnt RGrVAE

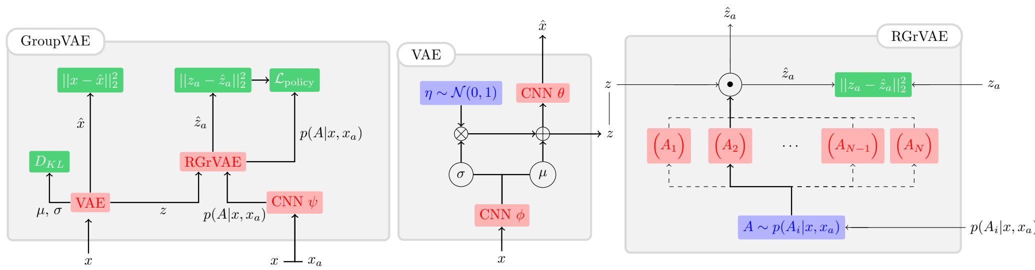


Figure 2: Schematic diagram for RGroupVAE and components. Dashes denote possible selections in the policy. • denotes matrix vector multiplication.
- Learnable module. - Loss. - Operation without parameters

Table 3: Comparison of RGrVAE to ForwardVAE for latent space and disentanglement metrics.

(a) Reconstruction errors for post action observation ($\mathbf{x}_a$), latent ($\mathbf{z}_a$) and representation α.

	ForwardVAE	RGrVAE
$\lvert\lvert\hat{\mathbf{x}}_a - \mathbf{x}_a\rvert\rvert_1$	$\mathbf{0.011}\pm 0.004$	0.016 ± 0.004
$\lvert\lvert\hat{\mathbf{z}}_a - \mathbf{z}_a\rvert\rvert_1$	$\mathbf{0.034}\pm 0.023$	0.100 ± 0.029
$\lvert\lvert\hat{\alpha} - \alpha\rvert\rvert_1$	$\mathbf{0.009}\pm 0.013$	0.012 ± 0.040
Independence	0.961 ± 0.013	0.955 ± 0.014

(b) Disentanglement Metrics for Flatland RGrVAE.

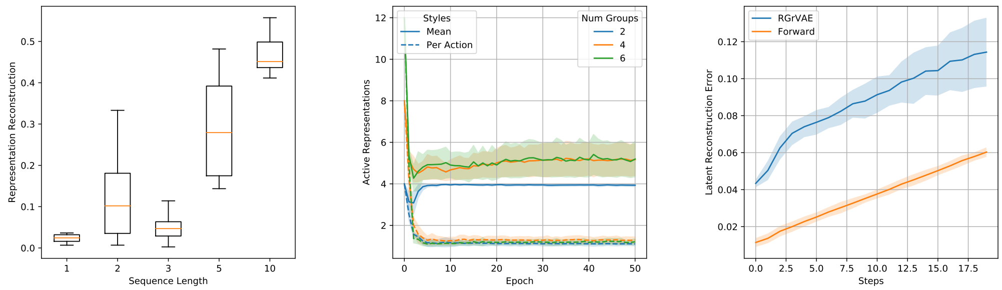
Metric	ForwardVAE	RGrVAE
Beta	$\mathbf{1.000}_{\pm .001}$	$\mathbf{1.000}_{\pm .001}$
MIG	$0.025_{\pm .007}$	$\mathbf{0.105}_{\pm .041}$
DCI	$\mathbf{0.935}_{\pm .009}$	$0.695_{\pm .090}$
Mod	$\mathbf{0.974}_{\pm .000}$	$0.918_{\pm .038}$
SAP	$0.293_{\pm .002}$	$\mathbf{0.636}_{\pm .073}$
FL	$\mathbf{0.329}_{\pm .005}$	$\mathbf{0.3434}_{\pm .017}$
Indep	$\mathbf{0.961}_{\pm .013}$	$\mathbf{0.955}_{\pm .014}$

actions are given in appendix section C. Table 3a reports reconstruction errors against ForwardVAE. RGrVAE achieves similar error rates to ForwardVAE. The latent reconstruction, showing the largest error is still far below the baselines (with the exception of CC-VAE, which performs much worse on the independence score). This proves the basic premise that RGrVAE can induce linear disentangled representations without action supervision. We will now further demonstrate this through disentanglement metrics.

Classical Disentanglement Metrics Table 3b compares ForwardVAE and RGrVAE with classical disentanglement metrics. We can see that despite removing the requirement of action supervision, RGrVAE performs similarly to ForwardVAE on the Beta, Modularity, FL and Independence metrics. The SAP score is more comparable to baseline methods, implying that despite the high independence and low cyclic reconstruction errors, RGrVAE learns relies more on single dimensions than ForwardVAE, the DCI disentanglement is also lower and the MIG larger, which are both consistent with this finding. Despite this, the DCI disentanglement is comparatively much larger than baselines, which shows, with the Beta, FL and independence scores that RGrVAE does not encode factors in solely single dimensions, instead capturing a linear disentangled representation.

