Towards Robust Metrics For Concept Representation Evaluation

Mateo Espinosa Zarlenga[1*], Pietro Barbiero[1*], Zohreh Shams[1, 2*], Dmitry Kazhdan[1], Umang Bhatt[1, 3], Adrian Weller[1, 3], Mateja Jamnik[1]

[1] University of Cambridge
[2] Babylon Health
[3] The Alan Turing Institute
{me466, pb737, zs315, dk525, usb20, aw665}@cam.ac.uk, mateja.jamnik@cl.cam.ac.uk

Abstract

Recent work on interpretability has focused on concept-based explanations, where deep learning models are explained in terms of high-level units of information, referred to as concepts. Concept learning models, however, have been shown to be prone to encoding impurities in their representations, failing to fully capture meaningful features of their inputs. While concept learning lacks metrics to measure such phenomena, the field of disentanglement learning has explored the related notion of underlying factors of variation in the data, with plenty of metrics to measure the purity of such factors. In this paper, we show that such metrics are not appropriate for concept learning and propose novel metrics for evaluating the purity of concept representations in both approaches. We show the advantage of these metrics over existing ones and demonstrate their utility in evaluating the robustness of concept representations and interventions performed on them. In addition, we show their utility for benchmarking state-of-the-art methods from both families and find that, contrary to common assumptions, supervision alone may not be sufficient for pure concept representations.

Introduction

Addressing the lack of interpretability of deep neural networks (DNNs) has given rise to explainability methods, most common of which are feature importance methods (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017) that quantify the contribution of input features to certain predictions (Bhatt et al. 2020). However, input features may not necessarily form the most intuitive basis for explanations, in particular when using low-level features such as pixels. *Concept-based explainability* (Kim et al. 2018; Ghorbani et al. 2019; Koh et al. 2020; Yeh et al. 2020; Ciravegna et al. 2021) remedies this issue by constructing an explanation at a concept level, where concepts are considered intermediate, high-level and semantically meaningful units of information commonly used by humans to explain their decisions. Recent work, however, has shown that concept learning (CL) models may not correctly capture the intended semantics of their representations (Margeloiu et al. 2021), and that their learnt concept representations are prone to encoding *impurities* (i.e., more information in a concept than what is intended) (Mahinpei et al. 2021). Such phenomena may have severe consequences for how such representations can be interpreted (as shown in the misleading attribution maps described by Margeloiu et al. (2021)) and used in practice (as shown later in our intervention results). Nevertheless, the CL literature is yet to see concrete metrics to appropriately capture and measure these phenomena.

In contrast, the closely-related field of *disentanglement learning* (DGL) (Bengio, Courville, and Vincent 2013; Higgins et al. 2017; Locatello et al. 2019, 2020b), where methods aim to learn intermediate representations aligned to disentangled factors of variation in the data, offers a wide array of metrics for evaluating the quality of latent representations. However, despite the close relationship between concept representations in CL and latent codes in DGL, metrics proposed in DGL are built on assumptions that do not hold in CL, as explained in our Background and Motivation Section, and are thus inappropriate to measure the aforementioned undesired phenomena in CL.

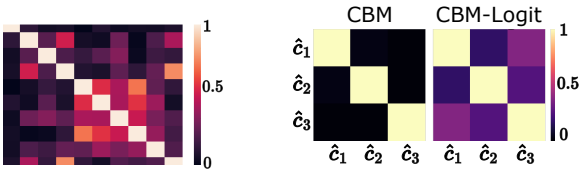
In this paper, we show the inadequacy of current metrics and introduce two novel metrics for evaluating the purity of intermediate representations in CL. Our results indicate that our metrics can be used in practice for quality assurance of such intermediate representations for:

1. Detecting impurities (i) concealed in soft representations, (ii) caused by different model capacities, or (iii) caused by spurious correlations.
2. Indicating when concept interventions are safe.
3. Revealing the impact of supervisions on concept purity.
4. Being robust to inter-concept correlations.

Background and Motivation

Notation In CL, the aim is to find a low-dimensional intermediate representation $\hat{\mathbf{c}}$ of the data, similar to latent codes $\hat{\mathbf{z}}$ in DGL. This low-dimensional representation corresponds to a matrix $\hat{\mathbf{c}} \in \hat{C} \subseteq \mathbb{R}^{d \times k}$ in which the i-th column constitutes a d-dimensional representation of the i-th concept, assuming that the length of all concept representations can be made equal using zero-padding. Under this view, elements in $\hat{\mathbf{c}}_{(:,i)} \in \mathbb{R}^d$ are expected to have high values (under some reasonable aggregation function) if the i-th concept is considered to be activated. As most CL methods assume $d = 1$, for succinctness we use $\hat{c}_i$ in place of $\hat{\mathbf{c}}_{(:,i)}$ when $d = 1$. Analogously, as each latent code $\hat{\mathbf{z}}_{(:,i)}$ aims to encode an

*These authors contributed equally.



(a) (b)

Figure 1: (a) Absolute correlation of the top-10 concepts with the highest label mutual information in CUB dataset. (b) Two identically-trained CBMs with almost identical accuracies yet different levels of inter-concept correlations.

independent factor of variation (i.e., concept) z_i in DGL, we use $\hat{\mathbf{c}}$ for both learnt concept representations and latent codes. Similarly, we refer to both ground truth concepts and factors of variations as $\mathbf{c} \in C \subseteq \mathbb{R}^k$.

