UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Evaluating Disentangled Representations

Sepliarskaia, A.; Kiseleva, J.; de Rijke, M.

Publication date
2019
Document Version
Submitted manuscript

Link to publication

Citation for published version (APA):
Sepliarskaia, A., Kiseleva, J., & de Rijke, M. (2019). *Evaluating Disentangled Representations*. (v1 ed.) ArXiv. https://arxiv.org/abs/1910.05587

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:10 Oct 2026

Evaluating Disentangled Representations

Anna Sepliarskaia a.sepliarskaia@uva.nl
University of Amsterdam
Amsterdam, The Netherlands

Julia Kiseleva Julia.kiseleva@microsoft.com
Microsoft Research AI
Seattle, WA, USA

Maarten de Rijke derijke@uva.nl
University of Amsterdam
Amsterdam, The Netherlands

Editor:

arXiv:1910.05587v1 [cs.LG] 12 Oct 2019

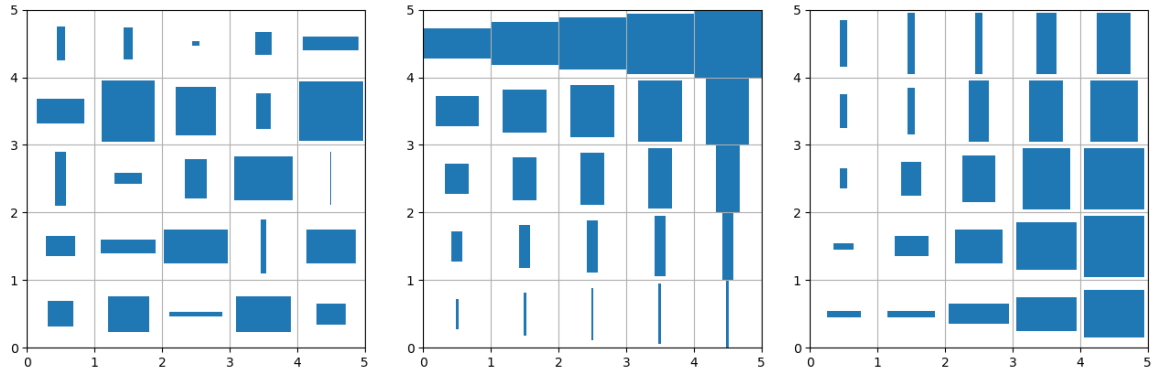
Abstract

There is no generally agreed upon definition of disentangled representation. Intuitively, the data is generated by a few factors of variation, which are captured and separated in a disentangled representation. Disentangled representations are useful for many tasks such as reinforcement learning, transfer learning, and zero-shot learning. However, the absence of a formally accepted definition makes it difficult to evaluate algorithms for learning disentangled representations. Recently, important steps have been taken towards evaluating disentangled representations: the existing metrics of disentanglement were compared through an experimental study and a framework for the quantitative evaluation of disentangled representations was proposed. However, theoretical guarantees for existing metrics of disentanglement are still missing. In this paper, we analyze metrics of disentanglement and their properties. Specifically, we analyze if the metrics satisfy two desirable properties: (1) give a high score to representations that are disentangled according to the definition; and (2) give a low score to representations that are entangled according to the definition. We show that most of the current metrics do not satisfy at least one of these properties. Consequently, we propose a new definition for a metric of disentanglement that satisfies both of the properties.

Keywords: representation learning, disentangled representation, auto encoder

1. Introduction

Algorithms for learning representations are crucial for a variety of machine learning tasks, including image classification (Vincent et al., 2008; Hinton and Salakhutdinov, 2006) and image generation (Goodfellow et al., 2014; Makhzani et al., 2015). These algorithms build a low-dimensional representation for each sample in a dataset, which is called a *latent representation*. Interpretable factors that describe every sample from the dataset are called *generative factors* of the dataset. While there is no single formalized notion of disentangled representation, the key intuition is that a disentangled representation should capture and separate the generative factors (Bengio et al., 2013; Higgins et al., 2018). For example, we show a dataset containing rectangles of different shapes (see Fig. 1a). There are two



(a) Example of a dataset containing rectangles with width w and height h from the uniform distribution $U(0, 1)$

(b) Change in the elements of the dataset caused by a change in the disentangled latent representation. Horizontally, the change is caused by a change in the first hidden factor; vertically, the change is caused by another hidden factor.

