Student Abstract: Sleep-Like Replay Reduces Loss-Landscape Sharpness to Improve Generalization

Krishi Chawda[1,2], **Jean Erik Delanois**[1], **Giri Krishnan**[3], **Maxim Bazhenov**[1]
[1] Department of Medicine, University of California, San Diego
[2] Department of Computer Science & Engineering, University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92092
[3] ARTISAN, Georgia Institute of Technology
North Avenue, Atlanta, GA 30332
krchawda@ucsd.edu, jdelanois@ucsd.edu, giri@gatech.edu, mbazhenov@ucsd.edu

Abstract

One of the central challenges in deep learning is that models trained on new tasks often overfit and lose the ability to generalize. This issue arises because gradient descent often converges to solutions in regions of the loss landscape that are sharp near their minima. High sharpness leads to rapid performance loss when test data are perturbed or statistically shifted. Although sharpness has been linked to generalization, few methods directly target it to improve generalization. Here we demonstrate that an unsupervised, sleep-like replay algorithm identifies low loss regions with lower sharpness leading to improvement in generalization to distortions, including Gaussian and salt-and-pepper noise. Our study identifies loss-function sharpness as a unifying measure for generalizable learning and robustness, and points to new principles for designing resilient AI systems.

Introduction

Although artificial neural networks (ANNs) now rival human performance on many tasks, they often underperform when test data differ even slightly from the training distribution (Geirhos et al. 2019). This lack of generalization raises two practical issues. First, ANNs are typically trained on curated datasets that capture idealized content, yet real-world inputs are frequently disturbed or noisy and not seen during training. Second, ANNs are vulnerable to adversarial examples - inputs crafted to fool the model while remaining nearly imperceptible (Szegedy et al. 2014). These issues limit real-world applicability and introduce security risks.

In human and animal brains, replay during sleep plays a central role in stabilizing long-term memory, preventing interference, and generalizing knowledge (Stickgold 2005; Rasch and Born 2013). Although replay is well established in neuroscience and AI algorithms such as Sleep Replay Consolidation (SRC) (Tadros et al. 2022; Delanois et al. 2023; Kubo, Delanois, and Bazhenov 2025) were developed to mimic it, we still lack a clear understanding of the fundamental changes unsupervised replay induces in synaptic weights - both in biological networks and ANNs.

In this study, we examine the problem through the lens of the loss landscape. Sharpness - how sensitive the loss is to small parameter perturbations - has been linked to both generalization and robustness. Yet little work has tested how sharpness influences the effects of distortions or how replay mechanisms like SRC reshape the landscape. We show that, for models trained on clean data, distortions progressively increase loss gradients and destabilize learning, whereas SRC moves the model to the low loss regions with lower sharpness, making it less sensitive to distortions.

By linking sleep-inspired replay to loss-landscape geometry, our study points to principles for building AI systems that learn continuously, stably, and generalize to real-world conditions, while also illuminating roles of biological sleep.

Algorithm

Image Distortions

To assess robustness, we trained a small CNN on MNIST data and tested it on systematically corrupted inputs, including salt-and-pepper noise (0.1–0.6) and speckle noise (0.25–1.0). Separate datasets were created for each distortion type and level. We evaluated accuracy and sharpness to examine how distortions affect the loss landscape.

Sleep Replay Consolidation (SRC)

To implement unsupervised sleep-like replay (Tadros et al. 2022), the trained ANN was converted to a spiking neural network (SNN) with identical architecture. The SNN was driven by Poisson-distributed binary inputs whose rates reflect per-pixel averages from prior training data. During the sleep phase, local Hebbian plasticity increased synaptic weights when presynaptic activity preceded postsynaptic firing, and decreased them when postsynaptic spikes occurred without presynaptic support. Finally, SNN was mapped back to an ANN and classification accuracy was evaluated.

Evaluation

Sharpness was estimated as the spread of attainable losses in a small ℓ_2 neighborhood around the trained parameters. Formally, sharpness can be defined as (Foret et al. 2021)

$$S(\theta;\varepsilon) = \max_{\|\Delta\|_2\le\varepsilon} \mathcal{L}(\theta+\Delta) - \min_{\|\Delta\|_2\le\varepsilon} \mathcal{L}(\theta+\Delta).$$

In practice, we approximated this quantity by evaluating losses after perturbing parameters in the positive and negative gradient directions at fixed radius ε. This captures the

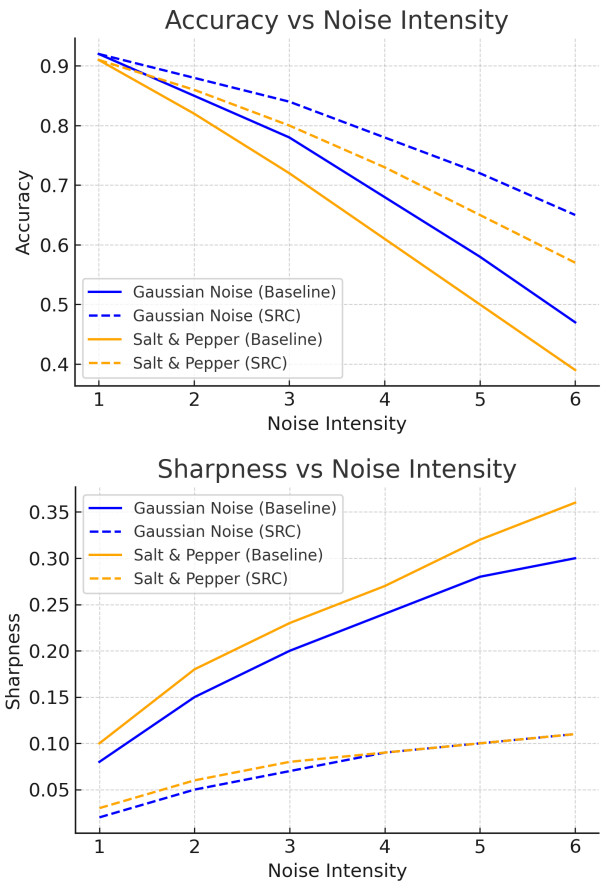

Figure 1: Accuracy (top) and sharpness (bottom) under distortions. Solid lines - baseline; dashed lines - SRC.

local variability of the loss landscape rather than the increase from a single reference point. For distortions, we compared $S(\theta;\varepsilon)$ (and accuracy) for baseline and SRC models across increasing corruption intensities.

Results

The model trained on clean images was tested on data corrupted with varying levels of Gaussian noise or salt-and-pepper noise. Increasing distortion reduced accuracy, which was partially offset by applying SRC (Figure 1, top). Sharpness increased with distortion level (Figure 1, bottom), indicating that the loss gradients grow as the data became more corrupted. Importantly, applying SRC significantly reduced sharpness at all distortion levels. Figure 2 summarizes these results by averaging accuracy and sharpness across all levels of distortions. On average we observed 10% increase in accuracy and about a threefold decrease in sharpness after SRC.

Conclusions

We show that SRC improves generalization - without access to distorted data - by altering synaptic weights to reach solutions with a flatter loss landscape around the minimum.

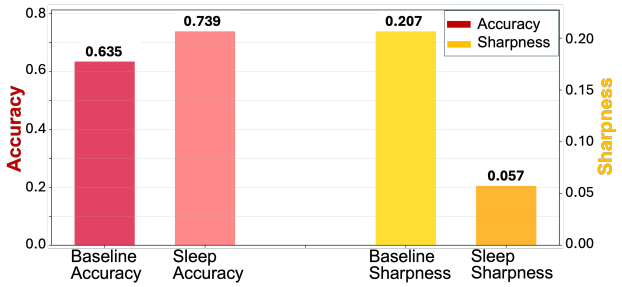

Figure 2: Sharpness and Accuracy averaged across distortions. SRC increased accuracy and decreased sharpness.

Our study points to a plausible synaptic-weight–dynamics strategy the brain may employ during sleep to generalize episodic memories into semantic knowledge. Applied to ANNs, sleep-like replay improves performance in a fully unsupervised manner, requires no additional data, and can be applied to already trained models.

Acknowledgments

Supported: NSF (2323241, 2223839), NIH (RFNS132913).

References

Delanois, J. E.; Ahuja, A.; Krishnan, G.; Tadros, T.; and Bazhenov, M. 2023. Improving Robustness of Convolutional Networks Through Sleep-Like Replay. In *2023 22nd IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE.

Foret, P.; Kleiner, A.; Mobahi, H.; and Neyshabur, B. 2021. Sharpness-Aware Minimization for Efficiently Improving Generalization. In *Proceedings of the International Conference on Learning Representations (ICLR)*. Published as a conference paper at ICLR 2021.

Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F. A.; and Brendel, W. 2019. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*. ICLR 2019.

Kubo, Y.; Delanois, J. E.; and Bazhenov, M. 2025. Toward Lifelong Learning in Equilibrium Propagation: Sleep-like and Awake Rehearsal for Enhanced Stability. *arXiv preprint arXiv:2508.14081*.

Rasch, B.; and Born, J. 2013. About sleep's role in memory. *Physiological Reviews*, 93(2): 681–766.

Stickgold, R. 2005. Sleep-dependent memory consolidation. *Nature*, 437(7063): 1272–1278.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*. ICLR 2014; original preprint arXiv:1312.6199.

Tadros, T.; Krishnan, G. P.; Ramyaa, R.; and Bazhenov, M. 2022. Sleep-like unsupervised replay reduces catastrophic forgetting in artificial neural networks. *Nature Communications*, 13(1): 7742.