Interactive Disentanglement: Interactive Concept Learning via Prototype Representations

Anonymous CVPR submission

Paper ID *****

Abstract

Learning visual concepts from raw images without strong supervision is a challenging task. In this work, we show the advantages of prototype representations for understanding and revising the latent space of neural concept learners. For this, we introduce the interactive Concept Swapping Network (iCSN), a novel method for learning concept-grounded representations via weak supervision and implicit prototype representations. By swapping the latent representations of paired images the iCSN learns to bind conceptual information to specific prototype slots. This semantically grounded and discrete latent space of iCSN is advantageous for human understanding and human-machine interaction. We support our claims by conducting experiments on a novel data set for Elementary Concept Reasoning (ECR) that focuses on learning visual concepts shared by single objects.

1. Introduction

TODO: Still very preliminary

TODO: remark how we use interpretable/ what we mean by this TODO: definition of concepts used in this work –¿ see wikipedia article on concepts

A constraint is worth 1000 images.

"Interpretability via Interactions"

Imagine asking your 3-year-old self the following question: What do a green triangle and a red triangle have in common? Your younger version could most certainly answer the question with ease and perform this kind of comparison for most given sets of objects (assuming you can appropriately convey the question). Furthermore, upon making mistakes it would also react to your feedback.

For machines, however, solving this noticeably simple task is still error-prone. Therefore, recent approaches aim to get insight into the decision process and allow user feedback to improve the models' outcomes. Specifically, Neuro-Symbolic models are becoming more fashionable due to their promising capabilities for interpretability and interaction. Common approaches use deep learning models to process raw inputs, e.g., images, into symbolic representations that symbolic-based algorithms can then process further. Teaching those models to learn the relevant concept sets via supervision comes with extensive knowledge about the intermediate symbolic representation for the user. This knowledge makes it easier to interact and revise the model and allows for downstream tasks like planning or reasoning. However, a supervised approach to learning the symbolic representations requires a strong prior on the set of relevant concepts. The difficulty remains in learning human-interpretable symbolic spaces with reduced supervision allowing for more flexible symbolic spaces.

These properties are of particular interest when less powerful priors are available while human interaction and interpretability are required. TODO: add citations and check soundness of my rewrite [7, 9, 26]

One way to model the symbolic representations is to use a prototypical latent space, i.e., a set of representations constraining the model to a fixed number of concepts. To minimize an error function, the model has to identify the concepts that explain the most variability of the data, i.e., the most important concepts shared by the training examples. TODO: paragraph on prototype learning from psychology studies

In this paper we investigate the advantages of prototype representations in learning human-understandable, symbolic latent representations and consequently creating human-revisable concept learners. Particularly, we consider the setting of weak supervision as a more realistic scenario. We answer the question whether a machine can learn concepts from data without explicit feedback on the identifiable concepts. More specifically, we derive the following subquestions:

- Is weak supervision, i.e., no explicit concept supervision, sufficient for training a concept learner that provides discrete symbols to specific concepts?
- How can human users understand the symbolic space

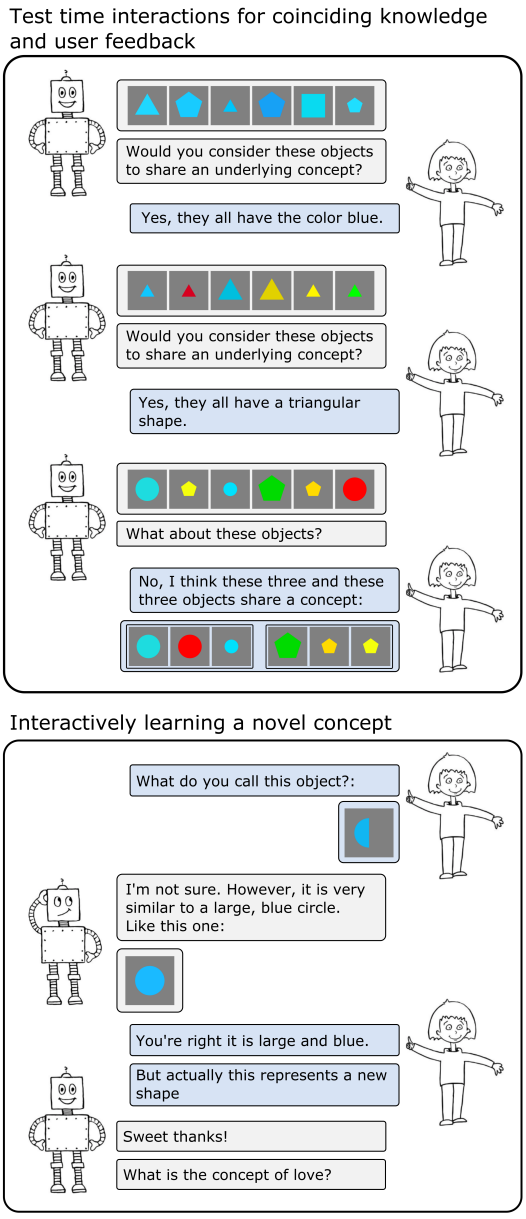

Figure 1. The trained model can now probe the human user if the concepts that it has extracted from the weaker supervision coincide with the knowledge of the user. TODO: add prototype to explanation here

of such a concept learner?

- And finally, does human interaction with the learner allows for more refined concepts or learning of novel concepts?

TODO: check revised questions. I tried to formulate them in a more concrete way. Also moving away from whether we can create such a learner to the capabilties of such.

The modularity of the model has key advantages in a continual knowledge learning setting, making it easy to add additional prototype representation for representing a novel attribute, but also allowing to add a novel category MLP in case a novel category should be extracted from the data.

We argue it is easier to explain and reason with discrete, prototypical representations and explanations, rather than presenting the interpolation spaces of a distribution. The model that is introduced is termed Concept Swapping Network (CSN).

In our work we show the advantages of using prototype representations in terms of (1) learning an understandable and communicatable latent concept representation, (2) revising the concept representations via interactions with a human user and (3) updating the concept representations in a continual learning fashion (a la ...).

Through the competition via softmax for the prototype codes the model learns to bind similar information to a single neural representation. In this way the continuous latent space of the encoder is discreteized via the read out layers and the prototypes and the continuous space is divided into meaningful subspaces as communicated via the interactive weak supervision/ match pairing. This is related to prototype learning observed in the psychology literature.

To sum up, this work makes the following contributions:

- Learn discrete concepts through weak supervision,
- Through the ease of interaction allow a human user to understand the concept space post learning
- The user can interactively revise and update the concept representations

2. Related Work

Concept Learning: The concept-based approaches we focus on aim to identify human-understandable explanations that models use to make predictions. The most common methods integrate a bottleneck into the model that predicts desired high-level concepts. The model then communicates these concepts to the user and as input for a downstream classification task. While these classification-based methods require learning, there also exist algorithmic solutions based on clustering [12]. Applications for identified concepts include explainability of the model's predictions, representation learning, as well as a better generalization for few-shot tasks that can reuse the learned representations [5].

To communicate the concepts to the user, Ciravegna et al. [7] introduce logic explained networks that provide human-understandable explanations through first-order logic formulas. Another way is to impose strong priors on the concept, assuming them to be user-defined and therefore also a priori human-interpretable [22]. Quantification of the concept importance then acts as a feature to solve a downstream classification task. To over-

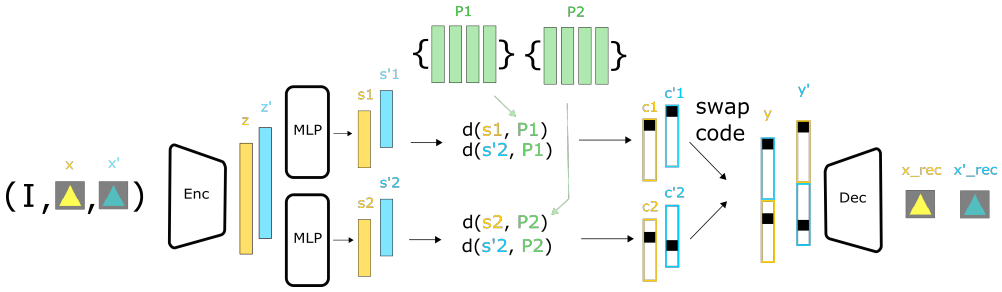

Figure 2. **Concept Swapping Network (Concept Swapping Network (CSN)).** TODO: old figure, must update

come the scalability issues posed by an a priori definition of the concepts, leveraging datasets to learn the intermediate multi-label prediction has proven to be successful [1, 26, 28]. However, these methods are still constrained by label-introduced supervision. Losch et al. [26] propose an unsupervised approach that draws the bottleneck outputs towards a one-hot-encoding, eliminating the need for predefined concepts. Without any additional priors or implicit architecture choices, the learned concepts are still difficult to interpret by humans. Furthermore, the joint training biases the representations towards the downstream classification task and hinders the transfer to, e.g., more complex reasoning tasks.

Concept Representations: The term concept is rooted in psychology and which defines it as a discrimination of stimuli that belong to one concept from stimuli belonging to another one [2]. Most common methods to represent concepts are exemplars and prototypes. While the former model assumes that multiple examples are maintained in memory, the latter only assumes an average representation, i.e., the most typical or representative member of a category, over the exemplars, a so called prototype [30, 31, 35]. In the field of cognitive psychology the lines between exemplar and prototype representation are more blurred and their contribution to concept representations it is still an open problem with recent work hinting a simultaneous use of both representation [4, 11]. Smith et al. [33] found that organism receive more informative signals about category belongingness when averaging exemplar experiences into prototypes, while Jäkel et al. [20] leverage their insights from machine learning to show generalization ability of exemplar models of categorization. Nonetheless, there is evidence of the use and importance of prototypes in the human memory system [8, 13, 19]. TODO: maybe some take on de-/composition? IbbotsonT2009, ... Inspiration sparked by such findings gave rise to neural network based Prototype Learning Systems [37] that use prototype vectors as internal concept representations which can be converted into explainable visualizations via decoding [23] or finding the closest, i.e., most similar training example [**?**, 6]. To make classifications, the authors compute the similarity of the encoded input to the prototypes via a distance metric. Intuitively, prototypes than can be visualized provide a more informative explanation than intermediate classifications. However, these networks still require a classification task to train good representations.

Disentanglement: A closely related field is disentanglement. Though the term disentanglement rarely appears in the context of interpretability, its goal is to extract the independent underlying factors that are responsible for generating the data [29]. Recently, through the work of Locatello et al. [24] much of disentanglement research has shifted from unsupervised learning to the weakly supervised setting. Shu et al. [32] showed in their work that supervision via match pairing for a known subset of factors gives guarantees for disentanglement via their defined calculus of consistency and restrictiveness. We define concept learning as identifying the properties that describe an entity, such as the color or shape of an object, or the symptom of a disease in the hyperspectral reflectance of leaf tissue [34]. The resulting goals are therefore closely linked to the ones of disentanglement research. Since disentanglement methods commonly utilize reconstruction and implicit probabilistic priors, they pose many benefits when approaching representation learning with less supervision. Literature on disentanglement also extends to group-based disentanglement, allowing for grouping of the identified generative factors [3, 15, 16, 18, 36, 38]. Such a grouping offers advantages in the communication with users who need to cluster peculiarities based on their similarities to interpret concepts correctly. For example, a peculiarity *light red* is impossible to interpret without prior knowledge, while grouping it together with *dark red* gives a better understanding of the underlying concept of *color saturation*.

TODO: self-supervised learning for vision should be covered by disentanglement literature, unless you want to mention it to justify the swapping + recon idea?
TODO: literature on interactive learning?-¿ maybe from StammerSK2021?
TODO: prototype literature from psychology
TODO: go through mattermost channel and dump all relevant papers we found there here

3. Concept Swapping Networks

In this section we explain how to learn implicit prototype representations via the CSN. For an overview see Fig. 2.

In this work we use the weakly-supervised pair matching setting [32] for interactively and iteratively learning properties from data. Our model consists of a deterministic autoencoder structure. However, rather than reconstructing from the encoding directly we apply several read-out encoders, each responsible for reading out the relevant information from the entangled latent space that corresponds to the information of one category, e.g. color. In this way each read-out encoder ideally outputs a latent vector containing the information of a single basic concept of an object. How this is enforced will be discussed below. Each split latent vector is then compared via a softmax-weighted dot-product to a codebook of prototype vectors, with each superordinate concept hereby possessing its own prototype codebook. The distance codes over all categories are discretized via a weighted softmax, enforcing competition and thereby the binding of semantic information to specific prototype vector. Finally the discrete distance codes are concatenated and passed to a decoder, which reconstructs the image from this sparse code.

Prototype-based concept architecture. Assume an input $x_i \in X$, whereby $X := [x_1, ..., x_N] \in \mathbb{R}^{N \times M}$. For simplicity we denote here with x_i an entire image, however x_i can also contain the representation of a subregion of the image extracted via a preprocessing step (e.g. only represent an object). Each x_i contains several attributes such as color, shape and size. Specifically, we refer to each realisation of these attributes, e.g. a "blue color" or "triangular shape" as a (basic) concept, whereas the term "color" is referred to as category concept in this work (often called superordinate concept in the field of cognitive and psychological sciences [10]). Thus each image x_i has corresponding groundtruth basic concepts $c_i := [c_i^0, c_i^1, ..., c_i^J]$, with J the number of category/superordinate concepts. For simplicity we assume here that each superordinate concept contains the same number of basic concepts, K, though in practice this can differ.

The input encoder, h, receives the input image and encodes it into a latent representation, $h(x_i) = \hat{z}_i$, with $\hat{z} \in \mathbb{R}^{N \times Z}$. However, rather than reconstructing from $\hat{z}_i$ directly the CSN applies several read-out encoders, $m^j(\hat{z}_i) = \phi_i^j \in \mathbb{R}^{N \times Q}$. Each read-out encoder is hereby responsible for extracting that relevant information from the entangled latent space, $\hat{z}_i$, which corresponds to the information of a superordinate concept, e.g. color. We refer to each ϕ_i^j as a concept embedding. How the extraction of concept specific information is enforced will be discussed below. Importantly, ϕ_i^j does not require to be fully disentangled.

A central component of the CSN is a set of codebooks of prototype vectors, $\Theta := [\mathrm{P}^0, \mathrm{P}^1, ..., \mathrm{P}^J]$, with each codebook $\mathrm{P}^j := [p^{j0}, p^{j1}, ..., p^{jK}] \in \mathbb{R}^{K \times Q}$ consisting of an ordered set of randomly initialised prototype vectors. Each concept encoding, ϕ_i^j, is now assigned to one of the prototype vectors (slot) using a softmax over dot products between the concept encoding and the prototype slots of P^j:

$$s_i^{jl} = \frac{\exp(\phi_i^j \cdot p^j l / \sqrt{S})}{\sum_{k=1}^{K} \exp(\phi_i^j \cdot p^{jk} / \sqrt{S})}.$$

Thus giving a similarity score for each prototype slot of category j. To enforce further discretization and binding of specific concepts to prototype slots a further weighted softmax term is applied to the similarity scores, yielding $\Pi_i^{jk} = \sigma(\frac{1}{\tau} \cdot s_i^{jk})$, with $\sigma(\cdot)$ representing the softmax funciton, $\Pi_i^j \in \mathbb{R}^{Q \times K}$ and $\tau \in \mathbb{R}^+$. In our experiments we decrease τ stepwise so as to gradually enforce the binding of information. In the extreme case of a low τ, Π_i^j will resemble a one-hot vector, indicating which prototype slot of the category j the concept encoding ϕ_i^j is most similar to.

Finally, the weighted similarity scores of each category are concatenated to a single vector, $\hat{y}_i \in [0, 1]^{J \mathrm{x} K}$, which is passed to the decoder to reconstruct the image, $g(\hat{y}_i) = \hat{x}_i \in \mathbb{R}^M$.

Concept swapping and training procedure. Prior to training, i.e. after initialisation there is no semantic knowledge bound to the prototype slots, each is just as meaningless as the other. The concept semantics found in a trained CSN is learnt via a weakly-supervised training procedure and a simple swapping trick.

A human user trains this model via pairs of data samples that share a property of a known category. We thus perform two parallel passes through the network an hereby receive distance codes for each category for each data sample. In order to now enforce the binding of information of specific attributes to an internal representation the model swaps the codes of the shared categories. This has an attractive semantics in that, if two independent entities share a common property, exchanging the internal representation of that specific property from one entity with the corresponding internal representation of the second entity should not result in a loss of information in representing each entity, respectively.

The model is finally trained solely on reconstructing the input data from the distance codes, in comparison to previous works that enforce learning semantic knowledge via an additional consistency loss TODO: cite. By decreasing the softmax temperature in a step-wise fashion one can enforce the binding to information to specific prototype vectors such that each attribute of a category is mapped to an exclusive prototype vector. Through this discretization process the decoder learns to produce reconstructions that correspond to the prototypical representations that are present in the data. For example, given a pair of images of blue objects that vary in the shade of blue, our model would learn

to map the color information of these objects to the same prototype, thus learning the prototypical blue that is present within the data. This is a key difference to the various VAE approaches, which try to learn the continuous latent space of the underlying factors.

Finally, rather than discarding the decoder, the model can now argue why it believes specific properties to be present, by presenting the prototypical example of each property that the input sample is closest to. By querying these prototypical reconstructions the human user can confirm if the predicted properties make sense or possibly detect unwanted model behaviour.

Interacting with CSNs Interactive Concept Swapping Network (iCSN)

Philosophical stuff The modularity of the model has key advantages in a continual knowledge learing setting, making it easy to add additional prototype representation for representing a novel attribute, but also allowing to add a novel category MLP in case a novel category should be extracted from the data.

There are further key advantages interaction-wise. Through the read-out structure of the initial latent space, which can be pretrained unsupervisedly such that it should contain all the relevant information that describe the data, a human user can iteratively add a read-out layer to add extracting the relevant information as queried for by paired samples.

Training only via reconstruction of all images. Semantic of learning via swapping implicitly leads to binding of relevant information, i.e. model learns to extract information from one image that it can use to represent parts of the second image.

Through the competition via softmax for the prototype codes the model learns to bind similar information to a single neural representation. In this way the continuous latent space of the encoder is discreteized via the read out layers and the prototypes and the continuous space is divided into meaningful subspaces as communicated via the interactive weak supervision/ match pairing. This is related to prototype learning observed in the psychology literature.

4. Elementary Concept Reasoning Data Set

As we are interested in an object-centric concept learning we provide a novel benchmark data set, called Elementary Concept Reasoning data set (ECR). This data set consists of RGB images of 2D geometric objects on a constant background. The objects can vary in shape (circle, triangle, square and pentagon), size (large and small) and color (red, green, blue, yellow). Notably, a uniform jitter is added to each color, thus producing various shades of each color.

Furthermore, the ECR data set consists of pairs of images following the match pairing setup of [32]. In this way two images are paired if the objects in the individual images share at least one common property, but at most two properties.

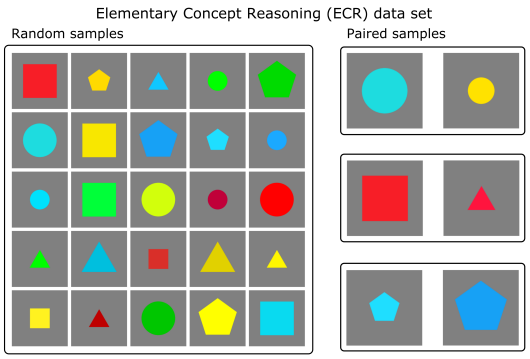

Figure 3. Samples of Elementary Concept Reasoning (ECR) data set. Each image of the data set (left) depicts a centered 2D object that possess three different properties: color, shape and size. Uniform color jitter is added to the object colors. Each image is paired with another image whose object shares between one and two concepts with the first object (right). (Best seen in color.)

Fig. 3 shows example images of ECR with randomly sampled images of the data set on the left, exemplifying the shape, size and color properties and variability of the objects. The images on the right present possible pairs of the data set. An important feature of ECR is that although there exist various shades of all colors, they all map to four discrete color names. Notice e.g., the color difference of the two paired blue objects in the right part of the figure. Although both objects present different shades of blue they are paired to represent the same distinct shape and color concept.

The ECR data set consists of a training and validation set, with the training set containing 5000 pairs of images and the validation set containing 2000 images.

5. Results

In this section we show the advantages of implicit semantic binding of concept representations via Concept Swapping Networks. We begin by showing the sparsity and semantics of the latent space learnt via CSN. Next, we show that due to the discretized latent space the model can communicate with a human user what the concepts are that it has learned to extract from the data. Next we show with a case study that via simple feedback a human user can easily revise the model's latent concept space. Lastly, we show that our model can naturally bind novel concepts into its concept space via human interactions.

5.1. Experimental Details

For our experiments we compared CSNs to several baseline methods. Specifically, we trained a variational autoencoder as in [25] (VAE) using the arithmetic mean of the

encoder distributions of [17]. As the experiments are run with the match pairing setting, in contrast to [25], rather than heuristically estimating which factors are being shared, we also provide the identifiers of shared factors when using this method for a fair comparison. The second baseline is the same setup as for the VAE model, however using a categorical distribution (Cat-VAE) via the Gumbel-softmax trick [21,27], rather than a normal distribution. Also here an averaging over encoder distributions is performed as in [17]. Lastly, we also provide results from training a β-VAE [14] in an unsupervised setting, i.e. without match pairing. We denote with VAE and Cat-VAE match pairing trained variaitonal autoencoders, in contrast to the β-VAE which was trained unsupervisedly. TODO: add information on number of variables for each model

The Concept Swapping Networks were trained via a simple reconstruction loss as detailed above. Lastly, the softmax temperature over distances to concepts was regularly decreased every 1000 epochs over $\tau \in [2, 0.5, 0.1, 0.01, 0.001, 0.0001, 0.00001]$. We trained the CSNs with several configurations. The standard setting, denoted simply as CSN below corresponds to an overestimated number of codes per superordinate concept, i.e. six prototypical codes for shape, size and color. CSN (w. fb.) denotes the case when additional human feedback has been given to the latent concept space of the standard setting. Finally CSN (w. pk.) denotes the case where the exact number of prototype codes was given prior to training, i.e. four shape, four color and two size prototype codes.

We trained all methods with each five random seed initializations and present the mean and standard deviations of these runs for all metrics. Further hyperaparameters and architectural details can be found in the Appendix.

5.2. Semantics in the latent Space

Reduced Code Variance. In order for a human user to understand and importantly interact via the latent space of a model there must exist a certain consistency in the representation of individual concepts. In other words the representation of the color blue should be consistent and specific to this concept and thus the latent representation for the blue color of one object should be the same or at least very similar compared to that of a second blue object. If this is not the case it remains difficult for a human user to identify learnt concepts and possibly interact on these.

We therefore begin by investigating the variance of latent representations of each method in consideration of the ground truth concept labels. Thus, given the ground truth multi-label information of each validation image and the factor identifier that the model had been trained with, how large is the variance in the code of each model. For this the latent codes of all validation images were stored and grouped according the the ground truth labels, such that all

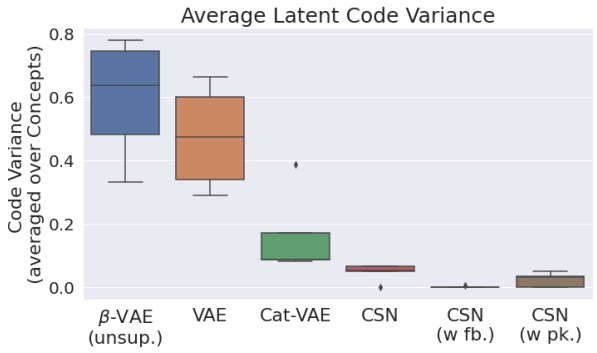

Figure 4. Average latent code variance given groundtruth concept labels for different model types and training settings. The models compared: unsupervisedly trained β-VAE, match pairing trained VAE and categorical VAE, the novel CSN method with an overestimated number of concepts, CSN with additional feedback (w. fb.) on the learnt concept space and CSN with prior knowledge (w. pk.) (exact number of concepts). Note here a lower variance is desirable.

the latent codes of all blue objects of the validation set were grouped. Given the shared factor id that was given during training to identify the color factor the variance was computed over all blue object latent codes TODO: need math notation here. This variance was calculated and finally averaged over all concepts, giving the average code variance per concept representation for an individual model.

