Can Topology Predict What Neural Networks Forget?

A Preliminary Investigation into Topological Signatures of Knowledge Persistence

Axion Deep Labs

Research Programs — EXP-01 (PERSIST) | Preliminary Report | February 2026

Abstract

Neural networks suffer from catastrophic forgetting. When trained on new information, they often overwrite previously learned knowledge. Despite decades of research into mitigation strategies, no prior work has investigated whether the geometric structure of learned representations predicts their vulnerability to being overwritten. We introduce a topological approach. Using persistent homology, we characterize the loss landscape around converged weight configurations and measure whether topological depth correlates with knowledge retention during sequential task training.

In preliminary experiments on Split CIFAR-100, we find that a Vision Transformer, ViT-Small, produces a loss landscape with nearly twice the topological persistence of a ResNet-18, H_0 equals 4254 versus 2151. Correspondingly, it exhibits slower and more gradual forgetting, retaining measurable accuracy twenty times longer during sequential training.

While results from two architectures do not constitute proof, the direction of this finding is consistent with our hypothesis. Networks that carve deeper topological structure into their loss landscape appear more resistant to catastrophic forgetting. If this relationship holds across additional architectures, it would provide both a diagnostic tool, predicting forgetting before it happens, and a foundation for topological regularization. This would enable training networks to learn in geometrically protected regions of parameter space.

To our knowledge, this is the first study connecting persistent homology of neural network loss landscapes to catastrophic forgetting.

1. The Problem

Every neural network deployed today is, in a fundamental sense, frozen. Once training is complete, introducing new knowledge typically comes at the cost of what the model already knows. This phenomenon was first described by McCloskey and Cohen in 1989 as *catastrophic interference* and is now widely referred to as *catastrophic forgetting*. The behavior is not subtle. A network trained to recognize fifty object categories, when subsequently trained on fifty new ones, does not accumulate knowledge into a unified set of one hundred. Instead, it adapts to the new categories while its performance on the original set collapses. The system does not integrate. It replaces. What appears to

be learning is often a zero-sum exchange in parameter space, where acquiring new competence destabilizes prior structure.

Importantly, this is not a problem that disappears with scale. Larger models, deeper architectures, and greater parameter counts do not fundamentally resolve the issue. A model with millions or even billions of parameters can forget just as decisively as a small network trained on a laptop. The intuition that capacity alone should allow separate tasks to coexist has not held up in practice. Instead, gradient-based training pushes parameters toward solutions optimized for the current objective, often traversing regions of the loss landscape that overwrite configurations supporting previous tasks. The instability is built into the dynamics of how these systems are trained.

A number of mitigation strategies have been proposed, and each has advanced the field, yet all operate by managing the symptoms rather than eliminating the cause. Replay-based methods store and revisit past examples so that older tasks remain present during optimization. Regularization approaches such as elastic weight consolidation attempt to identify parameters important to prior tasks and penalize changes to them. Architectural methods like progressive networks allocate new capacity for new tasks to prevent interference. These techniques can slow forgetting, redistribute it, or compartmentalize it, but they do not remove the underlying tension between plasticity and stability. The network remains fundamentally prone to destructive updates when objectives shift.

Humans, by contrast, appear to learn cumulatively. When a person studies calculus, they do not forget how to walk. When they acquire a new skill, earlier skills remain accessible. Even when knowledge fades, it rarely disappears instantaneously as a direct consequence of learning something new. This contrast suggests that the critical variable is not simply the number of parameters or neurons, but the structural organization of how knowledge is encoded. Biological systems seem to embed information in ways that allow new representations to form without erasing old ones. If artificial networks fail to do the same, the limitation may lie in the geometry and topology of their learned representations.

Seen from this perspective, catastrophic forgetting is not merely an optimization inconvenience. It is evidence that current neural networks organize knowledge in a fragile manner. Understanding the structural properties that make representations stable or unstable may be essential for building systems that truly learn over time. If we can identify what distinguishes representations that survive sequential training from those that collapse under it, we may move from patching the symptom to addressing the cause.

2. The Hypothesis

We propose that the *topological structure* of a neural network's loss landscape predicts its resistance to catastrophic forgetting. Specifically, we hypothesize that networks which carve deeper and more persistent topological features into their loss landscape during training are more resistant to forgetting when trained sequentially on new tasks. In this framing, resistance to forgetting is not merely a consequence of parameter count, architectural scale, or regularization strength. It is a structural property of the geometry that optimization sculpts in weight space.

The intuition is geometric. The loss landscape is the high-dimensional surface defined by model error as a function of its parameters, a perspective explored in foundational work on neural network optimization and visualization by Goodfellow et al. (2015) and Li et al. (2018). During training, gradient descent drives parameters into basins of low loss. Yet these basins differ in shape and internal structure. Some are shallow and weakly defined, while others are wide, deep, and geometrically intricate. When new training begins, the optimization process perturbs the parameters and reshapes

the local landscape. If a solution lies in a shallow basin, relatively small updates can displace the model into a configuration that no longer supports prior performance. If the solution resides in a deeper and more structured region, it may exhibit greater stability under perturbation. This view resonates with prior findings linking basin geometry, flatness, and curvature to generalization and stability properties in neural networks (Hochreiter and Schmidhuber, 1997; Keskar et al., 2017).

Persistent homology, a principal tool in topological data analysis, offers a rigorous and scale-invariant method for quantifying such geometric structure (Edelsbrunner and Harer, 2010; Ghrist, 2008). Rather than examining curvature at a single resolution, persistent homology tracks the birth and death of topological features across a filtration. It identifies connected components through H_0, loops through H_1, and higher-dimensional voids such as H_2, measuring how long each feature persists across scales. Features that persist over wide ranges are interpreted as structurally significant, whereas short-lived features are treated as topological noise. Because persistence is invariant under continuous deformation, it is particularly well suited for studying the highly nonconvex and irregular landscapes that arise in deep learning.

Topological data analysis has already been applied to neural systems. It has been used to study learned representations and data manifolds (Carlsson, 2009; Naitzat et al., 2020), and has been applied to loss landscapes directly, including work by Ballester and Araujo (2020) examining topological signatures of optimization geometry. In parallel, catastrophic forgetting has been studied extensively since its original characterization in 1989, producing a large body of research on replay-based methods, regularization approaches such as elastic weight consolidation (Kirkpatrick et al., 2017), and architectural strategies such as progressive networks (Rusu et al., 2016).

Yet these two lines of inquiry have remained separate. Topological data analysis has been applied to loss landscapes, and catastrophic forgetting has been examined for decades, but no work has directly asked the structural question at their intersection: *does the topological depth of learned representations predict their resistance to being overwritten?* By unifying these domains, we aim to move beyond surface-level mitigation strategies and toward a geometric account of why some learned solutions persist while others collapse under sequential training.

3. Method

3.1 Benchmark: Split CIFAR-100

We evaluate our hypothesis using Split CIFAR-100, a widely adopted benchmark in continual learning research. CIFAR-100 consists of 60,000 color images of size 32 by 32 pixels distributed across 100 object categories, with 600 images per class. Following standard protocol, we partition the dataset into two disjoint tasks. Task A contains classes 0 through 49 with 25,000 training images, while Task B contains classes 50 through 99 with 25,000 training images. Each task includes 2,500 test images. This split enforces a strict sequential learning scenario in which the model is first optimized on Task A and then exposed to Task B without access to prior task data. The setup isolates catastrophic forgetting in its most direct form and avoids confounds introduced by rehearsal or task mixing strategies commonly used in continual learning studies.

3.2 Architectures

ResNet-18 (He et al., 2016) serves as the convolutional baseline. It is an 18-layer deep residual network adapted for 32 by 32 inputs using a 3 by 3 initial convolution and no max pooling. The model contains approximately 11 million parameters. ResNet represents the canonical hierarchical

convolutional paradigm in which information flows locally through progressively abstract feature maps, with spatial inductive biases encoded by design.

ViT-Small represents the transformer paradigm applied to vision. The model consists of 4 transformer encoder layers, 4 attention heads, and 256-dimensional embeddings, with a patch size of 4, yielding 64 patches from each 32 by 32 image. The model contains approximately 3 million parameters. Unlike convolutional networks, Vision Transformers rely on global self-attention, allowing every spatial location to directly interact with every other location at each layer. This difference in information integration mechanism provides a structurally distinct comparison for evaluating geometric properties of the learned loss landscape.

3.3 Training Protocol

Both architectures are trained on Task A to convergence using stochastic gradient descent with momentum set to 0.9, a cosine learning rate schedule, and weight decay of 5×10^{-4}. ResNet-18 uses an initial learning rate of 0.1, while ViT-Small uses 0.01 to accommodate the different optimization dynamics of attention-based architectures, consistent with prior transformer training practice (Dosovitskiy et al., 2021). Each model is trained for 100 epochs. Final weight checkpoints at convergence are saved and used as the reference points for subsequent loss landscape analysis. All other training hyperparameters are held constant where architecture permits to ensure comparability.

3.4 Loss Landscape Sampling

To characterize the local geometry of each trained model, we follow the loss landscape visualization methodology of Li et al. (2018). Around the converged weight vector, we construct a two-dimensional slice of parameter space by generating two random direction vectors. These directions are normalized using filter normalization, meaning that for each filter or neuron, the perturbation magnitude is scaled to match the norm of the original weights. This ensures that perturbations are proportional to parameter scale and comparable across architectures with different layer structures and weight distributions.

We evaluate the loss on a 25 by 25 grid spanning the range from -1.0 to 1.0 along each direction, yielding 625 evaluation points. At each grid coordinate, we compute the cross-entropy loss on the Task A test set. The resulting grid defines a discretized approximation of the local loss surface surrounding the converged solution.

3.5 Persistent Homology

We compute persistent homology on the sampled loss surface using a lower-star filtration defined over an 8-connected grid graph. Each grid point is treated as a vertex with filtration value equal to the loss at that location. Edges between neighboring vertices appear at the maximum loss value of their two endpoints, forming a simplicial complex whose topology evolves as the filtration threshold increases.

Persistent homology is computed using Ripser (Bauer, 2021) with sparse distance matrices. We analyze H_0, representing connected components, and H_1, representing loops in the surface structure. For each topological feature, we measure its lifetime as the difference between its birth and death filtration values. Our primary summary statistic is total persistence, defined as the sum of all feature lifetimes within a homology dimension. Total persistence captures the aggregate topological depth of the landscape and provides a quantitative measure of geometric structure that is robust to small perturbations.

3.6 Forgetting Measurement

To measure catastrophic forgetting, the model trained to convergence on Task A is expanded with a new classification head supporting all 100 classes. The backbone weights are retained, and training proceeds sequentially on Task B using stochastic gradient descent at one-tenth of the original learning rate. This reduced rate stabilizes training while still allowing substantial parameter updates.

Task A test accuracy is evaluated at steps 100, 500, 1,000, 5,000, 10,000, and 25,000 during Task B training. No replay buffers, no regularization constraints, and no architectural isolation mechanisms are employed. This naive sequential training protocol exposes the intrinsic susceptibility of each architecture to catastrophic forgetting and allows us to examine whether measured topological depth correlates with retention dynamics.

4. Preliminary Results

4.1 Topological Features

Table 1 summarizes the topological statistics of the loss landscape at convergence on Task A for both architectures.

Metric	ResNet-18	ViT-Small
Task A Test Accuracy	82.0%	62.2%
H■ Total Persistence	2,151.5	4,254.2
H■ Feature Count	624	624
H■ Max Lifetime	5.21	11.07
H■ Total Persistence	0.0	0.0
Loss Range	[0.86, 5.21]	[0.86, 11.07]

Table 1. Topological features of the loss landscape at convergence on Task A. ViT-Small produces nearly 2× the total H■ persistence despite lower accuracy and fewer parameters.

Despite achieving lower Task A accuracy, 62.2 percent compared to 82.0 percent, the Vision Transformer produces a loss landscape with substantially greater measurable topological depth. Total H_0 persistence for ViT-Small is 4,254.2, almost exactly double the 2,151.5 measured for ResNet-18. Because total persistence aggregates the lifetimes of all connected components across the filtration, this result indicates that the transformer converges to a basin with significantly more persistent geometric structure.

The maximum lifetime of a single H_0 feature is also doubled, 11.07 for ViT-Small compared to 5.21 for ResNet-18. This suggests that the deepest connected basin surrounding the ViT solution extends across a much wider range of filtration values. In geometric terms, the transformer appears to settle into a solution region that is not merely low in loss, but structurally deeper when measured through scale-invariant topological persistence.

Neither architecture produced nonzero H_1 persistence at the current grid resolution. This means that no loop structures were detected in the sampled landscape slice. There are two possible interpretations. The first is that the local minima are genuinely smooth and simply connected in the sampled neighborhood. The second is that the 25 by 25 grid is too coarse to resolve higher-dimensional structure. Prior loss landscape work has shown that apparent smoothness can depend heavily on sampling resolution (Li et al., 2018). Future experiments will evaluate finer grids and alternative

filtrations to determine whether higher-order topological features emerge.

What is most important at this stage is not higher-dimensional structure, but the clear quantitative gap in H_0 persistence. The transformer produces nearly twice the total topological persistence despite having fewer parameters and lower classification accuracy. This dissociation suggests that topological depth is not reducible to raw performance or model size. It reflects a structural property of the learned basin itself.

4.2 Forgetting Dynamics

Table 2 reports Task A accuracy and retention ratio during sequential training on Task B.

Step	ResNet-18 Task A Acc	ResNet-18 Retention	ViT-Small Task A Acc	ViT-Small Retention
0	82.0%	100%	62.2%	100%
100	0.2%	0.2%	6.0%	9.6%
500	0.0%	0.0%	6.1%	9.8%
1,000	0.0%	0.0%	4.2%	6.8%
5,000	0.0%	0.0%	1.7%	2.7%
10,000	0.0%	0.0%	0.8%	1.3%
25,000	0.0%	0.0%	0.1%	0.2%

Table 2. Task A accuracy and retention during sequential Task B training. ResNet-18 forgets instantly; ViT-Small degrades gradually over thousands of steps.

The forgetting dynamics differ dramatically across architectures. ResNet-18 undergoes immediate catastrophic forgetting. Within 100 training steps on Task B, Task A accuracy collapses from 82.0 percent to 0.2 percent, effectively a complete erasure. By step 500, accuracy is exactly 0.0 percent, below random chance for 50 classes, which would be 2.0 percent. The original representation is not gradually degraded. It is overwritten almost instantly.

ViT-Small exhibits a markedly different trajectory. After 100 steps, it retains 6.0 percent Task A accuracy, thirty times higher than ResNet-18 at the same point. At step 500, retention remains at 6.1 percent. Even after 10,000 steps of Task B training, measurable retention persists at 0.8 percent. The decay curve is progressive rather than abrupt. Knowledge is eroded over time rather than destroyed immediately.

Critically, the architecture that exhibited nearly twice the total H_0 persistence also demonstrated dramatically slower forgetting. The model that carved a deeper and more persistent basin into its loss landscape retained prior knowledge longer under identical sequential training conditions. While this comparison involves only two architectures and does not establish causality, the directional alignment between topological depth and retention is striking.

The central empirical observation is therefore structural. The network that formed a deeper geometric basin proved more resistant to destructive updates. The model that converged to a shallower basin was rapidly displaced. This correspondence is consistent with our hypothesis that topological depth in the loss landscape predicts vulnerability to catastrophic forgetting.

5. Interpretation

Two data points do not constitute proof, and we are explicit about that limitation. This study compares only two architectures under a single benchmark and training protocol. However, the direction of the result aligns precisely with the proposed hypothesis, and the magnitude of the observed difference is substantial. The gap in total H_0 persistence is nearly twofold, and the divergence in forgetting dynamics is not incremental but categorical. One architecture undergoes near-immediate collapse, while the other degrades gradually over thousands of training steps. This is not a marginal fluctuation that can be casually attributed to sampling noise or minor hyperparameter sensitivity. It is a structural contrast.

The result is also architecturally revealing. ViT-Small contains roughly three million parameters compared to approximately eleven million in ResNet-18, and it achieves lower Task A accuracy at convergence. Yet it produces a loss landscape with nearly twice the total topological persistence and demonstrates markedly slower forgetting. This decoupling is important. If topological depth were merely a proxy for model capacity or raw performance, we would expect the larger and more accurate model to exhibit greater persistence. Instead, the opposite occurs. The metric appears to capture something structurally distinct about how knowledge is encoded in parameter space.

One plausible explanation lies in the representational geometry induced by self-attention. In transformer architectures, every spatial position can directly attend to every other position at each layer. This creates a densely interconnected representational structure in which features are globally integrated. Convolutional networks, by contrast, aggregate information locally through spatially constrained receptive fields and hierarchical composition. While residual connections enable deeper signal flow, the inductive bias remains fundamentally local. It is conceivable that dense global integration distributes task-relevant information across a broader portion of parameter space, embedding it within a more interconnected basin. In contrast, locally concentrated filters may encode task-specific structure in narrower regions that are more easily displaced by subsequent gradient updates. Under this interpretation, attention-based models may naturally carve deeper and more persistent geometric structure into the loss landscape because knowledge is distributed rather than compartmentalized.

If the observed relationship between topological depth and forgetting resistance generalizes across additional architectures and datasets, two practical applications follow.

> **First, topological persistence could serve as a diagnostic tool.** After training on an initial task, practitioners could compute total persistence on a sampled loss landscape slice to estimate vulnerability to catastrophic forgetting before deploying a model in a sequential or continual learning environment. This would provide a structural risk assessment grounded in geometry rather than post-hoc performance degradation.
>
> **Second, topological persistence could be incorporated directly into the training objective as a regularization signal.** A topological regularizer could penalize reductions in total persistence or explicitly reward the formation of deeper, longer-lived topological features. Such a mechanism would encourage optimization toward geometrically stable regions of parameter space. Instead of merely minimizing loss, the network would be incentivized to settle into structurally protected basins that are more resistant to perturbation from future tasks. In practical terms, the goal would be to shape the learning dynamics so that models carve deep, stable basins in weight space rather than shallow configurations that are easily overwritten.

While preliminary, these findings suggest that catastrophic forgetting may be partially predictable from the geometry of the learned solution itself. If so, the problem shifts from being purely algorithmic to

fundamentally structural, opening a new line of inquiry at the intersection of optimization, topology, and continual learning.

6. What's Next

This preliminary report establishes the experimental framework, defines the measurement pipeline, and presents initial empirical evidence supporting the topology–forgetting hypothesis. The current results demonstrate directional alignment between topological persistence and retention under sequential training, but they represent only the first stage of a broader investigation. The next phases of work are designed to test robustness, improve measurement fidelity, and establish statistical validity across architectural families.

Additional architectures are currently in progress. We are expanding the study to include ResNet-50 as a deeper convolutional baseline, wider ResNet variants to disentangle depth from width effects, and an LSTM-based recurrent model to introduce a fundamentally different temporal inductive bias. Evaluating three or more additional architectures will allow formal Spearman rank correlation analysis between total topological persistence and retention rate. Rank correlation is particularly appropriate here because the hypothesis predicts monotonic alignment rather than strict linear scaling. A statistically significant correlation across diverse architectures would constitute the first rigorous quantitative validation of the proposed relationship.

We also plan to increase loss landscape sampling resolution. The absence of H_1 features in the current experiments may reflect genuine local smoothness, but it may also be a discretization artifact of the 25 by 25 grid. Prior loss landscape studies have shown that qualitative geometric structure can depend strongly on sampling density. We will therefore evaluate 51 by 51 and 101 by 101 grids using optimized sparse computation and memory-efficient persistent homology pipelines. Higher resolution sampling will allow us to determine whether loop structures or higher-order features emerge when the landscape is examined more finely.

Stochastic variation must also be addressed systematically. Both model training and random direction sampling in parameter space introduce variability. To quantify uncertainty, each experiment will be repeated across five independent random seeds. This will enable reporting of confidence intervals for total persistence, maximum lifetime, and forgetting curves. Statistical aggregation will allow us to distinguish structural effects from incidental variance.

If the correlation between topological depth and retention remains stable under these controls, we will proceed to the intervention phase. In Phase 5, we will implement a topological regularizer designed to preserve geometric structure during sequential training. The proposed formulation is

$$L_{topo} = \lambda \cdot \max(0, P_A - P_{current})$$

where P denotes total persistence and P_A represents persistence at convergence on the original task. This penalty discourages reductions in persistence during subsequent optimization, effectively resisting topological erosion of the learned basin. The objective is not merely to observe correlation but to test causality by actively shaping the loss landscape toward deeper and more stable regions of parameter space.

Pending results across at least five architectures with formal statistical testing, the target venue is NeurIPS or ICML within the continual learning track. Acceptance will require demonstrating reproducibility, architectural generality, and statistically significant correlation between geometric structure and forgetting dynamics. The long-term objective is to establish topological depth not as

an isolated metric, but as a principled structural lens through which continual learning stability can be understood and engineered.

References

[1] Ballester, R. and Araujo, X. (2020). On the interplay between topological data analysis and deep learning. *NeurIPS Workshop on Topological Data Analysis*.

[2] Bauer, U. (2021). Ripser: Efficient computation of Vietoris–Rips persistence barcodes. *Journal of Applied and Computational Topology*, 5(3), 391–423.

[3] Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society*, 46(2), 255–308.

[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A. et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*.

[5] Edelsbrunner, H. and Harer, J. (2010). *Computational Topology: An Introduction*. American Mathematical Society.

[6] Ghrist, R. (2008). Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1), 61–75.

[7] Goodfellow, I.J., Vinyals, O. and Saxe, A.M. (2015). Qualitatively characterizing neural network optimization problems. *International Conference on Learning Representations*.

[8] He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

[9] Hochreiter, S. and Schmidhuber, J. (1997). Flat minima. *Neural Computation*, 9(1), 1–42.

[10] Keskar, N.S., Mudigere, D., Nocedal, J., Smelyanskiy, M. and Tang, P.T.P. (2017). On large-batch training for deep learning: Generalization gap and sharp minima. *International Conference on Learning Representations*.

[11] Kirkpatrick, J., Pascanu, R., Rabinowitz, N. et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13), 3521–3526.

[12] Kumaran, D., Hassabis, D. and McClelland, J.L. (2016). What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends in Cognitive Sciences*, 20(7), 512–534.

[13] Li, H., Xu, Z., Taylor, G., Studer, C. and Goldstein, T. (2018). Visualizing the loss landscape of neural nets. *Advances in Neural Information Processing Systems*.

[14] McCloskey, M. and Cohen, N.J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24, 109–165.

[15] Naitzat, G., Zhitnikov, A. and Lim, L.-H. (2020). Topology of deep neural networks. *Journal of Machine Learning Research*, 21(184), 1–40.

[16] Otter, N., Porter, M.A., Tillmann, U., Grindrod, P. and Harrington, H.A. (2017). A roadmap for the computation of persistent homology. *EPJ Data Science*, 6(1), 1–38.

[17] Rusu, A.A., Rabinowitz, N.C., Desjardins, G. et al. (2016). Progressive neural networks. *arXiv preprint arXiv:1606.04671*.

[18] Tononi, G. and Cirelli, C. (2014). Sleep and the price of plasticity: From synaptic and cellular homeostasis to memory consolidation and integration. *Neuron*, 81(1), 12–34.