Concept Learning In supervised CL, access to concept labels for each input, in addition to task labels, is assumed. Supervised CL makes use of (i) a concept encoder function $g : X \mapsto \hat{C}$ that maps the inputs to a concept representation; and (ii) a label predictor function $f : \hat{C} \mapsto Y$ that maps the concept representations to a downstream task's set of labels $\mathbf{y} \in Y \subseteq \mathbb{R}^L$. Together, these two functions constitute a *Concept Bottleneck Model* (CBM) (Koh et al. 2020). A notable approach that uses the bottleneck idea is *Concept Whitening* (CW) (Chen, Bei, and Rudin 2020) which introduces a batch normalisation module whose activations are trained to be aligned with sets of binary concepts. Unlike supervised CL, in unsupervised CL concept annotations are not available and concepts are discovered in an unsupervised manner with the help of task labels. Two notable data modality agnostic approaches in this family are *Completeness-aware Concept Discovery* (CCD) (Yeh et al. 2020) and *Self-Explainable Neural Networks* (SENNs) (Alvarez-Melis and Jaakkola 2018). We refer to supervision from ground truth concepts in supervised CL as *explicit*, while supervision from task labels alone is referred to as *implicit*.

On the other hand, generative models (e.g., VAEs (Kingma and Welling 2014)) used in DGL assume that data is generated from a set of independent factors of variation $\mathbf{c} \in C$. Thus, the goal is to find a function $g(\cdot)$ that maps inputs to a disentangled latent representation. In the light of recent work (Locatello et al. 2019) showing the impossibility of learning disentangled representations without any supervision, as in β-VAEs (Higgins et al. 2017), recent work suggests using *weak* supervision for learning latent codes (Locatello et al. 2020a).

Shortcomings of Current Metrics Generally, the DGL literature defines concept quality in terms of disentanglement i.e., the more learnt concepts are decorrelated the better (see Appendix A.1 for a summary of DGL metrics)[1]. We argue that existing DGL metrics are inadequate to ensure concept quality in CL as they: (i) Assume that each concept is represented with a single scalar value, which is not the case in some modern CL methods such as CW. (ii) Fail to capture subtle impurities encoded within continuous representations (as demonstrated in our Experimental Section). (iii) May assume access to a tractable concept-to-sample generative process (something uncommon in real-world datasets). (iv) Assume that inter-concept dependencies are undesired, an assumption that may not be realistic in the real world where ground truth concept labels often are correlated. This can be observed in Figure 1a where concept labels in the Caltech-UCSD Birds-200-2011 (CUB) dataset (Wah et al. 2011), a widely used CL benchmark, are seen to be *highly correlated*.

Metrics in CL (Yeh et al. 2020; Kazhdan et al. 2020), on the other hand, mainly define concept quality w.r.t. the downstream task (e.g., *task predictive accuracy*), and rarely evaluate properties of concept representations w.r.t. the ground truth concepts (except for *concept predictive accuracy*). Nevertheless, two CL models can learn concept representations that yield similar task and concept accuracies but have vastly different properties/qualities. For example, Figure 1b shows a toy experiment in which two CBMs trained on a dataset with 3 independent concepts, where "*CBM*" uses a sigmoidal bottleneck and "*CBM-Logits*" uses logits in its bottleneck, generate concept representations with the same concept/task accuracies yet significantly different inter-concept correlations (details in Appendix A.2).

Measuring Purity of Concept Representations

To address the shortcomings of existing metrics, we propose two metrics that make no assumptions about (i) correlations between concepts, (ii) the underlying data-generating process, and (iii) the dimensionality of a concept's representation. Specifically, we focus on measuring the quality of a concept representation in terms of their "*purity*", defined here as whether the predictive power of a learnt concept representation over other concepts is similar to what we would expect from their corresponding ground truth labels. We begin by introducing the **oracle impurity score** (OIS), a metric that quantifies impurities *localised* within individual learnt concepts. Then, we introduce the **niche impurity score** (NIP) as a metric that focuses on capturing impurities *distributed* across the set of learnt concept representations.

Oracle Impurity

To circumvent the aforementioned limitations of existing DGL metrics, we take inspiration from (Mahinpei et al. 2021), where they informally measure concept impurity as how predictive a CBM-generated concept probability is for the ground truth value of other independent concepts. If the pre-defined concepts are independent, then the inter-concept predictive performance should be no better than random. To generalise this assumption beyond independent concepts, we first measure the predictability of ground truth concepts w.r.t. one another. Then we measure the predictability of learnt concepts w.r.t. the ground truth ones. The divergence between the former and the latter acts as an impurity metric, measuring the amount of undesired information that is encoded, or lacking, in the learnt concepts. To formally introduce our metric, we begin by defining a *purity matrix*.

[1] Appendices can be found in https://arxiv.org/abs/2301.10367.

Definition 0.1 (Purity Matrix)**.** Given a set of n concept representations $\hat{\Gamma} = \{\hat{\mathbf{c}}^{(l)} \in \mathbb{R}^{d\times k}\}_{l=1}^{n}$, and corresponding discrete ground truth concept annotations $\Gamma = \{\mathbf{c}^{(l)} \in \mathbb{N}^{k}\}_{l=1}^{n}$, assume that $\hat{\Gamma}$ and Γ are aligned element-wise: for all $i \in \{1, \cdots, k\}$, the i-th concept representation of $\hat{\mathbf{c}}^{(l)}$ encodes the same concept as the i-th concept label in $\mathbf{c}^{(l)}$. The Purity Matrix of $\hat{\Gamma}$, given ground truth labels Γ, is defined as a matrix $\pi\big(\hat{\Gamma}, \Gamma\big) \in [0,1]^{k\times k}$ whose entries are given by:

$$\pi\big(\hat{\Gamma}, \Gamma\big)_{(i,j)} := \text{AUC-ROC}\big(\big\{\big(\psi_{i,j}(\hat{\mathbf{c}}^{(l)}_{(:,i)}), c^{(l)}_{j}\big)\big\}_{l=1}^{n}\big),$$

where $\psi_{i,j}(\cdot)$ is a non-linear model (e.g., an MLP) mapping the i-th concept's representation $\hat{\mathbf{c}}_{(:,i)}$ to a probability distribution over all values concept j may take.

