Modeling and Exploring Visual Concepts in Deep Networks

Mhd Walid Al Jallad and Kristína Malinovská

Centre for Cognitie Science, DAI FMPI,
Comenius University in Bratislava, Slovakia
Mlynská dolina, 84248 Bratislava
Email: kristina.malinovska@fmph.uniba.sk

Abstract

Deep neural networks (DNNs) produce powerful visual features, yet with very limited options of interpretability, compositional reasoning, topographic coherence and the possibility of continual learning. The way how DNNs recognize entities in images is often very far away from human cognition, despite their amazing performance. Perturbations in the data can easily cause the system to fail (as well known from the adversarial images research). These phenomena can still be observed even with recent foundation models (DINOv2/v3, SigLIP, CLIP). Inspired by cortical topographic maps and prototype theory, we propose a differentiable, topology-preserving SOM layer over features of convolutional and foundation models as a unified interface for visual concept modeling and introduce the concept of auxiliary unsupervised loss function. The prospect of our approach is to enhance the existing models with (1) a navigable concept atlas with semantic topography mirroring inferior-temporal cortex organization; (2) compositional fuzzy reasoning over concept prototypes, where logical operations respect graded membership and neighborhood smoothness; and (3) continual open-world detection that enables incremental learning of new categories. Our proposal bridges disjoint lines of research—prototype interpretability, concept bottlenecks, neuro-symbolic reasoning, and continual learning—under a single biologically grounded computational primitive.

1 Introduction

Deep neural networks have become the dominant approach to visual perception, where recent self-supervised foundation models, such as DINOv2/v3, SigLIP, CLIP, are able to produce remarkably strong general-purpose features without requiring labeled data (Oquab et al., 2024; Siméoni et al., 2025; Zhai et al., 2023; Radford et al., 2021). Despite being transferable across distributions and tasks, these features are organized by the training objective rather than by anything resembling a concept. They afford no human-interpretable readout, no compositional reasoning interface, and no principled mechanism for absorbing new categories without forgetting existing ones. We argue that deep neural networks (DNN) are missing a topographic, prototype-based structure that learns from raw feature spaces in an unsupervised manner.

In order to explore the concepts of DNNs and promote emergence of more human-like concepts, inspiration can be taken from biological cognitive systems. First, biological learning combines both bottom-up error-driven and top-down unsupervised processes, as demonstrated by O'Reilly's XCAL model and Leabra framework (O'Reilly et al., 2024). Second, primate inferior-temporal cortex are shown to contain a coarse navigable map of object space (Bao et al., 2020), suggesting that cortical visual representations are explicitly topographic, a property that DDN feature spaces lack by default. In support of this claim, recent artificial neural networks where able to reproduce this organization only when an explicit topographic prior is added during training (Margalit et al., 2024). Third, it has been suggested that natural categories in the brain are prototype-like and hierarchically graded (Rosch, 1978), and recent neuroimaging shows that prototype and exemplar representations in fact coexist across distinct cortical regions during learning (Bowman et al., 2020). This suggests that prototype-based categories behave as fuzzy sets, affording logical composition. Finally, the mammalian brain exhibits efficient lifelong learning due to its reliance on hippocampal episodic memory interleaved with slow neocortical consolidation (McClelland et al., 1995).

Recently, the missing topographic structure in DNNs is slowly starting to get acknowledged by some foundation-models. For instance, the prototype heads of the DINO family have been shown to be redundant and unordered, with many prototypes converging on near-identical directions (Govindarajan et al., 2024). Additionally, DINOv3 introduces an ad-hoc spatial coherence prior to recover the kind of topographic smoothness that classical biologically-motivated representations posses by construction (Siméoni et al., 2025).

