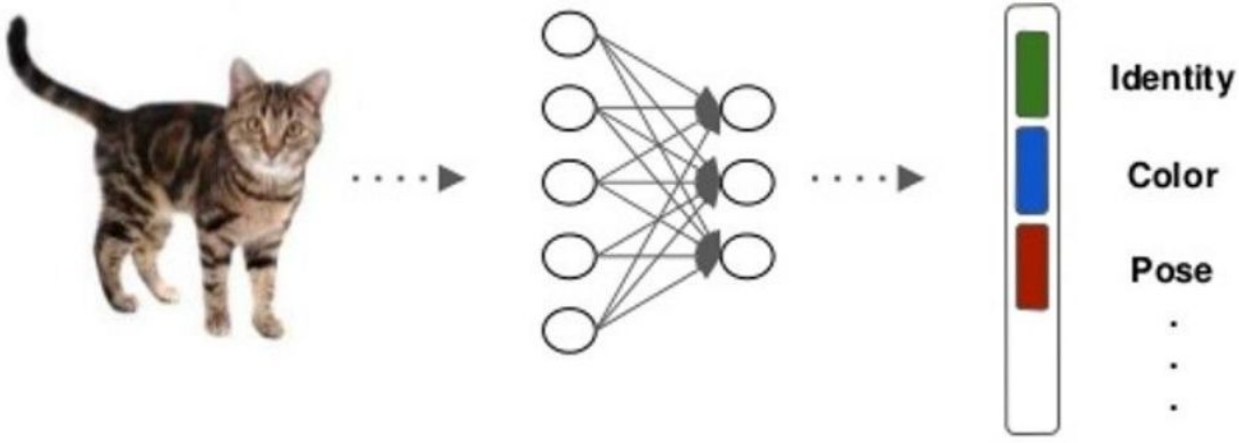
Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations

Tabisheva Anastasia

Disentangled representations

A set of latent components is said to be **disentangled** when each component is relatively sensitive to changes in a single aspect of the representations while being insensitive to the others.



Key contributions

- The unsupervised learning of disentangled representations is fundamentally impossible without inductive biases both on the considered learning approaches and the data sets
- 6 recent unsupervised disentanglement learning methods, 6 disentanglement measures from scratch and more than 12 000 models on 7 data sets
- **disentanglement_lib2** , a new library to train and evaluate disentangled representations

Key contributions

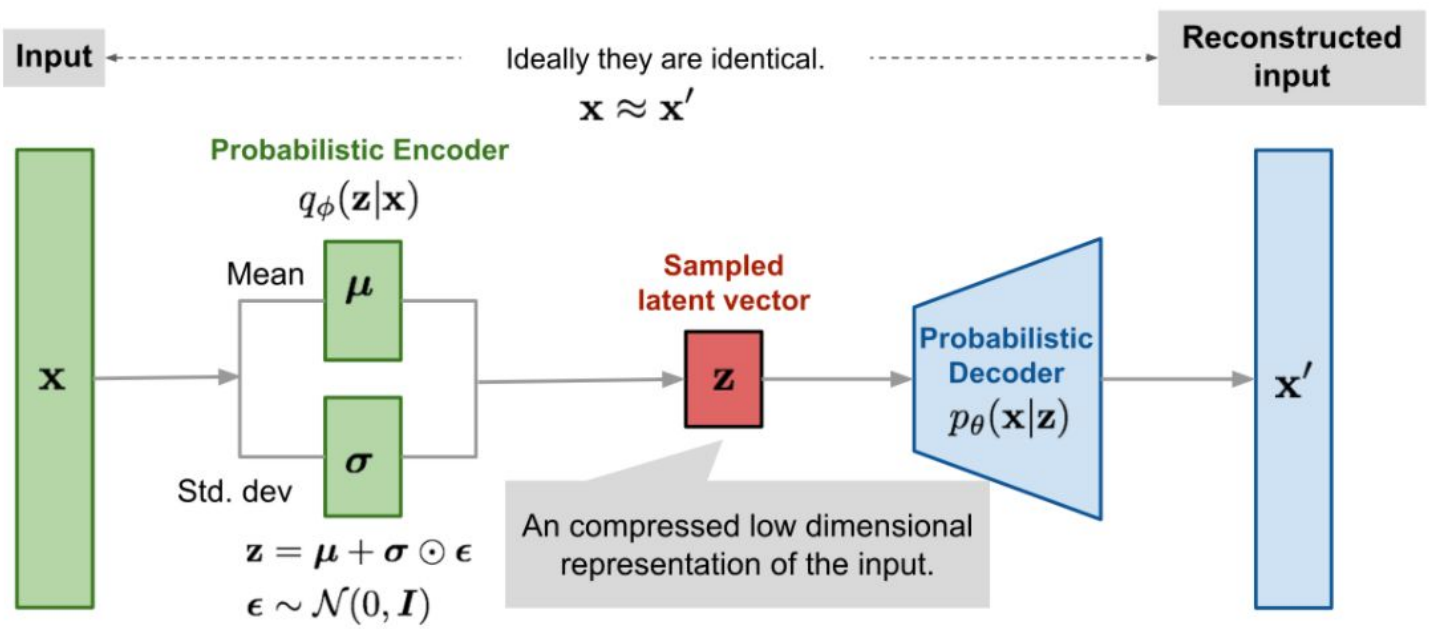
- Random seeds and hyperparameters seem to matter more than the model choice
- Can’t validate the assumption that disentanglement is useful for downstream tasks