The (i,j)-th entry of $\pi\big(\hat{\Gamma}, \Gamma\big)$ contains the AUC-ROC when predicting the ground truth value of concept j given the i-th concept representation. Therefore, the diagonal entries of this matrix show how good a concept representation is at predicting its aligned ground truth label, while the off-diagonal entries show how good such a representation is at predicting the ground truth labels of other concepts. Intuitively, one can think of the (i,j)-th entry of this matrix as a proxy of the mutual information between the i-th concept representation and the j-th ground truth concept. While in principle other methods could be used to estimate this mutual information (e.g., histogram binning), we choose the test AUC-ROC of a trained non-linear model primarily for its tractability, its bounded nature, and its easy generalisation to non-scalar concept representations. Furthermore, while in this work we focus on binary concepts, our definition can be applied to multivariate concepts by using the mean one-vs-all AUC-ROC score. See Appendix A.3 for implementation details and Appendix A.4 for a discussion on how the OIS is robust to the model selected for $\psi_{i,j}(\cdot)$.

This matrix allows us to construct a metric for quantifying the impurity of a concept encoder:

Definition 0.2 (Oracle Impurity Score (OIS))**.** Let $g : X \mapsto \hat{C} \subseteq \mathbb{R}^{d\times k}$ be a concept encoder and let $\Gamma_X := \{\mathbf{x}^{(i)} \in X\}_{i=1}^{n}$ and $\Gamma := \{\mathbf{c}^{(i)} \in \mathbb{N}^{k}\}_{i=1}^{n}$ be ordered sets of testing samples and their corresponding concept annotations, respectively. If, for any ordered set A we define $g(A)$ as $g(A) := \{g(a) \mid a \in A\}$, then the OIS is defined as:

$$\text{OIS}(g, \Gamma_X, \Gamma) := \frac{2\Big|\Big|\pi\big(g(\Gamma_X), \Gamma\big) - \pi\big(\Gamma, \Gamma\big)\Big|\Big|_F}{k}$$

where $||\mathbf{A}||_F$ represents the Frobenius norm of $\mathbf{A}$.

Intuitively, the OIS measures the total deviation of an encoder's purity matrix with the purity matrix obtained from using the ground truth concept labels only (i.e., the "oracle matrix"). We opt to measure this divergence using the Frobenius norm of their difference to obtain a bounded output which can be easily interpreted. Since each entry in the difference $\big(\pi\big(g(\Gamma_X), \Gamma\big) - \pi\big(\Gamma, \Gamma\big)\big)$ can be at most $1/2$, the upper bound of $\Big|\Big|\pi\big(g(\Gamma_X), \Gamma\big) - \pi\big(\Gamma, \Gamma\big)\Big|\Big|_F$ is $k/2$. Therefore, the OIS includes a factor of $2/k$ to guarantee that it ranges in $[0,1]$. This allows interpreting an OIS of 1 as a complete misalignment between $\pi\big(\Gamma, \Gamma\big)$ and $\pi\big(g(\Gamma_X), \Gamma\big)$ (i.e., the i-th concept representation can predict all other concept labels except its corresponding one even when concepts are independent). An impurity score of 0, on the other hand, represents a perfect alignment between the two purity matrices (i.e., the i-th concept representation does not encode any unnecessary information for predicting concept i).

Niche Impurity

While the OIS can correctly capture impurities that are localised within specific and individual concept representations, it is also possible that information pertinent to unrelated concepts is encoded *across multiple learnt representations*. To tractably capture such a phenomenon, we propose the *Niching Impurity Score* (NIS) inspired by the theory of niching. In ecology, a niche is considered to be a resource-constrained subspace of the environment that can support different types of life (Darwin 1859). Analogously, the NIS looks at the predictive power of subsets of disentangled concepts. In contrast with the OIS, the NIS is concerned with impurities encoded in *sets* of learnt concept representations rather than impurities in individual concept representations. The NIS efficiently quantifies the amount of shared information across concept representations by looking at how predictive disentangled subsets of concept representations are for ground truth concepts. We start by describing a *concept nicher*, a function that ranks learnt concepts by how much information they share with the ground truth ones. We then define a *concept niche* for a ground truth concept as a set of learnt concepts that are highly ranked by the concept nicher, while the set of concepts outside the niche is referred to as the *concept niche complement*. We conclude by constructing the NIS by looking at how predictable a ground truth concept is from its corresponding concept niche complement. The collective NIS of all concepts, therefore, represents impurities encoded across the entire bottleneck.

Definition 0.3 (Concept nicher)**.** Given a set of concept representations $\hat{C} \subseteq \mathbb{R}^{d\times k}$, we define a concept nicher as a function $\nu : \{1, \cdots k\} \times \{1, \cdots k\} \mapsto [0,1]$ that returns $\nu(i,j) \approx 1$ if the i-th concept $\hat{\mathbf{c}}_{(:,i)}$ is entangled with the j-th ground truth concept c_j, and $\nu(i,j) \approx 0$ otherwise.

Our definition above can be instantiated in various ways, depending on how entanglement is measured. In favour of efficiency, we measure entanglement using absolute Pearson correlation ρ, as this measure can efficiently discover (a linear form of) association between variables (Altman and Krzywinski 2015). We call this instantiation *concept-correlation nicher* (CCorrN) and define it as $\text{CCorrN}(i,j) := \big|\rho\big(\{\hat{\mathbf{c}}^{(l)}_{(:,i)}\}_{l=1}^{N}, \{c^{(l)}_{j}\}_{l=1}^{N}\big)\big|$.

If $\hat{\mathbf{c}}_{(:,i)}$ is not a scalar representation (i.e., $d > 1$), then for simplicity, we use the maximum absolute correlation coefficient between all entries in $\hat{\mathbf{c}}_{(:,i)}$, and the target concept label c_j as a representative correlation coefficient for the entire representation $\hat{\mathbf{c}}_{(:,i)}$. We then define a concept niche as:

Definition 0.4 (Concept niche)**.** The concept niche $N_j(\nu, \beta)$ for target concept j, determined by concept nicher $\nu(\cdot,\cdot)$ and

threshold $\beta \in [0, 1]$, is defined as $N_j(\nu, \beta) := \{i \mid i \in \{1, \cdots, k\} \text{ and } \nu(i, j) > \beta\}$.