Longer Action Sequences By applying a chosen action and reconstructing with the result, we have a new image pair (with the target) for which we can infer another action. Repeating this allows us to explore initial observation pairs which differ by more than the application of a single action. In the limit we would remove the requirement of having sequential data, however by [Caselles-Dupré et al., 2019], we know that we can no longer guarantee linear disentanglement with respect to any particular symmetry structure. Figure 3a reports the C_N α reconstruction error, where we see that larger steps results in gradually degraded estimation, as might be expected, however it is relatively consistent for a small number of steps.

Choosing the Number of Actions Recall from previous paragraphs, we introduced weight decay on the representations (towards the identity representation) in an effort to reduce the impact of over-representation preferring to use as few representations as possible. This is obviously desirable when



(a) (b) (c)

Figure 3: Exploring properties of RGrVAE. a) α reconstruction MSE. b) Estimated active number of representations in total (solid) and per action (dashed) over training for different total available representations. c) Latent reconstruction MSE for applying actions over a number of steps.

the exact symmetry structure of the dataset is unknown since it allows us to over-represent the actions and still learn their structure. In Figure 3b we plot an estimate of the number of total/per-action active representations ($N \approx e^h$ where h is the mean entropy or per action entropy). Over training we see the number of active representations decrease, towards 1 for each action but between 4 and 6 for the total. This total is higher than the ideal 4. Since for each action there is close to 1 active representation we expect it is increased by policies for multiple actions choosing the same minority representation some small fraction of the time.

Temporal Consistency A desirable property of representing actions is that they remain accurate over time - the representation isn't degraded as we apply new actions. Figure 3c reports the latent reconstruction error $||\hat{\mathbf{z}}_a - \mathbf{z}_a||_1$ as additional actions are applied. We can see that the quality of representations degrades gradually as we add steps and indeed is only marginally worse than the error achieved by ForwardVAE. Note that for this test, only the initial observation is provided to the model, unlike our experiment with longer action sequences, where predicted observations are fed back into it as it works towards a terminal observation.

6 Conclusion

Symmetry based disentangled representation learning offers a solid framework through which to study linear disentangled representations. Reflected in classical disentanglement metrics, we have shown empirically where such representations are beneficial, and how their presence is desirable. However, we find that even for simple problems, linear disentangled representations are not learnt by classical VAE baseline models, they require structuring the loss to favour them. Previous works achieved this through action supervision and imposing a reconstruction loss between post action latents and a learnt linear map (for each action) applied to the pre action latent. We introduced our independence metric to provide a means to measure the quality of linear disentangled representations, before introducing Reinforce GroupVAE, which offers a means to induce these representations without explicit knowledge of actions at every step by inferring the action that occurred. Without knowledge of the number or type of actions, we show that RGrVAE prefers to learn internal representations that reflect the true symmetry structure and ignore superfluous ones. We also find that it still performs when observations are no longer separated by just a single action, and can model short term action sequences.

7 Broader Impact

Representation learning as a whole does have the potential for unethical applications. Disentanglement if successfully applied to multi-object scenes could allow (or perhaps require, this is unclear) segmentation/separation of individual objects and their visual characteristics. Both segmentation and learning visual characteristics have numerous ethical and unethical uses on which we wont speculate. For our particular work, we don't believe understanding and encouraging linear disentangled representations has any ethical considerations beyond the broad considerations of disentanglement work as a whole. We do believe that routes towards reducing the degree of human annotation in data (such as our proposed model) is beneficial for reducing human bias, although this can introduced even by the choice of base data to train on. Unfortunately for our work (and most unsupervised work), explicit supervision allows for more rapid convergence and consequently a lower environmental impact, a topic which is of increasing concern especially for deep learning which leans heavily on power intensive compute hardware.

References

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in beta-vae. *arXiv preprint arXiv:1804.03599*, 2018.

Hugo Caselles-Dupré, Louis Annabi, Oksana Hagen, Michael Garcia-Ortiz, and David Filliat. Flatland: a lightweight first-person 2-d environment for reinforcement learning. *arXiv preprint arXiv:1809.00510*, 2018.

Hugo Caselles-Dupré, Michael Garcia Ortiz, and David Filliat. Symmetry-based disentangled representation learning requires interaction with environments. In *Advances in Neural Information Processing Systems*, pages 4608–4617, 2019.

Tian Qi Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 2610–2620, 2018.

Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. 2018.

Ashley Edwards, Himanshu Sahni, Yannick Schroecker, and Charles Isbell. Imitating latent policies from observation. In *International Conference on Machine Learning*, pages 1755–1763, 2019.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 2018.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *Iclr*, 2(5):6, 2017.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations, ICLR*, 2014.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*, 2017a.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*, 2017b.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv preprint arXiv:1811.12359*, 2018.