Grounded by our previous claims, we propose that a topology-preserving prototype layer, i.e., a differentiable self-organizing map (SOM) defined over a network's feature spaces, can serve as a unifying interface for visual concept modeling. We demonstrate

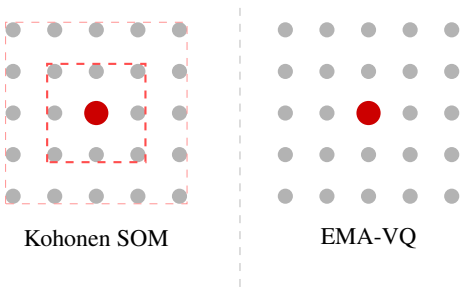

Fig. 1: The Kohonen self-organising map update with Gaussian neighbourhood (left) reduces to the EMA-VQ codebook update under zero neighbourhood width (right). The red node marks the best-matching unit; dashed rectangles indicate the neighbourhooh support. After (Irie et al., 2024).

proof of concept by drawing on prior work on a mean-teacher SOM auxiliary loss (Samporová, 2024), and we sketch three downstream affordances: (1) a global, navigable concept atlas with human-validated interpretability; (2) compositional fuzzy-logical reasoning over part-prototypes for top-down concept blending; and (3) continual open-world detection paired with associative memory.

2 Related Work

2.1 Self-organizing maps as a topology-preserving prototype layer

The classical primitive for topology-preserving prototype learning is the self-organizing map (Kohonen, 1990). A SOM is a two-dimensional grid of prototype vectors trained by repeated competitive selection of a best-matching unit and a neighborhood-weighted update which preserves the topological structure by ensuring prototypes that are close on the grid become close in the feature space. SOMs also exhibit density awareness, in that codebook (i.e. prototype vectors) density tracks the input density.

Recent work has shown that the exponential moving average vector-quantization (EMA-VQ) codebook update used in modern image tokenizers is a special case of the SOM update when the neighborhood approaches zero (Irie et al., 2024). Conversely, the SOM is a strict generalization of the discretization primitive that foundation tokenizers already deploy, since it adds a smooth continuous topological structure without sacrificing the discretization behavior (see Fig. 1)

The integration of SOMs into modern deep learning architectures is a small but a growing research direction. SOM-VAE (Fortuin et al., 2019) and DPSOM (Manduchi et al., 2020) embed SOMs in autoencoders for interpretable time-series representations; SOM-CPC adds SOMs to contrastive learning for high-rate physiological signals (Huijben et al., 2023); and ViT-SOM incorporates a SOM bottleneck into a vision transformer (Luo and Yuan, 2026). Additionally, prior work incorporated a SOM layer on top of the last hidden layer in a Mean Teacher (MT) architecture model (Samporová, 2024).

2.2 Approaches in visual concept modeling

Four lines of recent work address related aspects of the visual concept modeling problem. The first is regarding prototype-based methods. ProtoPNet (Chen et al., 2019) and its descendants attach a small set of trainable prototypes to a deep neural classifier, and explain predictions as similarity to prototypes. However, recent work shows that spatial alignment of pototype activatinos to image regions they intend to explain is poor, and that prototype interpretability methods provide no global geometry over the prototype set (Sacha et al., 2023). The fact that prototypes are returned as lists not as a naviagable structure.

Another instance of a missing global geometry is related to concept-bottleneck models (CBM). Discover-then-Name CBM (Rao et al., 2024) yields a dictionary of named concepts but no relational structure between them, offering per-concept activations rather than a continuous similarity surface. No global geometry of prototype methods are shared, in addition to the risk of leakage when concept activations values are continuous (Mahinpei et al., 2021).

In terms of continual learning, models such as RanPAC (McDonnell et al., 2024) and FeCAM (Goswami et al., 2024) use prototype-style representations for exemplar free continual learning over frozen foundation features achieving strong results but the prototypes themselves carry no internal organization.

Furthermore, neurosymbolic approaches like PROTO-LTN (Martone et al., 2022) ground logical operations on prototype activations within a logic-tensor-network framework. The logic, however, is imposed without any topological grounding since there is no notion that nearby prototypes should compose more smoothly than distant ones.