From this, the Niche Impurity (NI) measures the predictive capacity of the complement of concept niche $N_i(\nu, \beta)$, referred to as $\neg N_i(\nu, \beta) := \{1, \cdots, k\} \setminus N_i(\nu, \beta)$, for the i-th ground truth concept:

Definition 0.5 (Niche Impurity (NI))**.** Given a classifier $f : \hat{C} \mapsto C$, concept nicher ν, threshold $\beta \in [0, 1]$, and labeled concept representations $\{(\hat{\mathbf{c}}^{(l)}, \mathbf{c}^{(l)})\}_{l=1}^{n}$, the Niche Impurity of the i-th output of $f(\cdot)$ is defined as $\text{NI}_i(f, \nu, \beta) := \text{AUC-ROC}\big(\{(f|_{\neg N_i(\nu,\beta)}(\hat{\mathbf{c}}^{(l)}_{(:,\neg N_i(\nu,\beta))}), c_i^{(l)})\}_{l=1}^{n}\big)$, where $f|_{\neg N_j(\nu,\beta)}$ is the classifier resulting from masking all entries in $\neg N_j(\nu, \beta)$ when feeding f with concept representations.

Although f can be any classifier, for simplicity in our experiments we use a ReLU MLP with hidden layer sizes $\{20, 20\}$ (see Appendix A.4 for a discussion on our metric's robustness to f's architecture). Intuitively, a NI of $1/2$ (random AUC of niche complement) indicates that the concepts inside the niche $N_i(\nu)$ are the only concepts predictive of the i-th concept, that is, concepts outside the niche do not hold any predictive information of the i-th concept. Finally, the *Niche Impurity Score* metric measures how much information apparently disentangled concepts are sharing:

Definition 0.6 (Niche Impurity Score (NIS))**.** Given a classifier $f : \hat{C} \mapsto C$ and concept nicher ν, the niche impurity score $\text{NIS}(f, \nu) \in [0, 1]$ is defined as the summation of niche impurities across all concepts for different values of β: $\text{NIS}(f, \nu) := \int_0^1 (\sum_{i=1}^{k} \text{NI}_i(f, \nu, \beta)/k) d\beta$.

In practice, this integral is estimated using the trapezoid method with $\beta \in \{0.0, 0.05, \cdots, 1\}$. Furthermore, we parameterise f as a small MLP, leading to a tractable impurity metric that scales with large concept sets. Intuitively, a NIS of 1 means that all the information to perfectly predict each ground truth concept is spread on many different and disentangled concept representations. In contrast, a NIS around $1/2$ (random AUC) indicates that no concept can be predicted by any concept representation subset.

Experiments

We now give a brief account of the experimental setup and datasets, followed by highlighting the utility of our impurity metrics and their applications to model benchmarking.

Datasets To have datasets compatible with both CL and DGL, we construct tasks whose samples are fully described by a vector of ground truth generative factors. Moreover, we simulate real-world scenarios by designing tasks with varying degrees of dependencies in their concept annotations. To achieve this, we first design a parametric binary-class dataset *TabularToy*(δ), a variation of the tabular dataset proposed by Mahinpei et al. (2021). We also construct two multiclass image-based parametric datasets: *dSprites*(λ) and *3dshapes*(λ), based on dSprites (Matthey et al. 2017) and 3dshapes (Burgess and Kim 2018) datasets, respectively. They consist of 3D samples generated from a vector consisting of $k = 5$ and $k = 6$ factors of variation, respectively. Both datasets have one binary concept annotation per factor of variation. Parameters $\delta \in [0, 1]$ and $\lambda \in \{0, \cdots, k-1\}$ control the degree of concept inter-dependencies during generation: $\lambda = 0$ and $\delta = 0$ represent inter-concept independence while higher values represent stronger inter-concept dependencies. For dataset details see Appendix A.6.

Baselines and Setup We compare the purity of concept representations in various methods using our metrics. We select representative methods from (i) *supervised CL* (i.e., jointly-trained CBMs (Koh et al. 2020) with sigmoidal and logits bottlenecks, and CW (Chen, Bei, and Rudin 2020) both when its representations are reduced through a MaxPool-Mean reduction and when no feature map reduction is applied), (ii) *unsupervised CL* (i.e., CCD (Yeh et al. 2020) and SENN (Alvarez-Melis and Jaakkola 2018)), (iii) *unsupervised DGL* (vanilla VAE (Kingma and Welling 2014) and β-VAE (Higgins et al. 2017)), and (iv) *weakly supervised DGL* (Ada-GVAE and Ada-MLVAE (Locatello et al. 2020a)). For each method and metric, we report the average metric values and 95% confidence intervals obtained from 5 different random seeds. We include details on training and architecture hyperparameters in Appendix A.6.

