THE ROLE OF DISENTANGLEMENT IN GENERALISATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Combinatorial generalization — the ability to understand and produce novel combinations of familiar elements — is considered a core capacity of humans that modern AI systems currently struggle with. Recently, it has been suggested that learning disentangled representations may help address this problem. It is claimed that such representations should be able to capture the compositional structure of the world which can then be combined to produce novel representations. That is, compositional representations are thought to support combinatorial generalisation. In this study, we systematically tested how the degree of disentanglement affects various forms of generalisation, including two forms of combinatorial generalisation that varied in difficulty. We trained variational autoencoders (VAEs) with different levels of disentanglement on an unsupervised task by excluding combinations of generative factors during training. At test time we ask the models to reconstruct the missing combinations in order to measure generalisation performance. Irrespective of the degree of disentanglement, we found that the models supported only weak combinatorial generalisation. Next, we tested our model in a more complex task which explicitly required independent generative factors to be controlled. Although this added pressure improved the level of disentanglement compared to normal VAEs, generalisation did not improve. While learning disentangled representations does improve interpretability and sample efficiency in downstream tasks, our results suggest that they are not sufficient for supporting more difficult forms of generalisation.

1 INTRODUCTION

Generalization to unseen data has been a key challenge for neural networks since the early days of connectionism, with considerable debate about whether these new models can emulate the kinds of behaviours that are present in humans (McClelland et al., 1986; Fodor & Pylyshyn, 1988; Smolensky, 1987; 1988; Fodor & McLaughlin, 1990). While the modern successes of Deep Learning do indeed point to impressive gains in this regard, human level generalisation still remains elusive (Lake & Baroni, 2018; Marcus, 2018). One explanation for this is that humans encode stimuli in a compositional manner, with a small set of independent and more primitive features (e.g., separate representations of size, position, line orientation, etc.) being used to build more complex representation (e.g., a square of a given size and position). The meaning of the more complex representation comes from the meaning of it's parts. Critically, compositional representations afford the ability to recombine primitives in novel ways: if a person has learnt to recognize squares and circles in context where all squares are blue and all circles are red, they can nevertheless also recognise red squares, even though they have never seen these in the training data. This ability to perform *combinatorial generalisation* based on compositional representations is thought to be a hallmark of human level intelligence (Fodor & Pylyshyn, 1988) (See McClelland et al. (1986) for a diverging opinion).

Recently it has been proposed that generalisation in neural networks can be improved by extracting disentangled representations (Higgins et al., 2017) from data using (variational) generative models (Kingma & Welling, 2013; Rezende et al., 2014). In this view, disentangled representations capture the compositional structure of the world (Higgins et al., 2018a; Duan et al., 2020), separating the generative factors present in the stimuli into separate components of the internal representation (Higgins et al., 2017; Burgess et al., 2018). It has been argued that these representations allow downstream models to perform better due to the structured nature of the representations (Higgins et al., 2017; 2018b) and to share information across related tasks (Bengio et al., 2014).

In this study we systematically tested whether and how disentangled representations support various forms of generalisation, including combinatorial generalization. We trained a set of VAEs and manipulated the degree of disentanglement of their internal representations. We used two methods to do this: by varying the β parameter on a simple reconstruction task in a β-VAE or by using a novel architecture to solve a more challenging task that requires combining generative factors of the input. We tested how generalisation achieved by these VAEs varied as a function of the degree of disentanglement. We found that models only supported the simplest versions of combinatorial generalization and the degree of disentanglement had no impact on the degree of generalization. These findings are consistent with the claim that disentangled representations are necessary but not sufficient for supporting combinatorial generalization.

1.1 Previous Work

Recent work on learning disentangled representations in unsupervised generative models has indeed shown some promise in improving the performance of downstream tasks (Higgins et al., 2018b; van Steenkiste et al.) but this benefit is mainly related to sample efficiency rather than generalisation. Indeed, we are only aware of two studies that have considered the importance of learned disentanglement for combinatorial generalization and they have used different network architectures and have reached opposite conclusions. Bowers et al. (2016) showed that a recurrent model of short-term memory tested on lists of words that required some degree of combinatorial generalization (recalling a sequence of words when one or more of words at test were novel) only succeeded when it had learned highly selective (disentangled) representations ("grandmother cell" units for letters). By contrast, Chaabouni et al. (2020) found that models with disentangled representations do not confer significant improvements in generalisation over entangled ones in a language modeling setting, with both entangled and disentangled representations supporting combinatorial generalization as long as the training set was rich enough (we return to this issue in the General Discussion). At the same time, they found that languages generated through compositional representations were easier to learn, suggesting this as a pressure to learn disentangled representations.

