Correcting Flaws in Common Disentanglement Metrics

Anonymous authors
Paper under double-blind review

Abstract

Disentangled representations are those in which distinct features, such as size or shape, are represented by distinct neurons. Quantifying the extent to which a given representation is disentangled is not straightforward; multiple metrics have been proposed. In this paper, we identify two failings of existing metrics, which mean they can assign a high score to a model which is still entangled, and we propose two new metrics, which redress these problems. First, we use hypothetical toy examples to demonstrate the failure modes we identify for existing metrics. Then, we show that similar situations occur in practice. Finally, we validate our metrics on the downstream task of compositional generalization. We measure the performance of six existing disentanglement models on this downstream compositional generalization task, and show that performance is (a) generally quite poor, (b) correlated, to varying degrees, with most disentanglement metrics, and (c) most strongly correlated with our newly proposed metrics. Anonymous code to reproduce our results is available at https://github.com/anon296/anon.

1 Introduction

Early proponents of neural networks argued that a significant advantage was their ability to form distributed representations, where each input is represented by multiple neurons, and each neuron is involved in the representation of multiple different inputs (Hinton et al., 1986). Compared to using a separate neuron for each input, distributed representations are exponentially more compact (Bengio, 2009). An extension of distributed representations is the idea of disentangled representations, where each neuron represents a single human-interpretable feature of the input data, such as colour, size or shape, that is not represented by other neurons. These features are often referred to as "generative factors" or "factors of variation". Intuitively, a disentangled vector representation is one in which a certain subset of neurons represents (for example) shape and shape only, another distinct subset represents size and size only, etc., and changing the size of the input but not the shape, will mean that the size neurons change their activation value, but the shape neurons remain unchanged. In the strongest case, each factor is represented by a single neuron so that, e.g., changing the colour of the object in the image would cause a single neuron to change its value while all other neurons remain unchanged. (We discuss further below the ambiguity as to whether this stronger condition is required.) Disentanglement (DE) was originally formulated by Bengio (2009) (see also Bengio et al. (2013); Bengio (2013)). More recently, following Higgins et al. (2016), there have been many unsupervised DE methods proposed based on autoencoders.

In this paper, we examine the commonly used metrics to assess disentanglement. Firstly, we show how they fail to pick up certain forms of entanglement, and that representations can score highly on such metrics while being entangled. Specifically, we expose two problems with existing metrics: that they incorrectly align ground-truth factors to neurons, as they do not require distinct variables to be assigned to distinct factors; and that they measure the strength of the relationship of features to individual neurons, which is fundamentally different from the relationship to sets of neurons, and hence can give undesirable results. We show that these problems occur in practice by examining the results of six different DE models trained on three different datasets.

To address these problems, we present two new DE metrics, based on the ability of a classifier to predict the generative factors from the encoded representation. If a representation is truly disentangled, then all

the relevant information should be contained in a single neuron (or possibly a few neurons, see discussion in Section 2), and so a classifier using only this/these neuron(s) should be just as accurate as one using all neurons, and one using all other neurons should be very inaccurate. Our first metric is the accuracy of the single-neuron classifier. Our second metric is the difference between the accuracy of a classifier using all neurons, vs one using all neurons but the single selected neuron.

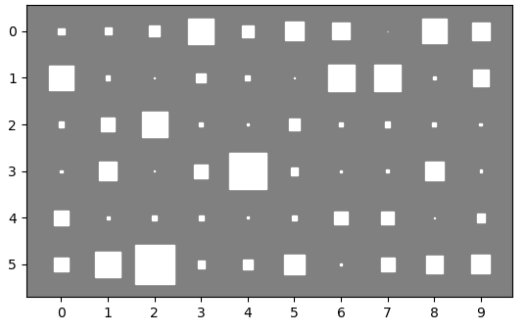
We also establish the superiority of our proposed metrics using a downstream compositional generalization task of identifying novel combinations of familiar features. Humans could recognize a purple giraffe, even if we have never seen one or even heard the phrase "purple giraffe" before, because we have disentangled the concepts of colour and shape, so could recognize each separately. The ability to form and understand novel combinations is a deep, important aspect of human cognition and is a direct consequence of humans being able to disentangle the relevant features of the objects we encounter. Our downstream task tests, for example, whether a network trained to identify blue squares, blue circles and yellow squares can, at test time, correctly identify yellow circles. If it had learned to disentangle colour from shape, then it could simply identify "yellow" and "circle" separately, each of which is familiar. We show that existing DE models generally perform poorly at this task, suggesting they are further from DE than previous analyses have implied. We also show that a high score on DE metrics is predictive of performance on this task, and that our proposed DE metrics are the most predictive in this respect. Our contributions are briefly summarized as follows.

- We identify and describe two shortcomings of existing DE metrics: (1) incorrect alignment of neurons to factors and (2) focusing on the importance of individual neurons instead of sets of neurons.
- We show, experimentally, that these problems with existing metrics occur in practice.
- We propose two alternative metrics, (1) single-neuron classification and (2) neuron knockout, that do not suffer from the problems that existing metrics suffer from.
- We validate our metrics on a downstream task of compositional generalization. We show empirically that, while existing models generally perform badly at recognizing novel combinations of familiar features (compositional generalization), their performance correlates with DE metrics, and correlates most strongly with our proposed metrics.

The rest of this paper is organized as follows. Section 2 discusses related work. Section 3 describes the shortcomings of existing DE metrics. Section 4 proposes our new metrics. Section 5 presents empirical results, including those on the downstream compositional generalization task. Section 6 then discusses some limitations and future work, and summarizes our contributions.

2 Related Work

Disentanglement Models. After the initial proposal by Bengio (2009), disentangled representations received new interest when Higgins et al. (2016) proposed β-VAE, as an adaption of the variational autoencoder (VAE) (Kingma and Welling, 2013). By taking the prior to have a diagonal covariance matrix, and increasing the Kullback-Leibler divergence (KL) loss weight, β-VAE encourages the model representations to have diagonal covariance too, which the authors claim enforces DE. Kumar et al. (2017) further encouraged a diagonal covariance matrix by minimizing the Euclidean distance of the model's covariance matrix from the identity matrix. Burgess et al. (2018) proposed to gradually increase the reconstruction capacity of the autoencoder by annealing the KL in the VAE loss. Chen et al. (2018) propose β-TCVAE, which minimizes the total correlation of the latent variables, approximated using Monte-Carlo-based on importance sampling. Kim and Mnih (2018) propose FactorVAE, which also minimizes total correlation, this time approximating using a discriminator network. Locatello et al. (2019) challenged earlier DE methods, proving that it is always possible for a model to learn an entangled representation that appears disentangled only on the available data, and presenting experiments that call into question whether disentangled representations lead to superior downstream performance, as previous works had claimed. Later, Locatello et al. (2020) claimed that including a small amount of supervision was sufficient to learn disentangled representations. More recently, other authors have pushed back against the claim that unsupervised DE is impossible, arguing that the priors



	blue	yellow
Square	0	0 or 1 with equal probability
Circle	0 or 1 with equal probability	1