Results and Discussion

In contrast to DGL metrics, our metrics can meaningfully capture impurities concealed in representations. We begin by empirically showing that our metrics indeed capture impurities encoded within a concept representation. For this, we prepare a simple synthetic dataset of ground-truth concept vectors $\mathcal{D} := \{\mathbf{c}^{(i)} \in \{0, 1\}^5\}_{i=1}^{3,000}$ where, for each sample $\mathbf{c}^{(i)}$, its j-th concept is a binary indicator $\mathbb{1}_{\tilde{c}_j^{(i)} \geq 0}$ of the sign taken by a latent variable $\tilde{c}_j^{(i)}$ sampled from a joint normal distribution $\tilde{\mathbf{c}}^{(i)} \sim \mathcal{N}(\mathbf{0}, \Sigma)$ (with $\tilde{\mathbf{c}}^{(i)} \in \mathbb{R}^5$). During construction, we simulate real-world co-dependencies between different concepts by setting Σ's non-diagonal entries to 0.25. To evaluate whether our metrics can discriminate between different levels of impurities encoded in different concept representations, we construct two sets of soft concept representations. First, we construct a baseline "*pure*" fuzzy representation $\hat{\mathbf{c}}^{(\text{pure})}$ of vector $\mathbf{c} \in \mathcal{D}$ by sampling $\hat{c}_j^{(\text{pure})}$ from $\text{Unif}(0.95, 1)$ if $c_j = 1$ or from $\text{Unif}(0, 0.05)$ if $c_j = 0$. Notice that each dimension of this representation preserves enough information to *perfectly predict* each concept's activation state without encoding any extra information. In contrast, we construct a perfectly "*impure*" soft concept representation $\hat{\mathbf{c}}^{(\text{impure})}$ by encoding, as part of each concept's fuzzy representation, the state of all other concepts. For this we partition and tile the sets $[0.0, 0.05]$ and $[0.95, 1.0]$ into $2^{5-1} = 16$ equally sized and disjoint subsets $\{[\text{off}_i, \text{off}_{i+1})\}_{i=0}^{15}$ and $\{[\text{on}_i, \text{on}_{i+1})\}_{i=0}^{15}$, respectively. From here, we generate an impure representation of ground truth concept vector $\mathbf{c} \in \mathcal{D}$ by sampling $\hat{c}_j^{(\text{impure})}$ from $\text{Unif}(\text{on}_{\text{bin}(\mathbf{c}_{-j})}, \text{on}_{\text{bin}(\mathbf{c}_{-j})+1})$ if $c_j = 1$ or from $\text{Unif}(\text{off}_{\text{bin}(\mathbf{c}_{-j})}, \text{off}_{\text{bin}(\mathbf{c}_{-j})+1})$ otherwise,

	OIS (↓)	NIS (↓)	SAP (↑)	MIG (↑)	R^4 (↑)	DCI Dis (↑)
Baseline Soft (%)	$\mathbf{4.69 \pm 0.43}$	$\mathbf{66.25 \pm 2.31}$	48.74 ± 0.41	99.93 ± 0.03	99.95 ± 0.00	$\mathbf{99.99 \pm 0.00}$
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: Comparison between our metrics (left of the middle line) and common DGL metrics (right) using hand-crafted soft concept representations and latent codes. We highlight statistically significant differences ($p < 0.05$) in scores between the baseline (i.e., "pure") and the impure representations. Furthermore, we use ↑/↓ to indicate that a metric is better if its score is higher/lower and compute all metrics over 5 folds. For statistical significance validation, we include p values (two-sided T-test).

where we use $\text{bin}(\mathbf{c}_{-j})$ to represent the decimal representation of the vector resulting from removing the j-th dimension of $\mathbf{c}$. Intuitively, each concept in this soft representation encodes the activation state of every other concept using different confidence ranges. Therefore, one can perfectly predict all concepts from a single concept's representation, an impossibility from ground truth concepts alone.

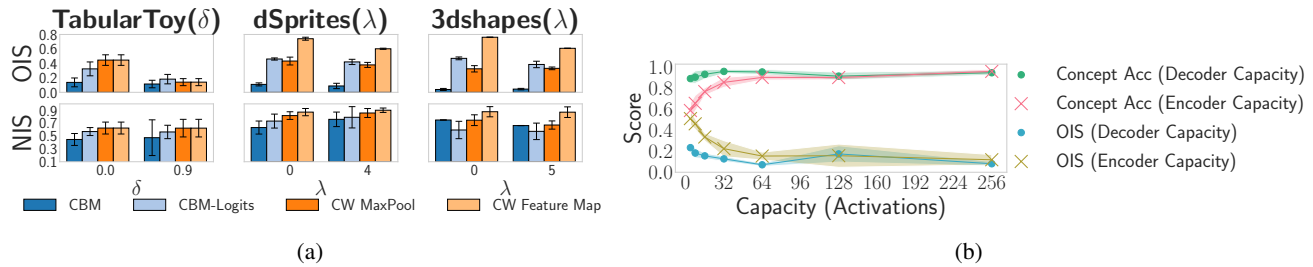
We hypothesize that if a metric is capable of accurately capturing undesired impurities within concept representations, then it should generate vastly different scores for the two representation sets constructed above. To verify this hypothesis, we evaluate our metrics, together with a selection of DGL disentanglement metrics, and show our results in Table 1. We include *SAP* (Kumar, Sattigeri, and Balakrishnan 2017), R^4 (Ross and Doshi-Velez 2021), *mutual information gap* (MIG) (Chen et al. 2018), and *DCI Disentanglement* (DCI Dis) (Eastwood and Williams 2018) as representative DGL metrics given their wide use in the DGL literature (Ross and Doshi-Velez 2021; Zaidi et al. 2020). Our results show that our metrics correctly capture the difference in impurity between the two representation sets in a statistically significant manner. In contrast, existing DGL metrics are incapable of clearly discriminating between these two impurity extremes, with DCI being the only metric that generates some statistically significant differences albeit the scores' differences are minimal (less than 0.5%). Surprisingly, although MIG is inspired by a similar mutual information (MI) argument as our OIS metric, it was unable to capture any meaningful differences between our two representation types. We believe that this is because to compute the MIG one requires an estimation of the MI which, being sensitive to hyperparameters, may fail to capture important differences. These results, therefore, support using a non-linear model's test AUC as a proxy of the MI. Further details can be found in Appendix A.7.

Our metrics can capture impurities caused by differences in concept representations and model capacities, as well as by accidental spurious correlations. Impurities in a CL model can come from different sources, such as differences in concept representations, as previously shown in Figure 1b, or architectural constraints (e.g., a CBM trained with a partial/incomplete set of concepts). Here, we show that impurities caused by differences in the nature of concept representations, as well as by inadequate model capacities and spurious correlations, can be successfully captured by our metrics and thus avoided.

