GlanceNets: Interpretabile, Leak-proof Concept-based Models

Emanuele Marconato[1,2], Andrea Passerini[1], and Stefano Teso[1]

[1] DISI, Università di Trento, Via Sommarive, 9 38123 Povo (TN)
name.surname@unitn.it
[2] Dipartimento di Informatica, Università di Pisa, Largo B. Pontecorvo, 3 56127 Pisa

Abstract. There is growing interest in concept-based models (CBMs) that combine high-performance and interpretability by acquiring and reasoning with a vocabulary of high-level concepts. A key requirement is that the concepts be interpretable. Existing CBMs tackle this desideratum using a variety of heuristics based on unclear notions of interpretability, and fail to acquire concepts with the intended semantics. We address this by providing a clear definition of interpretability in terms of alignment between the model's representation and an underlying data generation process, and introduce GlanceNets, a new CBM that exploits techniques from disentangled representation learning and open-set recognition to achieve alignment, thus improving the interpretability of the learned concepts. We show that GlanceNets, paired with concept-level supervision, achieve better alignment than state-of-the-art approaches while preventing spurious information from unintendedly leaking into the learned concepts.

Keywords: Explainable AI · Concept-based models · Interpretability · Disentanglement · Concept Leakage · Open Set Recognition.

1 Introduction

Concept-based models (CBMs) are an increasingly popular family of classifiers that combine the transparency of white-box models with the flexibility and accuracy of regular neural nets [Alvarez-Melis and Jaakkola, 2018, Li et al., 2018, Chen et al., 2019, Losch et al., 2019, Chen et al., 2020]. At their core, all CBMs acquire a vocabulary of concepts capturing high-level, task-relevant properties of the data, and use it to compute predictions and produce faithful explanations of their decisions [Rudin, 2019].

The central issue in CBMs is how to ensure that the concepts are *semantically meaningful* and *interpretable* for (sufficiently expert and motivated) human stakeholders. Current approaches struggle with this. One reason is that the notion of interpretability is notoriously challenging to pin down, and therefore existing CBMs rely on different heuristics—such as encouraging the concepts to be sparse [Alvarez-Melis and Jaakkola, 2018], orthonormal to each other [Chen et al., 2020], or match the contents of concrete examples [Chen et al., 2019]—with unclear properties and incompatible goals. A second, equally important

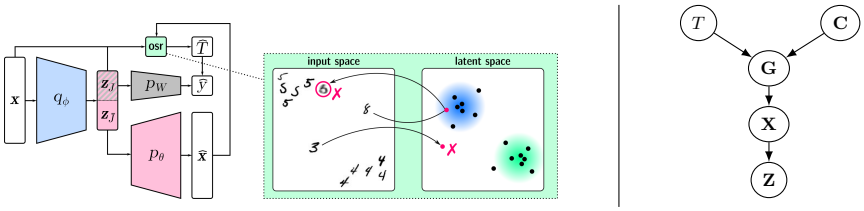

Fig. 1. Left: Architecture of GlanceNets showing the encoder q_ϕ, decoder p_θ, classifier p_W, and open-set recognition step. **Center**: GlanceNets prevent leakage by identifying and rejecting open-set inputs using a combined strategy, shown here for a model trained on digits "4" and "5" only: the "3" is rejected as its embedding falls far away from classes prototypes (colored blobs), while the "8" is rejected as its reconstruction loss is too large. **Right**: The data generation process.

issue is *concept leakage*, whereby the learned concepts end up encoding spurious information about unrelated aspects of the data, making it hard to assign them clear semantics [Mahinpei et al., 2021]. Notably, even concept-level supervision is insufficient to prevent leakage [Margeloiu et al., 2021], cf. Fig. 3.

Prompted by these observations, we define interpretability in terms of *alignment*: learned concepts are interpretable if they can be mapped to a (partially) interpretable data generation process using a transformation that preserves semantics. This is sufficient to unveil limitations in existing strategies, build an explicit link between interpretability and disentangled representations, and provide a clear and actionable perspective on concept leakage. Building on our analysis, we also introduce GlanceNets (aliGned LeAk-proof coNCEptual Networks), a novel class of CBMs that combine techniques from *disentangled representation learning* [Schölkopf et al., 2021] and *open-set recognition* [Scheirer et al., 2012] to actively pursue alignment – and guarantee it under suitable assumptions – and avoid concept leakage.

Contributions: Summarizing, we: (*i*) Provide a definition of interpretability as alignment that facilitates tapping into ideas from disentangled representation learning; (*ii*) Show that concept leakage can be viewed from the perspective of out-of-distribution generalization; (*iii*) Introduce GlanceNets, a novel class of CBMs that acquire interpretable representations and are robust to concept leakage; (*iv*) Present an extensive empirical evaluation showing that GlanceNets are as accurate as state-of-the-art CBMs while attaining better interpretability and avoiding leakage.

