AFRL-AFOSR-VA-TR-2022-0007

NETRIUS Deep Chaos

Mohammed Eslami
Netrias, LLC
1162 Gateway Dr
Annapolis, MD, 21409-4629
US

10/20/2021
Final Technical Report

Air Force Research Laboratory
Air Force Office of Scientific Research
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
20-10-2021	Final	15 Mar 2020 - 14 Sep 2021

4. TITLE AND SUBTITLE
NETRIUS Deep Chaos

5a. CONTRACT NUMBER
FA9550-20-C-0001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)
Mohammed Eslami

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Netrias, LLC
1162 Gateway Dr
Annapolis, MD 21409-4629
US

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AF Office of Scientific Research
875 N. Randolph St. Room 3112
Arlington, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/AFOSR RTA2

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-AFOSR-VA-TR-2022-0007

12. DISTRIBUTION/AVAILABILITY STATEMENT
A Distribution Unlimited: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Artificial intelligence, and deep learning in particular, is increasingly becoming a critical component in our national infrastructure. It is imperative that techniques used to train these models be consistent in their predictions. Fooling these systems with small perturbations to the input data is a well-known problem. Identifying and characterizing regions of sensitivity is less understood in the AI domain, but well studied in nonlinear dynamics mathematics. We seek to investigate whether deep neural networks are chaotic, and, if so, can they be stabilized?

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			FARIBA FAHROO
U	U	U	UU	9	**19b. TELEPHONE NUMBER** *(Include area code)* 426-8429

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
20210914	Final Report	3/15/2020	9/14/2021

4. TITLE AND SUBTITLE

Deep Chaos: Characterizing the Stability of Deep Learning Models

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
FA9550-20-C-0001		
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)

Eslami, Mohammed. Mischaikow, Konstantin. Kalies, William. Gamero, Marcio. Weston, Mark.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Netrias LLC 1162 Gateway Dr Annapolis MD 21409	12

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
Defense Advanced Research Projects Agency: Defense Sciences Office 675 N Randolph St Arlington, VA 22203	DARPA DSO	0001AM

12. DISTRIBUTION/AVAILABILITY STATEMENT

Distribution Statement A. Approved for public release; distribution is unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Artificial intelligence, and deep learning in particular, is increasingly becoming a critical component in our national infrastructure. It is imperative that techniques used to train these models be consistent in their predictions. Fooling these systems with small perturbations to the input data is a well-known problem. Identifying and characterizing regions of sensitivity is less understood in the AI domain, but well studied in nonlinear dynamics mathematics. We seek to investigate whether deep neural networks are chaotic, and, if so, can they be stabilized?

15. SUBJECT TERMS

Deep Learning, Machine Learning, Geometric Manifold, Stability

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
Unclassified	Unclassified	Unclassified	UU	8

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
Mark Weston	781-507-4340

PREVIOUS EDITION IS OBSOLETE.

STANDARD FORM 298 (REV. 5/2020)
Prescribed by ANSI Std. Z39.18

Deep Chaos: Quantifying the Sensitivity of Training Deep Learning Models

PI: Mohammed Eslami (Netrias, LLC)
Team: Konstantin Mischaikow (Rutgers University), Marcio Gameiro (Rutgers University), William Kalies (Florida Atlantic University), Hamed Eramian (Netrias, LLC)

Publication: Eslami, M., Eramian, H., Gameiro, K. Mischaikow and W. Kalies. *Extracting Global Dynamics of Loss Landscape in Deep Learning Models*. arXiv:2106.07683. Neural Information Processing Systems 2021. *Submitted.*

Project Objectives

Artificial intelligence, and deep learning in particular, is increasingly becoming a critical component in our national infrastructure - from scientific discovery, to automation and manufacturing, to fraud detection and cybersecurity. It is thus imperative that techniques used to train these models be consistent in their predictions. Fooling or hacking these systems with small perturbations to the input data or model is a well-known problem. These attacks can be thought of as AI's Stuxnet. Adversaries can attack both the data and the models, leaving the overall loss unchanged. Weaknesses in AI software cannot be identified automatically, leaving the Department of Defense at risk of depending on AI that can be fooled by our adversaries.