(c) Change in the elements of the dataset caused by a change in the entangled latent representation. Horizontally, the change is caused by a change in the first hidden factor; vertically, the change is caused by another hidden factor.

Figure 1: Example of a dataset and disentangled and entangled representations.

generative factors for this dataset: the length and width of the rectangles. In the disentangled latent representation of this dataset we can choose two latent factors. One of these factors is an invertible function of the length of the rectangles. Another factor is an invertible function of the width of the rectangles. In such a representation a change of one latent factor leads to a change only in one generative factor (see Fig. 1b). While in an entangled representation a change in one latent factor can lead to a change the length and width of the rectangles (see Fig. 1b).

Learning a disentangled representation is an important step towards better representation learning because a disentangled representation contains information about elements in a dataset in an interpretable and compact structure (Bengio et al., 2013; Higgins et al., 2018). Interpretability of the representation helps in tasks where users interact with a system, as they understand how it works and can provide informative feedback. Moreover, learning a disentangled representation helps for tasks where state-of-the-art machine learning-based approaches still struggle but where humans excel. Such scenarios include learning with knowledge transfer (Tommasi et al., 2010; Huang and Frank Wang, 2013; Pan et al., 2010), zero-shot inference (Lampert et al., 2009; Romera-Paredes and Torr, 2015) and supervised learning (Szegedy et al., 2013; Nguyen et al., 2015). The possible reason why people successfully solve these tasks is that they have a mental model that captures explanatory factors about the world. These factors are generalized and used in a new way when a person solves a new task. Algorithms that can obtain the same representation of the samples in a collection as humans will have a similar ability to generalize. Therefore, the development of an algorithm that learns disentangled representations has becomes an active area of research **????**. An important result in this area is the proof that learning unsupervised disentangled representations is only possible with inductive biases Locatello et al. (2018).

One successful way to include the bias using a few labels is to use the best model according to a given metric of disentanglement on the validation set **?**. Hence, in order to use this approach and to evaluate methods for learning disentangled representations, a metric of disentanglement is required. So far, several metrics of disentanglement have been proposed, but it is not clear which one should be preferred.

A number of important steps have been taken towards the formal evaluation of disentangled representations. For example, metrics of disentanglement have been compared through an experimental study on several datasets (Locatello et al., 2018). Another recent achievement is a framework for the evaluation of disentangled representations (Eastwood and Williams, 2018). Also, formal definition of disentangled representations using group theory has been proposed (Higgins et al., 2018). In this paper we continue this important line of research by providing an analysis of theoretical properties of disentanglement metrics.

To summarize, our key contributions are:

- We review existing metrics of disentanglement and discuss their fundamental properties.
- We propose a new metric of disentanglement with theoretical guarantees, and establish its fundamental properties.[1]

2. Background

2.1 Representation learning

Usually, a representation learning algorithm consists of two parts: an encoder and a decoder. An encoder is a function:

$$f_e : \mathbb{R}^d \to \mathbb{R}^N,\ \mathbf{c} = f_e(\mathbf{x}), \tag{1}$$

where $\mathbf{c}$ is a latent representation of the data sample $\mathbf{x}$. Typically, the dimension of the latent representation is much smaller than the dimension of the data. A decoder is a function:

$$f_d : \mathbb{R}^N \to \mathbb{R}^d,\ f_d(f_e(\mathbf{x})) \sim \mathbf{x}, \tag{2}$$

where $f_d(f_e(\mathbf{x}))$ should be close to $\mathbf{x}$. Thus, the latent representation should contain almost all the information that is contained in the original data.

2.2 Ground truth generative factors

Generative factors of a dataset are interpretable factors that describe the difference between any two samples from X. Consider, for example, a dataset containing rectangles of different shapes presented, as illustrated in Fig. 1a. The generative factors for this dataset are the length and width of the rectangles. Formally, the definition of the generative factor can be formulated as follows:

$$\exists g : \mathbb{R}^K \to \mathbb{R}^d \text{such that } \forall \mathbf{x} \in X \exists! \ \mathbf{z} : \mathbf{x} = g(\mathbf{z}), \tag{3}$$

where g is a generative process. The *ground truth* generative factors are generative factors that are given for a collection. This means that for each dataset sample $\mathbf{x} \in X$, the values of the generative factors $\mathbf{z} \in \mathbb{R}^K$ are known.

