Unsupervised Learning of Disentangled Representations

Zhiyuan Tang

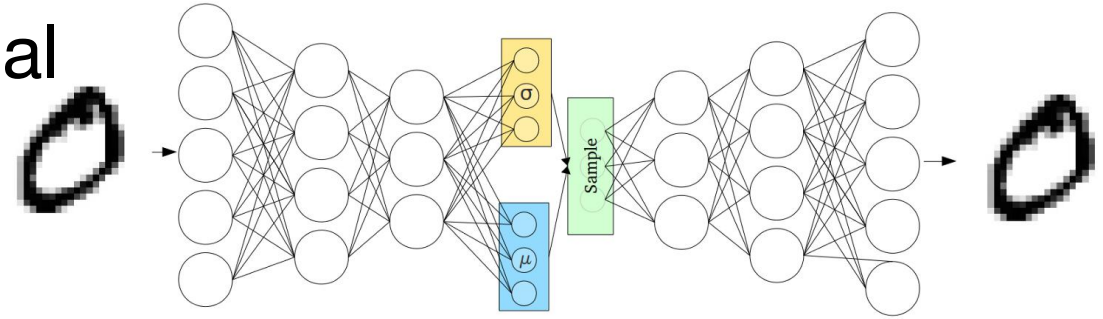
2020.2.17

Referred papers

- Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations, ICML 2019

Related concepts

- Disentangled representation
  - real-world data is generated by a few explanatory factors of variation
  - factorization
  - contain all the information present in x in a compact and interpretable structure
  - seperate dimensions, independent features
- Distributed vs Disentangled Representation
- High-dimensional vs low-dimensional
- Variational Autoencoders



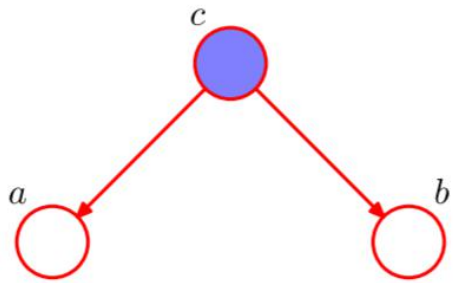
Two-step generative process

- P(z)
- P(x|z)

Inductive bias

The inductive bias (also known as learning bias) of a learning algorithm is the set of assumptions that the learner uses to predict outputs given inputs that it has not encountered.

- Maximum conditional independence
- Minimum cross-validation error
- Maximum margin
- Minimum description length
- Minimum features
- Nearest neighbors

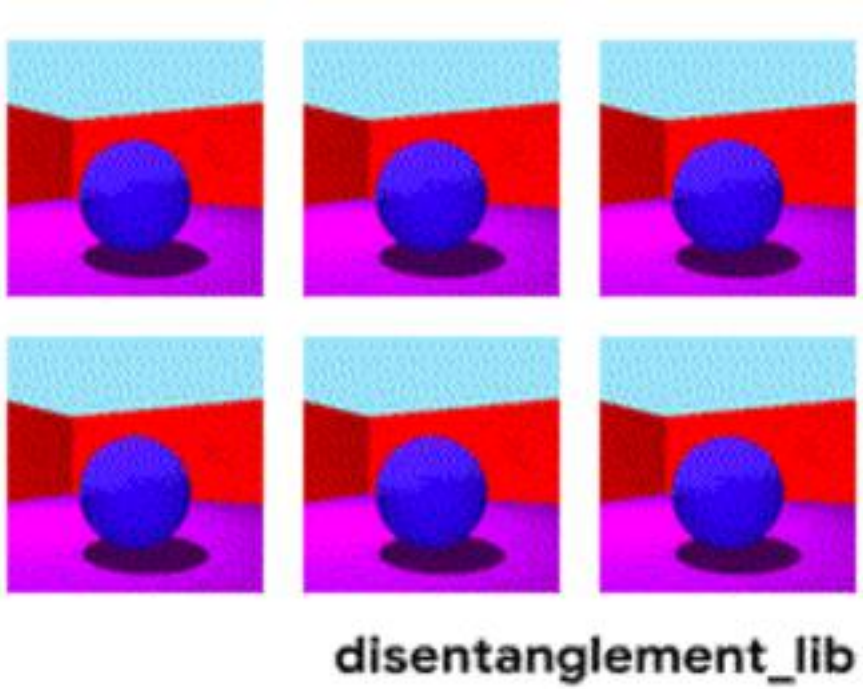


https://en.wikipedia.org/wiki/Inductive_bias

Impossibility

- The unsupervised learning of disentangled representations is fundamentally impossible without inductive biases both on the considered learning approaches and the data sets.