David Pfau, Irina Higgins, Aleksandar Botev, and Sébastien Racanière. Disentangling by subspace diffusion. *arXiv preprint arXiv:2006.12982*, 2020.

Robin Quessard, Thomas D Barrett, and William R Clements. Learning group structure and disentangled representations of dynamical environments. *arXiv preprint arXiv:2002.06991*, 2020.

Karl Ridgeway and Michael C Mozer. Learning deep disentangled embeddings with the f-statistic loss. In *Advances in Neural Information Processing Systems*, pages 185–194, 2018.

Oleh Rybkin, Karl Pertsch, Konstantinos G Derpanis, Kostas Daniilidis, and Andrew Jaegle. Learning what you can do before doing anything. *arXiv preprint arXiv:1806.09655*, 2018.

Pete Shinners. Pygame. `http://pygame.org/`.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

A Experimental Details

This section lists the experimental details for each figure in the main paper.

Tables 1, 2, 3 All results were run with the following parameters: learning rate 1e-3, 100 epochs, 4 latent dimensions, Adam optimiser, 3 repeats, MSE loss. Estimations $\hat{\alpha}$ were trained for 5 epochs using the Adam optimiser with MSE loss and learning rate 0.1. RGrVAE used a latent reconstruction weight $\gamma = 10$.

Table 4: β choices for each VAE model. † Capacity 5, 3500 step leadin

	Forward	RGrVAE	VAE	β-VAE	CC-VAE	FactorVAE	DIP-I	DIP-II
β	1	1	1	15	1000†	1	1	1

Figure 3 All results use the same setup as the previous tables with the following exception: Figure 3a, RGrVAE group structure is two cyclic representations alongside an identity representation, for each latent pair.

B Additional Experiments

Axis aligned alpha reconstruction The table below reports errors for reconstructing post action latents whilst restricted to purely axis aligned planes.

Table 1: Validation reconstruction MSE between predicted z (after applying cyclic representation) and true z (encoding post-action image) in each latent plane, alongside the expected distance between any two latent codes over the dataset. Standard VAEs never achieve low errors, cyclic representations are not present. Actions a: 0 - up, 1 - left, 2 - down, 3 - right.

	VAE						ForwardVAE					
a	0, 1	0, 2	0, 3	1, 2	1, 3	2, 3	0, 1	0, 2	0, 3	1, 2	1, 3	2, 3
0	0.711	0.718	0.714	0.718	**0.524**	0.718	**0.006**	0.525	0.525	0.529	0.525	0.525
1	**0.755**	0.913	0.898	0.918	0.912	0.918	0.475	0.475	0.478	0.474	0.476	**0.004**
2	0.770	0.786	0.770	0.784	**0.609**	0.785	**0.006**	0.529	0.529	0.531	0.528	0.529
3	**0.678**	0.878	0.871	0.881	0.875	0.881	0.476	0.475	0.487	0.479	0.476	**0.004**
	Independence: 0.791 Expected distance: 0.821						Independence: 0.926 Expected distance: 0.502					

Different RGrVAE Representations We present a brief exploration allowing more expressivity in RGrVAE internal representations. In the main paper, these representations were solely cyclic, where the phase angle α is the only learnable parameter. We now explore generic matrices such as those used by ForwardVAE. We report in Table 2 the disentanglement scores for the cyclic representations verses the generic matrix representations. Both methods achieve similar results, the major difference between them is in convergence rate. Figure 4 compares an estimated independence score over training for each method. The cyclic representations converge extremely quickly, whereas the matrix representations get stuck in local minima for long periods of time before eventually converging to the global minima.

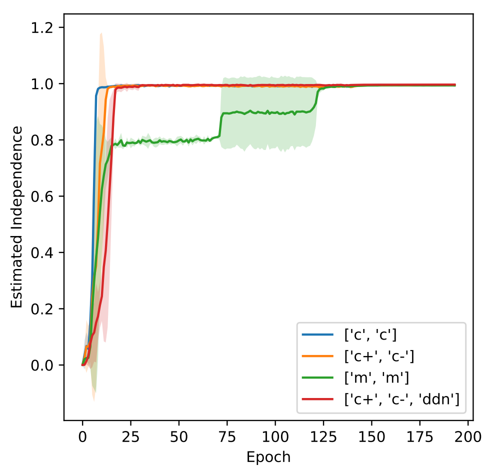
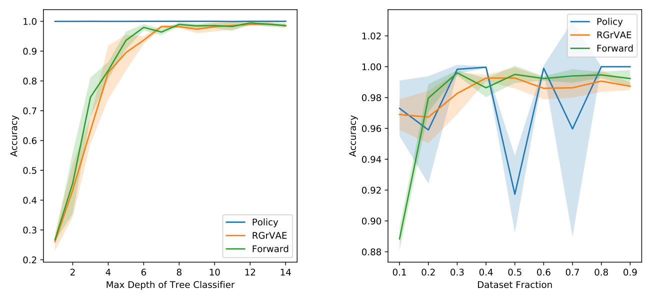


Figure 4: Estimated independence score over training for different internal RGrVAE representation structures. 'c' denotes standard cyclic representation. 'c+/-' denotes initialised cyclic representation to positive/negative. 'm' denotes generic 2D matrix representation. 'ddn' denotes (not learnable) representation of reflection.

