Causal Abstractions of NeSy models. Where are the concepts?

Tobias Lingenberg
tobias.lingenberg@unitn.it

Abstract

Understanding how neural models encode concepts internally is a central challenge in explainable AI. While Neuro-Symbolic (NeSy) models are designed to improve interpretability, they can still rely on reasoning shortcuts rather than learning meaningful abstractions. In this work, we analyze this phenomenon in DeepProbLog (DPL) models applied to the `MNIST-Addition` task, which requires both visual perception and reasoning. Using Causal Abstraction theory and Distributed Alignment Search (`DAS`), we investigate whether these models can be described by a high-level interpretable reasoning process and where they encode abstract concepts. Our findings reveal that architectural choices strongly influence the reliability of internal concept encodings, offering insights into which reasoning shortcuts may occur and into how abstract concept learning can be improved in NeSy models.

1 Introduction

Neuro-Symbolic (NeSy) AI aims to enhance deep learning models by integrating some prior knowledge and reasoning capabilities over high-level concepts in a rule-based manner (De Raedt et al., 2020). NeSy models are particularly promising for systematic generalization and compliance with predefined constraints, making them, in theory, ideal for high-stakes applications requiring transparency and control over the model's (in- and out-of-distribution) behavior (Marconato et al., 2023b).

Much of the promise of these models relies on the intermediate learned concepts being of high quality. However, recent studies like Marconato et al. (2023a,b); Li et al. (2024) suggest that NeSy predictors are not immune to reasoning shortcuts—unintended optima of the learning objective where models achieve high accuracy while leveraging spurious correlations rather than meaningful abstractions. Given this potential limitation, this work builds around the questions: Where do NeSy models that perform logical operations encode the underlying rules, and are they following the intended high-level reasoning?

To investigate this question, we leverage Causal Abstraction theory (Beckers and Halpern, 2019; Geiger et al., 2021) and Distributed Alignment Search (`DAS`, Geiger et al. (2023)) to interpret concept representations within NeSy models. Specifically, we study the `MNIST-Addition` task (Manhaeve et al., 2018; LeCun, 1998), where models must predict the sum of handwritten digits. We analyze different architectures, focusing on DeepProbLog (DPL, Manhaeve et al. (2018)) as NeSy models and purely neural baselines. By defining a causal abstraction model that aligns with human-like reasoning for this task, we evaluate how well various models encode the intended behavior in their intermediate representations and identify where reasoning shortcuts emerge.

The contributions are as follows: *(1)* Implementation of `DAS` for the `MNIST-Addition` task to analyze how DPL models and end-to-end neural networks align with the defined causal abstraction. *(2)* Investigation on the impact of disentanglement in the feature extractor for mitigating reasoning shortcuts. *(3)* Exploration on how different alignment hypothesis influence `DAS` and qualitatively assessment of different types of reasoning shortcuts that can occur in `MNIST-Addition`.

Project for Advanced Topics in Machine Learning and Optimization, Trento, April 3, 2025

2 Background

2.1 Causal Abstraction Theory

Causal Abstraction (Geiger et al., 2021) provides a framework for understanding whether a high-level computational model (e.g., a structured reasoning process) can be considered a valid simplification of a low-level neural model. The key idea is to look for an alignment between abstract concepts C in the high-level abstraction model and learned representations in the neural network. A high-level model is a causal abstraction A of a neural model M if and only if, under all possible interchangeable interventions, the outputs of A and M remain consistent. This means that replacing activations in the neural network for some set of neurons should produce predictable output changes, mirroring those expected in A.

Interchangeable Interventions are used to verify causal abstractions by replacing the activations of a set of neurons in one example (base input) with activations from another (source input) and observing whether the model's response aligns with the abstract model. This method allows researchers to associate specific neural activations with abstract concepts C in the causal model A.

2.2 Distributed Alignment Search

A classical alignment search process Geiger et al. (2021) consists of three key steps:

- Form an Alignment Hypothesis: Identify a set of neurons in M that potentially correspond to a variable in the high-level abstraction A.
- Counterfactual Testing: Perform interchangeable interventions on these neurons and check if the output changes as expected.
- Iterative Refinement: Aggregate results across multiple trials and refine the alignment hypothesis based on observed counterfactual behaviors.