Differences in concept representations: in Figure 2a, we show the impurities in (1) a CBM with a sigmoidal bottleneck (*CBM*) vs a CBM with logits in its bottleneck (*CBM-Logits*) and (2) a *CW* module with and without feature map reduction (*CW Feature Map* vs *CW Max-Pool-Mean*). Our metrics show that CBM-Logits and CW Feature Map are prone to encoding more impurities than their counterparts. This is because their representations are less constrained, as logit activations can be within any range (as opposed to $[0, 1]$ in *CBM*) and *CW Feature Map* preserves all information from its concept feature map by not reducing it to a scalar. The exception to this is the failure of NIS to detect impurities in CBM-Logits for 3dshapes(λ). We hypothesise that this is due to this task's higher complexity, forcing both CBM and CBM-Logits to distribute inter-concept information across all representations more than in other datasets.

Differences in model capacity: low-capacity models may be forced to use their concept representations to encode information outside their corresponding ground truth concept. To verify this, we train a CBM in TabularToy($\delta = 0$) whose concept encoder and label predictor are three-layered ReLU MLPs. We vary the capacities of the concept encoder or label predictor by setting their hidden layers' activations to $\{\text{capacity}, \text{capacity}/2\}$, while fixing the number of hidden units in their corresponding counterpart to $\{128, 64\}$. We then monitor the accuracy of concept representations w.r.t. their aligned ground truth concepts as well as their OIS. We observe (Figure 2b) that as the concept encoder and label predictor capacities decrease, the CBM exhibits significantly higher impurity and lower concept accuracy. Note that the concept encoder capacity has a significantly greater effect on the purity of the representations compared to the label predictor capacity. Measuring impurities in a systematic way using our metrics can therefore guide the design of CL architectures that are less prone to impurities.

Spurious correlations: we create a variation of dSprites($\lambda = 0$), where we randomly introduce spurious correlations by assigning each sample a class-specific background colour with 75% probability (see Appendix A.8 for details). We train two identical CBMs on dSprites($\lambda = 0$) and its corrupted counterpart. During training, we note that CBM-S (the CBM trained on the corrupted data) has a higher task validation accuracy than the other CBM (Figure 3), while having similar concept validation accuracies. Nevertheless, when we evaluate both models using a test set sampled from the original dSprites($\lambda = 0$) dataset, we see an interesting result: both models can predict ground truth concept labels with similarly high accuracy. However, unlike CBM, CBM-S struggles to predict the



(a) (b)

Figure 2: (a) Impurity scores, and their 95% confidence intervals, for concept representations in CBM and CW (in low and high concept inter-dependence) and a corresponding transformation which leads to higher impurities. (b) Effect of network capacity (i.e., number of hidden activations used in the concept encoder and label predictor) on a CBM's concept accuracy and OIS.

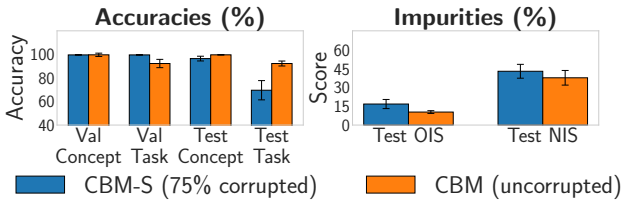


Figure 3: Impurities and validation/testing accuracies for CBMs trained on the original and corrupted dSprites($\lambda = 0$). We compute test scores in the uncorrupted data.

task labels. Failure of CBM-S to accurately predict task labels is remarkable as labels in this dataset are uniquely a function of their corresponding concept annotations and CBM-S is able to accurately predict concepts in the original dSprites($\lambda = 0$) dataset. We conjecture that this is due to the fact that concepts in CBM-S encode significantly more information than needed, essentially encoding the background colour in addition to the original concepts as part of their representations. To verify this, we evaluate the OIS and NIS of the concept representations learnt by both models and observe that, in line with our intuition, CBM-S indeed encodes significantly more impurity. Our metrics can therefore expose spurious correlations captured by CL methods which appear to be highly predictive of concept labels while underperforming in their downstream task.

Our metrics can indicate when it is safe to perform interventions on a CBM by giving a realistic picture of impurities. A major potential consequence of not being able to measure the impurities faithfully is that *concept interventions* (Koh et al. 2020), which allow domain experts to adjust the model by intervening and fixing predictions at the concept level, may fail: adjusting a concept $\hat{\mathbf{c}}_{(:,i)}$ may unintentionally impact the label predictor's understanding of another concept $\hat{\mathbf{c}}_{(:,j)}$ if representation $\hat{\mathbf{c}}_{(:,i)}$ encodes unnecessary information about concept c_j. To see this in practice, consider a CBM model and a CBM-Logits model both trained to convergence on dSprites($\lambda = 0$), and both achieving fairly similar task and concept accuracies (Figure 5). We then perform interventions at random on their concept representations as follows: in CBM, where concept activations represent probabilities, we intervene on the i-th concept representation by setting $\hat{c}_i$ to the value of its corresponding ground truth concept c_i. In CBM-Logits, as in (Koh et al. 2020), we intervene on the i-th concept by setting it to the 5%-percentile of the empirical distribution of $\hat{c}_i$ if $c_i = 0$, and we set it to the 95%-percentile of that concept's distribution if $c_i = 1$. Interestingly, our results (Figure 5) show that random interventions cause a significant drop in task accuracy of CBM-Logits while leading to an increase in accuracy in CBM. Looking at the impurities of these two models, we observe that although the CBM-Logits model has better accuracy, both its OIS and NIS scores are considerably higher than those for the CBM model, explaining why interventions had such undesired consequences.

To rule out the difference in intervention mechanism as the cause of these results, we train two CBM-Logits with the same concept encoder capacities but with different capacities in their label predictors and observe the same phenomena as above: performance degradation upon intervention, which occurs in the case of the lower capacity model, is evidence for higher OIS and NIS scores compared to that of the higher capacity model. Further details about this experiment are documented in Appendix A.10.