Downstream Tasks We evaluate on the downstream task of Caselles-Dupré et al. [2019], learning an inverse model. We find that RGrVAE representations have the same qualities as it's ForwardVAE counterparts and we can also perform the same tasks using the policy network. For a dataset of

Table 2: Disentanglement metric scores for the different internal representation choices.

Metric	Cyclic	Cyclic Initialised	Matrices	Cyclic + Reflection
Beta	$1.000_{\pm.000}$	$1.000_{\pm.000}$	$1.000_{\pm.000}$	$1.000_{\pm.000}$
MIG	$0.096_{\pm0.030}$	$0.055_{\pm0.028}$	$0.049_{\pm0.030}$	$0.085_{\pm0.020}$
DCI	$0.814_{\pm0.077}$	$0.747_{\pm0.070}$	$0.738_{\pm0.056}$	$0.770_{\pm0.108}$
Mod	$0.954_{\pm0.004}$	$0.953_{\pm0.015}$	$0.950_{\pm0.012}$	$0.952_{\pm0.017}$
SAP	$0.500_{\pm0.043}$	$0.559_{\pm0.019}$	$0.584_{\pm0.012}$	$0.540_{\pm0.160}$
FL	$0.346_{\pm0.025}$	$0.369_{\pm0.010}$	$0.356_{\pm0.025}$	$0.346_{\pm0.016}$
Indep	$0.905_{\pm0.068}$	$0.932_{\pm0.029}$	$0.908_{\pm0.045}$	$0.924_{\pm0.061}$



(a) Varying tree depth. (b) Varying training set size

Figure 5: Downstream Performance

observation pairs (x, x_a) which differ by action a, we aim to use linear disentangled representations to classify which action occurred in the observation. We compare ForwardVAE, RGrVAE and the policy distributions of RGrVAE.

For this we train a random forest classifier in accordance with Caselles-Dupré et al. [2019], and compare the accuracy with varied forest depth and varied data available to train the representations. From Figure 5a we see that ForwardVAE and RGrVAE allow similar classification performance with depth with regards to their representations, however the policy network outperforms both models with 100% accuracy for any depth of tree. This is unsurprising since the policy network in the ideal case will exactly be a classifier of this action.

From Figure 5b we see that ForwardVAE and RGrVAE have similar characteristics as we vary the number of training examples (out of 10000 total). The policy network performance is far more inconsistent on reduced training sets when compared to full training data.

Attention Attentional mechanisms offer alternate means to allow gradient through a distribution over choices. Instead of sampling the distribution and using policy gradients, attention forms a linear combination of the choices weighted by the distribution. In our case we are interested in learning the correct irreducible representations which requires each representation to learn exactly the correct cyclic phase angle. When we predict post-action latent codes through a linear combination of representations, we lose the guarantee that the gradient will point towards this solution. Since reinforce applies solely one representation exactly once, we are guaranteed that (if the policy is accurate and the latent structure is amenable) the gradient will point towards this solution. We find that the cyclic representation error $||\hat{\alpha} - \alpha|| = 0.157$ is far worse than the 0.012 error of RGrVAE. Furthermore, the independence score is $0.830_{\pm0.109}$ which is comparatively low compared to RGrVAE ($0.955_{\pm0.014}$) which larger deviation. These statistics showed us that the reinforcement method was a better candidate to learn linear disentangled representations.

C Figures

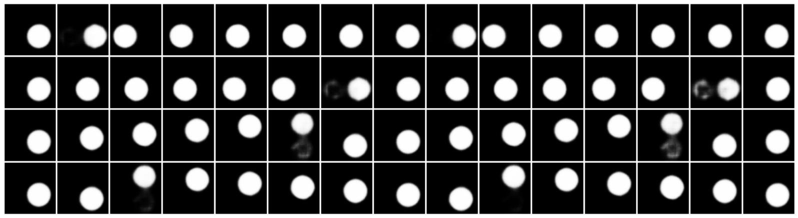


Figure 1: Actions learnt by RGrVAE corresponding with the environment actions up, down, left and right. Note the wrapping at boundaries which is expected by the symmetry structure.