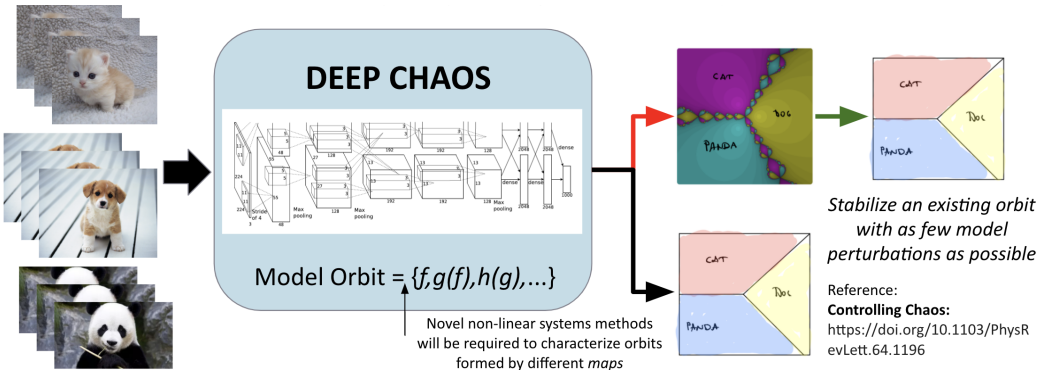


Figure 1: Deep Chaos seeks to automatically discover vulnerable regions, or complex basin boundaries, by formulating the optimization of a neural network as a dynamical system.

Identifying and characterizing regions of sensitivity, as well as the ability to control the sensitivity is less understood in the AI domain, but well studied in nonlinear dynamics mathematics. We seek to investigate whether deep neural networks are chaotic, or have complex basin boundaries, and, if so, can they be stabilized (Figure 1)? The primary challenge associated with modeling the optimization of neural networks as a dynamical system to automatically identify vulnerable regions is that of scale. Techniques that can provide complete characterizations of the evolution of dynamical systems are computationally expensive and require storage of the

entire space spanned by the model. In the case of deep neural networks, such as GPT-3, this is hundreds of billions of parameters.

Methods

The Deep Chaos effort focused on the development of a toolkit to characterize the loss landscape of neural networks. The loss landscape is defined as the high-dimensional surface of the loss function of the neural network. During training, a dynamical system, such as stochastic gradient descent, takes a set of initial conditions (architecture, starting weights, and training data), and traverses through the loss landscape to find the set of weights, or parameters, that minimizes the loss function.

Deep Chaos focused on the development of a toolkit for the Dynamical Organization of Deep Learning Loss Landscapes, or DOODL3. DOODL3 is a toolkit that formulates the training of neural nets as a dynamical system and analyzes the learning process to discover *how* a model learns (Figure 1). This analysis provides robustness guarantees of deep learning models and helps mitigate against regions of unstable or ineffective learning. These tools define parameter bounds that are unstable for deep learning models, a process that is analogous to similar robustness and stability analyses that we perform for other critical components of our national infrastructure, such as aircraft or the power grid. If these unstable regions are unavoidable during the learning process, these tools can inform methods that will ensure escape in a tractable amount of time to deploy robust and consistent deep learning models.