Figure 1: Left: Hinton diagram for β-TCVAE on 3dshapes. The size of the square at (i, j) shows the mutual information between factor i and neuron j. Existing metrics incorrectly align both factor 2 and factor 5 to neuron 2, whereas we correctly align factor 2 to neuron 1 instead. Right: Toy example of partially encoding two factors.

embodied in autoencoder architectures still allow unsupervised DE in practice (Roth et al., 2022; Rolinek et al., 2019). The models we test on include a range of both unsupervised and weakly supervised. Recent unsupervised models include (Klindt et al., 2020), a VAE-based model to learn disentangled representations from videos of natural scenes. Semi-supervised methods include WeakDE (Valenti and Bacciu, 2022), which adversarially pushes the latent distribution for each factor close to a prior computed using a small labelled subset. Some DE models are fully supervised, such as MTD (Sha and Lukasiewicz, 2021), which partitions the neurons into a subset for each factor, and defines several losses using the ground-truth labels. Some can operate either supervised or unsupervised, such as Parted-VAE (Hajimiri et al., 2021), which minimizes the Bhattacharyya distance (Bhattacharyya, 1946) from a multivariate normal.

Disentanglement Metrics. Higgins et al. (2016) made an early attempt to quantify disentanglement, by fixing one generative factor and varying others, then predicting which factor was fixed from the mean absolute difference of each neuron. Kim and Mnih (2018) replace mean absolute difference with the index of the lowest variance neuron. Ridgeway and Mozer (2018) decompose DE into modularity and explicitness, each with its own metric. They measure modularity as the deviation from an "ideally modular" representation, where every latent neuron has nonzero mutual information with exactly one generative factor. Explicitness is measured similarly to the metric of Kim and Mnih (2018), except using the mean of one-vs-rest classification and AUC-ROC. The SAP score (Kumar et al., 2017) calculates, for each factor, the R^2 coefficient with each latent dimension, and then takes the difference between the largest and second largest. A representation will score highly on SAP if, for each generative factor, one neuron is very informative and no other individual neuron is. A similar idea is employed by the mutual information gap (MIG) metric (Chen et al., 2018), except using mutual information instead of R^2. Eastwood and Williams (2018) decompose the task into disentanglement, completeness and informativeness (DCI). Then, they train a classifier (commonly a linear model or a gradient-boosted tree) to predict each generative factor from the latent factors, and estimate DCI from a measure of importance of each feature to each factor from the classifier. IRS (Suter et al., 2019) measures the maximum amount that a given neuron can be changed by changing a generative factor other than the one it corresponds to. Another recent metric is MED (Cao et al., 2022), which is the same as DCI, using mutual information as feature importance.

Definitions of Disentanglement. There is ambiguity in the literature as to whether DE requires that each factor is represented by a single neuron an no other neurons (strong DE), or a dedicated group of potentially several neurons, and no other neurons (weak DE). The original definition by Bengio (2009) (echoed by Higgins et al. (2016), Kim and Mnih (2018) and others) stipulates that "each neuron represents a single factor", which seems to allow that each factor is represented by multiple neurons. Other treatments of DE imply a stronger notion of the concept, namely, an injective function from factors to the unique neurons that represent them. This is implicit in the common metrics that map each factor to a single neuron, in the technique of latent traversals that aim to vary a single factor through its range of values by varying a single neuron, and in the descriptions of DE in other works, e.g., "learning one exclusive factor per dimension" (Pineau and Lelarge,

2018). Attempts at formal definitions, e.g. equivariance in group theory (Higgins et al., 2018), allow the possibility of many-to-one mappings in theory, but present examples and discuss benefits of DE with respect to one-to-one mappings. Weak DE is less compact than strong DE, and loses some interpretability benefits: it would be hard to tell which neurons represent a given factor if we have to check all *subsets* of neurons, of which there are exponentially many. Our proposed metrics focus on the strong notion, for the following reasons: (1) strong DE is generally assumed by many common DE metrics, (2) strong DE is assumed by the common practice of latent traversals, (3) weak DE loses interpretability as compared with strong DE. This does not mean that weak DE is without value, indeed some authors have argued that, for certain factors, it is more suitable represent using multiple neurons (Esmaeili et al., 2022). However, it is not possible for a single metric to measure both because they are different objectives. For example, if each factor is perfectly represented by a distinct set of 5 neurons, which don't overlap, then the representation shows perfect weak DE and should get a max score by a metric measuring weak DE. But it could be that none of these 5 neurons individually represent the corresponding factor (see Section 3.2), in which case the representation shows poor strong DE and should get a low score by a metric measuring strong DE. Thus, strong DE is valuable to measure because it has some advantages that weak DE does not (compactness, interpretability), and, as we have just argued, it requires a metric specifically dedicated to strong DE; there cannot be a metric that adequately measures both strong and weak DE at the same time. Therefore, there is value to a metric for strong DE, and that is what we provide in this paper.

3 Problems with Existing Disentanglement Metrics

3.1 Incorrect Alignment of Latent Variables

The majority of existing metrics are based on aligning the set of factors G with the set of neurons Z; that is, for each factor, finding the neuron that it is represented by. Each neuron is only supposed to represent a single factor, however, existing metrics simply relate each factor to the maximally informative variable, e.g., as measured by mutual information (Chen et al., 2018) or weight from a linear classifier (Kumar et al., 2017). This does not enforce the constraint of having distinct neurons for distinct features, it means that the same neuron could be selected as representing multiple different factors. For example, consider again the model trained on a dataset of blue squares, blue circles, yellow squares and yellow circles. Suppose such a model, call it M_1, has two latent variables, z_1 and z_2, the first is as shown on the right of Figure 1, and the second is random noise, unrelated to the inputs. Then z_1 encodes both colour and shape, each to an accuracy of 75%, whereas z_2 encodes both only to 50% (random guess). Thus, z_1 will be chosen as the representative of both colour and shape.

This situation is not merely theoretical, it often occurs in practice. One example is the left of Figure 1, which shows a Hinton diagram for β-TCVAE trained on 3dshapes (further examples of other datasets and models are given in the appendix). There, the existing approach of aligning each factor to the most informative neuron (e.g., by mutual information), incorrectly concludes that two different factors, size (factor 2) and colour (factor 5), should be aligned to the same neuron (neuron 2). In a Hinton diagram, the size of the square is proportional to the MI, and the square at (5,2) (using matrix indexing) the biggest in its row, so existing metrics align factor 5 to neuron 2. The square at (2,2) is also the biggest in its row, so existing metrics also align factor 2 to neuron 2. This is not good, because each neuron should only be aligned with one factor. Our method enforces aligning distinct factors to distinct neurons and so avoids this problem. It ends up aligning factor 2 to neuron 1 and factor 5 to neuron 2.