1. The comparison of methods for learning a disentangled representation is beyond the scope of this paper.

3. Metrics of Disentanglement of Representations

In this section, we analyze the properties of the metrics of disentangled representations proposed in (Higgins et al., 2017; Kim and Mnih, 2018; Eastwood and Williams, 2018; Chen et al., 2018; Kumar et al., 2017). To this end, we first repeat the definitions of disentanglement of representations that the proposed metrics reflect.

Definition 1 *(Higgins et al., 2017; Kim and Mnih, 2018; Eastwood and Williams, 2018)* A *disentangled representation* is a representation where a change in one latent dimension corresponds to a change in one generative factor while being relatively invariant to changes in other generative factors.

Definition 2 *(Locatello et al., 2018; Kumar et al., 2017)* A *disentangled representation* is a representation where a change in a single generative factor leads to a change in a single factor in the learned representation.[2]

A representation is *entangled* if it is not disentangled. These two definitions of disentanglement have important differences. Indeed, a representation in which several latent factors capture one common generative factor is disentangled by Def. 1, but entangled by Def. 2. On the other hand, a representation in which a latent variable captures multiple generative factors while there are no other latent variables that capture these generative factors, is entangled according to Def. 1, but disentangled by Def. 2. Consider, for example, the following latent representation of dimension 4 of the dataset containing rectangles of different shapes shown in Fig. 1a:

$$z_1 = x_1,\ z_2 = x_2^2,\ z_3 = y_1,\ z_4 = y_2^3,$$

where x is the length of a rectangle, while y is the width of a rectangle. It is disentangled by Def. 1, but it is entangled by Def. 2. Conversely, any one-dimensional latent representation of the same dataset would be disentangled by Def. 2, but not necessarily by Def. 1.

We note that the two definitions (Def. 1 and 2) depend on the concept of ground truth factors of variation. In this paper we assume that the ground truth factors of variation are predefined and given. For instance, in the example about rectangles illustrated in Example 1a, the length and width of the rectangles are selected as interpretable factors of variation. If, alternatively, one selects the rotation of these two factors as the ground truth factors of variation, then the disentangled representation should reflect the chosen rotation of the length and width of the rectangles.

Below we analyze existing disentanglement metrics using definitions Def. 1 and Def. 2: for each metric, we use a definition depending on which definition it should reflect according to the authors of the metric. We note that our methodology for evaluating metrics is "generous" for the metrics: we choose two basic properties for estimating the metric and the definition that is most favorable for the metric. The metrics that have been developed prior to this paper were developed to reflect different concepts and constraints (Def. 1 and Def. 2); hence, they should at least satisfy their own constraints, subject to their own definitions.

In particular, we analyze if the existing metrics satisfy the following properties:

2. This property of representations is also called *completeness* (Eastwood and Williams, 2018).

Property 1 *A metric gives a high score to all almost perfectly disentangled representations.*

Property 2 *A metric gives a low score for all entangled representations.*

3.1 BetaVAE and FactorVAE

BetaVAE (Higgins et al., 2017) and *FactorVAE* (Kim and Mnih, 2018) are metrics that rely on the following property of disentangled representations:

Property 3 *If one generative factor is fixed, while other generative factors are arbitrarily changed, then in the corresponding disentangled representation one latent factor should vary less than the others.*

3.1.1 Definition of BetaVAE

The algorithm that calculates *BetaVAE* (Higgins et al., 2017) consists of the following steps:
1. Choose a generative factor z_k.
2. Generate a batch of pairs of vectors for which the value of z_k within the pair is equal, while other generative factors are chosen randomly:

$$(\mathbf{p}_1 = \langle z_{1,1}, \ldots, z_{1,K} \rangle, \mathbf{p}_2 = \langle z_{2,1}, \ldots, z_{2,K} \rangle),\ z_{1,k} = z_{2,k}$$

3. Calculate the latent code of the generated pairs: $(\mathbf{c}_1 = f_e(g(\mathbf{p}_1)), \mathbf{c}_2 = f_e(g(\mathbf{p}_2)))$
4. Calculate the absolute value of the pairwise differences of these representations:

$$\mathbf{e} = \langle |c_{1,1} - c_{2,1}|, \ldots, |c_{1,N} - c_{2,N}| \rangle$$

5. The mean of these differences across the examples in the batch gives one training point for the linear regressor that predicts which generative factor was fixed.
6. BetaVAE is the accuracy of the linear regressor.