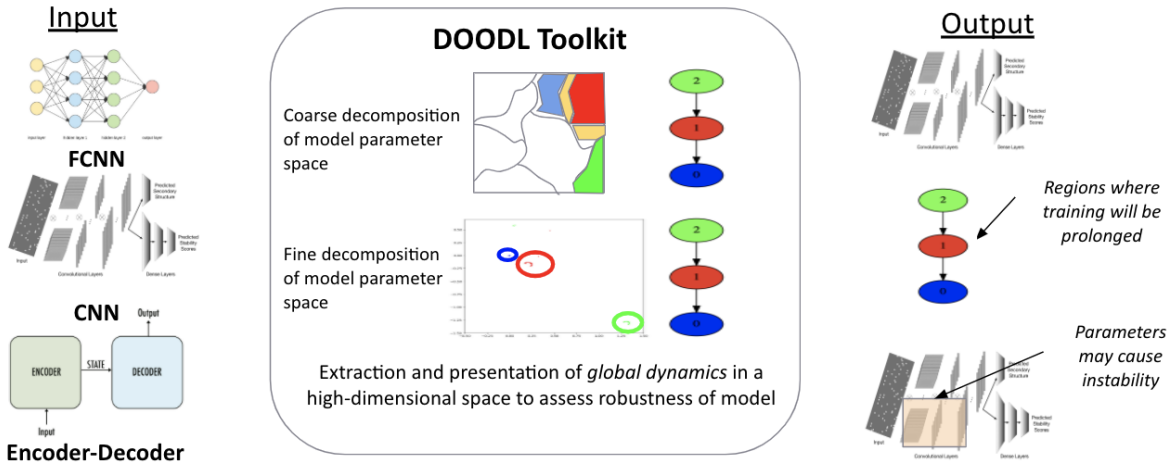


Figure 1: DOODL3 toolkit analyzes how a model learns at coarse and fine resolutions to identify regions in the process that may cause instability.

Techniques from topology and geometry can extract a system's regions of stability at both **coarse and fine levels** through analysis of its *global* dynamics. From a deep learning perspective, the global dynamics of a model is defined as how a family of models learn. Stable and unstable regions can be presented to human and/or automated controllers in a scalable, interpretable format. Analytical tools to extract dynamics from high-dimensional systems can be based on what we define as the minimal perspective: recurrent versus non-recurrent dynamics.

A state of the system is recurrent if for all future times the system returns to a nearby state. For a non-recurrent point there is no return. From the point of view of deep learning, this divides phase space into 3 regions:

A. The recurrent states that are associated with local minimizers of the objective function.
B. The non-recurrent states that converge to local minimizers of the objective function.
C. The recurrent states that are not local minimizers of the objective function along with non-recurrent states that converge to them.

If (A) has multiple distinct sets, then there are multiple solutions to an objective which strongly suggests the possibility of predictive inconsistencies, i.e. high generalization error. Each element of (B) is identified with a single set in (A), and the elements of (C) form boundaries separating the regions defined by (A) and (B) called *separatrices*. This provides information about what gets learned given the initial conditions. While the recurrent states in (C) are almost impossible to observe directly, i.e. through repeated training, their existence and structure determine the expected level of generalization error, sufficient time for learning, and robustness of transfer of learning algorithms. Two simple examples of the breakdown of phase space are shown in Figure 2. The simplest example of a recurrent set is a single fixed point to which all other points converge (Figure 2A). Since deep learning models are overparameterized, a recurrent set need not be a single fixed point and in fact is often a high dimensional manifold. Furthermore, there can be multiple recurrent sets that would then be made up of multiple manifolds (Figure 2B).

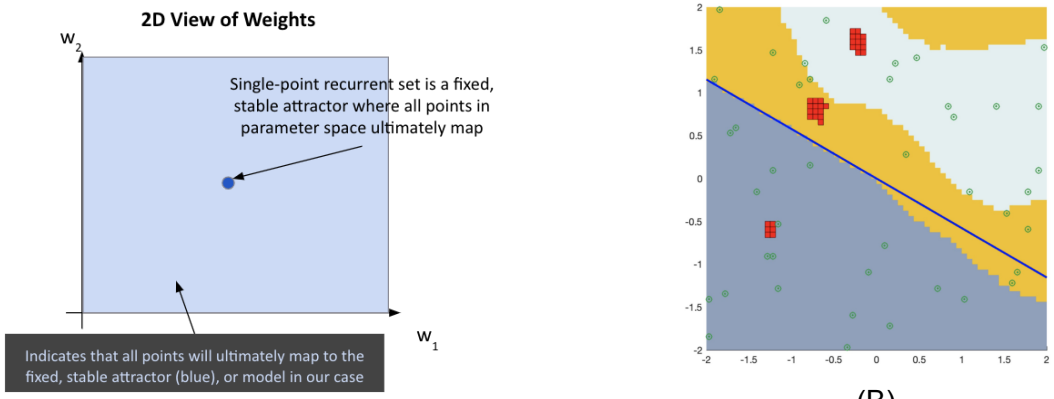


Figure 2: Example sets extracted from dynamical systems that are (A) a single fixed point and (B) two stable fixed points and an unstable separatrix.

To apply these techniques to deep learning models, DOODL3 takes a dataset, a neural network architecture, an optimizer, and a loss function as input and will analyze the neural network training dynamics. In its current implementation, DOODL3 varies initial weights using a uniform distribution in parameter space. It then takes the output weights, which it considers as the parameters, and decomposes the parameter space into an adaptive grid. Using the first and last epoch, it then builds a mapping from one box to the next, refining regions where there are recurrent sets. This is shown in Figure 3.

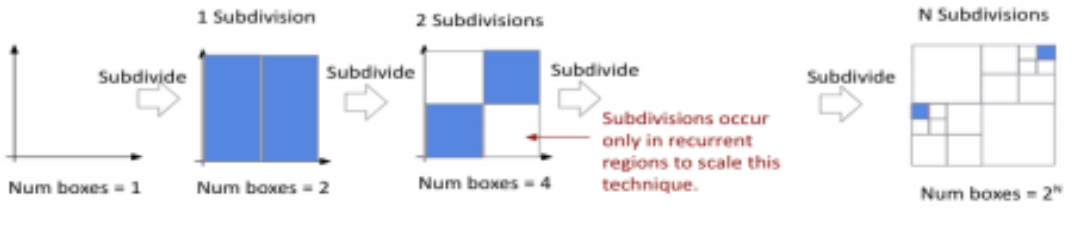
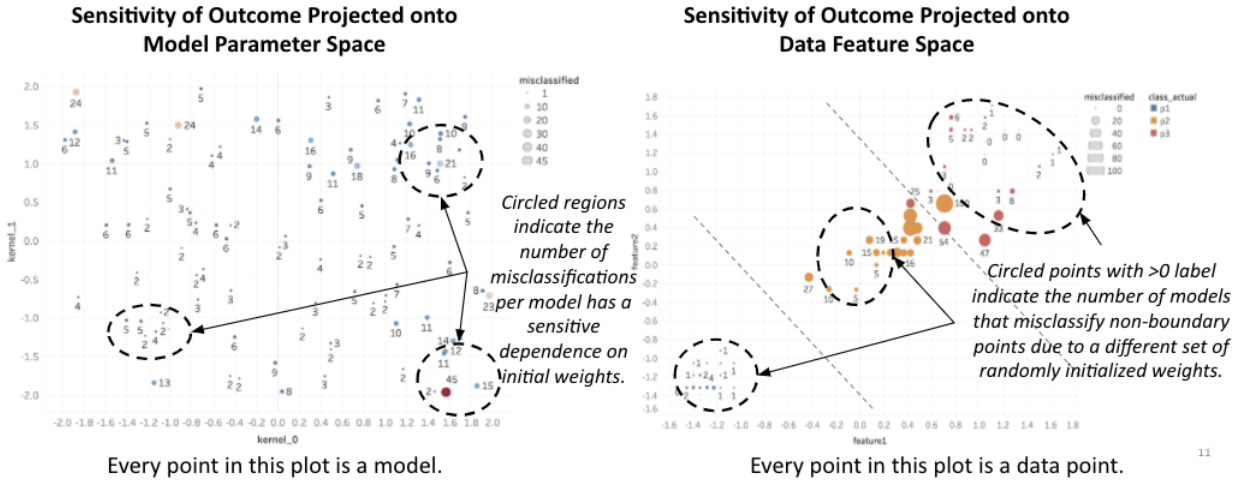


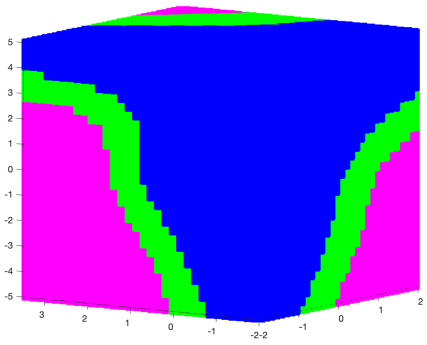
Figure 3: Multi-scale adaptive grid decomposition of parameter space to identify recurrent sets (blue).