The M1 would get a higher score than another, model call it M2, in which z_1 is as in the right of Figure 1, and z_2 represents shape to an accuracy of 70%[1]. Table 1 shows the scores produced on this toy hypothetical example by four common existing metrics: MIG, SAP, DCI and MED, as well as for our proposed metrics: SNC and NK. The existing metrics all give a higher score to M1, whereas SNC gives a higher score to M2 and NK gives an equal score to both. The calculations for these scores are given in Appendix A. What our metrics do is correct, as M2 is closer to the desired case where z_1 represents colour and colour only, and z_2 represents shape and shape only (WLOG on ordering of the neurons). In the first model, z_1 represents

[1]E.g, if the input is a square, then, the 70% of the time the neuron activation is 1 and 30% of the time it is 0, and the opposite if it is a circle

	M1	M2	change
MIG	18.9	11.9	decrease
SAP	25.0	15.0	decrease
DCI	58.3	47.1	decrease
MED	58.3	46.4	decrease
SNC	12.5	22.5	increase
NK	12.5	12.5	unchanged

Table 1: Scores of existing metrics (MIG, SAP, DCI and MED), and our proposed metrics (SNC and NK) for the toy model described in Figure 1, which we denote M1, and another in which the second neuron encodes colour to 70% accuracy, which we denote M2. Existing metrics incorrectly give M1 a higher score than M2. SNC gives M2 a higher score and NK gives both the same score. Note that IRS cannot be computed from the information in our example.

nothing and z_2 represents both, whereas in the second model z_1 represents shape and z_2 represents both. M1 needs z_2 to learn shape and z_1 to unlearn shape, M2 just needs z_1 to unlearn shape. See also the appendix in (Cao et al., 2022), which mentions a similar failing.

3.2 Conflating Representation by Individual Neurons vs by a Set of Neurons

The second problem with existing metrics is how they handle information distributed over multiple neurons. This problem manifests when measuring whether the information for one factor is restricted to just one variable, and is a shortcoming both of classifier-based and classifier-free metrics. Let $g_0, \ldots, g_n$ be the ground-truth generative factors, and $z = z_0, \ldots, z_m$, $m \geq n$ the corresponding neurons and suppose g_i has been aligned to z_i for each i. Further, let $z_{\neq i}$ be the set of neurons other than z_i. Now, consider XOR: $g_1 = z_2 \oplus z_3$. (We use discrete values for clarity, a continuous approximation could be $|z_2 - z_3|, z_2, z_3 \in [0, 1]$.) Almost all existing metrics, would fail here. Those based on linear classifiers (Kumar et al., 2017; Ridgeway and Mozer, 2018; Eastwood and Williams, 2018) would conclude that z_2 and z_3 are unrelated to g_1, because the strength of the linear relationship in XOR is zero. Those that use pairwise mutual information (Chen et al., 2018; Cao et al., 2022; Do and Tran, 2019) would also react incorrectly. They would find that $I(g_1; z_2) = I(g_1; z_3) = 0$, and so conclude that g_1 is not represented by $z_{\neq i}$, when in fact g_1 is perfectly represented by $z_{\neq i}$. In Section 5.1 we show these distributed entanglements occur in practice.

DCI can be used with a non-linear classifier, most commonly a gradient-boosted tree, but it still computes an importance score for each feature individually to each factor, so fails for a different but similar reason. Recall that DCI uses these importance scores to first compute an importance matrix R, where the (i, j)th entry is the importance of neuron i for predicting feature j. It then quantifies disentanglement as a combination of the normalized entropy across rows and columns. The failure of DCI to measure strong DE is shown by Theorem 3.1, using an extension of XOR to multiple feature values.

Theorem 3.1. *As the number of neurons and factors in the representation increases, the D and C components of DCI can, under a very broad class of feature importance measures, including gradient-boosted trees, assign a score that is arbitrarily close to perfect, even though the model is completely entangled in the sense that no neuron, by itself, contains any information about any generative factor.*

Proof. Let $z_0, \ldots z_{m-1}$ be a set of neurons encoding factors $g_0, \ldots, g_{n-1}$, with $m \geq n$. Let $k_0, \ldots, k_{n-1}$ denote the number of different values of each feature, that is, g_i can take on k_i different values, for each i. Assume WLOG that each neuron is normalized to the interval $[0, 1]$. This is equivalent to any unnormalized representation via a simple scaling of weights in the output of the encoder.

Let $b(z, k)$ denote the bin index of z after binning into k different bins:

$$b(z, k) = \lfloor (zk) \rfloor ,$$

and let $c(z, k)$ denote the scaled remaining portion of z that does not contribute to its binned value:

$$c(z, k) = kz - b(z, k) .$$

Then, suppose that the representation function is as follows:

$$g_i = b(z_i, k_i) + b(c(z_j, k_j), k_i) \mod k_i \,, \tag{1}$$

where $j = i + 1 \mod n$. Intuitively, this means that we divide the significant bits of z_i into two portions, the first chunk of significant bits is used to compute $b(z_i, k_i)$ and contributes towards g_i, the remaining chunk is used to compute $c(z_i, k_i)$ and contributes towards g_{i-1} (where $i - 1$ is taken with modulo arithmetic). For any values of $g_0, \ldots, g_{n-1}$, such a representation exists, by setting

$$\begin{aligned} b(z_i, k_i) =& g_i - b(c(z_j, k_j), k_i) \mod k_i \\ c(z_j, k_j) =& k_j z_j - b(z_j, k_j) \mod k_i \,. \end{aligned}$$

and then computing a solution for the system of $2n$ linear equations $2n$ unknowns. Together $b(z_i, k_i)$ and $c(z_i, k_i)$, along with k_i, one can use this solution to set values for z that represent the feature values in the way just described. This shows that such a representation always exists.

Clearly, $\forall l \neq i, j, I(z_l; g_i) = 0$, where I denotes mutual information, because (1) does not involve z_l at all. We now show that, when g_i is uniformly distributed, $I(z_i; g_i) = I(z_j; g_i) = 0$ as well, establishing that the representation is completely entangled in the sense that $I(z; g) = 0$ for all neurons z and factors g.

Remark 3.2. There is currently interest in applying disentanglement to data where the values of each feature are non-uniformly distributed, and we expect the following proof could be extended to cover such cases. However, the present form is enough to show the flaw in DCI. As well as the non-uniform case, a good metric should also of course be able to give the correct answer in the uniform case. In other words, giving the correct answer in the uniform case is a necessary but not sufficient condition for a good metric.