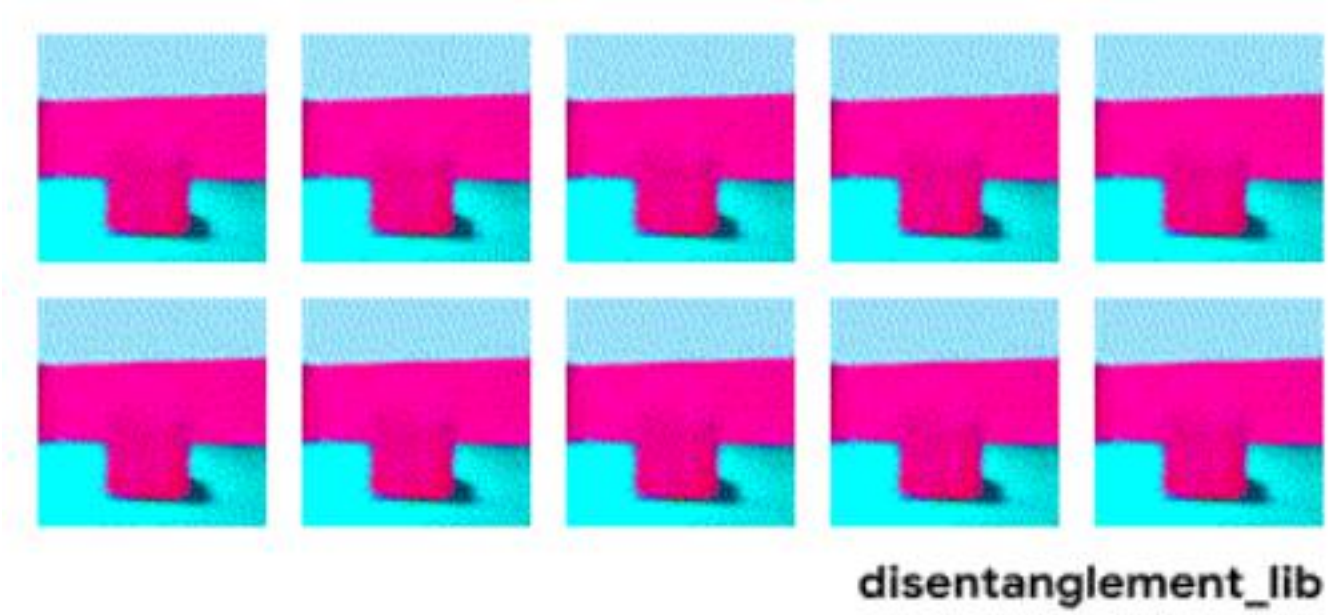
Disentanglement example



Visualization of the ground-truth factors of the Shapes3D data set: Floor color (upper left), wall color (upper middle), object color (upper right), object size (bottom left), object shape (bottom middle), and camera angle (bottom right).

https://ai.googleblog.com/2019/04/evaluating-unsupervised-learning-of.html

Disentanglement example



10-dimensional representation vector. The ground-truth factors wall and floor color as well as rotation of the camera are disentangled (see top right, top center and bottom center panels), while the ground-truth factors object shape, size and color are entangled (see top left and the two bottom left images).

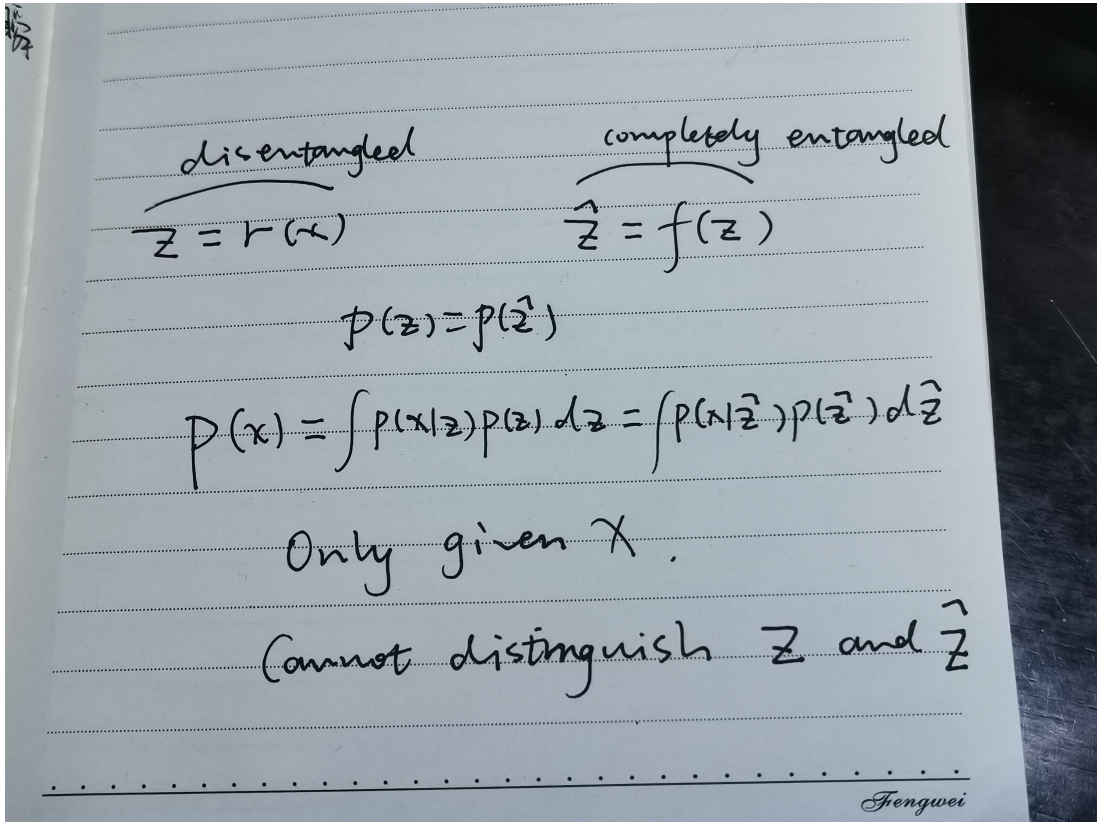
https://ai.googleblog.com/2019/04/evaluating-unsupervised-learning-of.html

Impossibility

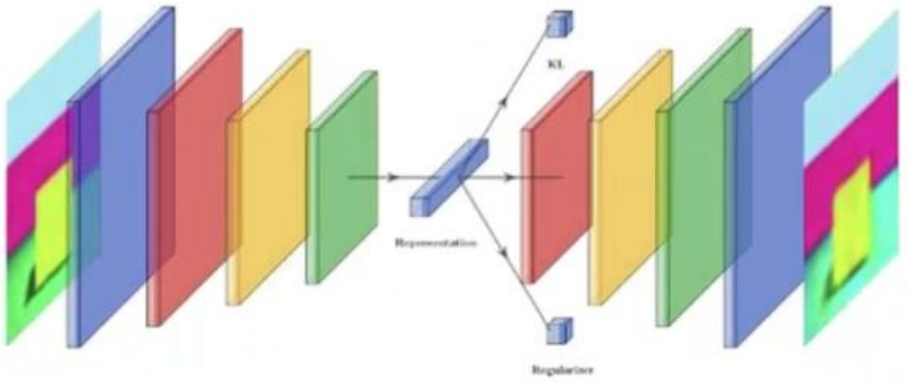
Theorem 1. *For $d > 1$, let $\mathbf{z} \sim P$ denote any distribution which admits a density $p(\mathbf{z}) = \prod_{i=1}^{d} p(\mathrm{z}_i)$. Then, there exists an infinite family of bijective functions $f : \mathrm{supp}(\mathbf{z}) \rightarrow \mathrm{supp}(\mathbf{z})$ such that $\frac{\partial f_i(\boldsymbol{u})}{\partial u_j} \neq 0$ almost everywhere for all i and j (i.e., $\mathbf{z}$ and $f(\mathbf{z})$ are completely entangled) and $P(\mathbf{z} \leq \boldsymbol{u}) = P(f(\mathbf{z}) \leq \boldsymbol{u})$ for all $\boldsymbol{u} \in \mathrm{supp}(\mathbf{z})$ (i.e., they have the same marginal distribution).*

Impossibility

Consider the commonly used "intuitive" notion of disentanglement which advocates that a change in a single ground-truth factor should lead to a single change in the representation. In that setting, Theorem 1 implies that unsupervised disentanglement learning is *impossible* for arbitrary generative models with a factorized prior[3] in the following sense: Assume we have $p(\mathbf{z})$ and some $P(\mathbf{x}|\mathbf{z})$ defining a generative model. Consider any unsupervised disentanglement method and assume that it finds a representation $r(\mathbf{x})$ that is perfectly disentangled with respect to $\mathbf{z}$ in the generative model. Then, Theorem 1 implies that there is an equivalent generative model with the latent variable $\hat{\mathbf{z}} = f(\mathbf{z})$ where $\hat{\mathbf{z}}$ is completely *entangled* with respect to $\mathbf{z}$ and thus also $r(\mathbf{x})$: as all the entries in the Jacobian of f are non-zero, a change in a single dimension of $\mathbf{z}$ implies that all dimensions of $\hat{\mathbf{z}}$ change. Furthermore, since f is deterministic and $p(\mathbf{z}) = p(\hat{\mathbf{z}})$ almost everywhere, both generative models have the same marginal distribution of the observations $\mathbf{x}$ by construction, i.e., $P(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z} = \int p(\mathbf{x}|\hat{\mathbf{z}})p(\hat{\mathbf{z}})d\hat{\mathbf{z}}$. Since the (unsupervised) disentanglement method only has access to observations $\mathbf{x}$, it hence cannot distinguish between the two equivalent generative models and thus has to be entangled to at least one of them.



Methods: VAE + regularizer



Beta-VAE, (Higgins et al., 2017);
Fix capacity of VAE bottleneck.
Annealed-VAE, (Burgess et al., 2017);
Progressively increase capacity of VAE bottleneck.
Factor-VAE, (Kim & Mnih, 2018);
Penalize Total Correlation with adversarial training.
Beta-TCVAE, Chen et al., 2018;
Penalize Total Correlation with Monte Carlo estimate.
DIP-VAE I and **II**, Kumar et al., 2018
Match moments with a disentangled prior.

Key idea: regularize such that the aggregated posterior factorizes

Metrics

Beta-VAE Metric, (Higgins et al., 2017);
Accuracy of a linear classifier that predicts the index of a fixed factor of variation.

Factor-VAE Metric, (Kim & Mnih, 2018);
Accuracy of a majority vote classifier that predicts the index of a fixed factor of variation.

Mutual Information Gap, (Chen et al., 2018);
Normalized gap in mutual information between the highest and second highest coordinate in r(x).

SAP score, (Kumar et al., 2018);
Average difference of the prediction error of the two most predictive latent dimensions for each factor.

DCI Disentanglement, (Eastwood & Williams, 2018);
Entropy of the predictive importance of each dimension of r(x).

Modularity, (Ridgeway & Mozer) 2018;
Measure if each dimension of r(x) depends on at most a factor of variation using their mutual information.

Can current methods enforce a uncorrelated aggregated posterior and representation?

