On the DCI Framework for Evaluating Disentangled Representations: Extensions and Connections to Identifiability

Cian Eastwood*[1,2], Andrei Liviu Nicolicioiu*[1], Julius von Kügelgen*[1,3], Armin Kekic[1], Frederik Träuble[1], Andrea Dittadi[1,4], Bernhard Schölkopf[1]

MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS
THE UNIVERSITY of EDINBURGH
UNIVERSITY OF CAMBRIDGE
DTU Denmark Technical University

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany; [2]University of Edinburgh, United Kingdom; [3]University of Cambridge, United Kingdom; [4]Technical University of Denmark, Copenhagen, Denmark.

Summary

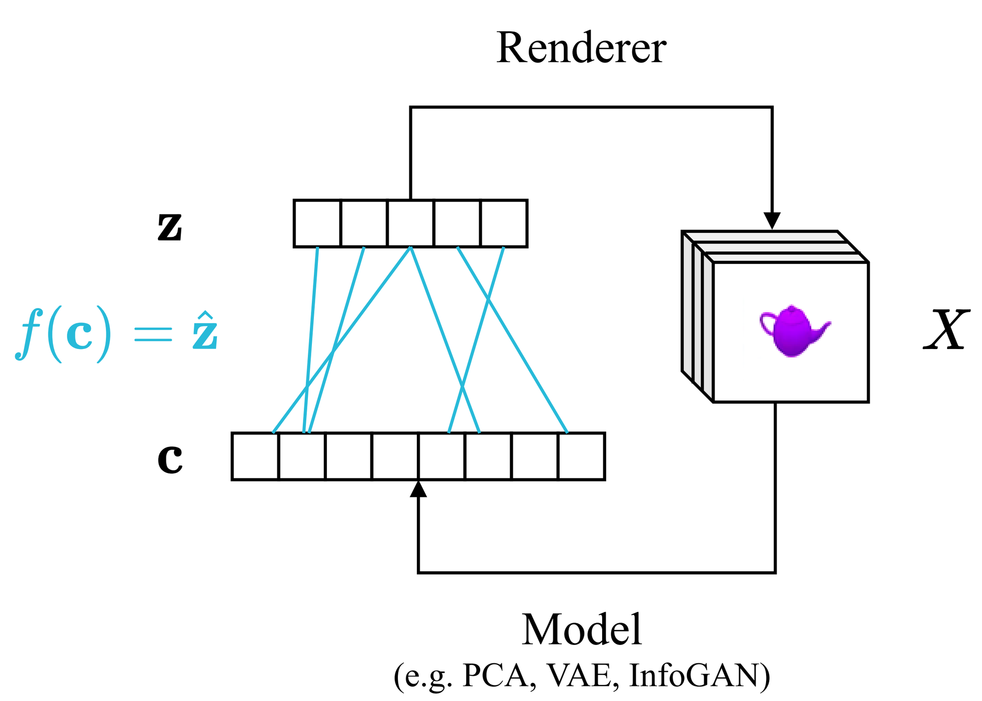
- We connect the DCI scores of Eastwood and Williams (2018, [1]) to two common notions of linear and nonlinear identifiability.
- We introduce a new complementary notion of disentanglement based on *the functional capacity required to use a representation*.

Notation & Background

- Data-generating factors: $\boldsymbol{z} \in \mathbb{R}^K$
- Observations: $\boldsymbol{x} = g(\boldsymbol{z}) \in \mathbb{R}^D$
- Representation or *code*: $\boldsymbol{c} = r(\boldsymbol{x}) \in \mathbb{R}^L$

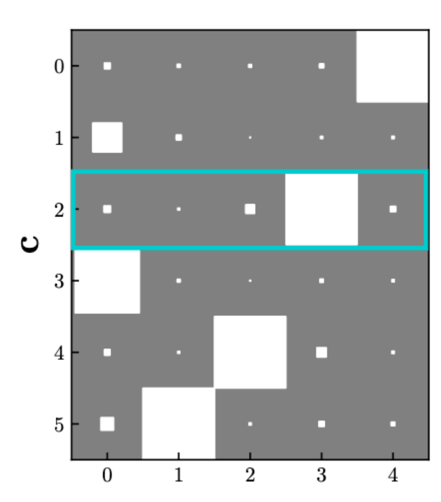
The DCI framework [1] quantitatively evaluates a representation or *code* $\boldsymbol{c}$ by:

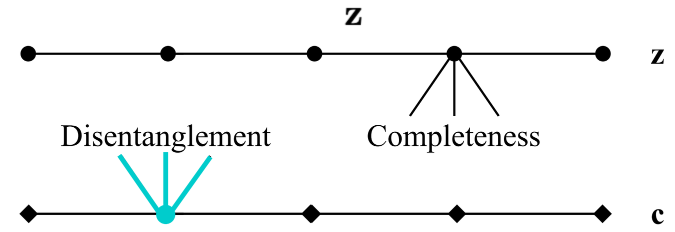
1. Training a probe f to predict $\boldsymbol{z}$ from $\boldsymbol{c}$, i.e., $\hat{\boldsymbol{z}} = f(\boldsymbol{c}) = f(r(\boldsymbol{x})) = f(r(g(\boldsymbol{z})))$;
2. Quantifying f's prediction error and deviation from an ideal one-to-one mapping.



Definition. $R \in \mathbb{R}^{L \times K}$ is called a *matrix of relative importances* of $\boldsymbol{c}$ for predicting $\boldsymbol{z}$ via $\hat{\boldsymbol{z}} = f(\boldsymbol{c})$ if R_{ij} captures some notion of the contribution of c_i to predicting z_j such that for all i, j: $R_{ij} \geq 0$ and $\sum_{i=1}^{L} R_{ij} = 1$.

- **Disentanglement (D)**
  - Degree to which c_i captures a single z_j
  - $D_i = 1 - H(\text{row 'distribution'})$
  - Higher is better: $[0, 1]$
- **Completeness (C)**
  - Degree to which z_j is captured by a single c_i
  - $C_j = 1 - H(\text{col. 'distribution'})$
  - Higher is better: $[0, 1]$
- **Informativeness (I)**
  - $I_j = 1 - \mathbb{E}[\ell(z_j, f_j(\boldsymbol{c}))]$
  - Higher is better: $[-\infty, 1]$





Connection to Identifiability

- Learning a data representation that recovers the underlying independent data-generating factors is closely related to blind source separation and widely-studied in independent component analysis (ICA) [2, 3, 4].
- Whether this goal is achieved up to acceptable ambiguities, subject to certain assumptions on the data-generating process, is typically formalised using the notion of *identifiability*.

Definition. Let $K = L$. We say that $\boldsymbol{c} = r(\boldsymbol{x}) = r(g(\boldsymbol{z}))$ identifies $\boldsymbol{z}$ up to

- *sign and permutation* if $\boldsymbol{c} = P\boldsymbol{z}$ for some signed permutation matrix P;
- *permutation and element-wise reparametrisation* if $\exists$ permutation π of $\{1, \dots, K\}$ and invertible scalar-functions $\{h_k\}_{k=1}^K$ s.t. $\forall j : c_j = h_j(z_{\pi(j)})$.

Proposition. If $D = C = 1$, then R is a permutation matrix.

Corollary. If additionally $\boldsymbol{z} = W^\top \boldsymbol{c}$ and $R = |W|$, then $\boldsymbol{c}$ identifies $\boldsymbol{z}$ up to permutation and sign.