2 Concept-based Models

Concept-based models (CBMs) comprise two key elements: (i) A learned vocabulary of k high-level concepts meant to enable communication with human stakeholders [Kambhampati et al., 2022], and (ii) a simulatable [Lipton, 2018] classifier whose predictions depend solely on those concepts. Formally, a CBM $f : \mathbb{R}^d \to [c]$, with $[c] := \{1, \ldots, c\}$, maps instances $\mathbf{x}$ to labels y by measuring

how much each concept activates on the input, obtaining an activation vector $\mathbf{z}(\mathbf{x}) := (z_1(\mathbf{x}), \ldots, z_k(\mathbf{x})) \in \mathbb{R}^k$, aggregating the activations into per-class scores $s_y(\mathbf{x})$ using a linear map [Alvarez-Melis and Jaakkola, 2018, Chen et al., 2019, 2020], and then passing these through a softmax, i.e.,

$$s_y(\mathbf{x}) := \sum_j w_{yj} z_j(\mathbf{x}), \quad p(y \mid \mathbf{x}) := \text{softmax}(\mathbf{s}(\mathbf{x}))_y. \tag{1}$$

Each weight $w_{yj} \in \mathbb{R}$ encodes the relevance of concept z_j for class y. The activations themselves are computed in a black-box manner, often leveraging pre-trained embedding layers, but learned so as to capture interpretable aspects of the data using a variety of heuristics, discussed below.

Now, *as long as the concepts are interpretable*, it is straightforward to extract human understandable local explanations disclosing how different concepts contributed to any given decision $(\mathbf{x}, y)$ by looking at the concept activations and their associated weights, thus abstracting away the underlying computations. This yields explanations of the form $\{(w_{yj}, z_j(\mathbf{x})) : j \in [k]\}$ that can be readily summarized[3] and visualized [Hase and Bansal, 2020, Guidotti et al., 2018]. Importantly, the score of class y is conditionally independent from the input $\mathbf{x}$ given the corresponding explanation, i.e., $s_y(\mathbf{x}) \perp\!\!\!\perp \mathbf{x} \mid \mathcal{E}(\mathbf{x}, y)$, ensuring that the latter is faithful to the model scores. GlanceNets inherit all of these features.

Heuristics for interpretability. Crucially, CBMs are only interpretable insofar as their concepts are. Existing approaches implement special mechanisms to this effect, often pairing a traditional classification loss (such as the cross-entropy loss) with an auxiliary regularization term [Alvarez-Melis and Jaakkola, 2018, Chen et al., 2019, 2020].

We are interested in particular to variants of concept bottleneck models (CBNMs) [Koh et al., 2020, Losch et al., 2019], which align the concepts using concept-level supervision, possibly obtained from a separate source, like ImageNet [Deng et al., 2009]. From a statistical perspective, this seems perfectly sensible: if the supervision is unbiased and comes in sufficient quantity, and the model has enough capacity, this strategy *appears* to guarantee the learned and ground-truth concepts to match.

Concept leakage in concept-bottleneck models. Unfortunately, concept-level supervision is *not* sufficient to guarantee interpretability. [Mahinpei et al., 2021] have demonstrated through simple examples that concepts acquired by CBNMs pick up spurious properties of the data. This phenomenon is known as *concept leakage*.

Intuitively, leakage occurs because in CBNMs the concepts end up unintentionally capturing distributional information about unobserved aspects of the input, failing to provide well-defined semantics. However, a clear definition of leakage is missing, and so are strategies to prevent it: a key contribution of our paper is showing that leakage can be understood from the perspective of domain shift and dealt with using open-set recognition [Scheirer et al., 2012].

[3] For instance, by pruning those concepts that have little effect on the outcome to simplify the presentation.

3 Interpretability and Leakage

The main issue with heuristics used by CBMs is that they are based on unclear notions of interpretability. In order to develop effective algorithms, we propose to view interpretability as a form of *alignment* between the machine's representation and that of its user. This enables us to identify conditions under which interpretability can be achieved, build links to well-understood properties of representations, and leverage state-of-the-art learning strategies.

Interpretability. We henceforth focus on the (rather general) generative process shown in Fig. 1 the observations $\mathbf{X} \in \mathbb{R}^d$ are caused by n generative factors $\mathbf{G} \in \mathbb{R}^n$, themselves caused by a set of confounds $\mathbf{C}$ (including the label Y [Schölkopf et al., 2012]). Notice that the generative factors *can* be statistically dependent due to the confounds $\mathbf{C}$, but as noted by [Suter et al., 2019], the total causal effect Peters et al. [2017] between G_i and G_j is zero for all $i \neq j$. The generative factors capture all information necessary to determine the observation [Reddy et al., 2022], so the goal is to learn concepts $\mathbf{Z} \in \mathbb{R}^k$ that recover them. The variable T will be introduced later on.

We posit that a (learned) representation is only interpretable if it supports *symbolic communication* between the model and the user, in the sense that it shares the same (or similar enough) semantics to the user's representation. The latter is however generally unobserved. Then, we make a second, critical assumption that *some* of the generative factors $\mathbf{G}_I \subseteq \mathbf{G}$ are interpretable to the user, i.e., they can be used as a proxy for the user's internal representation. Naturally, not all generative factors are interpretable [Gabbay et al., 2021], but in many applications some of them are, e.g., the hair color or noise size in CelebA [Liu et al., 2015].

Interpretability as alignment. Under this assumption, if the variables $\mathbf{Z}_J \subseteq \mathbf{Z}$ are *aligned* to the generative factors $\mathbf{G}_I$ by a map $\alpha : \mathbf{g} \mapsto \mathbf{z}_J$ that preserves semantics, they are themselves interpretable. Now, defining what a semantics-preserving map should look like is challenging, but constructing one is not: the identity is clearly one such map, and so are maps that permute the indices and independently rescale the individual variables. One desirable property is that α does not "mix" multiple G's into a single Z. E.g., if Z blends together head tilt, hair color, and nose size, users will have trouble pinning down what it means.This property can be formalized in terms of *disentanglement* [Eastwood and Williams, 2018, Suter et al., 2019, Schölkopf et al., 2021]. This is however insufficient: we wish the map between G_i and its associated factor Z_j to be "simple", so as to *conservatively* guarantee that it preserves semantics. This makes alignment strictly stronger than disentanglement.

Motivated by this, we say that $\mathbf{Z}_J$ is *aligned* to $\mathbf{G}_I$ if:

(i) There exists an injective map between indices $\pi : [n_I] \to [k]$, where $[n_I]$ identifies the subset of generative factors indexes in $\mathbf{G}_I$, such that, for all $i, i' \in [n_I]$, $i \neq i'$, and $j = \pi(i)$, it holds that fixing G_i is enough to fix Z_j regardless of the value taken by the other generative factors $G_{i'}$, and

(ii) The map α can be written as $\alpha(\mathbf{g}) = (\mu_1(g_{\pi(1)}), \ldots, \mu_n(g_{\pi(n_I)}))$, where the μ_i's are monotone functions. This holds, for instance, for linear transformations of the form $A(g_{\pi(1)}, \ldots, g_{\pi(n_I)})$, where $A \in \mathbb{R}^{n_I \times k}$ is a matrix with no non-zero off-diagonal entries. This second requirement can be relaxed depending on the application.

Measuring alignment with DCI. Disentanglement can be measured in a number of ways [Zaidi et al., 2020], but most of them provide little information about how simple the map α is. In order to estimate alignment, we repurpose DCI, a measure of disentanglement introduced by Eastwood and Williams [2018], by fitting a linear model from $\mathbf{z}_J$ to $\mathbf{g}_I$. Further details are included in the Supplementary Material.

Achieving alignment with concept-level supervision. It has been shown that disentanglement cannot be achieved in the purely unsupervised setting [Locatello et al., 2019]. This immediately entails that alignment is also impossible in that setting, highlighting a core limitation of [Alvarez-Melis and Jaakkola, 2018]. However, disentanglement can be attained if supervision about the generative factors is available, even only for a small percentage of the examples [Locatello et al., 2020a]. As a matter of fact, supervision is used in representation learning to achieve *identifiability*, a stronger condition than – and that entails both of – disentanglement *and* alignment [Khemakhem et al., 2020]. Thus, following CBNMs, we seek alignment by leveraging concept-level supervision.

Interpretability and concept leakage. Intuitively, concept leakage occurs when a model is trained on a data set on which (i) some generative factors $\mathbf{G}_V \subset \mathbf{G}$ vary, while the others $\mathbf{G}_F = \mathbf{G} \setminus \mathbf{G}_V$ are fixed, and (ii) the two groups of factors are statistically dependent. For instance, in the even vs. odd experiment of [Mahinpei et al., 2021], no training examples are annotated with concepts besides 4 and 5. CBNMs with access to supervision on $\mathbf{G}_V$ tend to acquire a latent representation that approximates these factors, and that because of (ii) correlates with the fixed factors $\mathbf{G}_F$.

In contrast with previous assessments [Mahinpei et al., 2021, Margeloiu et al., 2021], we notice that point (i) can be viewed as a special form of domain shift: the training examples are sampled from a ground-truth distribution $p(\mathbf{X}, \mathbf{G} \mid T = 1)$ in which $\mathbf{G}_F$ is approximately fixed, e.g., $p(\mathbf{G}_F \mid T = 1) = \delta(\mathbf{g}'_F)$ for some vector $\mathbf{g}'_F$, and the test set from a different distribution $p(\mathbf{X}, \mathbf{G} \mid T = 0)$ in which $\mathbf{G}_F$ is no longer fixed. Here, T is a random variable that selects between training and test distribution, see Fig. 3. Since regular CBMs have no strategy to cope with domain shift, they fail to adapt when this occurs.

Motivated by this, we propose then to tackle concept leakage by designing a CBM specifically equipped with strategies for detecting instances that do not belong to the training distribution using open-set recognition [Scheirer et al., 2012]. By estimating the value of the variable T at inference time, we are essentially predicting whether an input was sampled from a distribution similar enough to the training distribution, and therefore can be handled by a model learned on this distribution, or not. This strategy proves very effective in practice, as shown by our empirical evaluation (Section 5.2).

4 GlanceNets

GlanceNets combine a VAE-like architecture [Kingma and Welling, 2014, Rezende et al., 2014] for learning disentangled concepts with a prior and classifier designed for open-set prediction [Sun et al., 2020]. In order to accommodate for non-interpretable factors, the latent representation of GlanceNets $\mathbf{Z}$ is split into two: (i) k concepts $\mathbf{Z}_J$, aligned to the *interpretable* generative factors $\mathbf{G}_I$, that are used for prediction, and (ii) $\bar{k}$ *opaque* factors $\mathbf{Z}_{\bar{J}}$ that are only used for reconstruction. Specifically, a GlanceNet comprises an encoder $q_\phi(\mathbf{Z} \mid \mathbf{X})$ and a decoder $p_\theta(\mathbf{X} \mid \mathbf{Z})$, both parameterized by deep neural networks, as well as a classifier $p_W(Y \mid \mathbf{Z}_J)$ feeding off the interpretable concepts only. Following other CBMs, the classifier is implemented using a dense layer with parameters $W \in \mathbb{R}^{v \times k}$ followed by a softmax activation, i.e., $p_W(Y \mid \mathbf{z}_J) := \text{softmax}(W\mathbf{z}_J)$, and the most likely label is used for prediction. The overall architecture is shown in Fig. 1.

In contrast to regular VAEs, GlanceNets associate each class to a prototype in latent space through the prior $p(\mathbf{Z} \mid \mathbf{Y})$, which is conditioned on the class and modelled as a *mixture of gaussians* with one component per class. The encoder, decoder, and prior are fit on data so as to maximize the evidence lower bound, defined as [Kingma and Welling, 2019] $\mathbb{E}_{p_D(\mathbf{x},y)}[\mathcal{L}(\theta, \mathbf{x}, y; \beta)]$ with:

$$\begin{aligned}\mathcal{L}(\theta, \mathbf{x}, y; \beta) :=& \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x} \mid \mathbf{z}) + \log p_W(y \mid \mathbf{z}_J)] \\ &- \beta \cdot \mathsf{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) \,\|\, p(\mathbf{z} \,|\, y))\end{aligned} \tag{2}$$

Here, $p_D(\mathbf{x}, y)$ is the empirical distribution of the training set $D = \{(\mathbf{x}_i, y_i) : i = 1, \ldots, m\}$. The first term of Eq. (2) is the likelihood of an example after passing it through the encoder distribution.

The second term penalizes the latent vectors based on how much their distribution differs from the prior and encourages disentanglement. As mentioned in Section 3, learning disentangled representations is impossible in the unsupervised i.i.d. setting [Locatello et al., 2019]. Following [Locatello et al., 2020a], and similarly to CBNMs, we assume access to a (possibly separate) data set $\widetilde{D} = \{(\mathbf{x}_\ell, \mathbf{g}_{I,\ell})\}$ containing supervision about the *interpretable* generative factors $\mathbf{G}_I$ and integrate it into the ELBO by replacing the per-example loss $\mathcal{L}$ in Eq. (2) with:

$$\mathcal{L}(\theta, \mathbf{x}, y; \beta) + \gamma \cdot \mathbb{E}_{p_{\tilde{D}}(\mathbf{x},\mathbf{g})} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\Omega(\mathbf{z}, \mathbf{g})\right] \tag{3}$$

where $\gamma > 0$ controls the strength of the concept-level supervision. Following Locatello et al. [2020a], the term $\Omega(\mathbf{z}, \mathbf{g})$ penalizes encodings sampled from $q_\phi(\mathbf{z} \,|\, \mathbf{x})$ for differing from the annotation $\mathbf{g}$. We implement this term using the average cross-entropy loss $\Omega(\mathbf{z}, \mathbf{g}) := -\sum_k g_k \log \sigma(z_k) + (1 - g_k) \log(1 - \sigma(z_k))$, where the annotations g_k are rescaled to lie in $[0, 1]$ and σ is the sigmoid.

In order to tackle concept leakage, GlanceNets integrate the open-set recognition strategy of [Sun et al., 2020]. This strategy identifies out-of-class inputs by considering the class prototype $\mu_y := \mathbb{E}_{p(\mathbf{z}|y)}[\mathbf{z}]$ in $\mathbb{R}^k$ defined by the prior distribution and the decoder $p_\theta(\mathbf{x}|\mathbf{z})$. During training, GlanceNets use the training

data to estimate: (i) a distance threshold η_y, which defines a spherical subset in the latent space $\mathcal{Z}_y = \{\mathbf{z} : ||\mu_y - \mathbf{z}|| < \eta_y \}$ centered around the prototype of class y, and (ii) a maximum threshold on the reconstruction error η_{thr}. If new data points have reconstruction error above η_{thr} or they do not belong to any subset $\mathcal{Z}_y$, they are inferred as open-set instances, i.e., $\hat{T} = 0$. This procedure is illustrated in Fig. 1. In practice, we found that choosing the thresholds as to include the 95% of training examples to work well in our experiments.

Benefits and limitations. GlanceNets can naturally be combined with different VAE-based architectures for learning disentangled representations Esmaeili et al. [2019], including β-TCVAEs [Chen et al., 2018], InfoVAEs [Zhao et al., 2019], DIP-VAEs [Kumar et al., 2018], and JL1-VAEs [Rhodes and Lee, 2021]. Since our experiments already show substantial benefits for GlanceNets building on β-VAEs [Higgins et al., 2016], we leave a detailed study of these extensions to future work.

Like CBNMs, GlanceNets foster alignment by leveraging supervision on the interpretable generative factors [Locatello et al., 2020a], possibly derived from an external data set [Koh et al., 2020]. However, GlanceNets can be readily adapted to a variety of different kinds of supervision used for VAE-based models, including *partially* annotated examples [Gabbay et al., 2021], group information [Bouchacourt et al., 2018], pairings [Shu et al., 2020, Locatello et al., 2020b] and other kinds of weak supervision [Gabbay and Hoshen, 2019, Chen and Batmanghelich, 2020], as well as feedback from a domain expert [Stammer et al., 2021].

One limitation inherited from VAEs by GlanceNets is the assumption that the interpretable generative factors are disentangled from each other [Suter et al., 2019]. In practice, GlanceNets work even when this does not hold (as in our even vs. odd experiment, see Section 5.2). However, one direction of future work is to integrate ideas from hierarchical disentanglement [Ross and Doshi-Velez, 2021].

5 Empirical Evaluation

In this section, we present results on several tasks showing that GlanceNets outperform CBNMs [Chen et al., 2020] in terms of alignment and robustness to leakage, while achieving comparable prediction accuracy. The details on the hardware, architectures and hyperparameters are collected in the **Supplementary Material**, which can be found in the arxiv repository for the paper https://arxiv.org/abs/2205.15612 .

5.1 Evaluating Alignment

In a first experiment, we compared GlanceNets with CBNMs on three classification tasks for which supervision on the generative factors is available. In order to evaluate the impact of this supervision on the different competitors, we varied the amount of training examples annotated with it from 1% to 100%. For each increment, we measured prediction performance using accuracy, alignment and explicitness using the lasso variant of DCI.

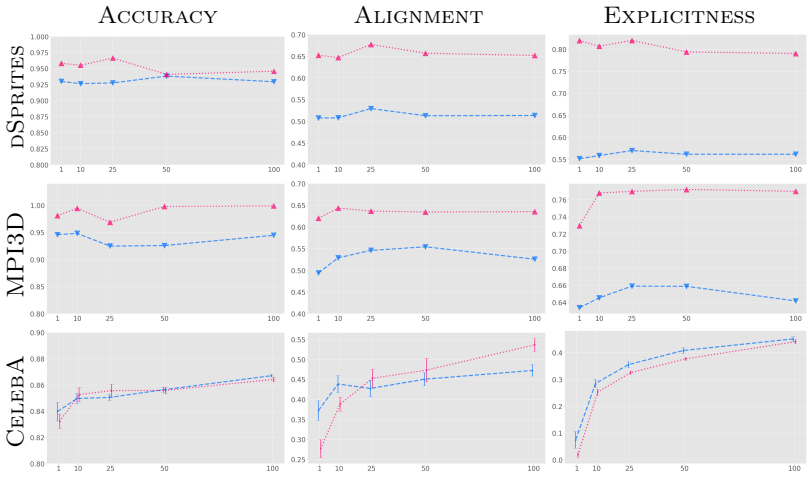

Fig. 2. GlanceNets are better aligned than CBNMs. Each row is a data set and each column reports a different metric. The horizontal axes indicate the % of training examples for which supervision on the generative factors is provided.

Data sets. We carried out our evaluation on two data sets taken from the disentanglement literature *dSprites* [Matthey et al., 2017] and *MPI3D* [Gondal et al., 2019], and a very challenging real world dataset, *CelebA-64* [Liu et al., 2015]. They all consist in 64×64 annotated images, with only one channel for dSprites and three for the others. For CelebA, since we are interested in measuring alignment, we considered only those 10 binary generative factors that CBNMs can fit well (in the Appendix). We also dropped all those examples for which hair color is not unique, obtaining approx. $127k$ examples. For dSprites and MPI3D, we used a random 80/10/10 train/validation/test split, while for CelebA we kept the original split Liu et al. [2015].

We generated the ground-truth labels y as follows. For dSprites, we labeled images according to a random but fixed linear separator defined over the *continuous* generative factors, chosen so as to ensure that the classes are balanced. For MPI3D and CelebA, we focused on the *categorical* factors instead. Specifically, we clustered all images using the algorithm of Huang [1997], for a total of 10 and 4 clusters for MPI3D and CelebA respectively, and then labeled all examples based on their reference cluster. This led to slightly unbalanced classes containing different percentages of examples, ranging from 5% to 16% in MPI3D and from 21% to 29% in CelebA.