Our metrics can provide insights on the impact of different degrees of supervision on concept purity. As models of various families benefit from varying degrees of supervision ranging from explicit supervision (supervised CL) to implicit (unsupervised CL), weak (weakly-supervised DGL) and no supervision at all (unsupervised DGL), different models are expected to learn concept representations of varying purity. The intuitive assumption is that more supervision leads to better and purer concepts. Here, we compare models from all families using our metrics and show that, contrary to our intuition, this is not necessarily the case. Within variants of CBM and CW, we choose CBM without logits and CW MaxPool-Mean, as they tend to encode fewer impurities (see Figure 2a). Furthermore, given the tabular nature of TabularToy(δ), we do not compare DGL methods in this task. Finally, for details on computing our metrics when an alignment between ground truth concepts and learnt representations is missing, see Appendix A.5.

In terms of task accuracy, the overall set of learnt concept representations is equally predictive of the downstream task across all surveyed methods (see Appendix A.9 for de-

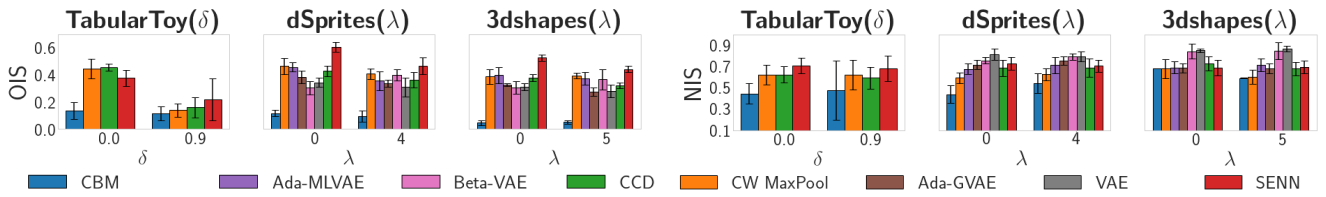


Figure 4: Evaluation of different models, deployed across various tasks, using our Oracle Impurity Score (left) and Niche Impurity Score (right) in two extreme cases of no correlation and high concept correlation for each dataset. For all DGL and unsupervised methods, we learn as many concepts/latent dimensions as known ground truth concepts k.

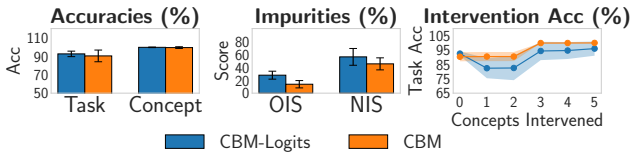


Figure 5: Intervening in CBMs with different bottleneck activation functions in dSprites($\lambda = 0$).

tails). However, as discussed previously, models with the same task accuracy can encode highly varied levels of impurities in their individual concept representations. Figure 4 (left) shows a comparison of impurities observed across methods using OIS. CBM's individual concepts consistently experience the least amount of impurity due to receiving explicit supervision, which is to be expected. Unexpectedly though, we observe that the same explicit supervision can lead to highly impure representations in CW. Indeed, CW impurities are on par or more than those of unsupervised approaches. Looking into implicit supervision, we observe that individual concepts in CCD and SENN do not correspond well to the ground truth ones. This indicates that the information about each ground truth concept is distributed across the overall representation rather than localised to individual concepts, leading to relatively high OIS. We attribute CCD's lower impurity, compared with SENN, to the use of a regularisation term that encourages coherence between concept representations in similar samples and misalignment between concept representations in dissimilar samples. More interestingly, however, both CCD and SENN encode higher impurities than some DGL approaches, despite benefiting from task supervision. Within DGL approaches, astonishingly no supervision in unsupervised DGL can result in purer individual concept representations than those of weakly-supervised DGL methods. This suggests that concept information may be heavily distributed in weakly-supervised DGL methods.

Moving from individual concepts, Figure 4 (right) shows a comparison of impurities observed across subsets of concept representations using our NIS metric. Similar to our OIS results, the overall set of concept representations in CBM shows the least amount of impurity. Unlike individual impurities, however, the overall set of concept representations in DGL methods shows a higher impurity than that of explicitly supervised approaches. This can be explained by the fact that DGL methods seem to learn representations that are not fully aligned with our defined set of ground truth concepts, yet when taken as a whole they are still highly predictive of individual concepts. This would lead to complement niches being highly predictive of individual ground truth concepts even when individual representations in those niches are not fully predictive of that concept itself, resulting in relatively high NIS scores and lower OIS scores. Furthermore, notice that weakly supervised DGL methods show a lesser niche impurity than unsupervised DGL methods, suggesting, as in Locatello et al. (2020a), that weakly-supervised representations are indeed more disentangled. We notice, however, that this decrease in impurity for weakly-supervised methods comes at the cost of their latent codes being less effective at predicting individual concepts than unsupervised latent codes (see Appendix A.11). Finally, within methods benefiting from explicit supervision, the overall set of learnt concepts in CCD has fewer impurities than that of SENN, which was similarly observed with individual concepts above.

Our metrics are robust to concept correlations. As seen in Figure 4, the preserved method ranking using our metrics in settings with and without correlations confirms our metrics' robustness to concept correlations.

Conclusion

Impurities in concept representations can lead to models accidentally capturing spurious correlations and can be indicative of unexpected behaviour during concept interventions, which is crucial given that performing safe interventions is one of the main motivations behind CBMs. Despite this importance, current metrics in CL literature and the related field of DGL fail to fully capture such impurities. In this paper, we address these limitations by introducing two novel robust metrics that can circumvent several limitations in existing metrics and correctly capture impurities in learnt concept representations. Indeed, for the first time, we can systematically compare the purity of concepts in CL and DGL and show that, contrary to common assumptions, more explicit supervision does not necessarily translate to better concept quality in terms of purity. More importantly, beyond comparison, our experiments show the utility of these metrics in designing and training more reliable and robust concept learning models. Therefore, we envision them to be an integral part of future tools developed for the safe deployment of concept-based models in real-world scenarios.

