Towards Robust Metrics For Concept Representation Evaluation

Mateo Espinosa Zarlenga*, **Pietro Barbiero***, **Zohreh Shams***, Dmitry Kazhdan, Umang Bhatt, Adrian Weller, Mateja Jamnik

Overview

- Concept-based explainability increases trustworthiness and model transparency by conditioning tasks on high-level units of information, or concepts (e.g. “has whiskers”).
- Concept learning models have been shown to be prone to **encoding impurities** in their representations. However, **appropriate metrics to measure such phenomena are lacking**.
- We propose **novel metrics** for evaluating the purity of concept representations and show their utility in evaluating the robustness of concept representations and benchmarking SOTA methods from concept learning (CL) and the related field of disentanglement learning (DGL).

Why do we need new metrics?

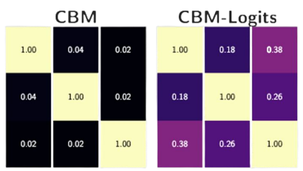

Figure 1: Existing metrics cannot capture impurities that have been empirically found in concept learning methods such as CBMs [1].

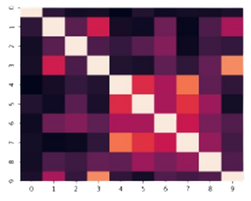

Figure 2: The above phenomena is mostly due to the fact that existing metrics assume that concepts are disentangled, an unrealistic assumption in the real world, as evidenced by the correlation of concepts in the CUB dataset [2].

	OIS (↓)	NIS (↓)	SAP (↑)	MIG (↑)	R^4 (↑)	DCI Dis (↑)
Baseline Soft (%)	**4.69 ± 0.43**	**66.25 ± 2.31**	48.74 ± 0.41	99.93 ± 0.03	99.95 ± 0.00	**99.99 ± 0.00**
Impure Soft (%)	22.58 ± 2.34	72.36 ± 1.26	48.83 ± 0.53	99.93 ± 0.04	99.95 ± 0.00	99.50 ± 0.01
p-value	7.38×10^{-5}	3.24×10^{-3}	7.89×10^{-1}	9.26×10^{-1}	9.76×10^{-1}	3.66×10^{-9}

Table 1: Existing metrics in the disentanglement learning literature are unable to discriminate between concept representations which are pure (baseline) and impure.

Measuring Concept Quality as Concept Impurity

- We measure a concept representation’s quality by quantifying **how much ”unnecessary”** information it encodes about a set of ground-truth concepts.
- **We let “unnecessary” information** be any information encoded in a learnt concept **beyond what is observed in its respective ground-truth concept.**
- The proposed metrics capture both **localized and distributed impurities.**

Oracle Impurity Score (OIS)

Measures impurities localized within a **single learnt concept** by:

- measuring how accurately we can predict a concept from a learnt concept representation, and
- evaluating how this diverges from what we expect from their corresponding ground-truth concepts.

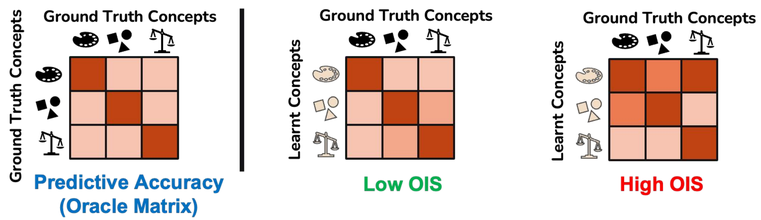

Figure 3: Low and high degree of divergence from ground truth concept oracle matrix constitutes “Low OIS” and “High OIS”, respectively.

Niche Impurity Score (NIS)

Measures impurities distributed across **multiple learnt concepts** by:

- looking at subsets of concept representations which are uncorrelated with a ground-truth concept, and
- measuring how predictive such a set is for the ground truth concept they are uncorrelated with.

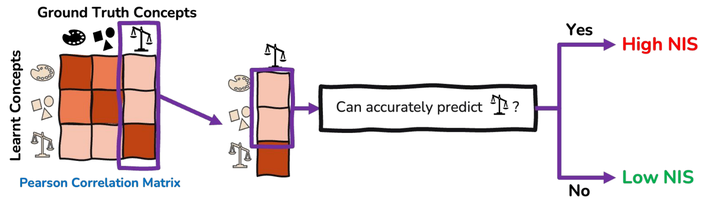

Figure 4: Diagram showing when the NIS can be high and low for a specific ground-truth concept.

Detecting Learnt Spurious Correlations

Our metrics can be used to detect impurities encoded in learnt concept representations due to spurious correlations in the training data.

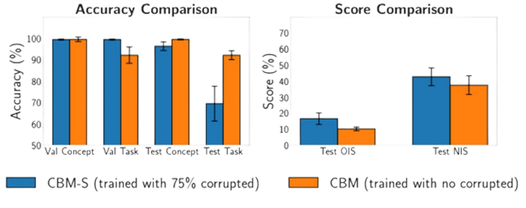

Figure 5: Impurity scores and validation/testing accuracies for CBMs trained on a spuriously corrupted dSprite task (CBM-S) and an uncorrupted task (CBM). The model trained on corrupted samples has significantly higher impurities, explaining why its test-time task performance is low despite its high test concept accuracy.

Effects of Impurities in Concept Interventions

Impurities can affect how test-time concept interventions fare in concept bottleneck models.

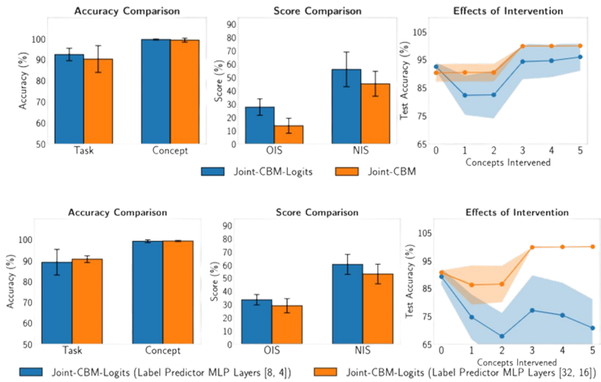

Figure 6: Intervention performance in two CBMs trained on dSprites when (a) one uses sigmoidal concept representations and the other does not and (b) when the models use different capacities .

Benchmarking Concept Learning Methods

We evaluate our metrics on concept representations learnt by a wide variety of methods and show that concept supervision alone may not be sufficient to ensure concept purity.

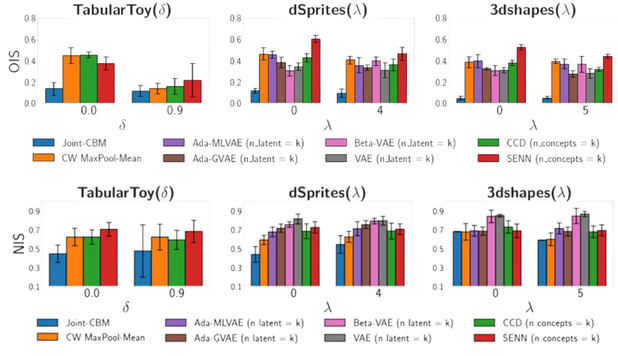

Figure 7: OIS and NIS scores for a variety of concept learning and disentanglement learning methods across multiple datasets with varying degrees of intra-concept correlations (controlled by their respective parameters in the x-axis).

References

[1] Koh, Pang Wei, et al. "Concept bottleneck models." *International Conference on Machine Learning*. PMLR, 2020.

[2] Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Code + Paper