3.1.2 Definition of FactorVAE

The idea behind FactorVAE (Kim and Mnih, 2018) is very similar to BetaVAE. The main difference between them concerns how a batch of examples is generated to obtain a variation of latent variables when one generative factor is fixed. Another difference is the classifier that predicts which generative factor was fixed using the variation of latent variables. *FactorVAE* can be calculated by performing the following steps:
1. Choose a generative factor z_k.
2. Generate a batch of vectors for which the value of z_k within the batch is fixed, while other generative factors are chosen randomly.
3. Calculate latent codes of vectors from one batch.
4. Normalize each dimension in the latent representation by its empirical standard deviation over the full data.
5. Take the empirical variance in each dimension of these normalized representations.
6. The index of the dimension with the lowest variance and the target index k provides one training point for the classifier.
7. FactorVAE is the accuracy of the classifier.

3.1.3 Facts about BetaVAE and FactorVAE

Fact 1 BetaVAE *and* FactorVAE *do not satisfy Property 1 when the disentangled representation is defined using Definition 1.*

Proof By Definition 1, in a perfectly disentangled representation there may be several latent factors that correspond to changes in the same generative factor. Consequently, these latent factors have variation 0 when the corresponding generative factor is fixed. That is why the classifier cannot distinguish between these latent factors and its accuracy is less than 1. Consequently, the BetaVAE and FactorVAE metrics return scores of less than 1 for a perfectly disentangled representation. ■

Fact 2 BetaVAE *does not satisfy Property 2 when the disentangled representation is defined using Definition 1.*

Proof As a proof, we give a counterexample. Let us consider all training points for a linear classifier with a fixed label. The classifier can learn to map some regularity in the values of features to the right class. However, there do not exist any constraints on this regularity. The classifier can learn to map samples with a value of 0 of some feature to the correct class. But the classifier can also learn to map samples with other patterns in the feature values to the correct class. Given this intuition, let us consider the following example. Suppose there are 3 generative factors from a uniform distribution and the dimension of the latent representation is 3. Assume that the latent variables are equal to the generative factors with the following probabilities:

$$p_1 = (0.5, 0.5, 0),\ p_2 = (0, 0.5, 0.5),\ p_3 = (0.5, 0, 0.5)$$

We generate 10,000 training points with a batch size of 128. The accuracy of the linear classifier is equal to 0.9967 in this case, but the latent representation is totally entangled. This shows that BetaVAE can assign high scores to entangled representations. ■

Fact 3 FactorVAE *does not satisfy Property 2 when the disentangled representation is defined using Definition 1.*

Proof First, let us analyze the algorithm that calculates the FactorVAE score. Suppose that for each generative factor z_j there is a latent variable c_{i_j} that correlates with z_j more than other variables in the latent code $\mathbf{c}$. In this case, when z_j is fixed and the batch size is large enough, the variation in the latent factor c_{i_j} will be smaller than the variation in other latent factors. Consequently, the classifier will have high accuracy. Given this intuition, let us consider the following example. Suppose there are 3 generative factors from a Gaussian distribution with $\mu = 0, \sigma = 1$, and each latent variable is a weighted sum of the generative factors:

$$c_1 = 0.5 \cdot z_1 + 0.4 \cdot z_2 + 0.5 \cdot z_3,\ c_2 = 0.4 \cdot z_1 + 0.5 \cdot z_2 + 0.5 \cdot z_3,\ c_3 = 0.4 \cdot z_1 + 0.4 \cdot z_2 + 0.6 \cdot z_3$$

We generate 10,000 training points with a batch size of 128. The FactVAE disentanglement score is equal to 1 in this case, but the representation is entangled. This shows that FactoVAE can assign high scores to entangled representations. ■

3.2 The DCI, SAP and MIG metrics

In this section, we describe metrics of disentangled representations that are calculated using a notion of informativeness between latent variables and generative factors.