Results

Our initial efforts have developed techniques that scale nonlinear systems analysis tools to operate on twice the number of dimensions used in traditional tools. We tested DOODL3 on two toy examples, a fully connected neural network (FCNN) and a convolutional neural network (CNN). The FCNN used the Iris dataset, trained a baseline model for 100 training cycles and then evaluated the inconsistency present in the predictions (Figure 4a and 4b). Selecting a vanilla gradient-descent algorithm as the dynamical system, we are able to identify and characterize the separatrix that acts as a boundary between different model solutions (Figure 4c). DOODL3 would consider the points in the green region to be unstable and would recommend to either start in the blue or pink to achieve consistent predictions.



(a) (b)



(c)

Figure 4: Deep neural network models with >100 random weight initializations and gradient descent-like optimization. (a) A 2D projection of a subset of the models that are trained and tested on the same data with a different set of initial weights showing inconsistent predictions for different random initial weights. (b) A 2D projection of the test dataset where the size of the circles indicates the number of models that misclassify the point. (c) Dynamics of the model show two stable regions (blue, purple) and one unstable (green) separatrix of the model. Initial weights in green must be avoided because it will probably be more expensive to compute a solution with these initial weights.

We then swept the number of layers, number of nodes, epochs, and batch size to generate training dynamics data for different architectures and model parameters. Analyzing the predictions for consistency, while the overall model performance always remained >95%, we see that the entropy of the predictions (the number of different predictions) is quite large (Figure 5). This implies that one should not just use accuracy as a metric for model performance, but entropy is a good measure of consistency in model predictions for the same test point.

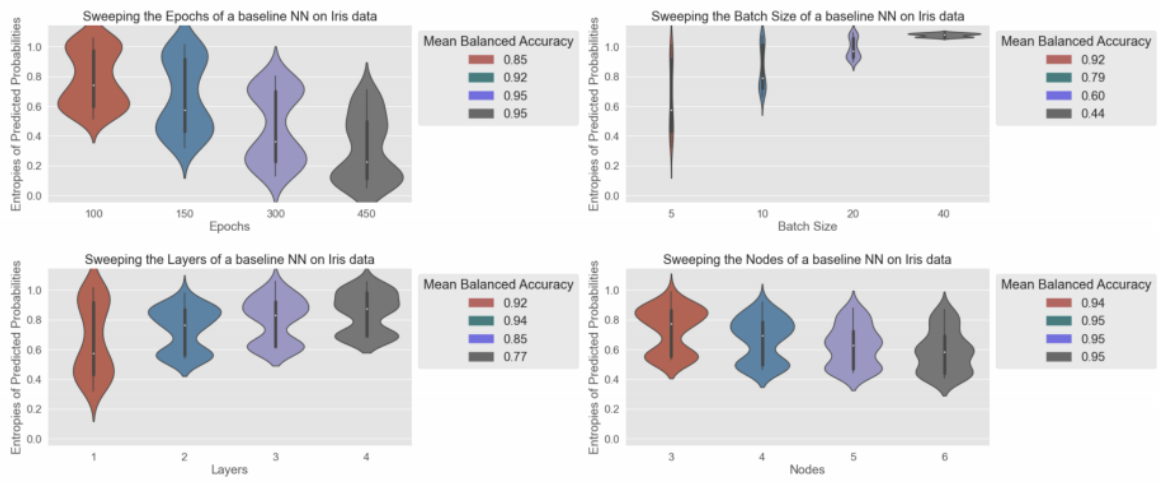


Figure 5: Distribution of entropy across every test point in iris dataset. A single layer, single node neural network trained for 150 epochs was used as the baseline. X-axis indicate additional hyperparameters and their ranges that were tested.

Feeding the data from the model, namely the parameters at the start and end points, to DOODL3, we see that when the model is trained for longer epochs, the separatrices (region between blue and green shade) begin to emerge (Figure 6).