If g_i is uniformly distributed on $\{0, \ldots, k_i - 1\}$, then one solution to the system of equations is where $b(z_i, k_i)$ and $b(c(z_j, k_j), k_i)$ are also uniformly distributed. This follows from the fact that, on any countable discrete group G, the convolution of two uniform probability measures is again uniform, and the fact that addition of two random variables, as in (1), corresponds to a convolution. In this case, the group is the integers modulo k_i. Thus, setting $p(b(z_i, k_i)) = b(c(z_j, k_j), k_i) = \mathcal{U}(k_i)$, gives, $p(g_i) = \mathcal{U}(k_i)$, where $\mathcal{U}(k)$ is the discrete uniform distribution on k elements. This gives the following:

$$\begin{aligned} p(g_i | z_i = z) =& p(b(z_i, k_i) = g_i - b(k_j c(z, k_j), k_i)) = \mathcal{U}(k) \, \forall z \in [0, 1] \\ p(g_i | z_j = z) =& p(b(z_i, k_i) = g_i - b(k_j c(z, k_j), k_i)) = \mathcal{U}(k) \, \forall z \in [0, 1] \,, \end{aligned}$$

where the second equality in each line follows from the fact that $b(k_j c(z, k_j), k_i)$ is a constant that doesn't depend on g_i, and that subtraction of a constant constitutes a bijection in the group of integers modulo k_i. Therefore, the conditional distribution of g_i given z_i or z_j is equal to the marginal distribution of g_i. In particular, the entropies are equal, giving zero mutual information:

$$\begin{aligned} I(g_i; z_i) =& H(g_i) - H(g_i | z_i) = \mathcal{U}(k_i) - \mathcal{U}(k_i) = 0 \\ I(g_i; z_j) =& H(g_i) - H(g_i | z_j) = \mathcal{U}(k_i) - \mathcal{U}(k_i) = 0 \,. \end{aligned}$$

This completes the first part of the proof. Now we must show that DCI assigns a high score to such a representation. In principle, DCI can use any measure of feature importance, however the following argument makes a very general assumption about the feature importance measure, which includes all but pathological feature importance measures, and so covers all cases of practical interest. Let R be the $m \times n$ matrix of feature importance, where R_{ji} is the importance of neuron j in predicting feature i. The assumption is that, if g_i is uniquely defined by an equation that does not involve z_j, and z_j gives no indication as to the value of g_i, then $R_{ji} = 0$. We can formalize "gives no indication as to the value of" as $I(z_j, g_i) = 0$ and, for any non-empty set of neurons $\bar{z}$, $I(g_i; z_j, \bar{z}) = I(g_i; \bar{z})$. This means

$$R = \begin{pmatrix} R_{1,1} & 0 & 0 & \ldots & 0 & 0 & R_{1,n-1} \\ R_{2,1} & R_{2,2} & 0 & \ldots & 0 & 0 & 0 \\ 0 & R_{3,2} & R_{3,3} & \ldots & 0 & 0 & 0 \\ 0 & 0 & R_{4,3} & \ldots & R_{n-2,n-3} & R_{n-2,n-2} & 0 \\ 0 & 0 & 0 & \ldots & 0 & R_{n-1,n-2} & R_{n-1,n-1} \end{pmatrix} .$$

DCI normalizes feature importance across all features, $P_{j,i} = R_{j,i} / \sum_{k=0}^{n-1} R_{j,k}$. Assuming the feature importances are not all zero for any feature, in which case DCI is undefined because of zero-division, both D and C are equal to one minus the entropy of some distribution over n elements, with either exactly one or exactly two elements given non-zero probability. The entropy of such a distribution is maximized when both non-zero elements are equal to $\frac{1}{2}$, giving entropy

$$-\sum_{i=0}^{n} p(x) \log p(x) = 2 \log_n 2 = \frac{2}{\log n} \,.$$

Thus, $D, C \geq 1 - 2/\log n$ and so

$$\lim_{n \to \infty} D = \lim_{n \to \infty} C = 1 \,.$$

This holds for all features individually, and so by symmetry, it holds on the weighted sum across all features. □

Remark 3.3. We do not need very many neurons before the lower bound on both D and C becomes significant. For only 16 neurons, $D, C \geq 0.5$, for 64 neurons, $D, C \geq 0.67$.

Intuitively, this proof shows that it is possible for n neurons to perfectly encode n factors, where each factor depends only on two neurons, but where knowing the value of any one neuron reveals nothing about the value of any factor. This situation constitutes a false positive for DCI. The representation is not disentangled at all, but DCI gives it a high score. The only assumption on the feature importance measure is that it assigns a score of 0 whenever the mutual information, even when conditioned on any number of other neurons, is zero. This is a weak assumption and should be met by any reasonable feature importance measure. While DCI to some extent approximates measuring both weak DE and strong DE, Theorem 3.1 shows that it can be an unreliable measure of strong DE, further motivating the search for reliable strong DE metrics.

Distributed entanglements are relevant not just to quantifying DE, but also to building DE models and losses, many of which optimize the informativeness of individual neurons and factors. We hope that redressing this mistake can lead to better DE models in future. Early methods, such as β-VAE, equated uncorrelated with independent, which is true only if the activation values conform to specific distributions, such as the multivariate normal. Later methods improved on this, focussing instead on information-theoretic measures of independence, but they still measure importance of each neuron individually, and so falsely equate pairwise independence with independence. Similarly, many DE models aim to optimize some version of the mutual information between individual neurons and features.

4 Proposed Metrics

4.1 Alignment Procedure

Both our metrics require aligning factor indices G to neuron indices Z. Unlike prior works, we align all factors simultaneously so that we can enforce aligning distinct factors to distinct neurons:

$$\underset{\{f:G \to Z \mid \text{ f is injective }\}}{\operatorname{argmax}} \sum_{g \in G} I(g; f(g)) \,, \tag{2}$$

where I denotes mutual information (this could be replaced with R^2 or any other measure of informativeness). This is an instance of the linear sum assignment problem, as it seeks an assignment of all features to unique neurons, where each assignment has a predifined cost. A optimal solution can be computed in $O(N^3)$, where N is the number of features, using the Kuhn-Munkres algorithm (Munkres, 1957). This gives a mapping in which no two factors are mapped to the same neuron, which better fits the notion of DE than previous approaches.

4.2 Single-neuron Classification (SNC)

Our first metric, SNC, uses z_i, after alignment, as a classifier, by dividing its values across the dataset into bins, where the bin sizes are chosen such that the number of data points in each bin is the same, equal to the

greatest common divisor (gcd) of the size of classes in the dataset. This is to allow the neuron to, potentially achieve 100% accuracy, by allow it to, potentially, match the ground truth distribution. For example, if 40% of examples are in class A, 30% in class B and 30% in class C, then we divide into bins such that each bin contains 10% of the data. If four bins are then aligned to class A, three to class B and three to class C, the predicted distribution would be the same as the ground truth distribution. In the case that gcd is very small, implying a very large number of bins, we suggest setting a minimum bin size and using that instead. Using a minimum bin size does not arise in our datasets and we have not tested it empirically.