Acknowledgements

The authors would like to thank our reviewers for their insightful comments on earlier versions of this manuscript. MEZ acknowledges support from the Gates Cambridge Trust via a Gates Cambridge Scholarship. PB acknowledges support from the European Union's Horizon 2020 research and innovation programme under grant agreement No 848077. UB acknowledges support from DeepMind and the Leverhulme Trust via the Leverhulme Centre for the Future of Intelligence (CFI), and from a JP Morgan Chase AI PhD Fellowship. AW acknowledges support from a Turing AI Fellowship under grant EP/V025279/1, The Alan Turing Institute, and the Leverhulme Trust via CFI. MJ is supported by the EPSRC grant EP/T019603/1.

References

Altman, N.; and Krzywinski, M. 2015. Points of Significance: Association, correlation and causation. *Nature methods*, 12(10).

Alvarez-Melis, D.; and Jaakkola, T. S. 2018. Towards robust interpretability with self-explaining neural networks. In *Neural Information Processing Systems (NeurIPS)*.

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828.

Bhatt, U.; Xiang, A.; Sharma, S.; Weller, A.; Taly, A.; Jia, Y.; Ghosh, J.; Puri, R.; Moura, J. M.; and Eckersley, P. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 648–657.

Burgess, C.; and Kim, H. 2018. 3D Shapes Dataset. https://github.com/deepmind/3dshapes-dataset/. Accessed: 2021-09-01.

Chen, R. T.; Li, X.; Grosse, R.; and Duvenaud, D. 2018. Isolating sources of disentanglement in variational autoencoders. *arXiv preprint arXiv:1802.04942*.

Chen, Z.; Bei, Y.; and Rudin, C. 2020. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12): 772–782.

Ciravegna, G.; Barbiero, P.; Giannini, F.; Gori, M.; Lió, P.; Maggini, M.; and Melacci, S. 2021. Logic Explained Networks. *CoRR*, abs/2108.05149.

Darwin, C. 1859. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. *London: John Murray*.

Eastwood, C.; and Williams, C. K. I. 2018. A Framework for the Quantitative Evaluation of Disentangled Representations. In *International Conference on Learning Representations (ICLR)*. OpenReview.net.

Ghorbani, A.; Wexler, J.; Zou, J. Y.; and Kim, B. 2019. Towards Automatic Concept-based Explanations. In *Neural Information Processing Systems (NeurIPS)*, 9273–9282.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *International Conference on Learning Representations (ICLR)*. OpenReview.net.

Kazhdan, D.; Dimanov, B.; Jamnik, M.; Liò, P.; and Weller, A. 2020. Now You See Me (CME): Concept-based Model Extraction. In *Conference on Information and Knowledge Management (CIKM) Workshops*, volume 2699 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C. J.; Wexler, J.; Viégas, F. B.; and Sayres, R. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, 2673–2682. PMLR.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In Bengio, Y.; and LeCun, Y., eds., *International Conference on Learning Representations (ICLR)*.

Koh, P. W.; Nguyen, T.; Tang, Y. S.; Mussmann, S.; Pierson, E.; Kim, B.; and Liang, P. 2020. Concept Bottleneck Models. In *International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, 5338–5348. PMLR.

Kumar, A.; Sattigeri, P.; and Balakrishnan, A. 2017. Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*.

Locatello, F.; Bauer, S.; Lucic, M.; Rätsch, G.; Gelly, S.; Schölkopf, B.; and Bachem, O. 2019. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In *International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, 4114–4124. PMLR.

Locatello, F.; Poole, B.; Rätsch, G.; Schölkopf, B.; Bachem, O.; and Tschannen, M. 2020a. Weakly-Supervised Disentanglement Without Compromises. In *International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, 6348–6359. PMLR.

Locatello, F.; Tschannen, M.; Bauer, S.; Rätsch, G.; Schölkopf, B.; and Bachem, O. 2020b. Disentangling Factors of Variations Using Few Labels. In *International Conference on Learning Representations (ICLR)*. OpenReview.net.

Lundberg, S. M.; and Lee, S. 2017. A Unified Approach to Interpreting Model Predictions. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 4765–4774.

Mahinpei, A.; Clark, J.; Lage, I.; Doshi-Velez, F.; and Pan, W. 2021. Promises and Pitfalls of Black-Box Concept Learning Models. *CoRR*, abs/2106.13314.

Margeloiu, A.; Ashman, M.; Bhatt, U.; Chen, Y.; Jamnik, M.; and Weller, A. 2021. Do Concept Bottleneck Models Learn as Intended? *CoRR*, abs/2105.04289.

Matthey, L.; Higgins, I.; Hassabis, D.; and Lerchner, A. 2017. dSprites: disentanglement testing sprites dataset (2017). https://github.com/deepmind/dsprites-dataset. Accessed: 2021-09-01.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any

Classifier. In *International Conference on data science and advanced analytics (DSAA)*, 1135–1144. ACM.

Ross, A.; and Doshi-Velez, F. 2021. Benchmarks, algorithms, and metrics for hierarchical disentanglement. In *International Conference on Machine Learning*, 9084–9094. PMLR.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The Caltech-UCSD birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Yeh, C.; Kim, B.; Arik, S. Ö.; Li, C.; Pfister, T.; and Ravikumar, P. 2020. On Completeness-aware Concept-Based Explanations in Deep Neural Networks. In *Neural Information Processing Systems (NeurIPS)*.

Zaidi, J.; Boilard, J.; Gagnon, G.; and Carbonneau, M.-A. 2020. Measuring disentanglement: A review of metrics. *arXiv preprint arXiv:2012.09276*.