Impossibility result

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 $f : \mathrm{supp}(\mathbf{z}) \to \mathrm{supp}(\mathbf{z})$ 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).*

Since the (unsupervised) disentanglement method only has access to observations x, it hence cannot distinguish between the two equivalent generative models and thus has to be entangled to at least one of them.

Variational Autoencoder Recap



$$\max_{\phi,\theta} \quad \mathbb{E}_{p(\mathbf{x})}[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))]$$

Experimental design. Considered methods

- **β-VAE** constrains the capacity of the VAE bottleneck

$$\mathbb{E}_{p(\mathbf{x})}[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))]$$

- **FactorVAE** penalizes the total correlation

$$\mathbb{E}_{p(\mathbf{x})}[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))] - \gamma D_{\mathrm{KL}}(q(\mathbf{z})\|\prod_{j=1}^{d} q(\mathbf{z}_j)).$$

Experimental design. Considered metrics

- **BetaVAE metric**

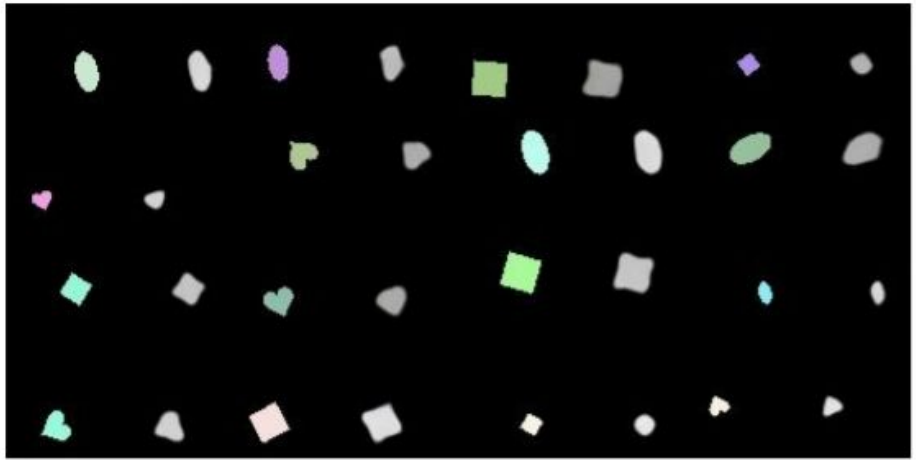
  Accuracy of a linear classifier that predicts the index of the fixed factor based on the coordinate-wise sum of absolute differences between the representation vectors in the two mini batches

- **Mutual Information Gap**

$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{H_{z_k}}\left(I(v_{j_k},z_k)-\max_{j\neq j_k}I(v_j,z_k)\right) \qquad j_k=\arg\max_j I(v_j,z_k).$$

