In Search of Meaning: Unveiling Interpretable Concepts in Neural Representation*

Tamara Bíla, Igor Farkaš

Centre for Cognitive Science, Department of Applied Informatics
Comenius University Bratislava, Slovakia
tamara.bila@fmph.uniba.sk

Abstract

In light of the increasing presence of artificial neural networks in both the public and private sectors, the demand and need for explainability is self-evident. In this work, we investigate several neural networks with varying depth trained for image classification. Our approach builds upon the previously developed Concept Activation Region (CAR) method in combination with hierarchical agglomerative clustering. By combining these two techniques simultaneously, we aim to achieve a multifaceted understanding of the internal representation structure of the target network.

1 Introduction

Interpretability of trained deep neural networks is essential for establishing trustworthiness, and hence allowing applicability in sensitive and high-stakes domains such as medicine or policy making. Moreover, understanding the reasoning of the model enables failure detection, debugging, diagnostics, and even reverse engineering solutions (Rauker et al., 2023). Evaluation of the quality of explanations is usually based on metrics such as accuracy, consistency, and user studies (Ali et al., 2023). Bereska and Gavves (2024) introduced a taxonomy that divides the explanation methods into: *behavioral* (focusing only on input–output relations); *attributional* (predictions are traced back to individual input features); *concept-based* (analyzing latent representations of high-level concepts at hidden layers) and *mechanistic* (identifying fundamental components of the model and their causal relationships).

Our work is concerned with *post hoc* concept-based explainability of visual neural networks by building auxiliary models and training them on the hidden representations of the target model. In particular, our approach identifies human understandable concepts and compares them with network-inherent clusters that occur in the latent space.

The concept extraction problem, when viewed as a learning task, may be approached just as any other machine learning problem in *supervised* or *unsupervised* setting. A prominent example of a *supervised ad hoc* method is the concept bottleneck model (Koh et al., 2020), where a layer is inserted into the target network and fine-tuned in order to align the concept directions along the available neurons (axes). On the other hand, popular *post hoc* algorithms include the concept activation vector (CAV) method (Kim et al., 2018) with the corresponding TCAV score, and concept activation regions (CARs) (Crabbé and van der Schaar, 2022), who replaced the linear concept classifier of CAV by support vector classifier with radial kernel, providing better overall concept accuracy. O'Mahony et al. (2023) employed an *unsupervised* method, hierarchical agglomerative clustering, in order to disentangle the concept directions of the top 100 neuron activating samples.

2 Combining supervised and unsupervised concept search

In this section, we describe the methods used to investigate the internal representations of convolutional neural network (CNN) image classifiers trained on the MNIST dataset. In our experiments, we applied discovery tools of *visual concepts* at each hidden layer of a neural network to analyse the evolution of learned representations. By a visual concept we mean a recurring, human recognisable pattern, or abstraction, that is present across input data samples (see, e.g., four concepts in Fig. 1). Our hypothesis is that target networks learn to represent concepts as regions or clusters in their latent space.

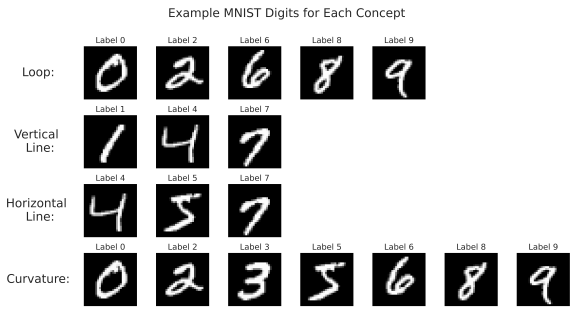

Fig. 1: Example instances of each class for the MNIST dataset containing the four concepts.

*This research was funded by Horizon Europe project TRAIL, no. 101072488 (T.B.) and by project VEGA 1/0373/23 (I.F.).