While this process is flexible and powerful, it becomes computationally expensive for complex neural networks M due to the vast number of possible alignments.

Distributed Alignment Search (`DAS` Geiger et al. (2023)) scales alignment search by formulating it as an optimization problem. Instead of doing brute-force (Geiger et al., 2021) or approximated (Beckers et al., 2019) search, `DAS` *learns* the alignment by introducing a rotation matrix R that restructures latent representations, uncovering interpretable structure. The key insight is that meaningful high-level concepts might not be aligned with the standard basis of neural activations but could emerge through an appropriate transformation of the latent space (leading to a distributed alignment).

`DAS` operates as follows: Instead of activation swapping in the original latent space, the representations are first rotated using R. The intervention is applied in the rotated space, and then the activations are transformed back using the inverse R^{-1}. The rotation matrix is optimized via stochastic gradient descent to maximize the alignment between counterfactual behaviors and expected outputs while ensuring R remains orthonormal. If we can find such a rotation R leading to a high Interchangeable Intervention Accuracy (IIA), we claim there exists an alignment of all concepts in A and M and thus M successfully implements the abstraction A.

By optimizing for alignment, `DAS` allows for scaling causal abstraction analysis to much larger models, as demonstrated by its application to state-of-the-art LLMs such as Alpaca (Wu et al., 2023). This potential of `DAS` being valuable for analyzing the largest and most powerful modern AI models, combined with its ability to identify potentially distributed concept representations, makes it a powerful tool for our study.

3 Methodology

3.1 Problem Description

We evaluate causal abstractions on the `MNIST-Addition` task, which combines visual perception and symbolic reasoning. Each input consists of a pair of handwritten digit images, and the output is their sum. The underlying causal structure is simple and known: the two digit concepts C_1 and C_2

must be inferred from the images and deterministically summed to produce the output Y. This makes the task well-suited for analyzing whether and how models internally encode the relevant concepts.

3.2 Definition of the Causal Abstraction

We define a high-level causal abstraction aligned with human reasoning. This abstraction assumes perfect digit perception and uses a simple directed acyclic graph (DAG) formalizing this process (Fig.1). It serves as the ground-truth abstraction model A in our analysis and is used to generate expected behavior for `DAS` evaluation.

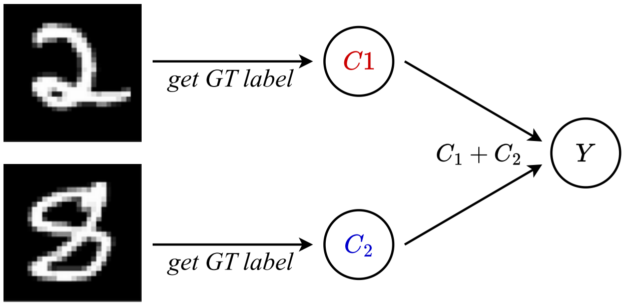

Figure 1: Causal abstraction model for `MNIST-Addition`. The concepts C_1 and C_2 represent the digit values; by taking the ground truth (GT), a perfect perceptual part is modeled. Y is the sum of the digit values. This simple causal model perfectly solves the `MNIST-Addition` task.

3.3 Model Architectures

We evaluate two categories of models: *(1)* a Neuro-Symbolic (NeSy) model based on DeepProbLog (DPL) and *(2)* a purely neural end-to-end baseline. For both categories, we investigate two architectural variants differing in the feature extraction strategy: *(i)* a *disentangled* architecture using an encoder repeated for each single digit, and *(ii)* a *joint* architecture with an encoder processing a pair of digit images together.

Disentanglement has been shown to reduce the occurrence of reasoning shortcuts (Marconato et al., 2023b). By enforcing separate processing of the two inputs, disentanglement constrains the model to treat each digit independently, encouraging more faithful concept representations.

The DPL model combines neural concept extraction with symbolic reasoning. The reasoning step is implemented by a probabilistic logic head to compute the sum:

$$\begin{aligned} p_\theta(c_1, c_2 \mid \mathbf{d}_1, \mathbf{d}_2) &= \begin{cases} f_\theta(\mathbf{d}_1) f_\theta(\mathbf{d}_2) & \text{(Disentangled)} \\ f_\theta(\mathbf{d}_1, \mathbf{d}_2) & \text{(Joint)} \end{cases} \\ p_\theta(y \mid \mathbf{d}_1, \mathbf{d}_2) &= \sum_{(c_1,c_2)=(0,0)}^{(9,9)} u_{\mathsf{K}}(y \mid c_1, c_2)\, p_\theta(c_1, c_2 \mid \mathbf{d}_1, \mathbf{d}_2) \end{aligned} \tag{1}$$

Here, $u_{\mathsf{K}}(y \mid c_1, c_2)$ encodes the summation constraint defined by the prior knowledge K. The concepts are implicitly represented in the neural predictor f_θ. An overview of the DPL pipeline with both encoder variants is shown in Fig. 2.

For comparison, we also train purely neural baselines that directly predict the sum Y from the digit images without an explicit reasoning component, using the same disentangled and joint feature extraction setups.

3.4 Implementation Details of DAS

`DAS` searches for a linear transformation R that aligns the latent space of our different model architectures M with the concept variables C_1 and C_2 (see Fig. 1). We target the latent layer immediately after the feature extractor, before the reasoning step in DPL.

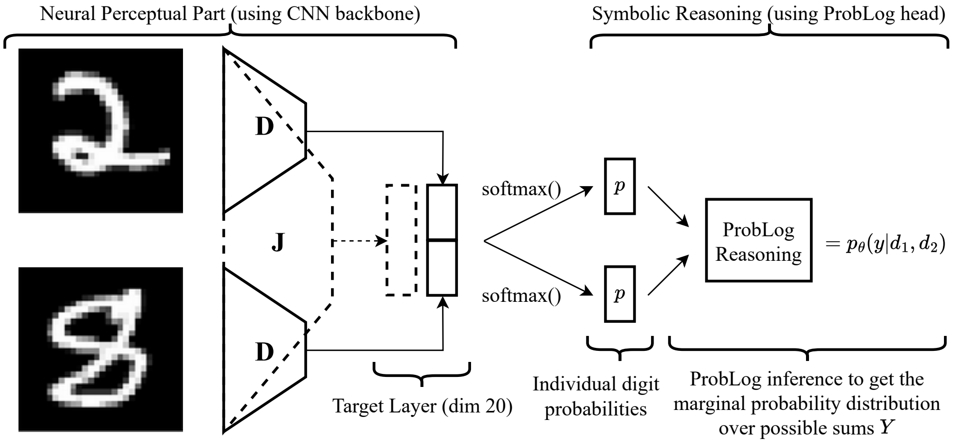
Neural Perceptual Part (using CNN backbone)
Symbolic Reasoning (using ProbLog head)
D
J
D
softmax()
softmax()
p
p
ProbLog Reasoning
$= p_\theta(y|d_1, d_2)$
Target Layer (dim 20)
Individual digit probabilities
ProbLog inference to get the marginal probability distribution over possible sums Y

Figure 2: DPL pipeline with two feature extraction variants: Disentangled (D) and Joint (J).

By default we define the *alignment hypothesis* to partition the 20-dimensional latent space such that the first 10 dimensions correspond to C_1 and the last 10 dimensions to C_2. Additionally, we experiment with alternative hypotheses that include "None" dimensions, assumed to be unrelated to the concepts, to explore *where* the model actually encodes most of the concept information.

DAS optimizes the rotation matrix R to maximize the *Distributed Interchange Intervention (DII)* score, which measures how well the model's internal representations implement the abstraction under counterfactual interventions. Specifically, DII records the best achieved Interchangeable Intervention Accuracy (IIA) on the validation set during optimization. A high DII score (e.g., $> 95\%$) indicates that the model's latent representations can be aligned to match the causal abstraction.

4 Results

4.1 Verification of DAS Implementation

To ensure the correctness of our DAS implementation, we first evaluate it on a handcrafted latent space where the desired alignment is explicitly known. As expected, DAS achieves a perfect DII score of 100%, confirming its ability to recover correct causal abstractions when present. Conversely, when applied to randomly initialized DPL and purely neural networks, the DII score remains low ($\leq 9\%$), indicating that no alignment is found in untrained activations.

4.2 Model Architecture Comparisons

All models achieve high accuracy ($> 95\%$) on the MNIST-Addition task, indicating that they successfully solve the problem. However, only the DPL with a disentangled encoder (DII $\mathbf{99.24}\%$), DPL with a joint encoder (DII $\mathbf{98.37}\%$), and the purely neural model with a disentangled encoder (DII $\mathbf{97.41}\%$) correctly implement the causal abstraction A (Fig. 1). The purely neural model with a joint encoder fails to implement A, achieving only $\mathbf{45.42}\%$.

To have a more fine-grained analysis on what is happening, we present confusion matrices for labels Y, concepts C, and concept pairs, as well as visualizations of the learned rotation matrix R and concept contribution plots, indicating which dimensions of the original latent vector “belong” more to C_1 and C_2.

DPL (disentangled) This model behaves as expected and aligns nearly perfectly with the abstraction. The rotation matrix (see Fig. 3) reveals, that only the upper right and lower left block is occupied with non-zero values. This confirms, that the initial alignment hypothesis, where the first 10 dimensions correspond to C_1, and the last 10 to C_2, was already correct.

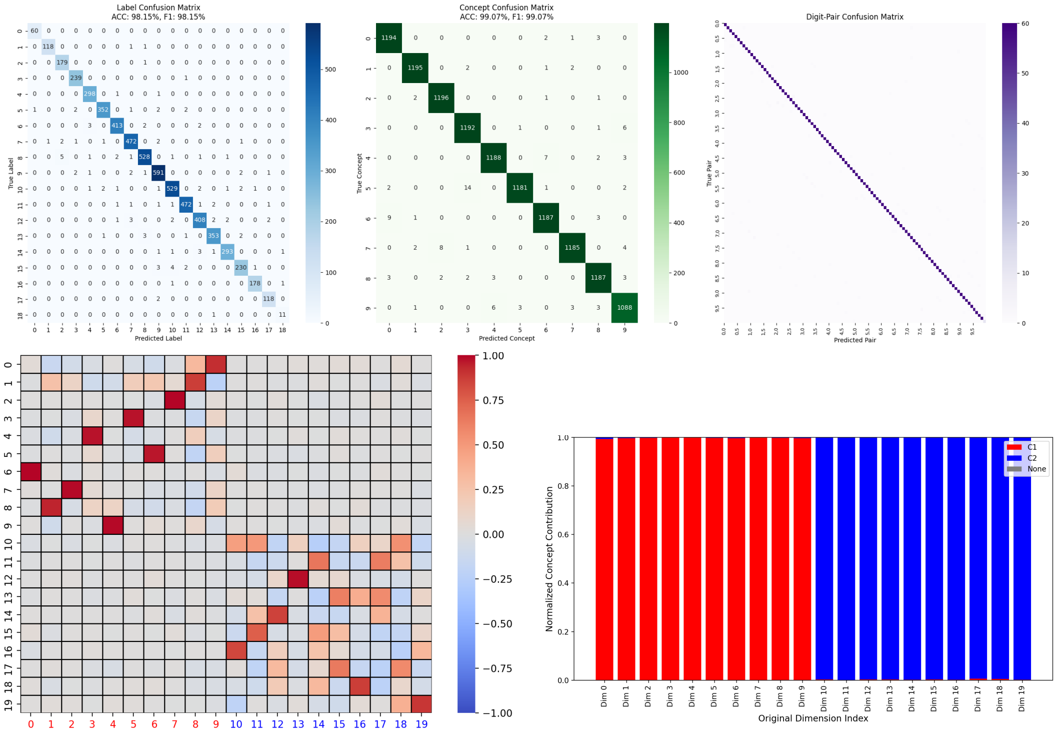

Figure 3: Analysis of the DPL (disentangled) model. **Top:** Confusion matrices showing performance for final labels and concepts. **Bottom Left:** Heatmap visualization of the learned rotation matrix R. **Bottom Right:** Concept contribution per dimension of the original latent space.

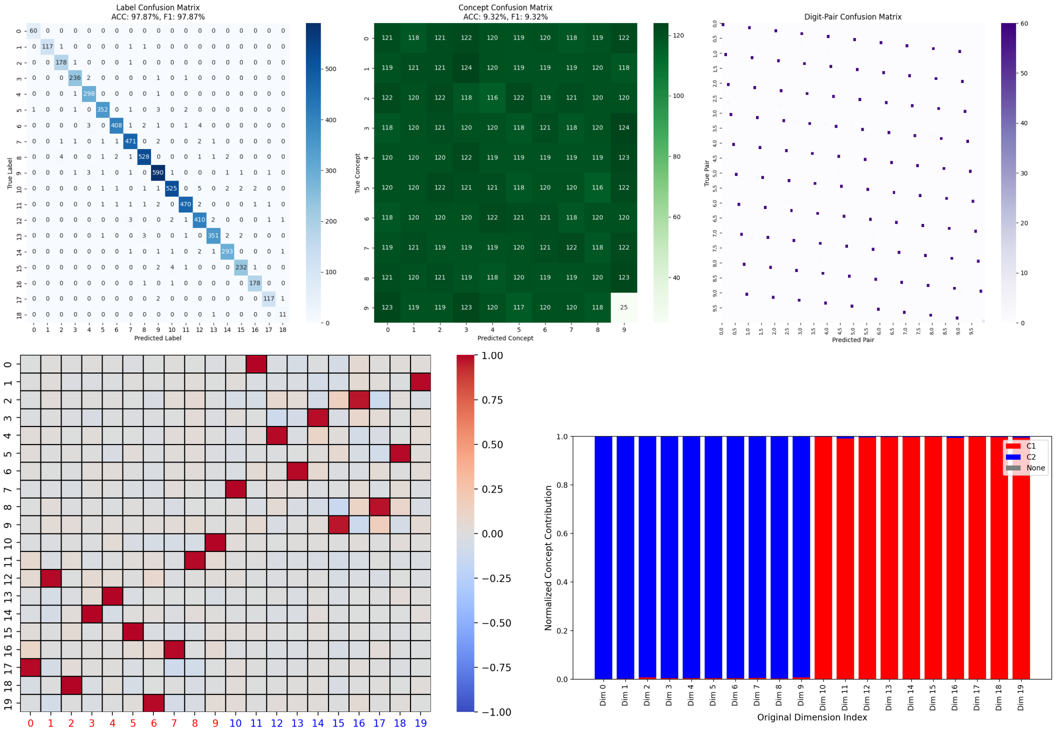

Figure 4: Analysis of the DPL (joint) model. **Top:** Confusion matrices showing performance for final labels and concepts. **Bottom Left:** Heatmap visualization of the learned rotation matrix R. **Bottom Right:** Concept contribution per dimension of the original latent space.

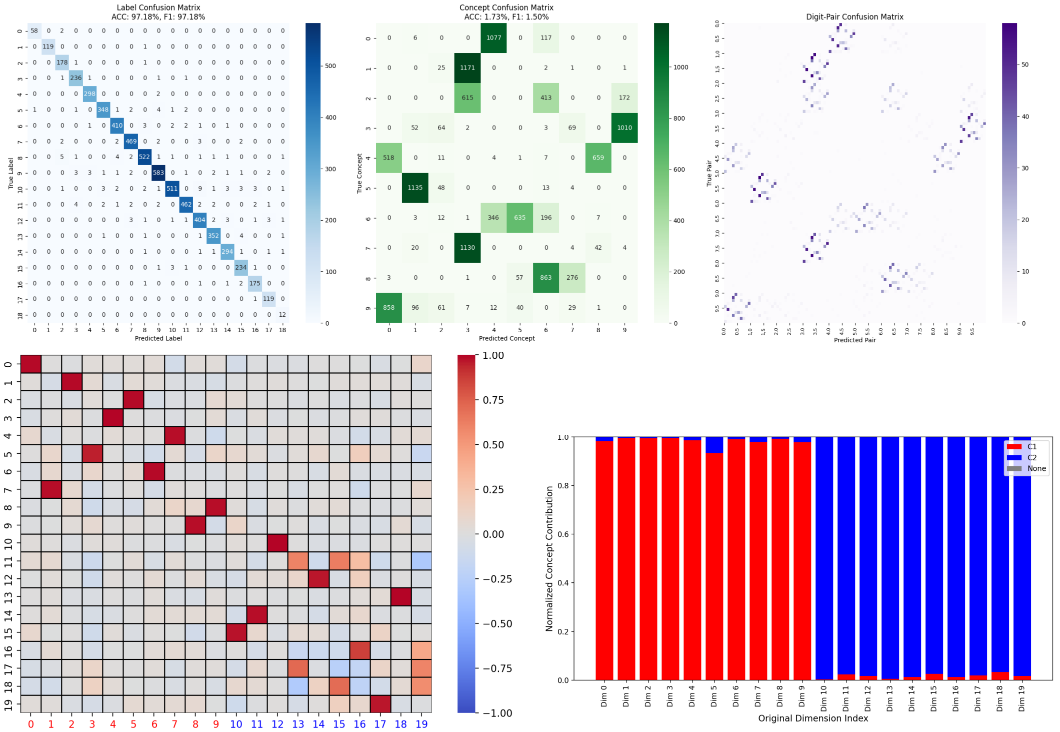

Figure 5: Analysis of the purely neural model (disentangled). **Top:** Confusion matrices showing performance for final labels and concepts. **Bottom Left:** Heatmap visualization of the learned rotation matrix R. **Bottom Right:** Concept contribution per dimension of the original latent space.

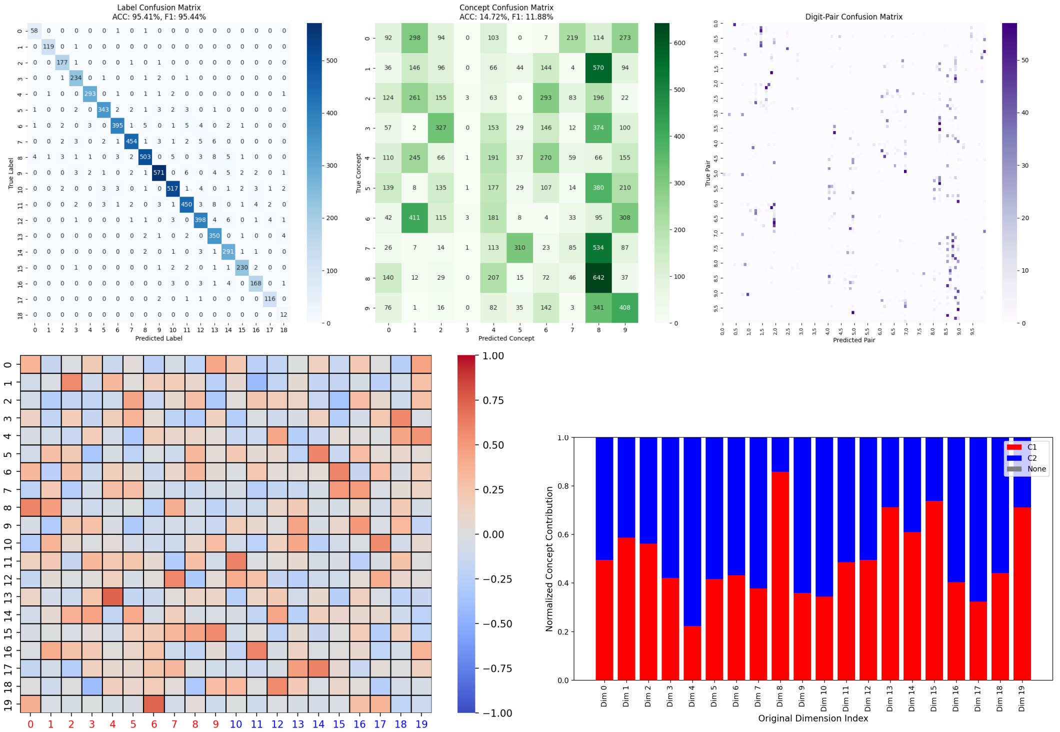

Figure 6: Analysis of the purely neural model (joint). **Top:** Confusion matrices showing performance for final labels and concepts. **Bottom Left:** Heatmap visualization of the learned rotation matrix R. **Bottom Right:** Concept contribution per dimension of the original latent space.

DPL (joint) This model in Figure 4 also implements the abstraction according to the DII score. But, only visible through the visualization of R and the confusion matrices, we can observe a first type of reasoning shortcut: the model has learned a flipped representation where the first 10 dimensions encode C_2 instead of C_1.

Purely neural model (disentangled) Also this model achieves a clear separation of C_1 and C_2, but the exact encoding remains complex (Fig. 5), as reflected in the confusion matrix patterns.

Purely neural model (joint) Here we observe clear results. This model fails to align with the abstraction, yielding only a maximum DII score of 45%. Consequently, R has no meaningful interpretation (Fig. 6). We just know, that the concepts C_1 and C_2 are not separable and the model does not internally follow the causal abstraction.

4.3 Varying Alignment Hypothesis Size

We systematically analyze the impact of varying the number of neurons assigned to each concept, revealing a plateau effect in DII scores. Figure 7 shows that, particularly for the neural model with a disentangled encoder, performance saturates when using 6 to 10 dimensions per concept. This suggests that only a subset of the latent space is effectively used for concept encoding.

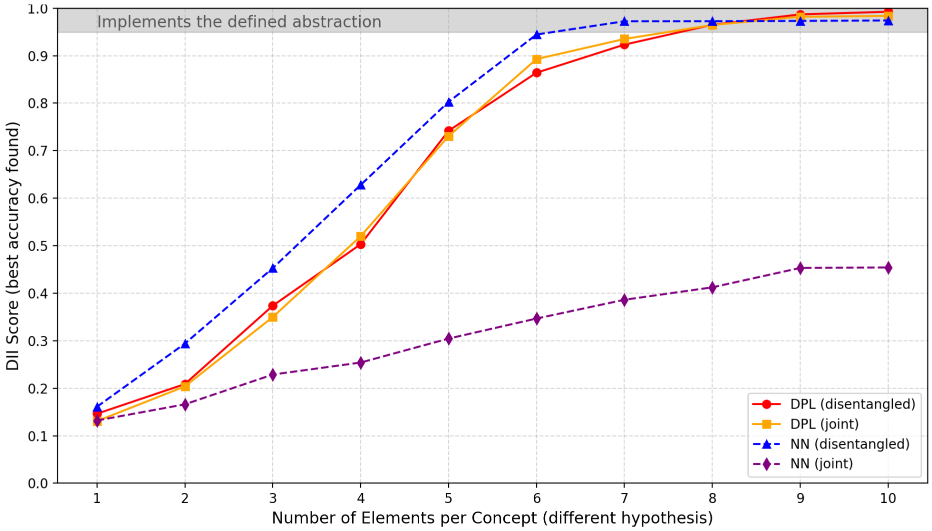

Figure 7: DII scores for varying alignment hypothesis sizes (1–10 elements per concept).

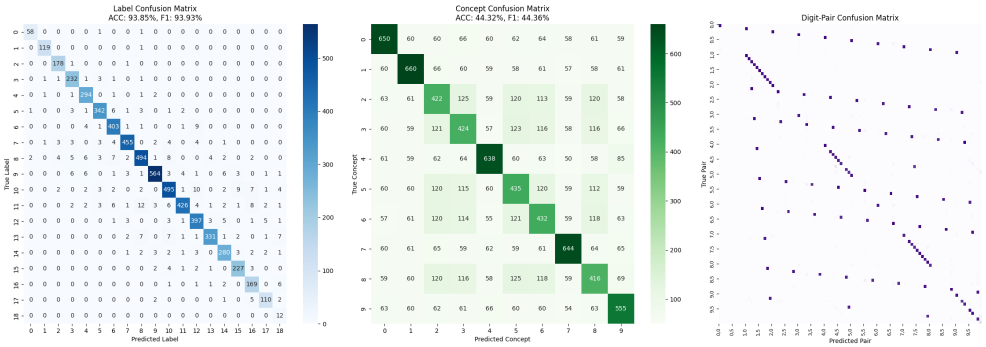

Figure 8: Runs that shows concept mixing as reasoning shortcut in DPL (joint).