3.2.1 DCI: Disentanglement, Completeness and Informativeness

Eastwood and Williams (2018) propose to use a metric of disentangled representations, which we call DCI, that is calculated as follows:

1. First, the *informativeness* between c_i and z_j is calculated. To determine the informativeness between c_i and z_j, Eastwood and Williams (2018) suggest training K regressors. Each regressor f_j predicts z_j given $\mathbf{c}$ ($\hat{z_j} = f_j(\mathbf{c})$) and can provide an importance score $P_{i,j}$ for each c_i. The normalized importance score obtained by regressor f_j for variable c_i is used as the informativeness between c_i and z_j:

$$I_{i,j} = \frac{P_{i,j}}{\sum_{k=0}^{k=K} P_{i,k}}.$$

2. For each latent variable its score of disentanglement is calculated as follows:

$$H_K(I_i) = 1 + \sum_{k=1}^{K} I_{i,k} \log_K I_{i,k}.$$

3. The weighted sum of the obtained scores of disentanglement for the latent variables is DCI:

$$\text{DCI}(\mathbf{c}, \mathbf{z}) = \sum_i \left(\rho_i \cdot H_K(I_i)\right), \text{ where } \rho_i = \frac{\sum_j P_{i,j}}{\sum_{ij} P_{i,j}}. \tag{4}$$

3.2.2 SAP score: Separated Attribute Predictability

Kumar et al. (2017) provide a metric of disentanglement that is calculated as follows:

1. Compute a *matrix of informativeness* $I_{i,j}$, in which the ij-th entry is the linear regression or classification score of predicting the j-th generative factor using only the i-th variable in the latent representation.
2. For each column in the matrix of informativeness $I_{i,j}$, which corresponds to a generative factor, calculate the difference between the top two entries (corresponding to the top two most predictive latent factors). The average of these differences is the final score, which is called the SAP:

$$\text{SAP}(\mathbf{c}, \mathbf{z}) = \frac{1}{K} \sum_k \left(I_{i_k,k} - \max_{l \neq i_k} I_{l,k} \right), \; i_k = \arg\max_i I_{i,k}.$$

3.2.3 MIG: Mutual Information Gap

Chen et al. (2018) propose a disentanglement metric, Mutual Information Gap (MIG), that uses mutual information between the j-th generative factor and the i-th latent variable as a notion of informativeness between them. The *mutual information* between two variables c and z is defined as

$$I(c; z) = H(z) - H(z|c),$$

where $H(z)$ is the entropy of the variable z. Mutual information measures how much knowing one variable reduces uncertainty about the other. A useful property of mutual information is that it is always non-negative $I(c; z) > 0$. Moreover, $I(c; z)$ is equal to 0 if and only if c and z are independent. Also, mutual information achieves its maximum if there exists an invertible relationship between c and z. The following algorithm calculates the MIG score:

1. Compute a *matrix of informativeness* $I_{i,j}$, in which the ij-th entry is the mutual information between the j-th generative factor and the i-th latent variable.
2. For each column of the score matrix $I_{i,j}$, which corresponds to a generative factor, calculate the difference between the top two entries, and normalize it by dividing by the entropy of the corresponding generative factor. The average of these normalized differences is the MIG score:

$$\text{MIG}(\mathbf{c}, \mathbf{z}) = \frac{1}{K} \sum_{k} \frac{I_{i_k,k} - \max_{l \neq i_k} I_{l,k}}{H(z_k)}, \;\; i_k = \arg\max_{i} I_{i,k}.$$

3.2.4 Facts about DCI, SAP and MIG

Fact 4 *DCI does not satisfy Property 1 when the disentangled representation is defined using Definition 1.*

Proof We argue that using entropy as a score of disentanglement of one latent variable is not correct. Indeed, a score of disentanglement of c_i should be high when c_i reflects one generative factor well, while it reflects other generative factors equally poorly. However, since the distribution may be close to uniform for these generative factors, the entropy is large. Let us provide an example that is built on this observation. Suppose there are 11 generative factors, and 11 is the dimension of the latent representation. Each latent factor c_i captures primarily a generative factor z_i:

$$I_{i,i} = 0.8, \;\; I_{i,k} = 0.02, \;\; k \neq i.$$

Then, the DCI score is 0.6, so the DCI assigns a small score to a disentangled representation. ■

Fact 5 *SAP does not satisfy Property 1 when the disentangled representation is defined using Definition 2.*

Proof We think that it is incorrect to use the R^2 score of linear regression as informativeness between latent variables and generative factors. Indeed, a linear regression cannot capture non-linear dependencies. Thus, informativeness, which is calculated using the R^2 score of a linear regression, may be low if each generative factor is a non-linear function of some

latent variable. Let us give an example that is built on this observation. Suppose there are 2 generative factors from the uniform distribution $U([-1,1])$ and the dimension of the latent representation is 2. Let us assume the latent variables are obtained from the generative factors according to the following equations:

$$c_1 = z_1^{15},\ c_2 = z_2^{15}$$

For this perfectly disentangled representation, we generate 10,000 examples and obtain the SAP score equal to 0.32. It proves that SAP can assign a low score to disentangled representations. ■

Fact 6 *MIG satisfies Property 1 when the disentangled representation is defined using Definition 2.*

Proof Indeed, in a disentangled representation each generative factor is primarily captured in only one latent dimension. This means that for each generative factor z_j, there is exactly one latent factor c_{i_j} for which z_j is a function of c_{i_j}: $z_j \sim f(c_{i_j})$. Therefore,

$$I_{i_j,j} = H(z_j) - H(z_j|c_{i,j}) \sim H(z_j),$$

whereas for other latent variables $I_{k,j} = I(c_k, z_j) \sim 0$. Consequently, according to MIG, the score of disentanglement of each generation factor is close to 1:

$$\frac{I_{i_j,j} - \max_{k \neq i_j} I_{k,j}}{H(z_j)} \sim 1. \tag{5}$$

Therefore, the average of these scores is also close to 1. This shows that MIG score always assigns a high score to disentangled representations. ■

Fact 7 *DCI does not satisfy Property 2 when the disentangled representation is defined using Definition 1.*

Proof We give a counterexample, which is built on the fact that the weighted sum in Eq. 4 can be large if only one latent variable is disentangled, while the other latent variables are entangled and do not capture any information about generative factors. Suppose there are 2 generative factors and the dimension of the latent representation is 2, and the matrix of informativeness is the following:

$$P_{0,0} = 1, P_{0,1} = 0, P_{1,1} = 0.09, P_{1,0} = 0.01.$$

In this case, the DCI score is 0.957. This counterexample shows that the DCI score can be close to 1 for entangled representations. ■

Fact 8 *SAP does not satisfy Property 2 when disentangled representation is defined using Definition 2.*

Proof A high SAP score indicates that the majority of generative factors is captured linearly in only one latent dimension. However, the SAP metric does not penalize the existence of several latent factors that capture the same generative factor non-linearly. Let us consider the following example. Suppose there are 2 generative factors from the uniform distribution $U([-1, 1])$, and the dimension of the latent representation is 3. Let us assume that the latent factors are obtained from the generative factors according to the following equations:

$$c_1 = z_1,\ c_2 = z_1^{25} + z_2^{25},\ c_3 = z_2.$$

For this latent representation, a change in each generative factor leads to a change in several latent factors, but the SAP score is equal to 0.98. This shows that the SAP score can be close to 1 for entangled representations. ■

Fact 9 *MIG satisfies Property 2 when the disentangled representation is defined using Definition 2.*

Proof

A high MIG score indicates that the majority of generative factors is captured in only one latent dimension. Consequently, a change in one of the generative factors entails a change primarily in only one latent dimension. ■

A summary of the results of our analysis is given in Table 1.

Table 1: Summary of facts about proposed metrics of disentangled representations.

Metric	Satisfies Property 1	Satisfies Property 2	Definition used
BetaVAE	No	No	Definition 1
FactorVAE	No	No	Definition 1
DCI	No	No	Definition 1
SAP	No	No	Definition 2
MIG	Yes	Yes	Definition 2

4. A New Metric of Disentanglement, DCIMIG

While previous metrics are either based on Def. 1 or on Def. 2, we believe that both of these definitions should be satisfied simultaneously for a disentangled representation. This means that we consider the following definition of disentangled representation:

Definition 3 A *disentangled representation* is a representation satisfying the following properties. In a disentangled representation, we can choose a subset of latent variables: $c' = \{c_{i_1}, \ldots, c_{i_K}\}$, which satisfy Def. 1 and Def. 2. Moreover, a disentangled representation should contain nearly all information about generative factors, i.e., it should have a high degree of *informativeness* (Eastwood and Williams, 2018).

Based on Def. 3, we propose a new metric of disentanglement of representation, called DCIMIG. The previously introduced MIG metric is a good starting point, because it is the only metric that gives high scores for disentangled representations and low scores for entangled representations when using Def. 2. However, the MIG metric has some drawbacks. The MIG metric does not penalize latent representations in which latent factors capture several generative factors. Consequently, the MIG metric can assign large scores to representations that are entangled according to Def. 1. Moreover, the MIG metric does not capture the informativeness of latent representation as it equally penalizes latent representations for not capturing informative generative factors and for not capturing non-informative generative factors. That is why we propose a new metric, *DCIMIG*, that captures the Disentanglement, Completeness (see footnote 2) and Informativeness of a representation using the Mutual Information Gap.