For balanced data, this gives K bins of equal size, where K is the number of classes in g_i. However, this method does not require the classes to be balanced. We then align these bins with the K ground-truth classes, and compute the chance-adjusted accuracy. Similar to (2), the accuracy can be done with the Kuhn-Munkres algorithm. Let $X = (x)_{1 \leq i \leq N}$ denote the data, $c : X \to \{0, \ldots, K-1\}$ specify the ground truth labels, and and $b : X \to \{0, \ldots, K-1\}$ specify the bin index after alignment. Then, the metric score is the accuracy a on a single neuron, adjusted for chance accuracy r, where r is the accuracy obtained by randomly guessing labels in proportion to their true frequency:

$$SNC = \max(0, \frac{a-r}{1-r}) \,,$$

where

$$a = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(b(x_i) = c(x_i)) \text{ and} \tag{3}$$

$$r = \frac{1}{N^2} \sum_{j=1}^{K} (\sum_{i=1}^{N} \mathbb{1}(c(x_i) = j)^2) \,. \tag{4}$$

If the representation is disentangled, then this accuracy should be high, ideally as high as that of a classifier that uses all neurons. For representations that encode g_i mostly via z_i, even training an MLP, a much more powerful model than the single-neuron classifier, should not be much more accurate on a held-out test set than the simpler model. This essentially quantifies the property that latent traversals aim to show qualitatively. In the terminology of (Ridgeway and Mozer, 2018), it is a measure of explicitness, except that they fit a linear classifier on all neurons, not just z_i. That is, if, for each g_i, there is some line in representation space such that the representation vector encodes g_i as the distance of the projection along that line, then this is regarded as disentangled. Using a single-neuron classifier, on the other hand, requires that line to be an axis, i.e., it requires all elements of the vector but one to equal zero. In the other direction, if one had these vectors such that projection along the ith represented g_i, then transforming z to be in terms of these vectors as basis would give neurons that individually represent factors, as in our case. Thus, allowing linear classifiers on all neurons is a strictly weaker condition, though similar to ours. We also show linear classifier scores, which are broadly similar.

Both our metrics, like all metrics we compare to, assume access to the ground truth feature labels for all data points.

4.3 Neuron Knockout (NK)

Our second proposed metric is inspired by the technique of gene knockout in genetics, which tests how relevant a given gene is to a given function, by removing the gene and measuring the loss in function. We test whether $z_{\neq i}$ contains any information about g_i, by training an MLP to predict g_i from $z_{\neq i}$. If the representation is disentangled, then this accuracy should be low, in comparison to an MLP that uses all neurons. Our second metric is $NK_i = Acc_z - Acc_{z_{\neq i}}$, where A_x denotes the accuracy of an MLP trained on neurons x to predict g_i. This is crucially different from most existing methods, which only measure the feature importance of each neuron individually, and so suffer from the problems articulated in Section 3.2. Such methods, e.g., cannot detect when $z_{\neq i}$ encode g_i via XOR, whereas our method can. The results from Section 5 show that such encodings occur in practice. Some existing works have used a similar idea to NK, (Sha and Lukasiewicz, 2021; Mathieu et al., 2016). These works partition the neurons, a priori, into two subsets, A and B, representing

		β-VAE	β-TCVAE	FactorVAE	PartedVAE	PartedVAE-ss	weakde
Dsprites	SNC	24.8 (3.20)	**41.3 (3.00)**	15.1 (1.78)	16.1 (4.54)	19.0 (4.80)	5.8 (0.75)
	linear	46.2 (2.54)	**49.5 (1.60)**	30.8 (2.43)	36.0 (2.58)	26.3 (10.03)	28.6 (2.00)
	MLP	89.4 (3.95)	86.2 (1.10)	77.1 (4.10)	70.7 (10.33)	48.5 (9.08)	**99.8 (3.50)**
	NK	**39.9 (1.11)**	31.3 (4.10)	32.9 (1.45)	34.9 (4.23)	14.5 (8.17)	5.9 (1.17)
3dshapes	SNC	19.9 (4.70)	20.1 (2.06)	16.9 (3.36)	43.8 (24.97)	**68.3 (9.55)**	15.4 (2.11)
	linear	**79.0 (1.60)**	76.3 (3.47)	65.2 (4.74)	70.8 (6.73)	65.4 (11.40)	73.3 (18.17)
	MLP	99.8 (3.50)	**99.9 (2.50)**	98.1 (7.24)	96.2 (6.83)	88.9 (14.53)	99.8 (0.68)
	NK	8.0 (0.20)	8.1 (0.11)	16.8 (2.94)	**60.2 (6.19)**	41.8 (8.07)	2.5 (0.18)
MPI3D	SNC	32.6 (5.63)	**35.2 (4.43)**	27.5 (3.63)	35.1 (3.29)	23.9 (1.77)	17.1 (0.39)
	linear	51.3 (2.73)	**52.1 (4.20)**	46.9 (2.52)	49.4 (2.32)	43.0 (5.11)	46.2 (1.57)
	MLP	**91.3 (3.36)**	83.5 (9.00)	88.6 (2.85)	65.5 (1.51)	60.6 (3.04)	83.4 (0.89)
	NK	19.8 (0.39)	17.8 (10.82)	**23.2 (0.41)**	9.5 (1.64)	3.3 (7.11)	6.0 (0.14)

Table 2: Mean score across five runs (std. dev. in parentheses) for our proposed metrics, SNC and NK. MLP and linear denote the chance-adjusted accuracy of an MLP and linear model, respectively, on trained all neurons (MLP). For each metric, for each dataset, the highest performing model is in **bold**.

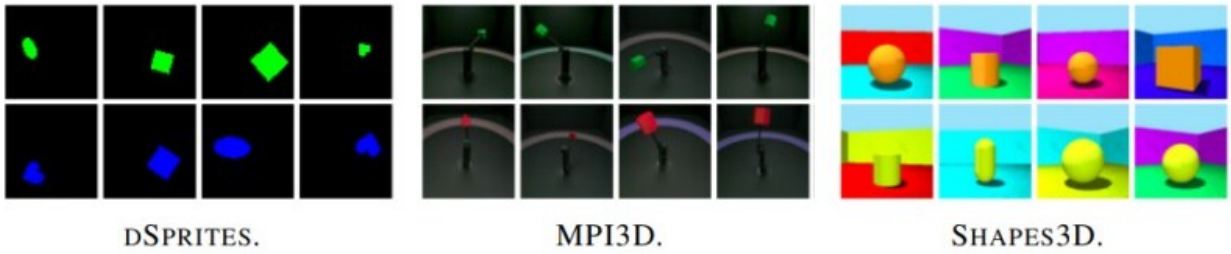


Figure 2: Examples of images from the three datasets we test on.

distinct factors a and b, respectively. Performance is then measured by training an MLP to predict b from A and a from B. This technique, however, is only applicable where the set of neurons has been partitioned during training by the use of labels, whereas ours is more broadly applicable to the case where we do not know which neurons represent which factors, because we include a method for identifying which neurons to knock out. Secondly, we measure the difference with an MLP trained on all neurons, and so can distinguish between a disentangled representation and one that simply contains no information about the input. For prior works, the latter would score highly, but for NK, it would not, as then $Acc_z \approx Acc_{z_{\neq i}} \approx 0$.

5 Experimental Evaluation

Datasets. We test our metrics and task on three datasets. **Dsprites** contains 737,280 black-and-white images with features (x, y)-coordinates, size, orientation and shape. **3dshapes** contains 480,000 images with features object/ground/wall colour, size, camera azimuth, and shape. **MPI3D** contains 103,680 images of objects at the end of a robot arm with features object colour, size and shape, camera height and azimuth, and altitude of the robot arm. All datasets have images of size 64×64. Figure 2 shows some example images from these three datasets.