We apply the CAR method to each layer of pre-trained target networks with three layers. CARs utilise a binary support vector classifier (SVC) with radial kernel for every concept from the concept set (see Fig. 1). The SVC is then trained on randomly sampled latent representations of the ℓ-th layer with concept-positive and concept-negative membership.

The next step consists of unsupervised (automatic) concept extraction via agglomerative clustering of the same latent embeddings without fixing the number of clusters. Embeddings in each layer of the target model were clustered, since we expect a different level of semantic depth when moving deeper into the network. Eventually, CARs were trained on top of discovered cluster sets.

With CARs (SVCs) trained for both concepts and clusters, their test accuracy was evaluated across all layers, and the mean over the concepts and the clusters computed separately. A comparison of the accuracy across all layers of the 3-layer deep CNN is depicted in Fig. 2. The results show a higher precision for SVCs trained to determine the membership of the cluster in contrast to concept membership classifiers. Moreover, the accuracy of both types of SVCs rises towards the output layers of the target network leading to network ability to classify the input images.

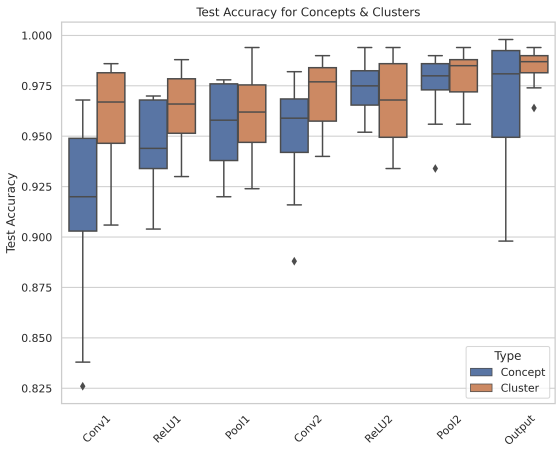

Fig. 2: Test accuracy of the CAR method averaged over predefined set of concepts as opposed to a set of discovered clusters, across all layers of a target CNN trained on MNIST.

3 Conclusion

As the pilot results suggest, higher performance of cluster classifiers is due to their better separability in the latent space, since a cluster naturally contains data points lying close to each other. In principle, representations of a single concept can be scattered throughout the latent space in an orderless manner, and their separability is the tested assumption rather than a given fact. Secondly, the increase in accuracy over the depth of the analyzed layer points to the manifold disentanglement hypothesis, proposed by Brahma et al. (2015) and tested by Pócoš et al. (2021), stating that the convoluted manifold representation of the classes decouples from the input towards the output of the network. Hence, it is meaningful to conclude that concepts which are inherently related to a certain class or a set of classes, gradually unfold towards the output layer. The future work will focus on complex datasets containing color images.

References

Ali, S. et al. (2023). Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99:101805.

Bereska, L. and Gavves, S. (2024). Mechanistic interpretability for AI safety - a review. *Transactions on Machine Learning Research*.

Brahma, P. et al. (2015). Why deep learning works: A manifold disentanglement perspective. *IEEE Transactions on Neural Networks and Learning Systems*, 27:1–12.

Crabbé, J. and van der Schaar, M. (2022). Concept activation regions: A generalized framework for concept-based explanations. In *Neural Information Processing Systems*.

Kim, B. et al. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning*.

Koh, P. W. et al. (2020). Concept bottleneck models. In *International Conference on Machine Learning*, volume 119, pages 5338–5348.

O'Mahony, L. et al. (2023). Disentangling neuron representations with concept vectors. In *Conference on Computer Vision and Pattern Recognition*, pages 3770–3775. IEEE.

Pócoš, S. et al. (2021). Assessment of manifold unfolding in trained deep neural network classifiers. In *Trustworthy AI - Integrating Learning, Optimization and Reasoning*, pages 93–103. Springer Nature.

Rauker, T. et al. (2023). Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. In *Conference on Secure and Trustworthy Machine Learning*, pages 464–483. IEEE.