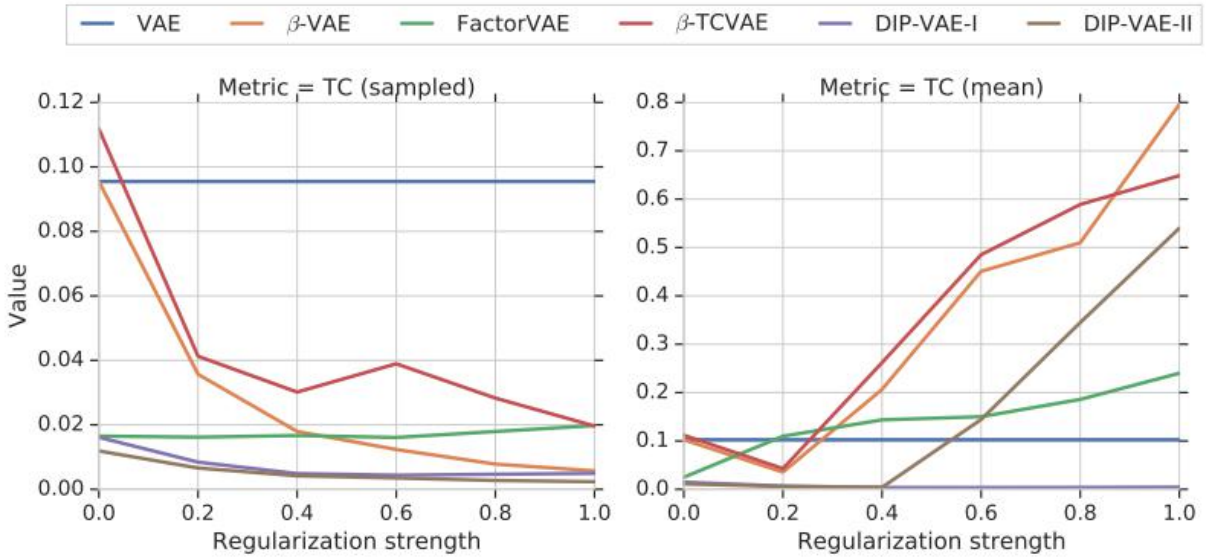


Figure 1. Total correlation based on a fitted Gaussian of the sampled (left) and the mean representation (right) plotted against regularization strength for Color-dSprites and approaches (except AnnealedVAE). The total correlation of the sampled representation decreases while the total correlation of the mean representation increases as the regularization strength is increased.

Mean vector of the Gaussian encoder as the representation
vs
Sample from the Gaussian encoder

It is not clear whether a factorizing aggregated posterior also ensures that the dimensions of the mean representation are uncorrelated.

How much do the disentanglement metrics agree?

- All disentanglement metrics except Modular ity appear to be correlated. However, the level of correlation changes between different data sets.

How important are different models and hyperparameters for disentanglement?

- how disentanglement is affected by the model choice, the hyperparameter selection and randomness (in the form of different random seeds).

How important are different models and hyperparameters for disentanglement?

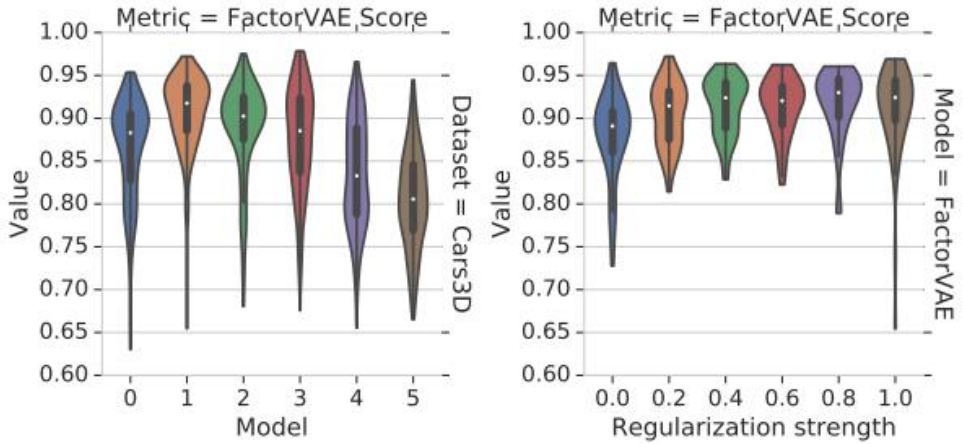


Figure 3. (left) FactorVAE score for each method on Cars3D. Models are abbreviated (0=β-VAE, 1=FactorVAE, 2=β-TCVAE, 3=DIP-VAE-I, 4=DIP-VAE-II, 5=AnnealedVAE). The variance is due to different hyperparameters and random seeds. The scores are heavily overlapping. (right) Distribution of FactorVAE scores for FactorVAE model for different regularization strengths on Cars3D. In this case, the variance is only due to the different random seeds. We observe that randomness (in the form of different random seeds) has a substantial impact on the attained result and that a good run with a bad hyperparameter can beat a bad run with a good hyperparameter.

The disentanglement scores of unsupervised models are heavily influenced by **randomness** (in the form of the random seed) and the choice of the **hyperparameter** (in the form of the regularization strength). The **objective** function appears to have less impact.

Are there reliable recipes for model selection?

- Hyperparameter selection
  - No consistent model, object function, hyperparameter
- Model selection based on unsupervised scores.
  - reconstruction error, the KL divergence, ELBO, ...
  - unlikely to be successful in practice
- Hyperparameter selection based on transfer.
  - Transfer of good hyperparameters between metrics and data sets does not seem to work as there appears to be no unsupervised way to distinguish between good and bad random seeds on the target task.

Are these disentangled representations useful for downstream tasks in terms of the sample complexity of learning?

- recover the true factors of variations from the learned representation using either multi-class logistic regression (LR) or gradient boosted trees (GBT).
- the lack of concrete examples of useful disentangled representations necessitates that future work on disentanglement methods should make this point more explicit.