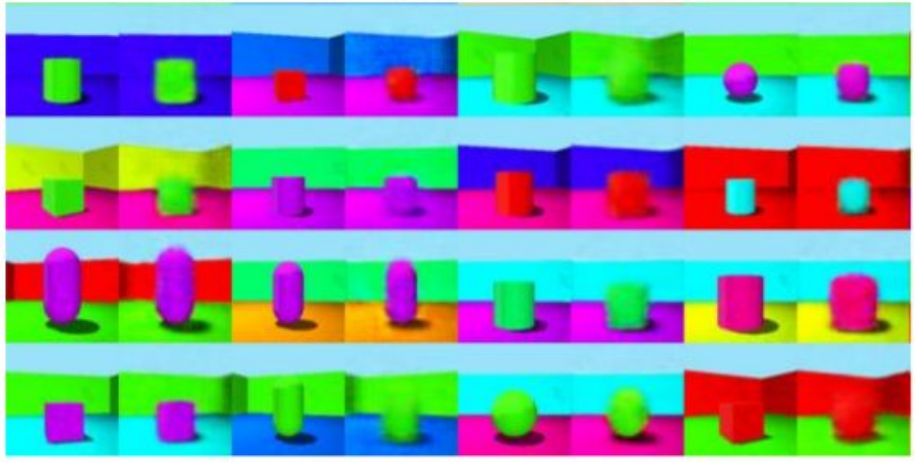
Experimental design. Inductive biases

Encoder	Decoder
Input: $64 \times 64 \times$ number of channels	Input: $\mathbb{R}^{10}$
4×4 conv, 32 ReLU, stride 2	FC, 256 ReLU
4×4 conv, 32 ReLU, stride 2	FC, $4 \times 4 \times 64$ ReLU
4×4 conv, 64 ReLU, stride 2	4×4 upconv, 64 ReLU, stride 2
4×4 conv, 64 ReLU, stride 2	4×4 upconv, 32 ReLU, stride 2
FC 256, F2 2×10	4×4 upconv, 32 ReLU, stride 2
	4×4 upconv, number of channels, stride 2

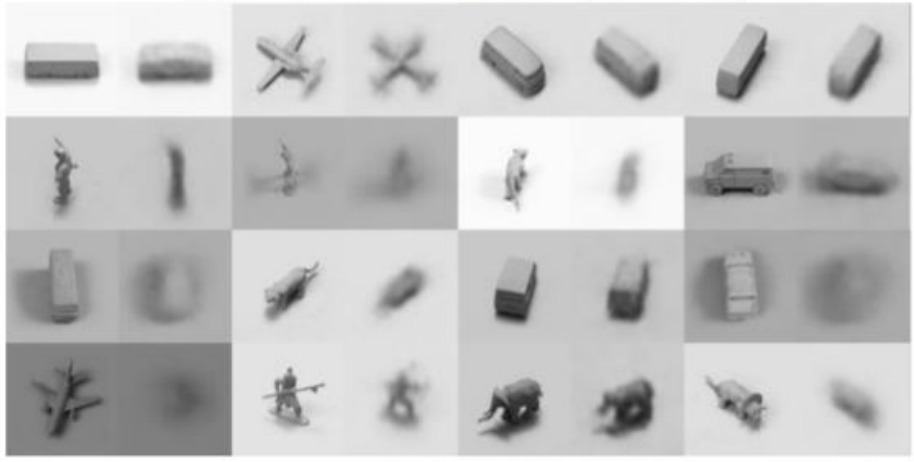


(c) FactorVAE trained on Color-dSprites.

(d) FactorVAE trained on Scream-dSprites.



(e) AnneaeledVAE trained on Shapes3D.



(f) β-TCVAE trained on SmallNORB.

Key experimental results

How important are different models and hyperparameters for disentanglement?