3 Our Proposal

We propose a self-organizing map primitive that supports semi-supervised training of DNNs by extending them to include a topology-preserving prototype layer that is trained over the distribution of feature vectors using competitive selection of best-matching unit and a neighborhood-weighted update. This primitive should be differentiable end-to-end and exhibit three main properties: (1) *topology preservation*—the neighborhood term should ensure that proximity on the grid, representing a conceptual structure, mirrors proximity in

feature space; (2) *density awareness*—prototype density tracks the input density in such a way that rare regions of the feature space receive proportionally fewer prototypes; and (3) *discrete sparse activations*, demonstrated by the best-matching unit SOM mechanism, where only a small subset of neurons activate for each input, rather than all neurons being continuously active. Structurally, this is closer to hard concept-bottleneck models than to concept-embedding models that are less transparent due to leakage of information across distributed concept representations (Mahinpei et al., 2021).

3.1 MT-SOM

A concrete instantiation of the SOM primitive at small scale appears in a prior work on a mean-teacher SOM auxiliary loss (Samporová, 2024). The Mean Teacher framework (Tarvainen and Valpola, 2017) maintains a student network and a teacher network whose parameters are an exponential moving average over the student's network. It adds to the supervised loss a *consistency cost* between student and teacher predictions on unlabeled inputs. In MT-SM, this consistency cost is reformulated as a distance over a self-organizing map laid over the last convolutional layer of the network. Thus, minimizing the overall loss requires both that student and teacher features map to nearby prototypes on the SOM grid, and that the features themselves remain close to those prototypes.

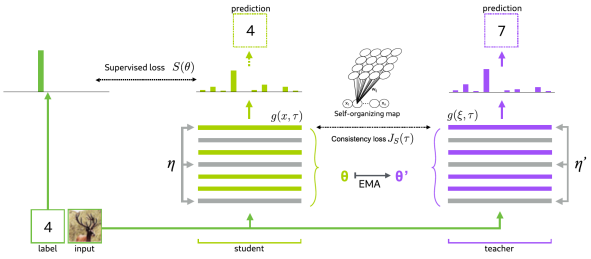

Fig. 2: Illustration of MT-SOM pipeline. Adopted from (Samporová, 2024)

On CIFAR-10, MT-SOM produces a clearly organized 10×10 prototype grid in which semantically similar classes settle on neighboring grid cells (Figure 3), and yields modest accuracy gains over the unaugmented MT baseline (Samporová, 2024). The result is small in scale but illustrates the topology preservation and density-aware prototype placement properties of our proposed primitive.

3.2 Downstream affordances

We sketch three downstream uses of the proposed primitive (see Figure 4). The first concerns a navigable concept atlas. Existing prototype-interpretability methods do not produce a 2D topology where spatial proximity

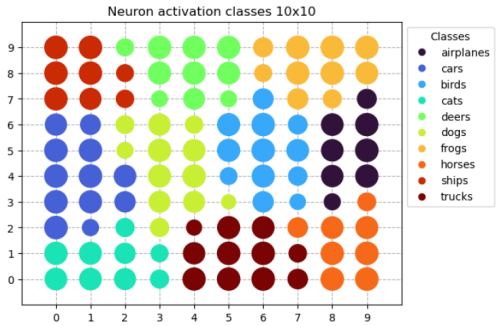

Fig. 3: CIFAR-10 MT-SOM neuron-activation grid, adopted from (Samporová, 2024).

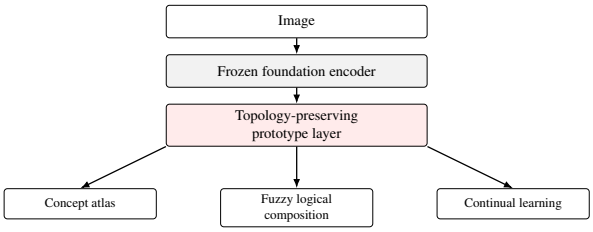

Fig. 4: The proposed pipeline. Images are passed through a frozen foundation encoder; the resulting features are routed through a topology-preserving prototype layer—a differentiable self-organising map—whose grid serves as a shared substrate for the three downstream affordances.