4.4 Reasoning Shortcuts in DPL Models

We further investigate reasoning shortcuts in DPL models by analyzing behavior over 10 differently initialized training runs. When using a disentangled encoder, the model reliably separates concepts and avoids shortcuts. However, with a joint encoder, we observe two notable behaviors:

Concept flipping: In all runs, the model internally swaps C_1 and C_2. While this does not affect DII scores under the current abstraction, it may be problematic in downstream applications where concept integrity is required.

Concept mixing: In 2 out of 10 runs, concepts are not clearly separable anymore, leading to lower DII scores (Fig. 8). This suggests that the joint encoder sometimes fails to learn a clean abstraction and instead encodes entangled representations, being harmful for generalization and downstream applications building on the concepts.

DAS proves useful in detecting and interpreting reasoning shortcuts. While concept flipping is undetectable through DII scores alone, visualization techniques reveal the issue. In contrast, concept mixing is directly reflected in a drop in DII score, signaling an inadequate internal representation.

These results highlight the importance of disentanglement in NeSy models. We hypothesize that insights gained using DAS can be improved ever further, when defining different types of abstraction models (e.g. one that has concepts containing the specific value of a digit) to provide even deeper insights into concept encoding.

5 Conclusion

In this work, we investigated where and how NeSy models encode concepts by leveraging Causal Abstraction theory and DAS in the MNIST-Addition task. Our analysis reveals that while all models achieve high accuracy, their internal concept representations differ significantly depending on architectural choices. DeepProbLog (DPL) models with a disentangled encoder faithfully align with the intended causal abstraction, supporting human-like reasoning. In contrast, models with a joint encoder exhibit reasoning shortcuts—such as concept flipping and mixing—that reduce interpretability and reliability in downstream applications.

An interesting insight is that concept information is not evenly distributed across latent dimensions. Purely neural models tend to encode concepts in only a subset of their dimensions, whereas DPL models mitigate this effect to some extent by promoting a more structured representation.

Despite the strengths of DAS, its current formulation has limitations in capturing finer-grained details, such as the precise encoding of individual digit values. Future work could explore alternative abstraction models that cover different reasoning processes and potentially expose additional types of reasoning shortcuts. Additionally, further visualization techniques, such as mapping concept activations back onto the input image space (Singla and Feizi, 2021), could deepen our understanding of internal concept encoding in NeSy models.

References

Beckers, S., Eberhardt, F., and Halpern, J. Y. (2019). Approximate causal abstraction.

Beckers, S. and Halpern, J. Y. (2019). Abstracting causal models.

De Raedt, L., Dumančić, S., Manhaeve, R., and Marra, G. (2020). From statistical relational to neuro-symbolic artificial intelligence. *arXiv preprint arXiv:2003.08316*.

Geiger, A., Lu, H., Icard, T., and Potts, C. (2021). Causal abstractions of neural networks.

Geiger, A., Wu, Z., Potts, C., Icard, T., and Goodman, N. D. (2023). Finding alignments between interpretable causal variables and distributed neural representations.

LeCun, Y. (1998). The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*.

Li, Z., Liu, Z., Yao, Y., Xu, J., Chen, T., Ma, X., and Lü, J. (2024). Learning with logical constraints but without shortcut satisfaction. *arXiv preprint arXiv:2403.00329*.

Manhaeve, R., Dumancic, S., Kimmig, A., Demeester, T., and De Raedt, L. (2018). Deepproblog: Neural probabilistic logic programming. *Advances in neural information processing systems*, 31.

Marconato, E., Bontempo, G., Ficarra, E., Calderara, S., Passerini, A., and Teso, S. (2023a). Neuro-symbolic continual learning: Knowledge, reasoning shortcuts and concept rehearsal.

Marconato, E., Teso, S., Vergari, A., and Passerini, A. (2023b). Not all neuro-symbolic concepts are created equal: Analysis and mitigation of reasoning shortcuts.

Singla, S. and Feizi, S. (2021). Salient imagenet: How to discover spurious features in deep learning? *arXiv preprint arXiv:2110.04301*.

Wu, Z., Geiger, A., Icard, T., Potts, C., and Goodman, N. (2023). Interpretability at scale: Identifying causal mechanisms in alpaca. *Advances in neural information processing systems*, 36.