The resulting code variance over the β-VAE, VAE, Cat-VAE and CSN runs can be seen in Fig. 4. Note that a reduced code variance is desirable and gives an indication of ow well a concept has been mapped to a distinct representation. The results of Fig. 4 indicate that in fact the variance of the latent space is much more reduced using the CSN method than the baseline methods. This gives an indication that the latent concept space of a CSN is more consistent than that of the baseline methods. Obviously reduced code variance is not a sufficient feature for concept consistency and human understandability, e.g. a model that has learnt to map all concepts to a single representation will have a zero latent code variance. To investigate the latent concept space further in the next section we turn to linear probing.

Probing the Concept Space. Similar to works of the self-supervision community TODO: cite we investigate the latent code of each model via linear probing. For this we performed single label classification via a decision tree (DT) as well as logistic regression (LR) model on the latent codes of each model. For this the multi-label ground truth labels were converted to single labels, e.g. large blue squares were mapped to a single integer, resulting in 32 individual class labels. Both the DT and LR models were trained on the predicted latent codes of a separate training set, prior to classifying the latent codes of the validation set. Details on the

	DT	LR
β-VAE (unsup.)	88.98 ± 11.53	73.86 ± 18.07
VAE (weakly)	92.18 ± 6.26	$\circ$ $\mathbf{97.8 \pm 1.09}$
Cat-VAE (weakly)	71.8 ± 11.23	90.5 ± 13.39
CSN	$\circ$ $\mathbf{95.43 \pm 8.79}$	95.4 ± 8.83
CSN (w. pk.)	80.11 ± 11.72	80.11 ± 11.72
iCSN	$\bullet$ $\mathbf{99.82 \pm 0.36}$	$\bullet$ $\mathbf{99.82 \pm 0.36}$

Table 1. Linear probing via decision tree (DT) and logistic regression (LR) on the latent codes of different model types. The best ("$\bullet$") and runner-up ("$\circ$") results are bold. The classification accuracies were computed on a held out test set.

model details can be found in the Appendix.

The results can be seen in Tab. 1 showing the average accuracy and standard deviation over the five random initializations for β-VAE, VAE, Cat-VAE and CSN. One can observe that the latent code of CSNs can be classified quite easily with linear models, being surpassed only by the match pairing trained VAE. Indeed the unsupervisedly trained β-VAE performs than the weakly-supervised models. Interestingly, although also using discrete latent representations Cat-VAE performs worse than CSN. On inspection we noticed that several Cat-VAE runs converged to suboptimal states.

Explaining learnt concepts. Through the swapping approach of CSN the model implicitly learns to bind specific conceptual information to an individual prototype code. In this way semantic information is discretely and consistently bound to specific representations. After training with weak supervision it is very recommendable for a machine and human user to discuss over the concepts that the machine has learnt, whether these coincide with the knowledge of the user or, possibly if a revision is necessary to correct for suboptimal concept distributions.

Through the implicit semantic binding of visual concepts to discrete symbols a CSN can automatically, at test time, group together novel images according that share a single concept, according to the models representations, and inquire a human user if indeed the presented images share a common underlying concept (*cf.* Fig. 1 (top)). In this way it can also ask what the human user might call the concept.

Revising the latent space. The greater advantage, however lies therein that a human user can easily identify suboptimal concept encodings (*cf.* Fig. 1 (top)) and, upon identifying the correctly learned or falsely learned concepts, simply interact via the semantic and discrete concept space of the CSN, via logical rules. Due to the discretized latent space of CSNs a human user can communicate via one-hot encodings. For example a possible form of interaction would be to merge concepts that were distributed during training of a CSN over several prototype codes to a single one. On the other hand a user can simply state that a specific prototype slot should not be used.

By giving simple feedback to the CSN runs of Tab. 1, we could strongly improve the concept space in terms of low variance and linear probing accuracy. Specifically, for all runs we identified the prototype slots that in the majority of samples were used for binding to a specific concept. As the model was provided with an overestimated number of prototype slots it could, e.g. learn to bind the color blue over several slots. Simply by adding feedback such as $\neg$slot1, causing the network to bind the concept that had among others been bound to slot1 to other slots. In addition, a single run had converged to a suboptimal space, where pentagons and circles were encoded as a same shape concept. In this case we gave feedback to all pentagon samples of the training set to bind to an otherwise empty prototype slot.

Interactively learning novel subordinate concept.

Due to the semantically-binded, discrete latent representation, CSNs have an interesting property for online learning settings, i.e. where a machine comes in contact with novel concepts which it has not encountered before. Due to the decoder structure of the current CSN thresholding the reconstruction error can be used as a heuristic for identifying whether the model has a good representation of a novel sample. In case the model is being presented a novel concept it can utter this uncertainty to the human user and present what it considers the sample to present.

In terms of teaching a machine a novel subordinate concept like a novel shape attribute, "halfcircle" (*cf.* Fig. 1 (bottom)) all a user has to do is identify a so far unbound prototype slot and via a l2 loss encourage the binding to this slot. To prevent the forgetting of already learnt concepts, the CSN can 'self-supervise' itsself in the sense that it can predict all other concepts of all novel samples and use its own predictions for preventing forgetting.

Tab. 2 (top) shows the linear model accuracies on the latent space of CSN models (denoted as *prior*) that had been presented a data set that contained the novel halfcircle concept. After providing the simple user feedback to bind the shape information to a so far empty prototype slot (denoted as *posterior*) the accuracy strongly increases.

Interactively learning novel superordinate concept.

In the previous experiment, we observed how to add the knowledge of a novel subordinate concept, by binding the novel concept to a specific prototype slot via a simple l2 loss. Next we showcase how to easily add a novel superordinate concept, e.g. a novel higher order concept of spots on ECR objects.

For this setting we make use of a variation of the ECR data set. Specifically, in this version roughly half of the objects contain a small white spot in their middle. The other half depict a solid color as in the original ECR data set.

	DT	LR
Novel Subordinate Concept		
CSN (prior)	$80.95 \pm -$	$80.2 \pm -$
CSN (post.)	$100 \pm -$	$100 \pm -$
Novel Superordinate Concept		
CSN (prior)	$- \pm -$	$- \pm -$
CSN (post.)	$- \pm -$	$- \pm -$

Table 2. Linear probing via decision tree (DT) and logistic regression (LR) on the latent codes of CSN. The models were evaluated with a ECR data set containing a novel shape that was not seen during training (prior). A CSN model, given human feedback on the novel concept, can easily incorporate this information into its latent representations (post.). The classification accuracies were computed on a held out test set. TODO: use median instead of mean?

	DT	**LR**
Cat-VAE w. swapping	$- \pm -$	$- \pm -$
CSN w. averaging	$- \pm -$	$- \pm -$

Table 3. Linear probing via decision tree (DT) and logistic regression (LR) on the latent codes of different model types.

We now simulate an online learning setting in which the concept of spots is unimportant in the initial training phase, however reconsidered important in a second round of interaction. The beauty of the modularity of CSNs is that one can easily add a new read-out MLP that thus learns to extract the relevant information of a novel superordinate concept from the latent space z without the danger of needing to retrain the representation of all other concepts.

The results of this can be seen in Tab. 2 (bottom)... TODO: run experiments and finish text

5.3. Ablation Studies

We ran ablation studies on the importance of the concept swapping as well as the discretization procedure via prototype slots versus variational categorical distributions. For this we performed linear probing on the latent space of a pair-trained Cat-VAE with parameter swapping, rather than averaging, and CSN with averaging of the concept embeddings rather than swapping. The results can be found in Tab. 3 showing ...

6. Conclusion

A necessary future path to investigate is how to allow for continuous and discrete variables in the encoder-decoder structure, thus allowing for only certain discrete variables to bee learnt, however allowing for continuous variables that still allow to encode evrything necessary that is not covered by the discrete variables, but required for proper reconstruction. Additionally it would be very interesting to identify the factor ids between paired images within a causal framework, e.g. in an active environment a child can identify that the shade of red and blue objects changes when placed under a different light source, however their size and shape remain the same. "Interpretability via Interactions"

References

[1] Catarina Belém, Vladimir Balayan, Pedro Saleiro, and Pedro Bizarro. Weakly supervised multi-task learning for concept-based explainability. *arXiv preprint arXiv:2104.12459*, 2021. 3

[2] Marc D. Binder, Nobutaka Hirokawa, and Uwe Windhorst, editors. *Concept*, pages 841–841. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. 3

[3] Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. 3

[4] Caitlin R Bowman, Takako Iwashita, and Dagmar Zeithamova. Tracking prototype and exemplar representations in the brain across learning. *eLife*, 9:e59360, nov 2020. 3

[5] Kaidi Cao, Maria Brbic, and Jure Leskovec. Concept learners for few-shot learning. In *International Conference on Learning Representations*, 2020. 2

[6] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: Deep learning for interpretable image recognition. 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. 3

[7] Gabriele Ciravegna, Pietro Barbiero, Francesco Giannini, Marco Gori, Pietro Lió, Marco Maggini, and Stefano Melacci. Logic explained networks. *arXiv preprint arXiv:2108.05149*, 2021. 1, 2

[8] Stanley J. Colcombe and Robert S. Wyer. The role of prototypes in the mental representation of temporally related events. *Cognitive Psychology*, 44(1):67–103, 2002. 3

[9] Natalia Díaz-Rodríguez, Alberto Lamas, Jules Sanchez, Gianni Franchi, Ivan Donadello, Siham Tabik, David Filliat, Policarpo Cruz, Rosana Montes, and Francisco Herrera. Explainable neural-symbolic learning (x-nesyl) methodology to fuse deep learning representations with expert knowledge graphs: The monumai cultural heritage use case. *Information Fusion*, 79:58–83, 2022. 1

[10] Michael W Eysenck and Marc Brysbaert. *Fundamentals of cognition*. Routledge, 2018. 4

[11] Marcello Frixione. and Antonio Lieto. Prototypes vs exemplars in concept representation. In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development - KEOD, (IC3K 2012)*, pages 226–232. INSTICC, SciTePress, 2012. 3

[12] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. 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. 2

[13] William Heindel, Elena Festa, Brian Ott, Kelly Landy, and David Salmon. Prototype learning and dissociable categorization systems in alzheimer's disease. *Neuropsychologia*, 51, 06 2013. 3

[14] Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 6

[15] Haruo Hosoya. Group-based learning of disentangled representations with generalizability for novel contents. *arXiv preprint arXiv:1809.02383*, 2018. 3

[16] Haruo Hosoya. Group-based learning of disentangled representations with generalizability for novel contents. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 2506–2513. International Joint Conferences on Artificial Intelligence Organization, 7 2019. 3

[17] Haruo Hosoya. Group-based learning of disentangled representations with generalizability for novel contents. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 2506–2513. ijcai.org, 2019. 6

[18] Yordan Hristov, Alex Lascarides, and Subramanian Ramamoorthy. Interpretable latent spaces for learning from demonstration. In Aude Billard, Anca Dragan, Jan Peters, and Jun Morimoto, editors, *Proceedings of The 2nd Conference on Robot Learning*, volume 87 of *Proceedings of Machine Learning Research*, pages 957–968. PMLR, 29–31 Oct 2018. 3

[19] Paul Ibbotson and Michael Tomasello. Prototype constructions in early language acquisition. 1(1):59–85, 2009. 3

[20] Frank Jäkel, Bernhard Schölkopf, and Felix A. Wichmann. Generalization and similarity in exemplar models of categorization: Insights from machine learning. *Psychonomic Bulletin & Review*, 15(2):256–271, Apr 2008. 3

[21] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 6

[22] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2668–2677. PMLR, 10–15 Jul 2018. 2

[23] 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*, volume 32, 2018. 3

[24] 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 Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4114–4124. PMLR, 09–15 Jun 2019. 3

[25] Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 6348–6359. PMLR, 2020. 5, 6

[26] Max Maria Losch, Mario Fritz, and Bernt Schiele. Semantic bottlenecks: Quantifying and improving inspectability of deep representations. In *Pattern Recognition: 42nd DAGM German Conference, DAGM GCPR 2020, Tübingen, Germany, September 28–October 1, 2020, Proceedings 42*, pages 15–29. Springer, 2021. 1, 3

[27] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 6

[28] Diego Marcos, Ruth Fong, Sylvain Lobry, Rémi Flamary, Nicolas Courty, and Devis Tuia. Contextual semantic interpretability. In Hiroshi Ishikawa, Cheng-Lin Liu, Tomas Pajdla, and Jianbo Shi, editors, *Computer Vision – ACCV 2020*, pages 351–368, Cham, 2021. Springer International Publishing. 3

[29] Marius Memmel, Camila Gonzalez, and Anirban Mukhopadhyay. Adversarial continual learning for multi-domain hippocampal segmentation. In *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*, pages 35–45, Cham, 2021. Springer International Publishing. 3

[30] Daniel M. Oppenheimer, Joshua B. Tenenbaum, and Tevye R. Krynski. Chapter six - categorization as causal explanation: Discounting and augmenting in a bayesian framework. volume 58 of *Psychology of Learning and Motivation*, pages 203–231. Academic Press, 2013. 3

[31] Norbert M. Seel, editor. *Prototype*, pages 2714–2714. Springer US, Boston, MA, 2012. 3

[32] Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly supervised disentanglement with guarantees. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 3, 4, 5

[33] J. David Smith. Prototypes, exemplars, and the natural history of categorization. *Psychonomic bulletin & review*, 21(2):312–331, Apr 2014. 24005828[pmid]. 3

[34] Wolfgang Stammer, Patrick Schramowski, and Kristian Kersting. Right for the right concept: Revising neuro-symbolic concepts by interacting with their explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[35] J.R. Taylor. *Linguistic Categorization*. Oxford Textbooks in Linguistics. OUP Oxford, 2003. 3

[36] Julian Zaidi, Jonathan Boilard, Ghyslain Gagnon, and Marc-André Carbonneau. Measuring disentanglement: A review of metrics. *arXiv preprint arXiv:2012.09276*, 2020. 3

[37] Dagmar Zeithamova. *Prototype Learning Systems*, pages 2715–2718. Springer US, Boston, MA, 2012. 3

[38] Xinqi Zhu, Chang Xu, and Dacheng Tao. Where and what? examining interpretable disentangled representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5861–5870, 2021. 3