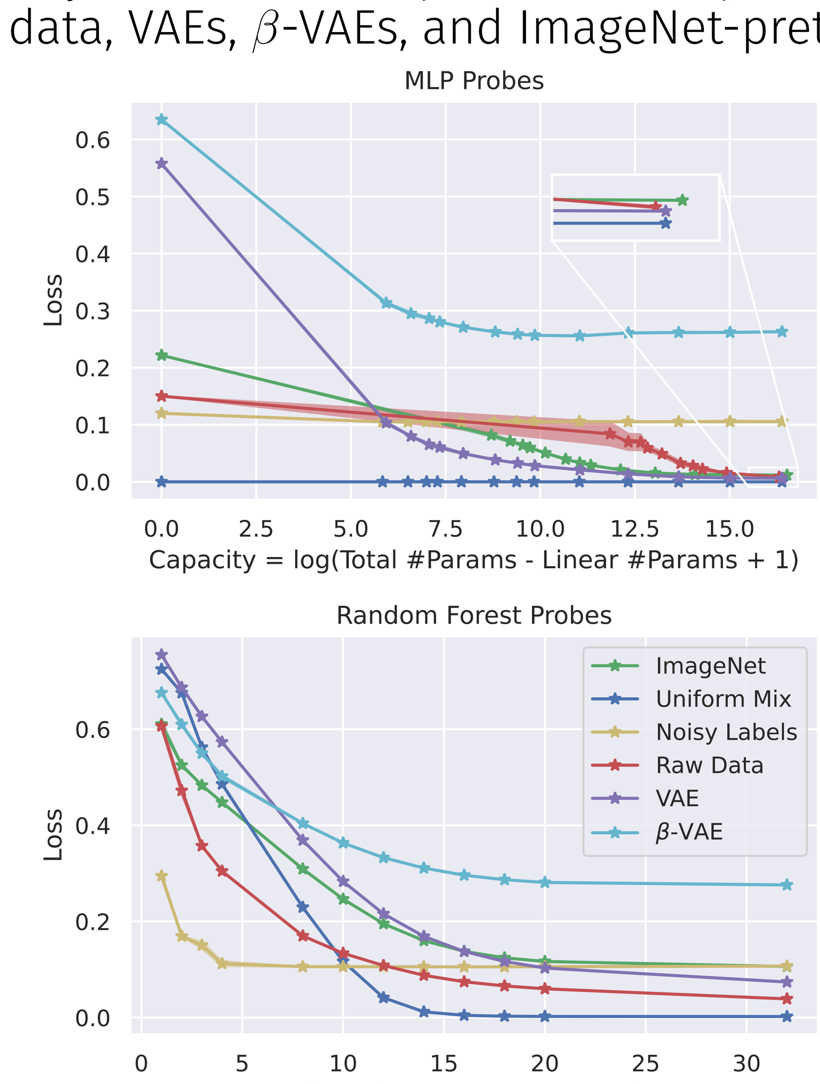
Corollary. Let $\boldsymbol{z} = f(\boldsymbol{c})$ with f an invertible function. If $D = C = 1$ and the feature importance matrix R satisfies $R_{ij} = 0 \iff \|\partial_i f_j\|_2 = 0$, then $\boldsymbol{c}$ identifies $\boldsymbol{z}$ up to permutation and element-wise reparametrisation.

Note. $\Longleftarrow$ holds for most feature importance measures, but $\Longrightarrow$ generally does not: for measures of *average* performance, a feature may not contribute on average, but still be used (sometimes helping, sometimes hurting).

Experiments

Dataset: `MPI3D-Real`. **Probes:** MLPs and Random Forest (RF).

Representations: synthetic (Noisy lebels and uniforn miximg of labels), raw data, VAEs, β-VAEs, and ImageNet-pretrained ResNet18.