equals semantic proximity. While each cell in a trained SOM grid carries its prototype vector, which ensures continuous semantic transitions across the grid. This affordance is the most direct translation of the inferior-temporal cortex topographic map (Bao et al., 2020) into a deep learning architecture. To validate such claim, we may adopt the HIVE protocol (Kim et al., 2022), which allows comparing concept-atlas readouts against ProtoPNet-style explanations (Sacha et al., 2023) and Discover-then-Name dictionaries (Rao et al., 2024).

The second affordance concerns compositional fuzzy reasoning. Prototype-based natural categories can be described as fuzzy sets (Rosch, 1978; Bowman et al., 2020), where membership is graded, and compositions of categories, such as AND, OR, NOT, should respect the gradedness. Over a SOM grid, we predict that nearby part-prototypes should compose more smoothly than distant ones, with neighborhood-weighted t-norms operating on activation patterns. For instance, for a query "unicorn", the activation pattern should be the intersection of a horse-prototype neighborhood and a horn-prototype neighborhood on the grid. The closet existing comparable model that achieves such compositional reasoning is PROTO-LTN (Martone et al., 2022), which grounds logical operations on prototype activations through logic tensor networks but imposes the logic without any topological structure.

The third affordance is enabling of the continual

learning. Open-world object detection requires the simultaneous capacity to learn new categories without catastrophic forgetting and to flag novel unknown categories (Zohar et al., 2023). Recent state-of-the-art continual learning methods that operate over frozen foundation features, use prototype-style representations but offer no mechanism for episodic recall (McDonnell et al., 2024; Goswami et al., 2024). We hypothesis that our topology-preserving prototype layer will offer a natural slow consolidation memory, where new categories settle into existing topographic structure, and rare regions of the grid absorb novel prototypes incrementally.

4 Conclusion and Future Work

In this paper we propose a single primitive, a topology-preserving prototype layer over frozen foundation features, that unifies four lines of recent work in visual concept modeling: prototype interpretability, concept-bottleneck explanation, neurosymbolic reasoning, and continual learning. Each component of our proposal has some inspiration in cognitive neuroscience. Overall, the primitive itself echoes the topographic functional organization of cortical maps (Bao et al., 2020), reflecting the correlation between spatial and semantic representations, and motivating the need of an interpretable and navigable concept atlas that is learned from the feature space of deep neural networks. Fuzzy reasoning draws on prototype theory (Rosch, 1978) and the finding by (Bowman et al., 2020) that prototype and exemplar representations coexist in the brain, motivating composition over graded prototype activations. Finally, continual learning follows complementary learning systems theory (McClelland et al., 1995).

Flowing this proposal, we will (1) design and benchmark the primitive on deep convolutional neural networks, starting from prior work of MT-SOM (Samporová, 2024) and then expanding the implementation to the attention models, namely the DINO family; (2) compare SOMs navigable concept atlas against other baseline models (e.g., ProtoPNet) utilizing interpretability frameworks; (3) evaluate fuzzy-logic composition over part-prototypes on synthetic compositional benchmarks; and (4) test the SOMs ability on continual learning tasks.

Acknowledgement

This paper was written at the Centre for Cognitive Science at DAI FMPI Comenius University Bratislava, with the support of a grant XYZ. We also thank for support to the Slovak Society for Cognitive SSKV[1].

[1] https://cogsci.fmph.uniba.sk/sskv/

References

Bao, P., She, L., McGill, M., and Tsao, D. Y. (2020). A map of object space in primate inferotemporal cortex. *Nature*, 583(7814):103–108.

Bowman, C. R., Iwashita, T., and Zeithamova, D. (2020). Tracking prototype and exemplar representations in the brain across learning. *eLife*, 9:e59360.

Chen, C., Li, O., Tao, C., Barnett, A. J., Su, J., and Rudin, C. (2019). This looks like that: Deep learning for interpretable image recognition.

Fortuin, V., Hüser, M., Locatello, F., Strathmann, H., and Rätsch, G. (2019). Som-vae: Interpretable discrete representation learning on time series.

Goswami, D., Liu, Y., Twardowski, B., and van de Weijer, J. (2024). Fecam: Exploiting the heterogeneity of class distributions in exemplar-free continual learning.

Govindarajan, H., Sidén, P., Roll, J., and Lindsten, F. (2024). On partial prototype collapse in the dino family of self-supervised methods.

Huijben, I. A., Nijdam, A. A., Overeem, S., Van Gilst, M. M., and Van Sloun, R. (2023). SOM-CPC: Unsupervised contrastive learning with self-organizing maps for structured representations of high-rate time series. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 14132–14152. PMLR.

Irie, K., Csordás, R., and Schmidhuber, J. (2024). Self-organising neural discrete representation learning à la kohonen. In Wand, M., Malinovská, K., Schmidhuber, J., and Tetko, I. V., editors, *Artificial Neural Networks and Machine Learning – ICANN 2024*, pages 343–362. Springer Nature Switzerland.

Kim, S. S. Y., Meister, N., Ramaswamy, V. V., Fong, R., and Russakovsky, O. (2022). HIVE: Evaluating the human interpretability of visual explanations. In Avidan, S., Brostow, G., Cissé, M., Farinella, G. M., and Hassner, T., editors, *Computer Vision – ECCV 2022*, pages 280–298. Springer Nature Switzerland.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480.

Luo, A. and Yuan, K. (2026). Simple self-organizing map with vision transformers. *IEEE Signal Processing Letters*, 33:331–335.

Mahinpei, A., Clark, J., Lage, I., Doshi-Velez, F., and Pan, W. (2021). Promises and pitfalls of black-box concept learning models.

Manduchi, L., Hüser, M., Vogt, J., Rätsch, G., and Fortuin, V. (2020). Dpsom: Deep probabilistic clustering with self-organizing maps.

Margalit, E., Lee, H., Finzi, D., DiCarlo, J. J., Grill-Spector, K., and Yamins, D. L. K. (2024). A unifying framework for functional organization in early and higher ventral visual cortex. *Neuron*, 112(14):2435–2451.e7.

Martone, S., Manigrasso, F., Lamberti, F., and Morra, L. (2022). Prototypical logic tensor networks (prototltn) for zero shot learning. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 4427–4433.

McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychol Rev*, 102(3):419–457.

McDonnell, M. D., Gong, D., Parveneh, A., Abbasnejad, E., and van den Hengel, A. (2024). Ranpac: Random projections and pre-trained models for continual learning.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. (2024). Dinov2: Learning robust visual features without supervision.

O'Reilly, R. C., Munakata, Y., Frank, M. J., Hazy, T. E., and Contributors (2024). *Computational Cognitive Neuroscience*. Online Book, 5th Edition, URL: `https://compcogneuro.org`.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Rao, S., Mahajan, S., Böhle, M., and Schiele, B. (2024). Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery.

Rosch, E. (1978). Principles of categorization. In Rosch, E. and Lloyd, B. B., editors, *Cognition and Categorization*, pages 27–48. Lawrence Elbaum Associates.

Sacha, M., Jura, B., Rymarczyk, D., Łukasz Struski, Tabor, J., and Zieliński, B. (2023). Interpretability benchmark for evaluating spatial misalignment of prototypical parts explanations.

Samporová, S. (2024). Auxiliary self-organization-based loss for semi-supervised learning. Master's thesis, Comenius University in Bratislava. Supervisor: K. Malinovská.

Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., and Bojanowski, P. (2025). Dinov3.

Tarvainen, A. and Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 1195–1204, Red Hook, NY, USA. Curran Associates Inc.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. (2023). Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11941–11952.

Zohar, O., Lozano, A., Goel, S., Yeung, S., and Wang, K.-C. (2023). Open world object detection in the era of foundation models.