Implementation Details. We test several popular DE models, β-VAE, FactorVAE and β-TCVAE, which are unsupervised, WeakDE, which is semisupervised, and PartedVAE, which can be trained unsupervised or semisupervised, and for which we test both settings. The details of these models are given in Section 2. Parted-VAE and Weak-DE are trained using the authors' public code for 100 epochs. Other models are trained using the library at https://github.com/YannDubs/disentangling-vae, using all default parameters

as specified in the code. The MLPs and linear classification heads are trained using Adam, learning rate .001, β_1=0.9, β_2=0.999, for 75 epochs. The MLP has one hidden layer of size 256. MTD is trained using the author's code (obtained privately) with all default parameters, for 10 epochs. All experiments were performed on a single Tesla V100 GPU on an internal compute cluster.

5.1 SNC and NK Results

Table 2 shows the results of our two proposed metrics, SNC and NK, on the three datasets described above. Each dataset includes a slightly different set of features, and displaying all features impairs readability, so we report the average across all features. Full results are given in the appendix.

The SNC accuracy is substantially lower than that of the full MLP. This suggests that each g_i is represented more accurately by a distributed, non-linear entangled encoding across all neurons, rather than just by z_i. Although an MLP is a more powerful model, this should not help test set accuracy unless there is relevant information in the input that it can leverage, i.e., distributed entanglements. The fact that MLP accuracy is much higher than SNC, suggests that such entanglements exist.

Distributed entanglements are also evidenced by the NK results. Here, there is often only a marginal drop in accuracy after removing the neuron that was supposed to contain all the relevant information, suggesting that much of the representation of g_i is distributed over $z_{\neq i}$. Consider, e.g., WeakDE on Dsprites. SNC=5.8, meaning (roughly) that the neuron aligned to each factor predicts with 5.8% accuracy. Because this is the most informative individual neuron, it means all other neurons, i.e. non-aligned neurons, predict that factor with accuracy <5.8%. Yet NK=5.9 and MLP=99.8, meaning all these non-aligned neurons predict with 99.8-5.9 = 93.9% accuracy. (NK is the difference between the full MLP and the MLP with the aligned neuron knocked out.) This constitutes a distributed entanglement: the MI for a set is high, while the MI for each individual neuron is very low. Thus, Table 2 substantiates the argument Section 3.2 about the importance of distributed entanglements. showing that they are not merely a theoretical possibility, but that they also occur in practice.

In Appendix F, Table 7 shows the results of existing metrics on these models and datasets. In general, the scores of other metrics, especially DCI, are higher than those of our metrics, which fits with our theoretical arguments that they miss certain types of entanglement.

5.2 Downstream Task: Compositional Generalization

Compositional generalization (CG) is the ability to combine familiar, learned concepts in novel ways. Here, we quantify CG using the same method as Xu et al. (2022). That is, we test whether the representations produced by a model can be used to correctly classify novel combinations of familiar features. The CG ability of machine learning models has mostly been studied in the context of language (Baroni, 2020). However, recently, a number of authors have observed the connection between CG and DE (Zheng and Lapata, 2021; Montero et al., 2020; 2022; Esmaeili et al., 2019; Zhang et al., 2022; Higgins et al., 2016). Disentangled models should be capable of performing CG, because they can represent each component separately and independently, whereas if there is entanglement between the different features, then the novel combination is out of distribution, and so the model will likely struggle to classify it correctly. Following Xu et al. (2022), we

1. randomly sample values for two features, e.g., shape and size,
2. form a test set of points with those two values for those two features, e.g., all points with size=0 and shape='square', and a train set of all other points,
3. train the VAE (or supervised model) on the train set,
4. encode all data with the VAE encoder,
5. train and test an MLP with one output per feature, to predict the feature values from the encodings.

Table 3 shows results for the task of classifying novel combinations of familiar features. As well as the models from Section 5.1, we also report results for MTD, a fully supervised method (Sha and Lukasiewicz, 2021).

		Dsprites			3dshapes			mpi3d		
		shape	size	both	shape	size	both	shape	size	both
β-VAE	CG	0.00	61.55	0.00	33.00	67.70	14.75	89.00	3.66	2.34
	linear CG	0.00	0.00	0.00	13.70	24.60	0.00	60.56	0.00	0.00
	normal test set	82.35	66.24	**89.44**	96.30	96.90	**99.85**	90.53	77.44	**91.29**
βTCVAE	CG	0.00	49.56	0.00	29.40	77.50	17.73	87.87	0.00	0.00
	linear CG	0.00	4.63	0.01	4.00	58.30	0.77	75.55	0.00	0.00
	normal test set	82.32	66.22	**86.21**	96.50	96.80	**99.89**	89.22	72.91	**83.47**
FactorVAE	CG	0.00	37.28	0.86	0.86	6.36	6.36	89.36	0.47	0.00
	linear CG	0.00	0.00	0.31	10.40	33.80	0.00	57.59	0.00	0.00
	normal test set	80.85	65.08	**77.05**	89.50	95.70	**98.09**	89.52	73.94	**88.59**
PartedVAE	CG	0.00	29.61	0.00	0.00	19.11	18.52	81.28	91.67	0.00
	linear CG	0.00	0.00	0.00	6.80	84.20	6.61	70.10	91.67	0.00
	normal test set	58.33	63.83	**70.65**	93.50	95.90	**96.24**	72.36	91.67	**65.49**
PartedVAE-ss	CG	0.00	31.83	0.00	0.00	30.33	30.33	72.30	0.00	0.00
	linear CG	0.00	0.00	0.00	3.70	33.40	0.00	60.80	0.00	0.00
	normal test set	39.47	38.48	**48.48**	76.70	94.20	**88.93**	71.20	25.70	**60.61**
WeakDE	CG	0.00	41.79	0.00	0.00	7.83	7.83	87.10	0.00	0.00
	linear CG	0.00	0.00	0.00	0.00	40.40	0.00	37.70	0.00	0.00
	normal test set	74.66	65.58	**79.01**	96.50	96.90	**99.81**	82.90	50.79	**83.36**
MTD	CG	0.00	0.03	0.00	0.00	0.50	0.00	23.30	0.00	0.00
	normal test set	99.66	100.00	**99.56**	100.00	100.00	**100.00**	99.80	95.01	**95.36**

Table 3: MLP classification accuracy for novel combinations of familiar features, denoted 'CG' and 'CG linear' respectively, and classification accuracy when the test set is chosen randomly, denoted 'normal test set'.

Due to the huge number of feature value combinations, it is not feasible to test all of them in enough detail to obtain reliable results. We restrict attention to shape and size, use five feature combinations for Dsprites, six for MPI3D and eight for 3dshapes, and report the average. Appendix G contains the full results for all metrics on all features, datasets and models. The "normal test set" setting uses the same method except divides the train and test sets randomly. We adjust for chance agreement as $\max(0, (a - r)/(1 - r))$, where a is the model accuracy, and r is the chance agreement. Even restricting our attention to a single combination of feature types, our experimental results are already extensive involving ~ 200 VAE models, and > 1000 classification heads. An empirical study of multiple feature types would require 10-100x more, which would be a valuable future contribution, but is outside the scope of the present work.