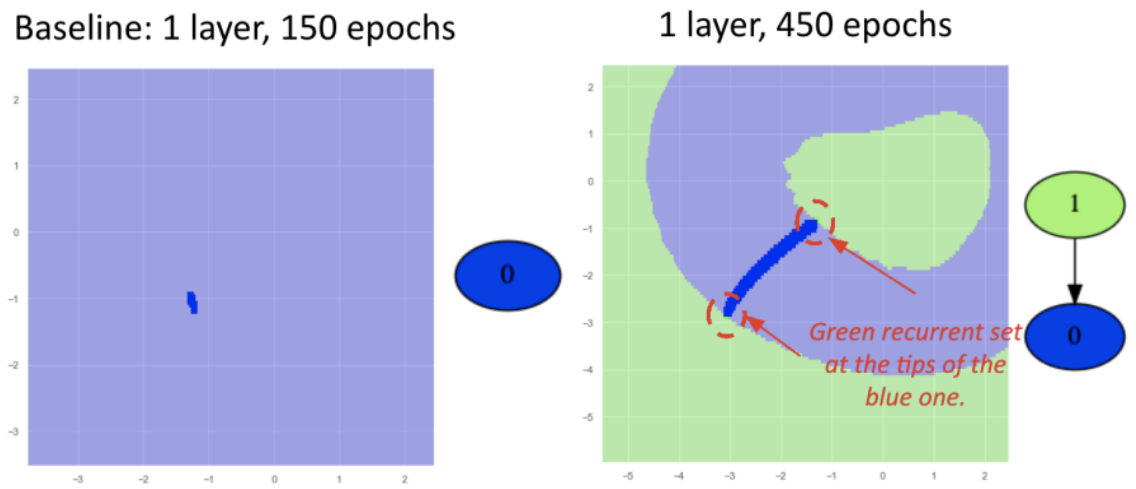


Figure 6: Starting point for regions in green are likely places that will require longer training time.

Applying the same analyses on the CNN, we see a more interesting pattern of the decomposed parameter space (Figure 7). We then selected a set of points in the green and blue regions and found that, on average, models in the blue region had higher accuracy than those within the green region. We believe this is because points in the green region need to be run for longer epochs, on average, to achieve similar results to that of the blue region.

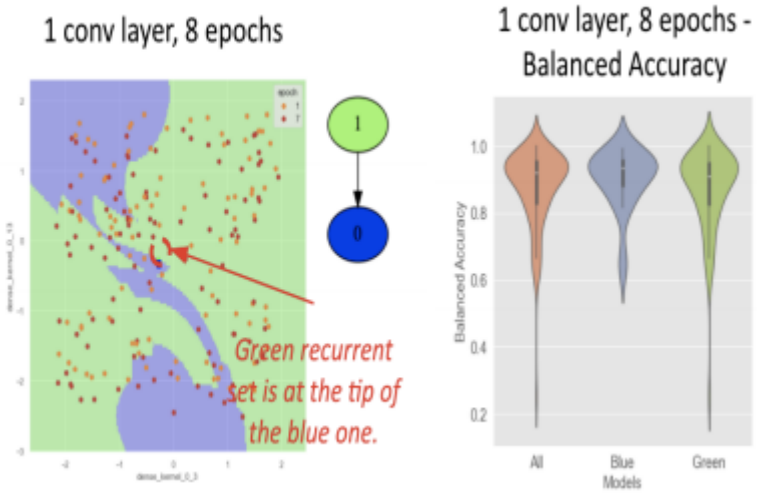


Figure 7: Recurrent set regions for a CNN again show green regions will likely need longer training times. The balanced accuracy distributions provide evidence to the hypothesis that models with initial weights that start in the blue region converge more quickly.

Technical Feasibility

While much progress has been made on this effort, to fully realize the benefits of DOODL3 to the deep learning community, additional research is required. Three areas that require further research for successful development of this approach are:

1. **Decomposition of Model Parameter Space:** Grid decompositions are defined and refined over R^N, where *N* is the number of parameters, grows combinatorially with added dimension, which is computationally and memory inefficient. To scale this technique for deep learning models, we propose to revolutionize dynamical systems analysis techniques with multi-scale *spline* decompositions of parameter space. The multiple scales will allow a model to refine regions of interest for further analyses (Figure 3).

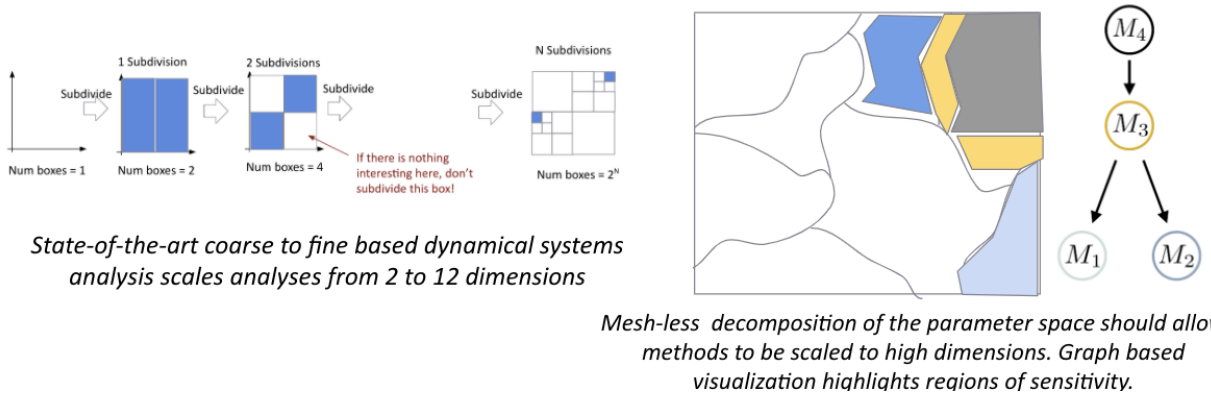


Figure 3: From grids to splines to decompose parameter space of a model.

2. **Surrogate maps and Guided-adaptive Sampling of Parameter Space:** Synergistic with the decomposition thrust, surrogate maps to labels sets and adaptively selecting points within regions to refine sets needs to be objectively characterized. As an example, the dynamics of the learning space can be modeled with a Gaussian Process or *k-nearest neighbors*. Objective metrics are needed to determine the quality of these maps to model the dynamics. Furthermore, not all parameters of a neural network are equally important. Techniques to identify important parameters coupled with coarse dynamic representations can inform the sampling of the space to adaptively refine the construction of the model's dynamics (Figure 4).

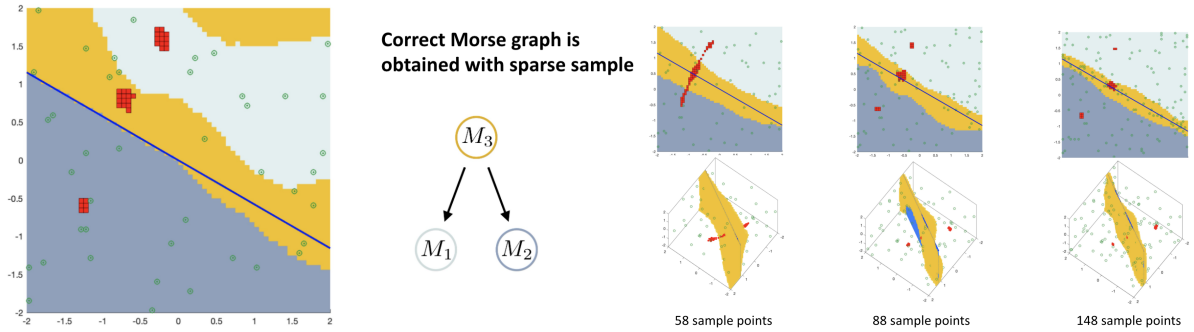


Figure 4: Guided and adaptive sampling of points to refine extracted global dynamics (sepatrix between two distinct stable regions).

3. **Computation of Algebraic topology to Characterize Dynamics:** The combinatorial computations described provide an abstract representation of the global dynamics of the model. As such, the combinatorial information in and of itself provides limited guarantees about the detailed structure of the learning process. To access this additional information requires the use of algebraic topology (homology theory) and in particular the Conley index. This is an algebraic topological extension of Morse theory that can be computed based on the combinatorial information (and its derivation) for the learning process and provides specific information about its dynamics. In particular it can be used to rigorously identify important structures for the non-recurrent dynamics (connecting orbits) and the recurrent dynamics (fixed points, oscillations, and even chaotic dynamics).

Appendix: Data and Code Reproduction

All data used and generated, code, and documentation for DOODL3 can be found in the open source repository: https://github.com/netrias/DOODL3