Results and discussion. The results of this first experiment are reported in Fig. 2. All models were tested with as many latent components as the number of supervised generative factors for each dataset. The behavior of both competitors on dSprites and MPI3D was extremely stable, owing to the fact that these data sets cover an essentially exhaustive set of variations for all generative factors, so we report their hold-out performance on the test set. Since for CelebA the

variance was non-negligible, we ran both methods 7 times varying the random seed used to initialize the network and report the average performance across runs and its standard deviation.

In addition to alignment, we also report explicitness [Eastwood and Williams, 2018], which measures how well the linear regressor employed by DCI fits the generative factors. The higher, the better. Details on its evaluation are included in Suppl. Material.

The plots clearly show that, although the two methods achieve high and comparable accuracy in all settings, **GlanceNets** attain better alignment in all data sets and for all supervision regimes than **CBNMs**, with a single exception in CelebA using low values of supervision, for a total of 13 wins out of 15 cases. In all *disentanglement* data sets, there is a clear margin between the alignment achieved by GlanceNets and that of CBNMs: performances vary up to maximum of 15% in dSprites, and a minumum of 8% in MPI3D. In CelebA, the gap is evident with full supervision (almost 8% of difference in alignment), and GlanceNets still attain overall better scores in the 25% and 50% regime. On the other hand, performance are lower, but comparable, with 10% supervision. The case at 1% refers to an extreme situation where both CBNMs and GlanceNets struggle to align with generative factors, as is clear also from the very low explicitness. In dSprites and MPI3D, both GlanceNets and CBNMs quickly achieve very high alignment at 1% supervision, as expected Locatello et al. [2020a], whereas better results in CelebA are obtained with growing supervision. Furthermore, both models display similar stability on this data set, as shown by the error bars in the plot.

5.2 Evaluating Leakage

Next, we evaluated robustness to concept leakage in two scenarios that differ in whether the unobserved generative factors are disentangled with the observed ones or not, see Section 3. In both experiments, we compare GlanceNets with a CBNM and a modified GlanceNet where the open-set recognition component has been removed (denoted CG-VAE).

Leakage due to unobserved entangled factors. We start by replicating the experiment of Mahinpei et al. [2021]: the goal is to discriminate between even and odd MNIST images using a latent representation $\mathbf{Z} = (Z_4, Z_5)$ obtained by trained (with complete supervision on the generative factors) *only* on examples of 4's and 5's. Leakage occurs if the learned representation can be used to predict the remaining eight digits better than random. During training, we use digit labels for conditioning the prior $p(\mathbf{Z} \mid \mathbf{Y})$ of the GlanceNet.

Fig. 3 (a, b) illustrates the latent representations of the training and test set output by a GlanceNet: since the two digits are mutually exclusive, the model has learned to map all instances along the (z_4, z_5) diagonal. This is where open-set recognition kicks in: if an input is identified as open-set, the GlanceNet rejects it. In all leakage experiments, we implement rejection by predicting a random label. Since MNIST is balanced, we measure leakage by computing the difference in accuracy between the classifier and an ideal random predictor, i.e.,

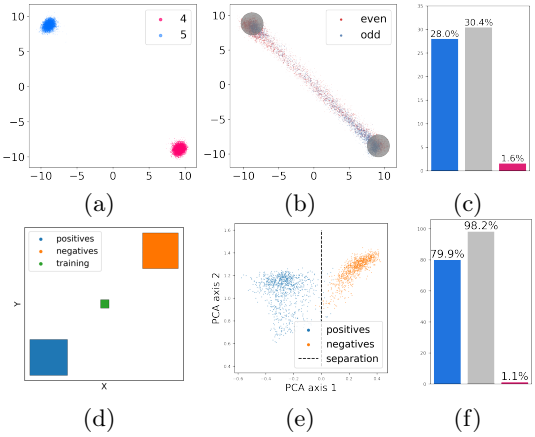

Fig. 3. GlanceNets are leak-proof. (a) MNIST training set embedded using GlanceNet; axes indicate z_4 and z_5 and color the concept label (4 and 5). (b) Latent representations of the test images, divided in even vs. odd. Every ball in light gray denotes the region $\mathcal{Z}_y$ for each class y. (c) Leakage % for CBNM, CG-VAE and GlanceNet. (d) dSprites: the variations over *pos_x* and *pos_y* for the training set, and for the test set, divided in positives vs. negatives. (e) PCA reduction for GlanceNet. (f) Leakage % for CBNM, CG-VAE and GlanceNet.

$2 \cdot |\text{acc} - \frac{1}{2}|$: the smaller, the better. The results, shown in Fig. 3 (c), show a substantial difference between GlanceNet and the other approaches. Consistently with the values reported in [Mahinpei et al., 2021], CBNMs are affected by a considerable amount of leakage, around 28%. This is not the case for our GlanceNet: most (approx. 85%) test images are correctly identified as open-set and rejected, leading to a very low (about 2%) leakage, 26% less than CBNMs. The results for CG-VAE also indicate that removing the open-set component from GlanceNets dramatically increases leakage back to around 30%.

Leakage due to unobserved disentangled factors. Next, we analyze concept leakage between *disentangled* generative factors using the dSprites data set. To this end, we defined a binary classification task in which the ground-truth label depends on *position_x* and *position_y* only. In particular, instances within a fixed distance from $(0, 0)$ are annotated as positive and the rest as negative, as shown in Fig. 3(a). In order to trigger leakage, all competitors are trained (using full concept-level supervision) on training images where *shape*, *size* and *rotation* vary, but *position_x* and *position_y* are almost constant (they range in a small interval around $(0.5, 0.5)$, cf. Fig. 3(d)). leakage occurs if the learned model can successfully classify test instances where *position_x* and *position_y* are no longer fixed.

For both competitors, we encode *shape* using a 3D one-hot encoding and *size* and *rotation* as continuous variables. During training, we use the *shape* annotation for conditioning the prior $p(\mathbf{Z} \mid \mathbf{Y})$ of the GlanceNet. The first two PCA components of the latent representations acquired by our GlanceNet are

shown, rotated so as to be separable on the first axis, in Fig. 3 (e): in particular, it is possible to separate positives from negatives based on the obtained representations in the five latent dimensions. As shown in Fig. 3 (f), this means that both CBNM and CG-VAE suffer from very large leakage, 80% and 98%, respectively. In contrast, open-set recognition allows GlanceNet to correctly identify and reject almost all test instances, leading to negligible leakage.

6 Related Work

Concept-based explainability. Concepts lie at the heart of AI [Muggleton and De Raedt, 1994] and have recently resurfaced as a natural medium for communicating with human stakeholders [Kambhampati et al., 2022]. In explainable AI, this was first exploited by approaches like TCAV [Kim et al., 2018], which extract local concept-based explanations from black-box models using concept-level supervision to define the target concepts. Post-hoc explanations, however, are notoriously unfaithful to the model's reasoning [Dombrowski et al., 2019, Teso, 2019, Sixt et al., 2020]. CBMs, including GlanceNets, avoid this issue by leveraging concept-like representations directly for computing their predictions. Existing CBMs model concepts using prototypes [Li et al., 2018, Chen et al., 2019, Rymarczyk et al., 2021, Nauta et al., 2021a] or other representations [Alvarez-Melis and Jaakkola, 2018, Koh et al., 2020, Losch et al., 2019, Chen et al., 2020], but they seek interpretability using heuristics, and the quality of concepts they acquire has been called into question [Nauta et al., 2021b, Hoffmann et al., 2021, Mahinpei et al., 2021, Margeloiu et al., 2021]. Our work shows that disentangled representation learning helps in this regard.

Disentanglement and interpretability. Interpretability is one of the main driving factors behind the development of disentangled representation learning [Bengio et al., 2013, Kulkarni et al., 2015, Chen et al., 2016]. These approaches however make no distinction between interpretable and non-interpretable generative factors and generally focus on properties *of the world*, like independence between causal mechanisms [Schölkopf et al., 2021] or invariances [Higgins et al., 2016]. Interpretability, however, depends on human factors that are not well understood and therefore usually ignored [Lipton, 2018, Miller, 2019]. The link between disentanglement and interpretability has never been made explicit. Importantly, in contrast to alignment, disentanglement does not require that the map between matching generative and learned factors preserves semantics. We remark that other VAE-based classifiers either do not tackle disentanglement or are unconcerned with concept leakage [van Steenkiste et al., 2019, Xu and Sun, 2016, Sun et al., 2020].

Disentanglement and CBMs. Neither the literature on disentanglement nor the one on CBMs have attempted to formalize the notion of interpretability or to establish a proper link between the latter and disentanglement. The work of Kazhdan et al. [2021] is the only one to compare techniques for disentangled representation learning and concept acquisition, however it makes no attempt at linking the two notions. Our work fills this gap.

Bibliography

David Alvarez-Melis and Tommi S Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 7786–7795, 2018.

Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: Deep learning for interpretable image recognition. *Advances in Neural Information Processing Systems*, 32: 8930–8941, 2019.

Max Losch, Mario Fritz, and Bernt Schiele. Interpretability beyond classification output: Semantic Bottleneck Networks. *arXiv preprint arXiv:1907.10882*, 2019.

Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1 (5):206–215, 2019.

Anita Mahinpei, Justin Clark, Isaac Lage, Finale Doshi-Velez, and Weiwei Pan. Promises and pitfalls of black-box concept learning models. In *International Conference on Machine Learning: Workshop on Theoretic Foundation, Criticism, and Application Trend of Explainable AI*, volume 1, pages 1–13, 2021.

Andrei Margeloiu, Matthew Ashman, Umang Bhatt, Yanzhi Chen, Mateja Jamnik, and Adrian Weller. Do concept bottleneck models learn as intended? *arXiv preprint arXiv:2105.04289*, 2021.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1757–1772, 2012.

Subbarao Kambhampati, Sarath Sreedharan, Mudit Verma, Yantian Zha, and Lin Guan. Symbols as a Lingua Franca for Bridging Human-AI Chasm for Explainable and Advisable AI Systems. In *Proceedings of Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3): 31–57, 2018.

Peter Hase and Mohit Bansal. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, 2020.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012.

Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf, and Stefan Bauer. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *International Conference on Machine Learning*, pages 6056–6065. PMLR, 2019.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. 2017.

Abbavaram Gowtham Reddy, L Benin Godfrey, and Vineeth N Balasubramanian. On causally disentangled representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

Aviv Gabbay, Niv Cohen, and Yedid Hoshen. An image is worth more than a thousand words: Towards disentanglement in the wild. *Advances in Neural Information Processing Systems*, 34, 2021.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

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

Julian Zaidi, Jonathan Boilard, Ghyslain Gagnon, and Marc-André Carbonneau. Measuring disentanglement: A review of metrics. *arXiv preprint arXiv:2012.09276*, 2020.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pages 4114–4124, 2019.

Francesco Locatello, Michael Tschannen, Stefan Bauer, Gunnar Rätsch, Bernhard Schölkopf, and Olivier Bachem. Disentangling factors of variations using few labels. In *International Conference on Learning Representations*, 2020a.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International conference on machine learning*. PMLR, 2014.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*. PMLR, 2014.

Xin Sun, Zhenning Yang, Chi Zhang, Keck-Voon Ling, and Guohao Peng. Conditional gaussian distribution learning for open set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13480–13489, 2020.

Diederik P Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

Babak Esmaeili, Hao Wu, Sarthak Jain, Alican Bozkurt, Narayanaswamy Siddharth, Brooks Paige, Dana H Brooks, Jennifer Dy, and Jan-Willem Meent. Structured disentangled representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2525–2534. PMLR, 2019.

Ricky TQ Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in vaes. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 2615–2625, 2018.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Balancing learning and inference in variational autoencoders. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):5885–5892, Jul. 2019. https://doi.org/10.1609/aaai.v33i01.33015885. URL https://ojs.aaai.org/index.php/AAAI/article/view/4538.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*, 2018.

Travers Rhodes and Daniel Lee. Local disentanglement in variational auto-encoders using jacobian l_1 regularization. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=8xyNqPvFZwC.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2016.

Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly supervised disentanglement with guarantees. In *International Conference on Learning Representations*, 2020.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, pages 6348–6359. PMLR, 2020b.

Aviv Gabbay and Yedid Hoshen. Latent optimization for non-adversarial representation disentanglement. *arXiv preprint arXiv:1906.11796*, 2019.

Junxiang Chen and Kayhan Batmanghelich. Weakly supervised disentanglement by pairwise similarities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3495–3502, 2020.

Wolfgang Stammer, Marius Memmel, Patrick Schramowski, and Kristian Kersting. Interactive disentanglement: Learning concepts by interacting with their prototype representations. *arXiv preprint arXiv:2112.02290*, 2021.

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

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Muhammad Waleed Gondal, Manuel Wuthrich, Djordje Miladinovic, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/d97d404b6119214e4a7018391195240a-Paper.pdf.

Zhexue Huang. Clustering large data sets with mixed numeric and categorical values. In *In The First Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 21–34, 1997.

Stephen Muggleton and Luc De Raedt. Inductive logic programming: Theory and methods. *The Journal of Logic Programming*, 19:629–679, 1994.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. *Advances in Neural Information Processing Systems*, 32:13589–13600, 2019.

Stefano Teso. Toward faithful explanatory active learning with self-explainable neural nets. In *Proceedings of the Workshop on Interactive Adaptive Learning (IAL 2019)*, pages 4–16, 2019.

Leon Sixt, Maximilian Granz, and Tim Landgraf. When explanations lie: Why many modified bp attributions fail. In *International Conference on Machine Learning*, pages 9046–9057. PMLR, 2020.

Dawid Rymarczyk, Łukasz Struski, Jacek Tabor, and Bartosz Zieliński. ProtoPShare: Prototypical Parts Sharing for Similarity Discovery in Interpretable Image Classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, page 1420–1430, 2021.

Meike Nauta, Ron van Bree, and Christin Seifert. Neural prototype trees for interpretable fine-grained image recognition. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pages 14933–14943, 2021a.

Meike Nauta, Annemarie Jutte, Jesper Provoost, and Christin Seifert. This looks like that, because... explaining prototypes for interpretable image recognition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 441–456. Springer, 2021b.

Adrian Hoffmann, Claudio Fanconi, Rahul Rade, and Jonas Kohler. This looks like that... does it? shortcomings of latent space prototype interpretability in deep networks. *arXiv preprint arXiv:2105.02968*, 2021.

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

Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. *Advances in neural information processing systems*, 28, 2015.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

Sjoerd van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. Are disentangled representations helpful for abstract visual reasoning? In *NeurIPS*, 2019.

Weidi Xu and Haoze Sun. Semi-supervised variational autoencoders for sequence classification. *ArXiv*, abs/1603.02514, 2016.

Dmitry Kazhdan, Botty Dimanov, Helena Andres Terre, Mateja Jamnik, Pietro Liò, and Adrian Weller. Is disentanglement all you need? comparing concept-based & disentanglement approaches. *arXiv preprint arXiv:2104.06917*, 2021.