The accuracy for identifying novel combinations of familiar features is generally low, often at the level of random guessing (i.e., 0 after adjusting for chance agreement). This is even true for MTD, the supervised model. In the 'normal test set" setting, on the other hand, every model is capable of classifying the unseen data accurately, which shows that it is the novelty of the combination that is degrading performance.

There is perhaps a danger that the MLP itself entangles the two features. For example, if yellow circles are excluded then, when classifying shape, the MLP could learn that whenever the "colour" neuron indicates "yellow", it should place low probability mass on "circle". We feel this is unlikely to affect results significantly, as the relationship between the "size" neuron and the value of shape would be highly irregular and present only for a small subset of data points. To make sure of this, we repeat the experiment using a linear classifier instead of an MLP, so this non-linear relationship could not be learnt. The performance of the linear classifier is as low or lower, almost never above a random baseline, which suggests the poor performance is not due to the MLP itself entangling the two factors.

There is a clear difference across datasets: results on Dsprites and MPI3D are essentially always at the level of random guessing, whereas those on 3dshapes are more promising, reaching nearly 30% (chance adjusted) for some models. PartedVae performs best, especially the semi-supervised variant, though perhaps surprisingly, the other recent semi-supervised method, WeakDE, performs less well.

	SNC (ours)	NK (ours)	MIG	SAP	IRS	D	C	I	DCI	MED
3dshapes	**.849**	.710	.800	.726	.668	.472	.819	.740	.717	.785
all datasets	**.850**	.716	.571	.311	.106	.457	.414	.535	.493	.566

Table 4: Correlation of DE metrics with accuracy on novel combinations. Best in bold, second best underlined.

Some prior works have claimed their model can meaningfully represent novel combinations: Higgins et al. (2016) display figures of reconstructed chairs with a round bottom for certain latent traversals and Esmaeili et al. (2019) present reconstructions for MNIST digits with certain combinations of digits and features (e.g., line thickness) excluded. Conflicting results were found by Montero et al. (2020; 2022), who claimed that ability to represent novel combinations was unrelated to the degree of DE. However, those prior works have mostly only examined the reconstructions by the decoder and so do not provide sufficient evidence to make claims about the internal representation. Being able to reconstruct an image accurately does not establish anything about the internal representation, it could just be the result of learning the identity function. The key question is not what the decoder reconstructs from the encoding or from a latent traversal or from a sampled latent vector, it is what representation the *encoder* produces. To some extent, the same could be said of our approach, as we evaluate the output of an MLP. However, this MLP is much smaller (over 100x) than the decoder, and is not trained end to end, so whatever interference it exerts will be significantly less than that of the decoder. Also, the experiment by Montero et al. (2020) attempted to quantify CG performance as the decoder's pixel loss, which has long been observed to be a poor measure of the quality of a generated image (Oprea et al., 2020; Higgins et al., 2016). Montero et al. (2022) does manage to assess the encoder, but this is only a qualitative assessment, so cannot determine quantitative correlation. On the other hand, classifying novel combinations, as we do, following Xu et al. (2022), is both quantitative, and is able to assess the encoder rather than the decoder. Unlike Higgins et al. (2016); Esmaeili et al. (2019), our experiments mostly show poor performance of existing DE models at CG. They also differ from Montero et al. (2020), as they reveal a correlation between DE and CG, with almost all metrics, and most strongly with our metrics.

5.3 Correlation with Compositional Generalization

Table 4 shows the Pearson correlation of our metrics and existing metrics, computed from the disentanglement_lib library, with CG performance. We restrict our metrics to just size and shape, because they are the relevant features for this task. We also tried restricting existing metrics to these two features only, without much change in results; see Appendix D. We show correlation on 3dshapes, and across all datasets. There is no insight to be gained from the correlation on Dsprites only or MPI3D only, since performance there is rarely above random, so all correlations will be essentially zero.

Although performance is generally low on CG, there is still enough variation to measure a meaningful correlation. Some models (PVAE on 3dshapes) achieve up to 30% even after chance adjusting, which is significantly higher than the random guessing of the lower-scoring models. A good DE metric should distinguish between models at the higher and lower ends of this range.

The first observation is that all metrics are at least weakly correlated with compositional generalization performance, with some showing a moderate to strong correlation. This contradicts Montero et al. (2020), who claimed to find no relationship between DE and CG. As argued in Section 5.2, our method for measuring CG performance is more indicative of the encoding quality. Additionally, we test six models, eight metrics and three datasets, whereas Montero et al. (2020) test only two models, one metric and two datasets. Later, the same authors conduct a more thorough investigation of CG in DE models (Montero et al., 2022). They show that the encoder often maps CG test examples to the wrong region of latent space, despite achieving high DCI score, and so conclude that DE does not offer much benefit to CG performance. However, the discrepancy between DCI and CG can also be explained by DCI overestimating DE, as in Theorem 3.1, and experimentally, as in Cao et al. (2022). The conclusion that our work supports is that DE and CG are indeed closely related, but that this relationship is obscured by the fact that, so far, DE has been overestimated, especially by DCI. That is, both DE and CG are low for all models, but they are correlated.

The strongest correlation is found with SNC. It is noteworthy that some existing metrics also score quite highly. This is more true on 3dshapes than across all datasets, and the correlation over all datasets may be more meaningful because it is computed over more points. Across all datasets, NK has the second strongest correlation and both SNC and NK are further ahead of the others. Comparing the different correlation coefficients using the Fischer transformation, shows SNC to be significantly ($p < 0.01$) greater than all other metrics except NK for the correlation across all datasets. All other comparisons, except NK with IRS across all datasets, do not rise to the level of statistical significance. See Appendix I for more details.

The consistent correlation between DE and CG evidenced by Table 4 also shows that there is more than enough variance in the CG scores to obtain a meaningful correlation. If all models were essentially at random guessing, correlation with any metric would be due only to chance, and we would expect correlations close to zero, but this is not the case. There is clearly enough variation in the CG results to identify a meaningful relationship to metric score, because the correlation is strong, far above zero, for most metrics, and far above statistical significance ($p < 0.01$) for our metrics (calculation in appendix). This further validates the suitability of CG as a downstream task to evaluate DE metrics.

SNC correlates more strongly than NK. This is expected, as SNC measures the extent to which information about the generative factors is encoded in a disentangled way, whereas NK penalizes representations that also encode information in a redundant, entangled way. NK is perhaps more relevant than SNC to interpretability, where we want to know that a given factor is encoded by a single neuron *only*, whereas SNC is more relevant for downstream performance. A similar comparison can be made between the components of DCI: C and I measure the extent to which the information is present in a disentangled way Thus, they are loosely analogous, in the context of DCI, to our SNC, and, like SNC, are a better predictor of downstream performance than D.

6 Discussion

Limitations and Future Work. One limitation of our work is the focus on only two features in combination. Our experimental results are already extensive involving ~ 200 VAE models, and > 1000 classification heads, but an interesting future work would be to perform the same analysis for a different feature combination. This could investigate, for example, whether novel combinations of simple features like (x, y) coordinates, are easier to recognize than novel size-shape combinations. Another limitation is the applicability of our metrics only to strong DE. As we have argued, strong DE and weak DE are fundamentally different objectives, and so we claim that they should be assessed by different metrics. This implies that it is not possible to design a metric that works for both forms. Instead, one could consider a modification of our metrics that work in the case of weak DE. Thirdly, NK, like all classifier-based metrics, requires training two classifiers for each feature. The classifiers are very small (<3000 parameters), and we find it to be significantly faster than DCI, which trains ensembles of decision trees, but still, NK requires some computational resources. Finally, the datasets we test on have only uniform and uncorrelated feature values. Future work should apply our metrics to datasets of non-uniform and correlated features, and also explore explicitly adding noise to the input.

Conclusion. In this paper, we identified two common flaws in existing disentanglement metrics: incorrect alignment of generative factors with neurons and conflating the importance of individual neurons with that of sets of neurons, and showed how these problems occur in real-world examples. We then introduced two new metrics, single-neuron classification and neuron knockout, which avoid these problems. Next, we proposed the classification of novel combinations of familiar features as a real-world downstream task against which to compare disentanglement metrics, and showed that our metrics are strongly predictive of performance on this task, more strongly than existing metrics.

Ethics Statement

We have considered the potential ethical implications of this work, and do not believe there are any concerns in this respect. The data used is all publicly available and impersonal, there is no use of human subjects, and there are no obvious malicious applications of our work. There are also no legal issues or issues are conflict with sponsorship.

Reproducibility Statement

In Section 4, we describe how our metrics are calculated. In Section 5, we describe the structure of the experiments we ran, and give implementation details and reference to publicly available code. Most importantly, we release all the code to reproduce our experiments in an anonymous repo https://github.com/anon296/anon.

References

Marco Baroni. Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B*, 375(1791):20190307, 2020.

Yoshua Bengio. Learning deep architectures for ai. *Foundations and Trends® in Machine Learning*, 2(1): 1–127, 2009.

Yoshua Bengio. Deep learning of representations: looking forward. In *Proceedings of the International Conference on Statistical Language and Speech Processing*, pages 1–37. Springer, 2013.

Yoshua Bengio, Grégoire Mesnil, Yann Dauphin, and Salah Rifai. Better mixing via deep representations. In *Proceedings of the International Conference on Machine Learning*, pages 552–560. PMLR, 2013.

A. Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 7(4):401–406, 1946.

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

Jinkun Cao, Ruiqian Nai, Qing Yang, Jialei Huang, and Yang Gao. An empirical study on disentanglement of negative-free contrastive learning. *arXiv preprint arXiv:2206.04756*, 2022.

Ricky TQ Chen, Xuechen Li, Roger B. Grosse, and David K. Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in Neural Information Processing Systems*, 31, 2018.

Kien Do and Truyen Tran. Theory and evaluation metrics for learning disentangled representations. *arXiv preprint ArXiv:1908.09961*, 2019.

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

Babak Esmaeili, Hao Wu, Sarthak Jain, Alican Bozkurt, Narayanaswamy Siddharth, Brooks Paige, Dana H. Brooks, Jennifer Dy, and Jan-Willem Meent. Structured disentangled representations. In *Proceedings of the The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2525–2534. PMLR, 2019.

Babak Esmaeili, Robin Walters, Heiko Zimmermann, and Jan-Willem van de Meent. Understanding optimization challenges when encoding to geometric structures. In *NeurIPS 2022 Workshop on Symmetry and Geometry in Neural Representations*, 2022.

Sina Hajimiri, Aryo Lotfi, and Mahdieh Soleymani Baghshah. Semi-supervised disentanglement of class-related and class-independent factors in VAE. *arXiv preprint ArXiv:2102.00892*, 2021.

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. In *Proceedings of the International Conference on Machine Learning*, 2016.

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.

Geoffrey E. Hinton et al. Learning distributed representations of concepts. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, volume 1, page 12. Amherst, MA, 1986.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *Proceedings of the International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint ArXiv:1312.6114*, 2013.

David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. *arXiv preprint ArXiv:2007.10930*, 2020.

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

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the International Conference on Machine Learning*, pages 4114–4124. PMLR, 2019.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *Proceedings of the International Conference on Machine Learning*, pages 6348–6359. PMLR, 2020.

Michael F Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. Disentangling factors of variation in deep representation using adversarial training. *Advances in Neural Information Processing Systems*, 29, 2016.

Milton Montero, Jeffrey Bowers, Rui Ponte Costa, Casimir Ludwig, and Gaurav Malhotra. Lost in latent space: Examining failures of disentangled models at combinatorial generalisation. *Advances in Neural Information Processing Systems*, 35:10136–10149, 2022.

Milton Llera Montero, Casimir JH Ludwig, Rui Ponte Costa, Gaurav Malhotra, and Jeffrey Bowers. The role of disentanglement in generalisation. In *International Conference on Learning Representations*, 2020.

James Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38, 1957.

Sergiu Oprea, Pablo Martinez-Gonzalez, Alberto Garcia-Garcia, John Alejandro Castro-Vargas, Sergio Orts-Escolano, Jose Garcia-Rodriguez, and Antonis Argyros. A review on deep learning techniques for video prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Edouard Pineau and Marc Lelarge. InfoCatVAE: representation learning with categorical variational autoencoders. *arXiv preprint ArXiv:1806.08240*, 2018.

Karl Ridgeway and Michael C. Mozer. Learning deep disentangled embeddings with the f-statistic loss. *Advances in Neural Information Processing Systems*, 31, 2018.

Michal Rolinek, Dominik Zietlow, and Georg Martius. Variational autoencoders pursue pca directions (by accident). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12406–12415, 2019.

Karsten Roth, Mark Ibrahim, Zeynep Akata, Pascal Vincent, and Diane Bouchacourt. Disentanglement of correlated factors via hausdorff factorized support. *arXiv preprint arXiv:2210.07347*, 2022.

Lei Sha and Thomas Lukasiewicz. Multi-type disentanglement without adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9515–9523, 2021.

Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf, and Stefan Bauer. Robustly disentangled causal mechanisms: validating deep representations for interventional robustness. In *Proceedings of the International Conference on Machine Learning*, pages 6056–6065. PMLR, 2019.

Andrea Valenti and Davide Bacciu. Leveraging relational information for learning weakly disentangled representations. *arXiv preprint arXiv:2205.10056*, 2022.

Zhenlin Xu, Marc Niethammer, and Colin A Raffel. Compositional generalization in unsupervised compositional representation learning: A study on disentanglement and emergent language. *Advances in Neural Information Processing Systems*, 35:25074–25087, 2022.

Yivan Zhang, Jindong Wang, Xing Xie, and Masashi Sugiyama. Equivariant disentangled transformation for domain generalization under combination shift. *arXiv preprint arXiv:2208.02011*, 2022.

Hao Zheng and Mirella Lapata. Disentangled sequence to sequence learning for compositional generalization. *arXiv preprint arXiv:2110.04655*, 2021.