4.1 Definition of DCIMIG

Following MIG, we create a matrix of informativeness $I_{i,j}$, in which the ij-th entry is the mutual information between the j-th generative factor and the i-th variable in the latent representation. The following steps for calculating DCIMIG differ from the steps suggested in MIG:

1. For each latent variable c_i, find the generative factor z_{j_i} that it reflects the most: $j_i = \arg\max_j I_{i,j}$.
2. Calculate the disentanglement for each latent variable: $D_i = I_{i,j_i} - \max_{k \neq j_i} I_{i,k}$.
3. For each generative factor z_j, find the most disentangled latent factor c_{k_j}, that reflects z_j: $k_j = \arg\max_{l \in \mathbb{I}_j} D_l$, where $\mathbb{I}_j = \{i : z_{j_i} = z_j\}$.
4. For each generative factor z_j, calculate the disentanglement score D_j^z, which is equal to D_{k_j} if there is at least one latent factor, that captures z_j, otherwise, it is 0.
5. Finally, the *disentanglement score* of a latent representation according to DCIMIG is the normalized sum of D_j^z:

$$\text{DCIMIG}(\mathbf{c}, \mathbf{z}) = \frac{\sum_{j=1}^{K} D_j^z}{\sum_{j=1}^{K} H(z_j)},$$

where $H(z_j)$ is the entropy of z_j.

4.2 Facts about DCIMIG

Fact 10 *DCIMIG satisfies Property 1 when the disentangled representation is defined using Def. 3.*

Proof Indeed, in a disentangled representation according to Def. 3, there is a subset c' of latent variables, in which each latent variable is sensitive to changes in one generative factor only. Moreover, for each generative factor z_j there is only one latent variable $c_{i_j} \in c'$ that captures the changes in z_j. Consequently, c_{i_j} is a function of z_j: $c_{i_j} = f_j(z_j)$, while the other latent factors are invariant to changes in z_j. This means that, $D_{i_j} = I_{i_j,j} - \max_{k \neq j} I_{i_j,k} = I_{i_j,j}$. Also, the disentangled representation should have a high degree of *informativeness*. Consequently, the latent variables in c' should capture all the information contained in z_j. But only $c_{j,i}$ contains some information about z_j. Therefore, $I_{i_j,j} = H(z_j)$, and $D_j^z = H(z_j)$.

Consequently, DCIMIG is equal to 1 in this case. ■

Fact 11 *DCIMIG satisfies Property 2 when the disentangled representation is defined using Def. 3.*

Proof When a representation is entangled for the majority of informative generative factors z', we cannot find a factor in the latent representation that reflects only this factor. There are 2 cases for the generative factors from z'. In the first case, there is no latent factor that captures the generative factor $z_j \in z'$. In that case, D_j^z is equal to 0. The second case is characterized by the fact that there is a latent factor that captures a generative factor, but this latent factor also captures other generative factors. In that case the disentangled score of this latent factor D_{i_j} is small, and consequently, D_j^z is small. ■

4.3 Difference between DCIMIG, DCI and MIG

While DCIMIG is similar to the DCI and MIG metrics it has important differences from them:

1. DCI does not penalize representations if there is a generative factor that is not captured by any latent variable; it penalizes representations that contains both disentangled latent variables and entangled ones. DCIMIG penalizes representations that do not capture some generative factors; the DCIMIG score is high if there is a subset of disentangled latent variables that captures all generative factors; DCIMIG does not penalize representations if there are other latent variables that are entangled.

2. MIG does not penalize representations if a latent variable captures several generative factors; if a representation only captures a subset of all the generative factors, then MIG assigns the same score no matter which subset is captured, whether the generative factors that have been captured have high or low entropy. In contrast, DCIMIG does distinguish between representations that capture different subsets of generative factors.

5. Related work

This paper is relevant to two research directions: the formulation of a notion of disentangled representation and the analysis of differences between proposed metrics of disentangled representations.

A definition of disentangled representation is presented by Higgins et al. (2018), who proposed to call a representation disentangled if it is consistent with transformations that characterized the dataset. In particular, Higgins et al. (2018) suggested that transformations that change only some properties of elements in the data set, while leaving other properties unchanged, give the structure of a data set. Desirable properties of a disentanglement metric were formulated by Eastwood and Williams (2018), which are disentanglement, completeness, and informativeness. Eastwood and Williams (2018) claimed that a good representation should satisfy all of these properties, namely (1) if a representation is good, then change in

one latent factor should lead to change in one generative factor, (2) a change in one generative factor should lead to a change in one latent factor, and (3) a latent representation should contain all information about the generative factors. Therefore, Eastwood and Williams (2018) proposed three metrics to satisfy each of the properties listed. However, the proposed metrics were not analyzed — a gap that we fill.

Several papers analyze the differences between metrics of disentanglement through experimental studies (Locatello et al., 2018; Chen et al., 2018). For example, Locatello et al. (2018) trained 12,000 models that cover the most prominent methods and evaluate these models using existing metrics of disentanglement. The study showed that the metrics are correlated, but the degree of correlation depends on the dataset. It is important to note that their experimental results are consistent with our theoretical findings: the BetaVAE (Higgins et al., 2017) and FactorVAE Kim and Mnih (2018) metrics are strongly correlated with each other; and the SAP (Kumar et al., 2017), MIG Chen et al. (2018), DCI (Eastwood and Williams, 2018) scores are also strongly correlated. Locatello et al. (2018) made an important step towards the evaluation of methods to create disentangled representations, however, the properties of the metrics were not analyzed theoretically. Chen et al. (2018) took a step in this direction, but only analyzed the BetaVAE, FactorVAE and MIG metrics. Chen et al. (2018) compared metrics by analyzing robustness to the choice of the hyperparameters during experiments. The experimental findings are quite similar to ours: BetaVAE is a very optimistic metric and assigns high scores to entangled representations.

To summarize, the key distinctions of our work compared to previous efforts are: (1) an in-depth analysis of previously proposed metrics of disentanglement, and (2) a proposal of a single metric of disentanglement that reflects all properties of previously proposed ones and has theoretical guarantees.

6. Conclusion

In recent years, several models have been developed to obtain disentangled representations (Yu and Grauman, 2017; Hu et al., 2017; Denton et al., 2017; Kim and Mnih, 2018). The importance of developing a reliable metric of disentanglement has been clearly stated by Kim and Mnih (2018). In this paper, we analyzed whether existing metrics of disentanglement are close to 1 when a representation is disentangled and whether the metrics are close to 0 when a representation is entangled. As the definition of disentanglement varies from paper to paper, for our analysis we use the definition of disentanglement, which is used by the authors of the corresponding metric. Surprisingly, we found that most of the existing metrics can either give a low score to a disentangled representation or a high score to an entangled representation.

In some papers, disentangled representations are defined as representations in which a change in one generative factor should lead to a change in a single latent factor. While in other papers, disentangled representations are defined as representations in which a change in one latent factor should lead to a change in one generative factor. We argued that both properties should be satisfied by a disentangled representation. Based on this richer definition of a disentangled representation, we propose a new metric of disentanglement, DCIMIG, that captures the Disentanglement, Completeness and Informativeness of a representation using

the Mutual Information Gap. We prove that DCIMIG assigns high scores to disentangled representation and low scores to entangled representations.

In future work, we plan to extend DCIMIG to the case where some generative factors form a subspace, and a disentangled representation should align with these subspaces instead of single generative factors.

References

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

Tian Qi Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. *arXiv preprint arXiv:1802.04942*, 2018.

Emily L Denton et al. Unsupervised learning of disentangled representations from video. In *Advances in Neural Information Processing Systems*, pages 4414–4423, 2017.

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

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

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

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

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

Qiyang Hu, Attila Szabó, Tiziano Portenier, Matthias Zwicker, and Paolo Favaro. Disentangling factors of variation by mixing them. *arXiv preprint arXiv:1711.07410*, 2017.

De-An Huang and Yu-Chiang Frank Wang. Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2496–2503, 2013.

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

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

Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Proceedings of the IEEE Computer Vision and Pattern Recognition Conference*, pages 951–958, 2009.

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

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.

Sinno Jialin Pan, Qiang Yang, et al. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*, pages 2152–2161, 2015.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Tatiana Tommasi, Francesco Orabona, and Barbara Caputo. Safety in numbers: Learning categories from few examples with multi model knowledge transfer. In *Proceedings of the IEEE Computer Vision and Pattern Recognition Conference*, pages 3081–3088, 2010.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1096–1103. ACM, 2008.

Aron Yu and Kristen Grauman. Semantic jitter: Dense supervision for visual comparisons via synthetic images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5571–5580. IEEE, 2017.