A number of recent papers have reported that VAEs can support some degree of combinatorial generalisation, but there is no clear understanding of whether and how disentangled representations played any role in supporting this performance. Esmaeili et al. (2019) showed that a model trained on the MNIST dataset could reconstruct images even when some particular combination of factors were removed during training, such as a thick number 7 or a narrow 0. The authors also showed that the model had learned disentangled representations and concluded that the disentangled representations played a role in the successful performance. However, the authors did not vary the degree of disentanglement in their models and, accordingly, it is possible that a VAE that learned entangled representations would do just as well. Similarly, Higgins et al. (2018c) have highlighted how VAEs that learn disentangled representations can support some forms of combinatorial generalisation when generating images from text. For example, their model could render a room with white walls, pink floor and blue ceiling even though it was never shown that combination in the training set. This is an impressive form of combinatorial generalisation but, as we show below, truly compositional representations should be able to support several other forms of combinatorial generalisations that were not tested in this study. Moreover, it is not clear what role disentanglement played in this successful instance of generalisation. Finally, Zhao et al. (2018) assessed VAE performance on a range of combinatorial generalization tasks that varied in difficulty, and found that the model performed well in the simplest settings but struggled in more difficult ones. But again, they did not consider whether learning disentangled representations was relevant to generalisation performance.

2 Methods

All of our experiments are based on the dSprites image dataset (Matthey et al., 2017), which contains more than 700k images that vary five generative factors – shape, scale, orientation and translation along two axes. To test combinatorial generalisation, we systematically excluded some combinations of these generative factors from the training data and tested reconstruction on these unseen values. For example, during training we might exclude a particular shape (with its corresponding value of scale and rotation) from appearing at a particular position. The model is trained on the rest of the

combinations and then tested on the excluded one. Applying this procedure in different ways allows us to systematically test the generalisation capacity of the models.

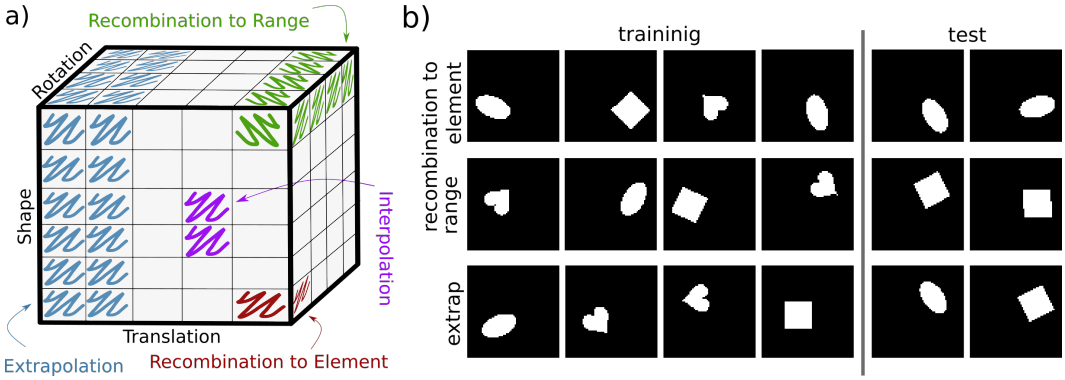

Figure 1: **Testing generalisation in image reconstruction.** (a) An illustration of different forms of generalisation based on three generative factors – shape, translation and rotation. The blank cells represent combinations that the model is trained on. Coloured cells represent novel test combinations that probe different forms of generalisation: Interpolation (purple), Recombination to Element (red), Recombination to Range (green) and Extrapolation (blue) – see main text for details. (b) Each row shows an example of training and test stimuli for testing a form of generalisation. In the top row, training set excludes ellipses in the bottom-right corner at less than 120° though they are present at the bottom-right corner at other rotations. In the middle row, training set excludes squares in the right side of the image though other shapes and rotations are present at this location and squares are seen at all other combinations of rotations and translations. In the bottom row, training set excludes all shapes on the right side of the image.

It is important to note that there are several forms of generalisation, varying in difficulty and only some of these can be classified as combinatorial generalisation. To systematically test a model's ability to generalise, we make the distinction between four different forms of generalisation in the dataset used in this study as illustrated in Figure 1:

- **Interpolation** (purple squares in Figure 1): This is the simplest form of generalisation. The model is tested on an unseen value of one of the generative factors, but it has been trained on neighbouring values. An example is a model trained on squares at rotations 20° and 30° but tested on squares presented at 25°. Interpolation is *not* a form of combinatorial generalisation as this does not involve combining values of different generative factors. Most deep learning systems are able to show this form of generalisation, and indeed, our models succeeded in all cases. And so we will not concern ourselves with it further.
- **Recombination to Element** (red squares in Figure 1): The model is trained on various values and combinations of generative factors, but one of the combinations (or small range of similar combinations) is left out. An example is a model that has never been trained on ellipses in the bottom-right corner at rotations less than 120°, though it has been trained on other rotations in this region. It has also been trained on ellipses at all rotations (including the ones less than 120°) at all other locations. This is an instance of combinatorial generalisation problem – a model that learns compositional representations for shape, translation and rotation should be able to combine observed values of shape (ellipses), translation (bottom-right) and rotation (0° to 120°) to generalise to the unseen images.
- **Recombination to Range** (green squares in Figure 1): The model has never seen one of the generative factors in combination with a range of values of another factor. An example is a model that has never been trained on squares on the right side of the image at *any* rotations, though it has been trained on shapes at all locations (including the right side of the image) at all other rotations and has also been trained on squares at all rotations on the left side of the image. This is again an instance of combinatorial generalisation as the model can generalise its observations of shapes (square), translation (right side) and rotations (all) to render unseen images. We expect a truly compositional system to be able to show this form

of generalisation because all the components of the novel observation have been previously seen by this system, even though the combination is new.

- **Extrapolation** (blue squares in Figure 1): This is the most challenging form of generalisation where models are tested on values of generative factors that are beyond the range of values observed in the training dataset. For example, testing a model on a spatial location, say the right side, even though it has never seen any shape being presented at this location. This is *not* a form of combinatorial generalisation as the model cannot combine observed values of generative factors to render images in the right side of the screen.

We tested the Recombination to element case by excluding a combination of values from the training set. We split each range of values of a generative factor into three bins, so that we had $3 \times 3 \times 3 \times 3 \times 3$ such combinations of bins for all five generative factors. During training, we removed one of these 243 combinations but presented the models the other 242.

For the Recombination to range case, we tested three different variants. First, we excluded a particular value of shape (for our experiments we used squares) when combined with all translation values beyond $x = 0.50$. The model sees other shapes at those positions during training and it sees squares on the left-hand side of the screen. Thus the models experiences both generative factor values independently and has to recombine them to produce a novel image at test time. We tested the same pattern of exclusions using scale and rotation, so that we excluded some but not all values of scales and rotation (in separate runs) for a given shape.

Finally, for the Extrapolation case, we excluded all combinations of generative factors where translation along the X-axis exceeded a cutoff value, x. We chose a set of different values: $x \in 0.16, 0.25, 0.50, 0.75$, where x is normalised in the range $[0, 1]$ (results shown in Figure 2 for $x = 0.50$). At test time the model needed to reconstruct images where translation along the x-axis, x, was greater than the cutoff value.

The models that we tested are all Variational Autoencoder architectures (VAEs, Kingma & Welling (2013); Rezende et al. (2014)), which we trained for 100 epochs using the β-VAE objective (Higgins et al., 2018b). The architectures are the ones found in Higgins et al. (2017) and Burgess et al. (2018) (Details in the Appendix). For each of these we tested three values of β on each condition: 1, 8 and 12. We used a batch size of 64 and a learning rate of $5e-4$ for the Adam optimizer (Kingma & Ba, 2017).

We measured the level of disentanglement using the framework introduced in Eastwood & Williams (2018). The procedure consists of using the latent representations generated for each image to predict the true generative factors using a regression model (in our case, Lasso regression; see Appendix). The level of disentanglement is quantified by their 'Overall disentanglement metric', which we call D-score here.

3 STANDARD UNSUPERVISED TASK EXPERIMENTS

We first test the models on the traditional unsupervised task for autoencoders, where they have to reconstruct the input image. Table 1 shows the results from this first set of experiments. The amount of disentanglement achieved by the models varied over a broad range and was a function of model architecture and the β-value used and the training set. The best disentanglement and reconstruction score on the test data for each condition is highlighted in bold. We found that models were able to perform the Recombination to Element generalisation but failed to generalise in the Recombination to Range and Extrapolation cases. Moreover, the amount of generalisation did not depend on the degree of disentanglement.

Figure 2 shows examples of model reconstructions for each of the conditions which help highlight the qualitative issues with their performance. More examples are shown in Appendix B and can be use to assess the reconstruction success qualitatively. A quantitative assessment can be made by comparing the negative log-likelihood of the training and test data for each condition (see Table 1). We observed that, in general, reconstruction accuracy was better for smaller values of β both during training and testing. This has been observed before and is a known issue encountered when increasing the value of β parameter (Hoffman & Johnson, 2016). For the Extrapolation condition, we observed that while training reconstruction was very good, generalisation failed due to the models

Table 1: Combinatorial generalisation for Experiment Set 1

	Dataset	Model	D-score	NLL (Train)	NLL (Test)
1	Recomb to Element	Higgins($\beta = 1$)	0.28	8.58	**6.20**
2	Recomb to Element	Higgins($\beta = 8$)	**0.76**	34.87	27.69
3	Recomb to Element	Burgess($\beta = 8$)	0.56	37.24	33.88
4	Recomb to Element	Burgess($\beta = 12$)	0.70	52.52	49.89
5	Recomb to Range	Higgins($\beta = 1$)	0.24	10.49	306.77
6	Recomb to Range	Higgins($\beta = 8$)	**0.46**	36.56	337.08
7	Recomb to Range	Burgess($\beta = 8$)	0.39	31.50	342.73
8	Recomb to Range	Burgess($\beta = 12$)	0.24	47.97	**287.41**
9	Extrapolation	Higgins($\beta = 1$)	0.29	28.99	3576.80
10	Extrapolation	Higgins($\beta = 8$)	0.63	68.66	4045.95
11	Extrapolation	Burgess($\beta = 8$)	0.51	45.40	**2810.54**
12	Extrapolation	Burgess($\beta = 12$)	**0.80**	117.60	3127.93

[a] D-score refers to the overall disentanglement score measured using the metric developed by Eastwood & Williams (2018). Larger score indicates more disentangled representations.
[b] Reconstruction success is measured in terms of the negative log-likelihood (NLL). Smaller score indicates better reconstruction.

not being able to translate the reconstructions properly. All of them seem to be unable to produce a shape beyond the translation values they saw during training. The shapes also appear distorted, and do not resemble the clear squares reconstructed in the training data.

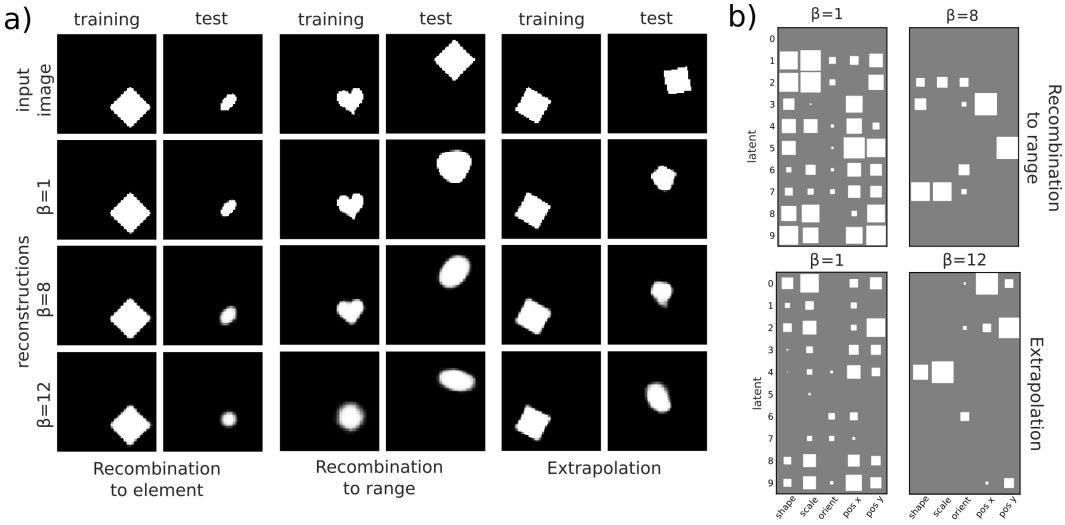

Figure 2: **Generalisation results in the unsupervised image reconstruction task** (a) Top row shows examples of input images and the three rows below show reconstructions by three different models. The columns show different conditions and we show a pair of training and test image under each condition. Left) Example from the Recombination to Element condition where the models did not see ellipses at a particular scale and rotation in the bottom right corner, yet they are able to reconstruct it. However, entangled models seem to be as good as disentangled ones. Middle) Example from the Recombination to Range condition where squares where not seen in the right side of the screen. Right) Example from the Extrapolation condition where models were never shown any shape on the right-hand side. All models fail to place the shape on the correct location and even failed to reproduce the shape and the scale. (b) Visualisation of disentanglement achieved by models at the end of training (following method used by Eastwood & Williams (2018)). In each panel, rows show latent variables and columns show the generative factors. The size of the square represents the relative importance of the latent variable for predicting the generative factor. For the model trained with $\beta = 12$ in the extrapolation condition, each generative factor is primarily represented by one latent variable indicating a high degree of disentanglement. The visualisation of the entire set of models is shown in Appendix B

The Recombination to Range condition shows another interesting qualitative difference between the entangled and disentangled models. All models failed to generalise again but in different ways. Entangled models tended to put a blob on the correct location, which allowed them to minimise loss in pixel space over a large set of test examples. In contrast, the models with higher level of disentanglement fell back to the most similar shape (in pixel space) that they had seen at that location.

Finally, the Recombination to Element condition was solved by all the models, regardless of disentanglement score. In fact, the entangled models tended to achieve better reconstructions as evidenced by the disentangled models with β=12 which had a hard time reconstructing ellipses at small scales and tended to just produce a circle instead.

The second panel in Figure 2 shows the coefficients computed by the disentanglement metric for two of the conditions. The size of each square denotes the importance of a latent (row) in predicting the corresponding generative factor (column). The higher the disentanglement, the sparser the matrices. It is clear that none of the models achieve perfect disentanglement, in which case there would be only as many squares as generative factors. Nevertheless, an examination of these matrices revealed that different models achieved a large range of disentanglement.

4 IMAGE COMPOSITION EXPERIMENTS

The limited combinatorial generalisation in the experiments above could be because of the limitations of the task rather than the models or their internal representations. Even though the models learn disentangled representations to some extent, it could be that the simple reconstruction task does not provide enough impetus for the model to learn how to combine these disentangled representations in a combinatorial manner. Therefore, in the second set of experiments we designed a variation of the standard unsupervised task that requires combining generative factors in order to solve the task.

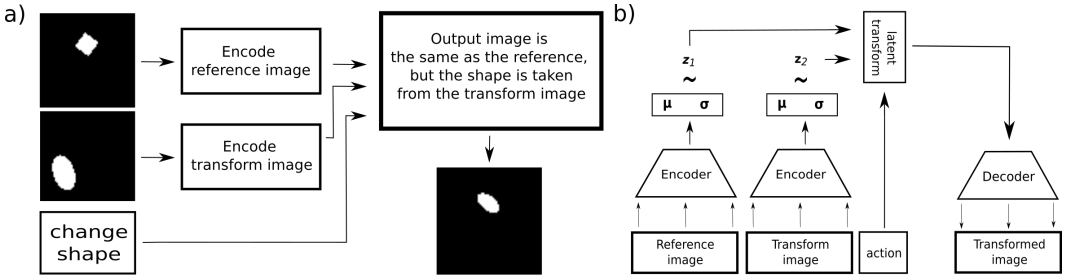

Figure 3: **Image composition task.** (a) An example of the composition task. In this case, the shape of the output must match the transform and the rest of the values must match the reference. (b) The general architecture used, based on the standard VAE. The model uses the same encoder on both images. Then a transform takes latent representation samples and combines them to produce a transformed representation. This is used to produce the transformed image.

This new task is illustrated in Figure 3(a). The input consists of two images and an action. The goal of the task is to take the first (*reference*) image and modify it so that it matches the second (*transform*) image along the dimension specified by the action. This action is coded using a one-hot vector. This design is based on the question answering task in Santoro et al. (2017) and the compositional one in Higgins et al. (2018c). We produced training and test sets for each condition by sampling reference-transform pairs along with an action uniformly from the generative factors. We ensured that this sampling respected the experiment restriction, so that the transformed image is not outside the current set.

The standard VAE is inadequate to solve this task. Therefore, we constructed a model with the architecture shown in Figure 3(b). This model first applies an encoder to both images, obtaining low-dimensional latent representations of each image. It then combines these latent representations with the action to obtain a transformed internal representation. There are several ways in which the

Table 2: Performance of the model in the second set of experiments

	Experiment	D-score	NLL (train)	NLL (test)
1	Extrapolation	0.82	31.73	19138.82
2	Recomb to range	0.71	50.10	346.10
3	Recomb to element	0.96	36.57	13.74

[a] D-score refers to the overall disentanglement score measured using the metric developed by Eastwood & Williams (2018). Larger score indicates more disentangled representations.
[b] Reconstruction success is measured in terms of the negative log-likelihood (NLL). Smaller score indicates better reconstruction.

input representations of the two images could be combined with the action. We tried three different methods: (i) using a standard MLP, (ii) element-wise interpolation between the two representations, with the interpolation coefficients determined by the action, and (iii) concatenating input representations with the actions and linearly combining the resultant vectors. We obtained qualitatively similar results with all three methods, but found that the method (iii) gave the best results, providing greatest accuracy of reconstruction as well as highest levels of disentanglement. Therefore, in the rest of the manuscript we describe the results obtained using this method. Once this transformed internal representation has been generated, it is decoded to obtain an output image. We use the same encoding and decoding modules as the ones used by Burgess et al. (2018).

The results for this model are shown in Table 2. The model managed to solve the task. In doing so, it also comes to rely on representations with a high level of disentanglement (see Figure 4(b)) even with low values for the β parameter. However, as was the case in the previous experiment, models failed to solve the more challenging generalisation tasks.

In Figure 3 we show some examples of the model's behaviour and its internal representations. The model failed in similar ways as the disentangled models in the previous experiment, confusing shapes when presented with unseen combinations. Even the Recombination to element case showed some failures (like in the example shown in Figure 3(a)) though the models were, in general, successful in this condition, as can be inferred by comparing the negative log-likelihoods for the training and test trials for this condition in Table 2.

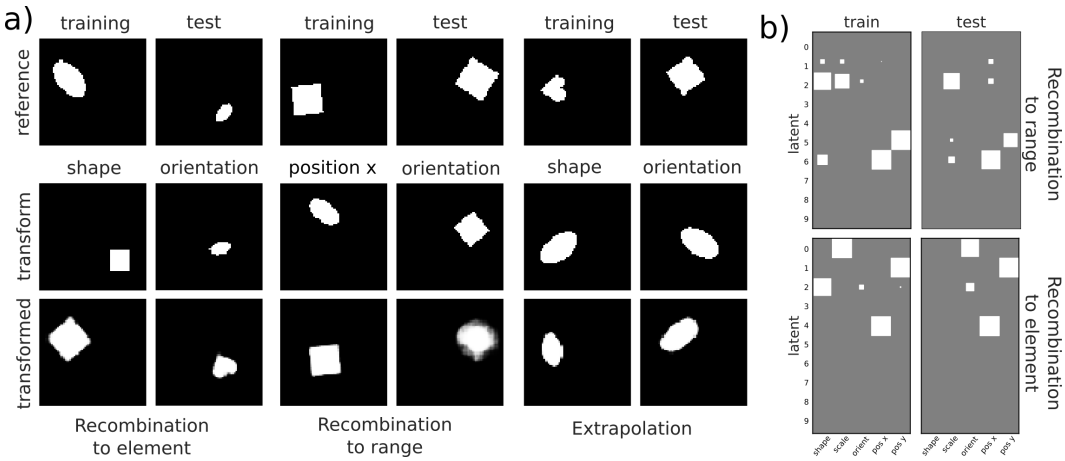

Figure 4: **Image composition generalisation results** (a) Each column shows an example trial which consists of a *reference* image, an *action*, a *transform* image and the *transformed* (output) image. We show examples of both training and test trials. Each of the training trials results in the correct (expected) transformed image, while each of the test trials shows a failure. (b) Visualisation of the degree of disentanglement for four conditions. The sparse representation reflects the high level of disentanglement achieved by these models in this task. Shape is not being predicted in the test case since there is only one missing for both conditions.

5 DISCUSSION

Learning disentangled representations has garnered significant interest in recent years. The main drive behind this interest is the lack of interpretability of modern unsupervised models and the desire for more flexible representations. Indeed, interpretability (Higgins et al., 2017) and better sample efficiency have been found when using these representations for downstream tasks in Reinforcement Learning (Higgins et al., 2018b) and abstract reasoning (van Steenkiste et al.).

If disentangled representations additionally possess the property of compositionality, it should allow these models to achieve combinatorial generalisation (Fodor & Pylyshyn, 1988). However, we found no evidence for this. Indeed highly disentangled and entangled representations were equally successful at interpolation and recombination-to-element generalisation, and both failed on recombination-to-range and extrapolation. This was the case even when we trained a VAE on an explicitly combinatorial task that led to the highest level of disentanglement. Thus, current approaches to learning disentangled representation are not sufficient to achieve true compositionality.

Our findings might seem to contradict previous reports showing success in combinatorial generalisation tasks. In Eslami et al. (2018), some success was reported when rendering novel 3D shapes with colours that had been previously seen on other shapes. And in Higgins et al. (2018c) it was reported that using a disentangled representation allowed the model to recombine observed shapes and colours in novel ways. These are indeed promising results but it has not been systematically tested to what extent these results support full combinatorial generalisation. We argue that both these examples of combinatorial generalisation correspond to the recombination-to-element case, and that the harder conditions remain out of the reach of current approaches. For example, consider again the SCAN model (Higgins et al., 2018c) which could render a room with {white suitcase, blue walls and magenta floor}, even though it was never shown this combination during training (see Figure 4 in Higgins et al 2018). But, unlike our training set, only two dimensions of variations were tested: colour and shape. The model had seen all the colours as well as shapes, and various combinations of colour and shape. Thus the novel concept, {white suitcase, blue walls and magenta floor} was recombination to a new element, rather than a whole range. Similarly, Chaabouni et al. (2020) showed that disentangled representations in a language model led to combinatorial generalisation as long as the "specific attribute values tend to occur with a large range of values from other attributes".

This raises the question as to why disentangled representations are not more effective in supporting generalisation. One possibility is that disentangled representations may indeed be a necessary first step. However they may not be sufficient to use the principle of compositionality to generate novel combinations. To do this, the model must additionally provide a mechanism of binding these representations in a way that maintains their independence. This point has previously been made in the context of connectionist representations by Hummel in Hummel (2000). In this view, the limited generalisation of VAEs may be due to the generic nature of the decoder, which codes various conjunctions of disentangled dimensions in a way that violates their independence.

An important goal for future research is to develop networks that support the more difficult forms of combinatorial generalisation and extrapolation. In fact there is already an active range of research in this direction that include networks with specialized modules (Santoro et al., 2017), or mechanisms (Mitchell & Bowers, 2020; Hummel & Biederman, 1992), or structured representations (Higgins et al., 2018c; Watters et al., 2019), or learning objectives (Vankov & Bowers, 2020) that may show greater success. But to date these models have not been systematically tested in the most challenging settings. We have highlighted the different types of generalisation that networks must support, and our systematic exploration of this space has shown that disentangled representations learned in VAEs are not enough. It will be interesting to see how these other approaches fare in the more difficult generalisation settings, and the role of disentanglement in any solutions.

REFERENCES

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation Learning: A Review and New Perspectives. *arXiv:1206.5538 [cs]*, April 2014. URL http://arxiv.org/abs/1206.5538. arXiv: 1206.5538.

Jeffrey S. Bowers, Ivan I. Vankov, Markus F. Damian, and Colin J. Davis. Why do some neurons in cortex respond to information in a selective manner? Insights from artificial neural networks. *Cognition*, 148:47–63, March 2016. ISSN 0010-0277. doi: 10.1016/j.cognition.2015.12.009. URL http://www.sciencedirect.com/science/article/pii/S0010027715301232.

Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE. *arXiv:1804.03599 [cs, stat]*, April 2018. URL http://arxiv.org/abs/1804.03599. arXiv: 1804.03599.

Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. Compositionality and generalization in emergent languages. *arXiv preprint arXiv:2004.09124*, 2020.

Sunny Duan, Loic Matthey, Andre Saraiva, Nicholas Watters, Christopher P. Burgess, Alexander Lerchner, and Irina Higgins. Unsupervised Model Selection for Variational Disentangled Representation Learning. *arXiv:1905.12614 [cs, stat]*, February 2020. URL http://arxiv.org/abs/1905.12614. arXiv: 1905.12614.

Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, pp. 15, 2018.

S. M. Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S. Morcos, Marta Garnelo, Avraham Ruderman, Andrei A. Rusu, Ivo Danihelka, Karol Gregor, David P. Reichert, Lars Buesing, Theophane Weber, Oriol Vinyals, Dan Rosenbaum, Neil Rabinowitz, Helen King, Chloe Hillier, Matt Botvinick, Daan Wierstra, Koray Kavukcuoglu, and Demis Hassabis. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, June 2018. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aar6170. URL https://science.sciencemag.org/content/360/6394/1204. Publisher: American Association for the Advancement of Science Section: Research Article.

Babak Esmaeili, Hao Wu, Sarthak Jain, Alican Bozkurt, N. Siddharth, Brooks Paige, Dana H. Brooks, Jennifer Dy, and Jan-Willem Meent. Structured Disentangled Representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2525–2534. PMLR, April 2019. URL http://proceedings.mlr.press/v89/esmaeili19a.html. ISSN: 2640-3498.

Jerry Fodor and Brian P. McLaughlin. Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition*, 35(2):183–204, May 1990. ISSN 0010-0277. doi: 10.1016/0010-0277(90)90014-B. URL http://www.sciencedirect.com/science/article/pii/001002779090014B.

Jerry A. Fodor and Zenon W. Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1):3–71, March 1988. ISSN 0010-0277. doi: 10.1016/0010-0277(88)90031-5. URL http://www.sciencedirect.com/science/article/pii/0010027788900315.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. pp. 13, 2017.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a Definition of Disentangled Representations. *arXiv:1812.02230 [cs, stat]*, December 2018a. URL http://arxiv.org/abs/1812.02230. arXiv: 1812.02230.

Irina Higgins, Arka Pal, Andrei A. Rusu, Loic Matthey, Christopher P. Burgess, Alexander Pritzel, Matthew Botvinick, Charles Blundell, and Alexander Lerchner. DARLA: Improving Zero-Shot Transfer in Reinforcement Learning. *arXiv:1707.08475 [cs, stat]*, June 2018b. URL http://arxiv.org/abs/1707.08475. arXiv: 1707.08475.

Irina Higgins, Nicolas Sonnerat, Loic Matthey, Arka Pal, Christopher P. Burgess, Matko Bosnjak, Murray Shanahan, Matthew Botvinick, Demis Hassabis, and Alexander Lerchner. SCAN: Learning Hierarchical Compositional Visual Concepts. *arXiv:1707.03389 [cs, stat]*, June 2018c. URL http://arxiv.org/abs/1707.03389. arXiv: 1707.03389.

Matthew D Hoffman and Matthew J Johnson. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NIPS*, volume 1, pp. 2, 2016.

J. E. Hummel. Localism as a first step toward symbolic representation. *Behavioral and Brain Sciences*, 23(4):480–481, December 2000. ISSN 0140-525X. doi: 10.1017/S0140525X0036335X.

John E Hummel and Irving Biederman. Dynamic binding in a neural network for shape recognition. *Psychological review*, 99(3):480, 1992.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, January 2017. URL http://arxiv.org/abs/1412.6980. arXiv: 1412.6980.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv:1312.6114 [cs, stat]*, December 2013. URL http://arxiv.org/abs/1312.6114. arXiv: 1312.6114.

Klaus Greff, Aaron Klein, Martin Chovanec, Frank Hutter, and Jürgen Schmidhuber. The Sacred Infrastructure for Computational Research. In Katy Huff, David Lippa, Dillon Niederhut, and M Pacer (eds.), *Proceedings of the 16th Python in Science Conference*, pp. 49 – 56, 2017. doi: 10.25080/shinma-7f4c6e7-008.

Brenden Lake and Marco Baroni. Generalization without Systematicity: On the Compositional Skills of Sequence-to-Sequence Recurrent Networks. In *International Conference on Machine Learning*, pp. 2873–2882. PMLR, July 2018. URL http://proceedings.mlr.press/v80/lake18a.html. ISSN: 2640-3498.

Gary Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.

Emile Mathieu, Tom Rainforth, N. Siddharth, and Yee Whye Teh. Disentangling Disentanglement in Variational Autoencoders. *arXiv:1812.02833 [cs, stat]*, June 2019. URL http://arxiv.org/abs/1812.02833. arXiv: 1812.02833.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. *dSprites: Disentanglement testing Sprites dataset*. 2017. URL https://github.com/deepmind/dsprites-dataset/.

James L McClelland, David E Rumelhart, PDP Research Group, and others. Parallel distributed processing. *Explorations in the Microstructure of Cognition*, 2:216–271, 1986. Publisher: MIT Press Cambridge, Ma.

Jeff Mitchell and Jeffrey S. Bowers. Harnessing the Symmetry of Convolutions for Systematic Generalisation. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, Glasgow, United Kingdom, July 2020. IEEE. ISBN 978-1-72816-926-2. doi: 10.1109/IJCNN48605.2020.9207183. URL https://ieeexplore.ieee.org/document/9207183/.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. *arXiv:1401.4082 [cs, stat]*, January 2014. URL http://arxiv.org/abs/1401.4082. arXiv: 1401.4082.

Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 4967–4976. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7082-a-simple-neural-network-module-for-relational-reasoning.pdf.

Paul Smolensky. The constituent structure of connectionist mental states: A reply to Fodor and Pylyshyn. *Southern Journal of Philosophy*, 26(Supplement):137–163, 1987. Publisher: Citeseer.

Paul Smolensky. *Connectionism, constituency, and the language of thought*. University of Colorado at Boulder, 1988.

S. Desroziers J. Kriss V. Fomin, J. Anmol and A. Tejani. High-level library to help with training neural networks in pytorch. https://github.com/pytorch/ignite, 2020.

Sjoerd van Steenkiste, Jürgen Schmidhuber, Francesco Locatello, and Olivier Bachem. Are Disentangled Representations Helpful for Abstract Visual Reasoning? pp. 14.

Ivan I. Vankov and Jeffrey S. Bowers. Training neural networks to encode symbols enables combinatorial generalization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375(1791):20190309, February 2020. doi: 10.1098/rstb.2019.0309. URL https://royalsocietypublishing.org/doi/10.1098/rstb.2019.0309. Publisher: Royal Society.

Nicholas Watters, Loic Matthey, Christopher P. Burgess, and Alexander Lerchner. Spatial Broadcast Decoder: A Simple Architecture for Learning Disentangled Representations in VAEs. *arXiv:1901.07017 [cs, stat]*, August 2019. URL http://arxiv.org/abs/1901.07017. arXiv: 1901.07017.

Shengjia Zhao, Hongyu Ren, Arianna Yuan, Jiaming Song, Noah Goodman, and Stefano Ermon. Bias and Generalization in Deep Generative Models: An Empirical Study. *arXiv:1811.03259 [cs, stat]*, November 2018. URL http://arxiv.org/abs/1811.03259. arXiv: 1811.03259.