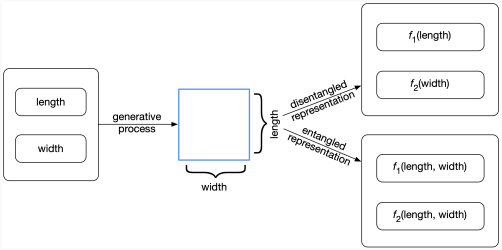
How Not to Measure Disentanglement

Anna Sepliarskaia Vienna University of Technology
Julia Kiseleva Microsoft Research
Maarten de Rijke University of Amsterdam

June 24, 2021

TECHNISCHE UNIVERSITÄT WIEN
Vienna | Austria

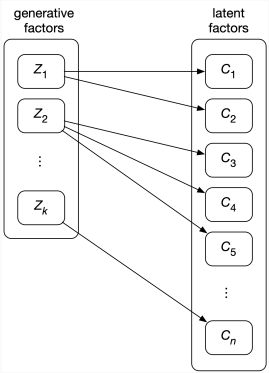
Disentangled Representations



- **NO** standard definition
- Separates generative factors
- Generative factors are interpretable factors

BetaVAE, FactorVAE, DCI

Characteristic 1 In a disentangled representation 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 of BetaVAE

1. Choose a generative factor z_r.

Definition of BetaVAE

1. Choose a generative factor z_r.
2. Generate a batch of pairs of vectors for which the value of z_r within the pair is equal, while other generative factors are chosen randomly

Definition of BetaVAE

1. Choose a generative factor z_r.
2. Generate a batch of pairs of vectors for which the value of z_r within the pair is equal, while other generative factors are chosen randomly
3. Calculate the latent code of the generated pair

Definition of BetaVAE

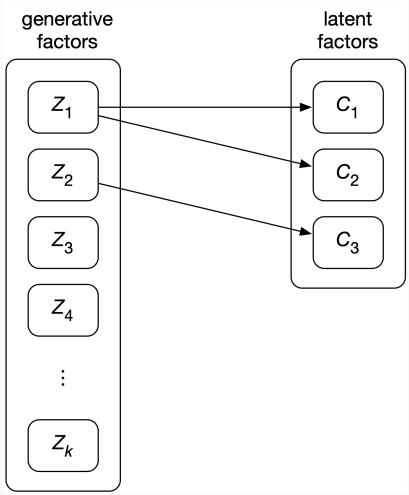
1. Choose a generative factor z_r.
2. Generate a batch of pairs of vectors for which the value of z_r within the pair is equal, while other generative factors are chosen randomly
3. Calculate the latent code of the generated pair
4. Calculate the absolute value of the pairwise differences of these representations

Definition of BetaVAE

1. Choose a generative factor z_r.
2. Generate a batch of pairs of vectors for which the value of z_r within the pair is equal, while other generative factors are chosen randomly
3. Calculate the latent code of the generated pair
4. Calculate the absolute value of the pairwise differences of these representations
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.

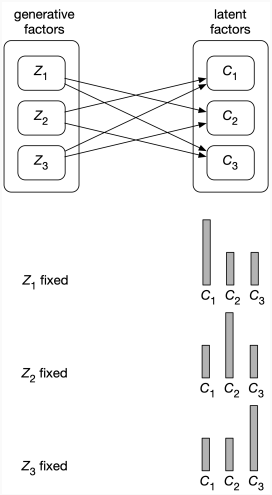
BetaVAE

Does a metric gives a high score to all representations that satisfy the characteristic 1



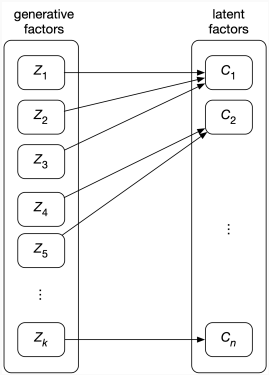
BetaVAE

Does a metric gives a high score to all representations that do **NOT** satisfy the characteristic 1



SAP and MIG metrics

Characteristic 2 In a disentangled representation a change in a single generative factor leads to a change in a single factor in the learned representation



Definition of MIG

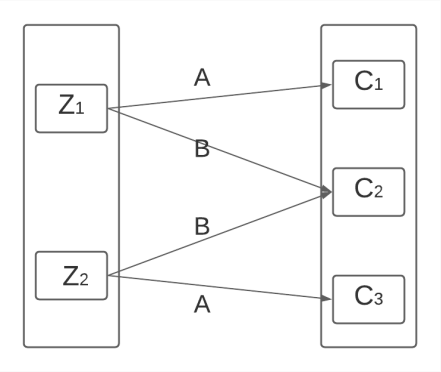
1. Compute a *matrix of informativeness* $I_{i,j}$, in which the ij-th entry is the mutual information between the j-th generative factor and the i-th latent variable.

Definition of MIG

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

Comparison Of Metrics

- $BetaVAE = 1$
- $FactorVAE = 1$
- $DCI = \frac{A}{A+B}$
- $MIG = |A - B|$



Coclusion

- *BetaVAE*, *FactorVAE*, *DCI* — characteristics 1
- *MIG*, *SAP* — characteristics 2
- All metrics besides *MIG* can give low score for disentangled representation
- All metrics besides *MIG* can give high score for entangled representation