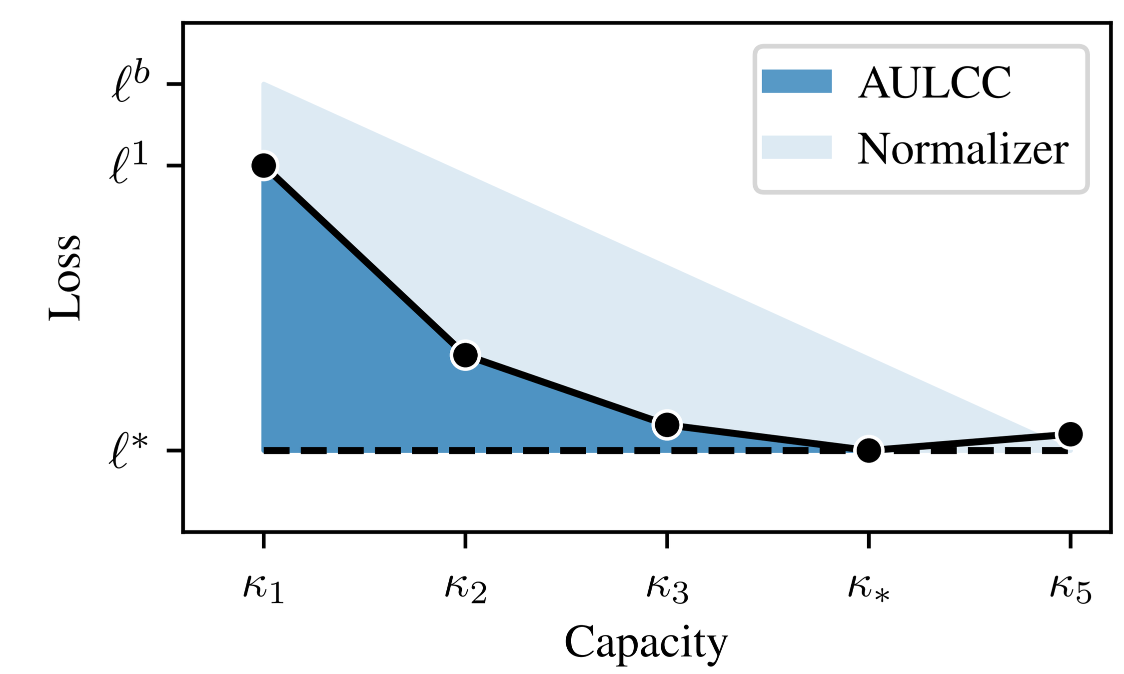
Representation	Probe	D	C	I	E	S
GT Labels $\boldsymbol{z}$	MLP*	1	1	1	1	1
Noisy labels	MLP*	1	1	0.9	1	1.0
	MLP	0.97	0.97	0.89	0.99	1.0
	RF	0.75	0.76	0.89	0.98	1.0
Uniform mix	MLP*	0	0	1	1	1.0
	MLP	0.13	0.22	1.0	1.0	1.0
	RF	0.17	0.21	1.0	0.72	1.0
VAE	MLP	0.15	0.14	0.99	0.71	0.7
	RF	0.10	0.10	0.93	0.65	0.7
β-VAE	MLP	0.26	0.38	0.74	0.81	0.7
	RF	0.22	0.25	0.72	0.85	0.7
ImgNet-pretr	MLP	0.16	0.10	0.99	0.82	0.01
	RF	0.35	0.20	0.89	0.78	0.01
Raw data	MLP	0.22	0.16	0.99	0.82	0.001
	RF	0.84	0.41	0.96	0.80	0.001

The Extended DCI-ES Framework

Probe-agnostic feature importances. D and C scores can be computed for arbitrary black-box probes f (e.g. MLPs) by using *predictor-agnostic* feature importance measures (e.g. SAGE [5]).

Explicitness (E):

- **Main idea:** the functional capacity required to recover $\boldsymbol{z}$ from $\boldsymbol{c}$ is an important but under-explored aspect of evaluating representations, e.g. recovering $\boldsymbol{z}$ from noisy observations $\boldsymbol{z}'$ is "easy" (low/linear capacity), but doing so from images is "hard" (high/nonlin. cap.).
- **Definition:** The ease-of-use or *explicitness* of a representation $\boldsymbol{c}$ for predicting z_j is quantified by the (normalized) *Area Under its Loss-Capacity Curve* (AULCC), which displays test loss against probe capacity.
- **Intuition:** large area means $\boldsymbol{c}$ was *hard-to-use* (required high capacity).



Size (S):

- **Motivation:** Increased representation size often improves other scores like I and E, so we report a measure of size to allow an analysis of the size-informativeness or size-explicitness trade-off.
- **Definition:**

$$S = \frac{K}{L} = \frac{\dim(\boldsymbol{z})}{\dim(\boldsymbol{c})}.$$

References

[1] Cian Eastwood and Christopher K I Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

[2] Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

[3] Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

[4] Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019.

[5] Ian Covert, Scott M Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33:17212–17223, 2020.