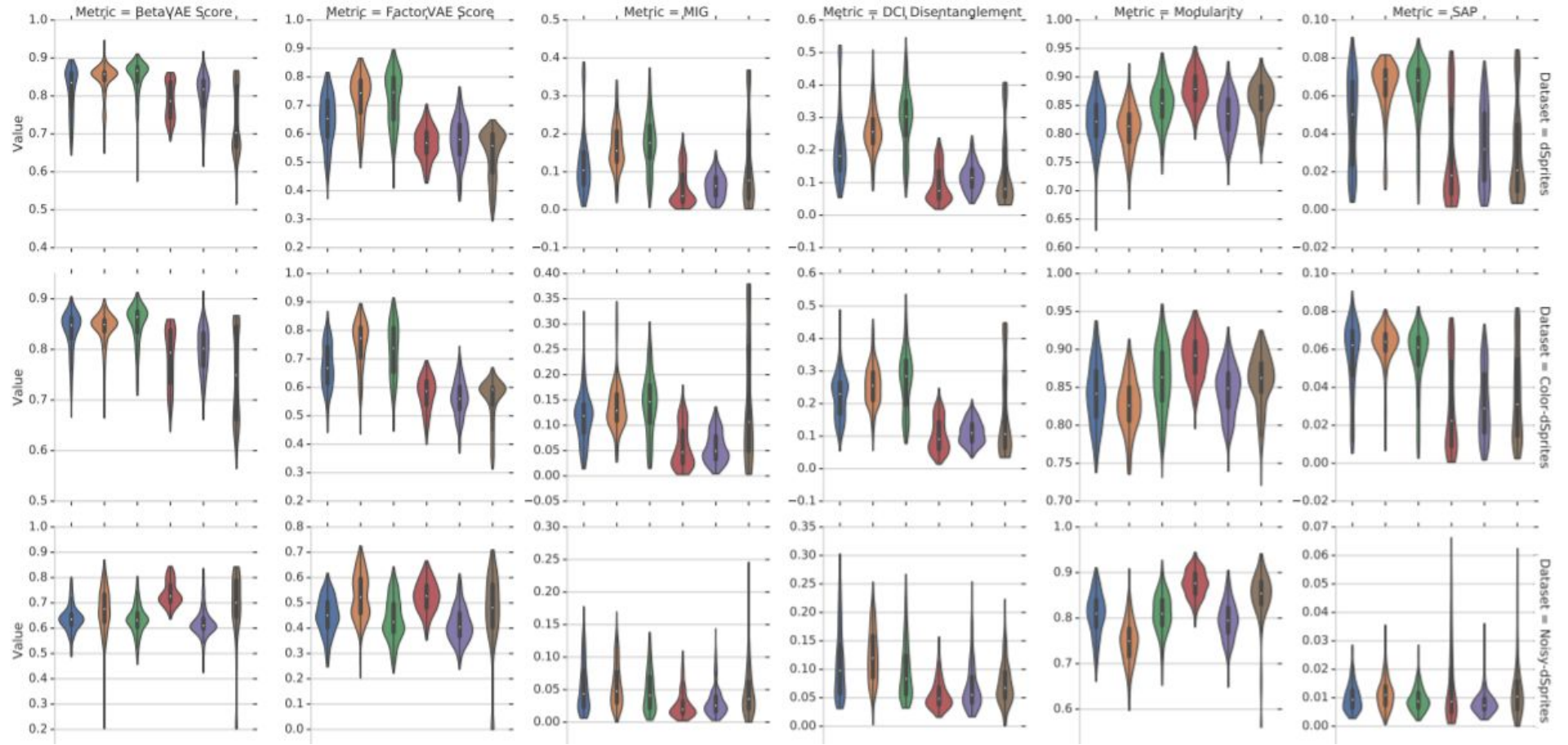
The disentanglement scores of unsupervised models are heavily influenced by randomness (in the form of the random seed) and the choice of the hyperparameter (in the form of the regularization strength). The objective function appears to have less impact.

Key experimental results

Are there reliable recipes for model selection?

- General recipes for hyperparameter selection
- Model selection based on unsupervised scores
- Hyperparameter selection based on transfer

	Random data set	Same data set
Random metric	54.9%	62.6%
Same metric	59.3%	80.7%


Metric = BetaVAE Score
Metric = FactorVAE Score
Metric = MIG
Metric = DCI Disentanglement
Metric = Modularity
Metric = SAP
Value
Dataset = dSprites
Dataset = Color-dSprites
Dataset = Noisy-dSprites

Key experimental results

Are there reliable recipes for model selection?

Unsupervised model selection remains an unsolved problem. 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.

Key experimental results

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

There is no evidence that models with higher disentanglement scores also lead to higher statistical efficiency.

Future research

- Unsupervised model selection persists as a key question
- The concrete practical benefits of enforcing a specific notion of disentanglement of the learned representations should be demonstrated
- Experiments should be reproducible on data sets of varying degrees of difficulty

Questions

1. Дайте определение распутанных скрытых представлений
2. Почему b-VAE лучше подходит для распутывания скрытых представлений, чем обычный VAE? Что оптимизирует b-VAE?
3. Какие основные выводы сделали авторы из своего исследования?

References

1. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations