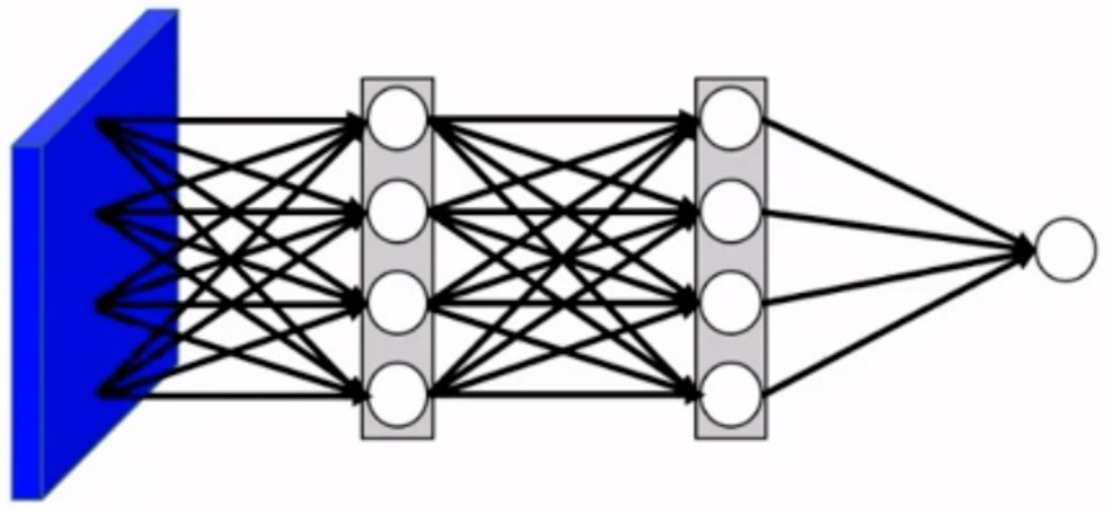
Reconstructing Training Data from Trained Neural Networks

Niv Haim, Gal Vardi, Gilad Yehudai, Ohad Shamir, Michal Irani

NeurIPS 2022

Problem Statement

- Given a trained binary image classifier, can we extract the data on which it was trained?

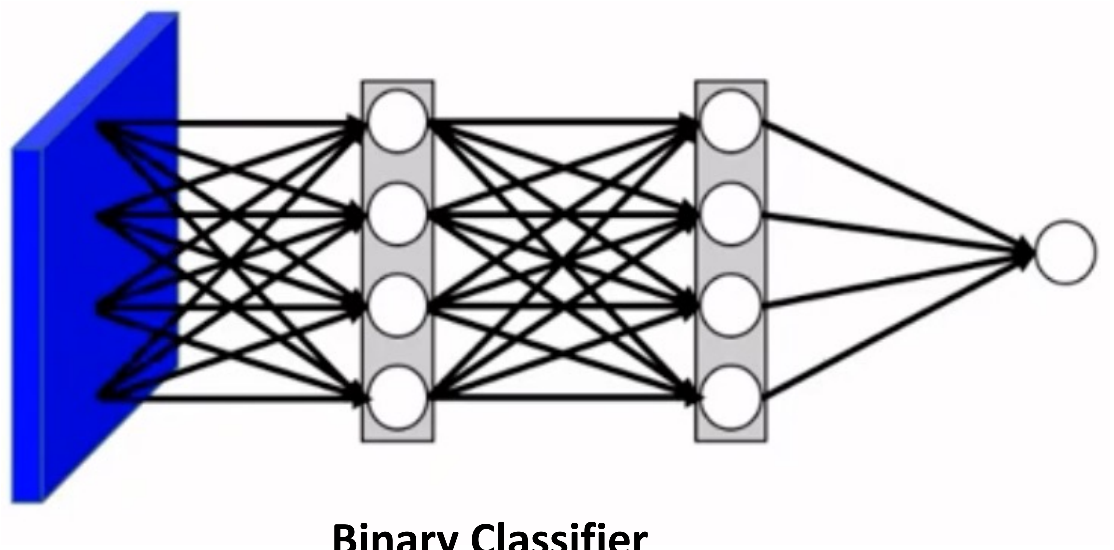


Binary Classifier

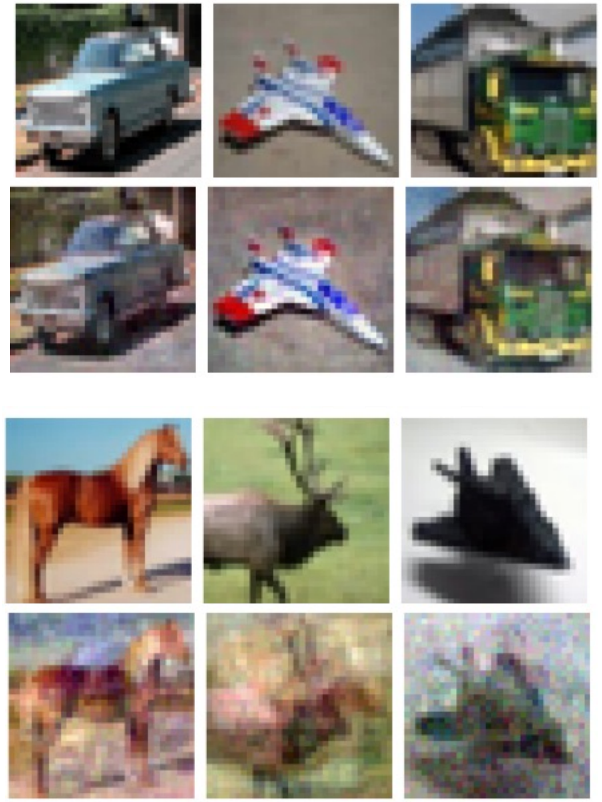
1. https://www.youtube.com/watch?v=sEsHNEEPegM&t=1s
2. Haim, Niv, et al. "Reconstructing training data from trained neural networks." *Advances in Neural Information Processing Systems* 35 (2022): 22911-22924.

Problem Statement

- Given a trained binary image classifier, can we extract the data on which it was trained?

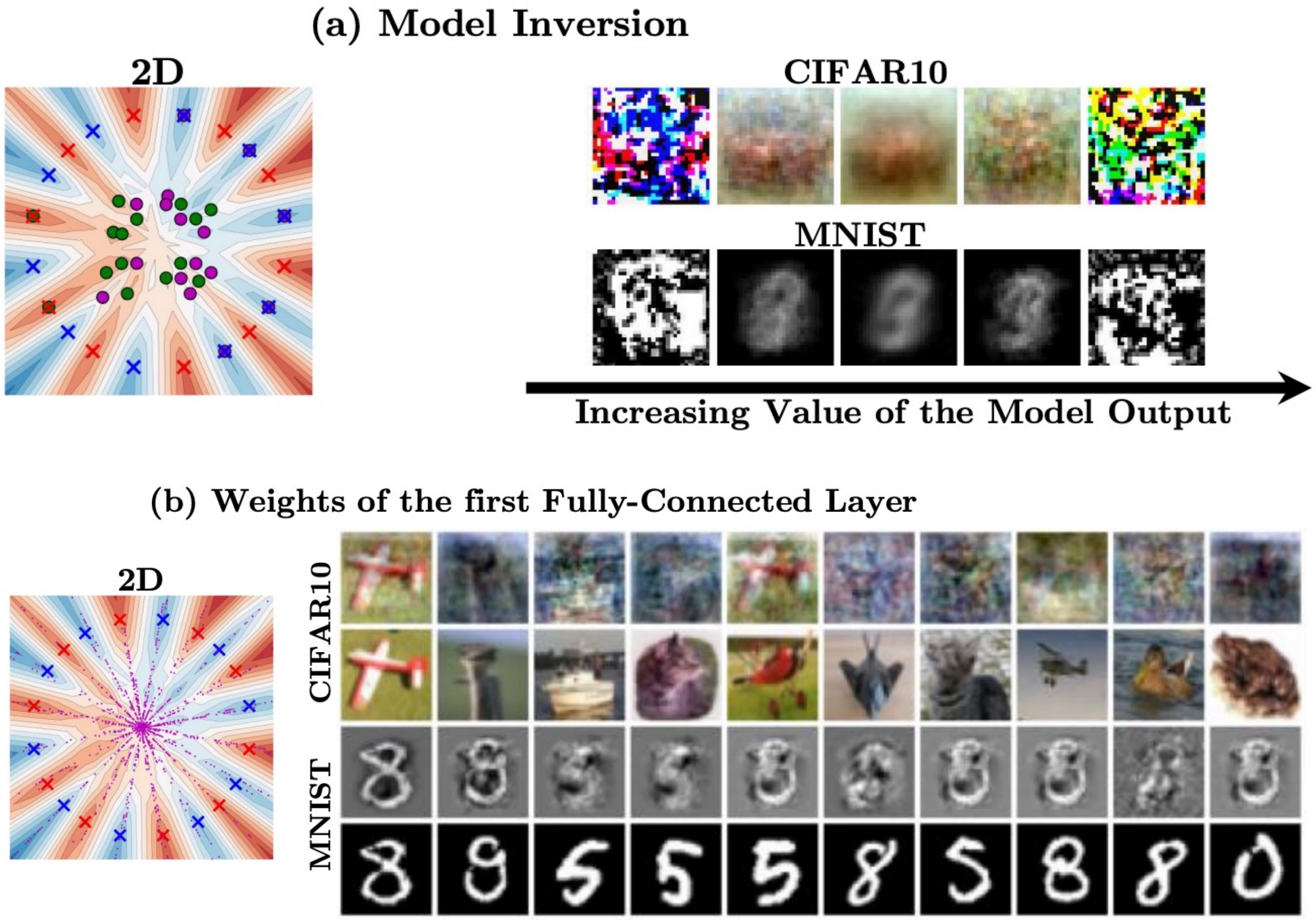


Binary Classifier



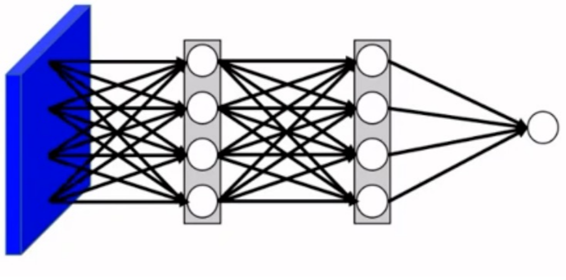
1. https://www.youtube.com/watch?v=sEsHNEEPegM&t=1s
2. Haim, Niv, et al. "Reconstructing training data from trained neural networks." *Advances in Neural Information Processing Systems* 35 (2022): 22911-22924.

Prior Works of Data Reconstruction



Reconstructing Training Data

- The main idea is to:
  1. Define a loss over the model
  2. Minimize the loss w.r.t. the input.

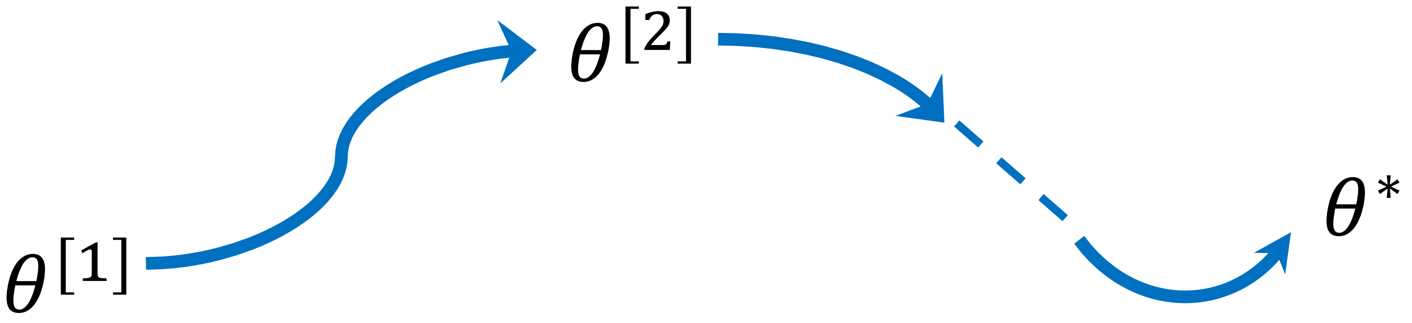
$\min_{x} \mathcal{L} \Big($  $\Big)$

1. https://www.youtube.com/watch?v=sEsHNEEPegM&t=1s
2. Haim, Niv, et al. "Reconstructing training data from trained neural networks." *Advances in Neural Information Processing Systems* 35 (2022): 22911-22924.

Implicit Bias of Gradient Descent

- Implicit bias:
  - If we train a neural network with binary cross entropy loss, its parameters will converge to a stationary point of certain margin-maximization problem.

- Or simply, the learnt parameters will satisfy a set of equations with respect to the trained data.



$$F(\theta^{*}, x) = 0$$

1. https://www.youtube.com/watch?v=sEsHNEEPegM&t=1s
2. Haim, Niv, et al. "Reconstructing training data from trained neural networks." *Advances in Neural Information Processing Systems* 35 (2022): 22911-22924.

Implicit Bias of Gradient Descent

Theorem 3.1 (Paraphrased from Lyu and Li [2019], Ji and Telgarsky [2020]) *Let $\Phi(\boldsymbol{\theta}; \cdot)$ be a homogeneous ReLU neural network. Consider minimizing the logistic loss over a binary classification dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ using gradient flow. Assume that there exists time t_0 such that $\mathcal{L}(\boldsymbol{\theta}(t_0)) < 1^{\ddagger}$. Then, gradient flow converges in direction to a first order stationary point (KKT point) of the following maximum-margin problem:*

$$\min_{\boldsymbol{\theta}'} \frac{1}{2} \left\|\boldsymbol{\theta}'\right\|^2 \quad s.t. \quad \forall i \in [n] \;\; y_i \Phi(\boldsymbol{\theta}'; \mathbf{x}_i) \geq 1 \,. \tag{1}$$

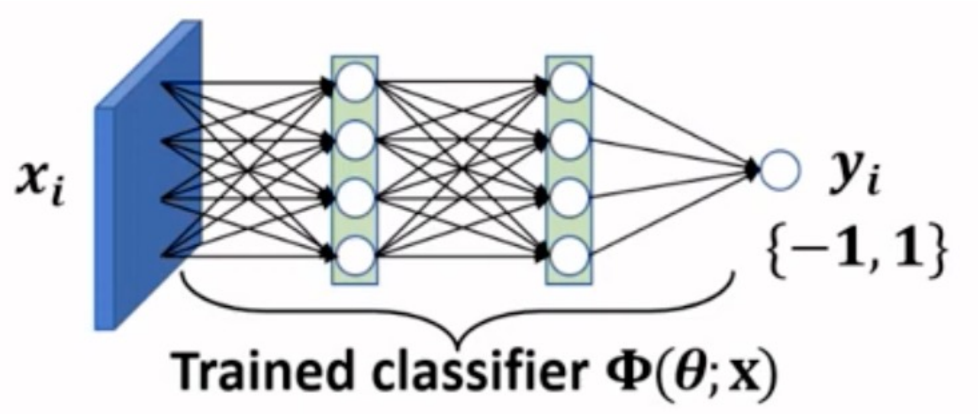
Moreover, $\mathcal{L}(\boldsymbol{\theta}(t)) \to 0$ as $t \to \infty$.

There are many possible directions of $\frac{\theta}{\|\theta\|}$ that classify the dataset correctly,

gradient flow converges only to directions that are KKT points of Problem(1)

Implicit Bias of Gradient Descent

- The model will converge to a solution θ^* which is specified by a set of equations.

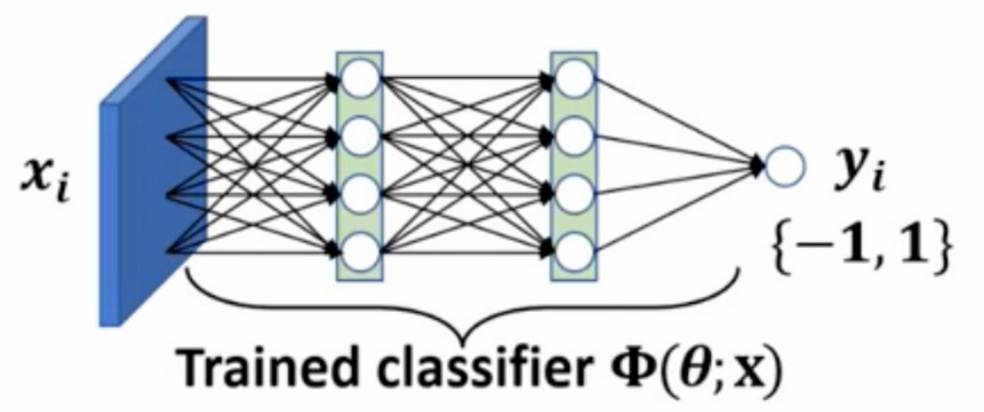


- During training, the parameters θ are learnt while the input x is fixed.
- During reconstruction, the parameters are fixed while inputs x are learnt

$$\tilde{\boldsymbol{\theta}} = \sum_{i=1}^{n} \lambda_i y_i \nabla_{\boldsymbol{\theta}} \Phi(\tilde{\boldsymbol{\theta}}; \mathbf{x}_i)$$
$$\forall i \in [n], \quad y_i \Phi(\tilde{\boldsymbol{\theta}}; \mathbf{x}_i) \geq 1$$
$$\lambda_1, \ldots, \lambda_n \geq 0$$
$$\forall i \in [n], \quad \lambda_i = 0 \text{ if } y_i \Phi(\tilde{\boldsymbol{\theta}}; \mathbf{x}_i) \neq 1$$

1. https://www.youtube.com/watch?v=sEsHNEEPegM&t=1s
2. Haim, Niv, et al. "Reconstructing training data from trained neural networks." *Advances in Neural Information Processing Systems* 35 (2022): 22911-22924.

Reconstructing Training Data - Algorithm



- Input: Trained classifier $\Phi(\theta; x)$
- Initialize $\{x_i, \lambda_i\}$ at random and $y_i \in \{-1,1\}$
- Optimize $\{x_i, \lambda_i\}$ to minimize loss

$$L_\lambda(\lambda_1, \ldots, \lambda_m) = \sum_{i=1}^{m} \max\{-\lambda_i, 0\}$$

L_{prior} : pixel $\in [-1, 1]$

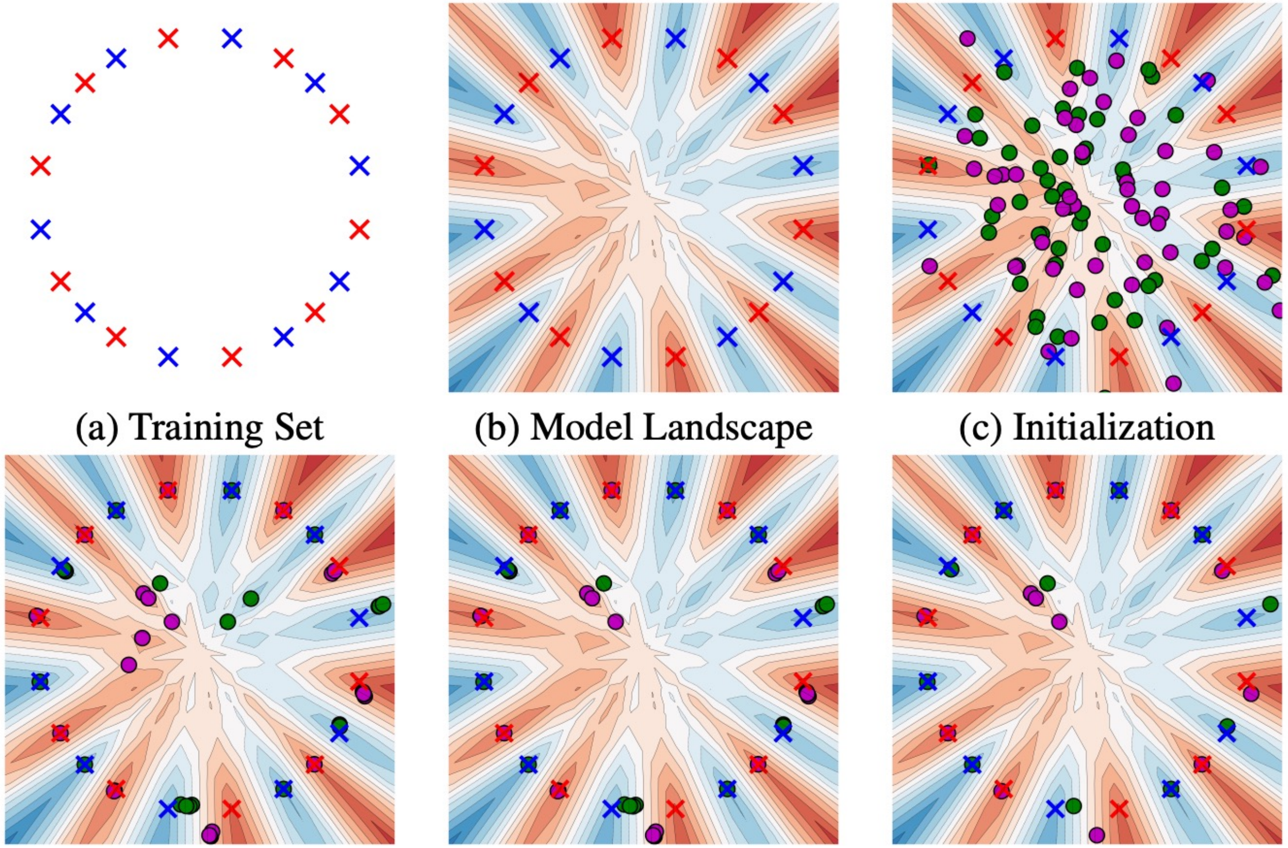
$$L_{\text{stationary}}(\mathbf{x}_1, \ldots, \mathbf{x}_m, \lambda_1, \ldots, \lambda_m) = \left\| \boldsymbol{\theta} - \sum_{i=1}^{m} \lambda_i y_i \nabla_{\boldsymbol{\theta}} \Phi(\boldsymbol{\theta}; \mathbf{x}_i) \right\|_2^2$$

$$L_{\text{reconstruct}}\left(\{\mathbf{x}_i\}_{i=1}^{m}, \{\lambda_i\}_{i=1}^{m}\right) = \alpha_1 L_{\text{stationary}} + \alpha_2 L_\lambda + \alpha_3 L_{\text{prior}}$$

1. https://www.youtube.com/watch?v=sEsHNEEPegM&t=1s
2. Haim, Niv, et al. "Reconstructing training data from trained neural networks." *Advances in Neural Information Processing Systems* 35 (2022): 22911-22924.

Experiments

- Toy 2D dataset
- 3 layers (1000 neurons each)
- 20 training samples
- 100 sampled points
- L_{prior} = 0



(a) Training Set (b) Model Landscape (c) Initialization

(d) Reconstruction (e) Removing Small λ's (f) Removing Duplicates

1. https://www.youtube.com/watch?v=sEsHNEEPegM&t=1s
2. Haim, Niv, et al. "Reconstructing training data from trained neural networks." *Advances in Neural Information Processing Systems* 35 (2022): 22911-22924.

Experiments

- Setting: Binary Classification
- Datasets: MNIST(Odd vs Even), CIFAR10 (Vehicle vs Animal)
- Model: MLP with 3 layers $(\mathrm{d} - 1000 - 1000 - 1)$
- Trained for 10^6 epochs with a learning rate of 0.01 and achieved zero training error.

Figure 3: Reconstructing training samples from two binary classifiers – one trained on 500 images with labels animals/vehicles (CIFAR), and the other trained on 500 odd/even digit images (MNIST). Train errors are zero, test accuracies are 88.0%/77.6% for MNIST/CIFAR

1. https://www.youtube.com/watch?v=sEsHNEEPegM&t=1s
2. Haim, Niv, et al. "Reconstructing training data from trained neural networks." *Advances in Neural Information Processing Systems* 35 (2022): 22911-22924.

Experiments

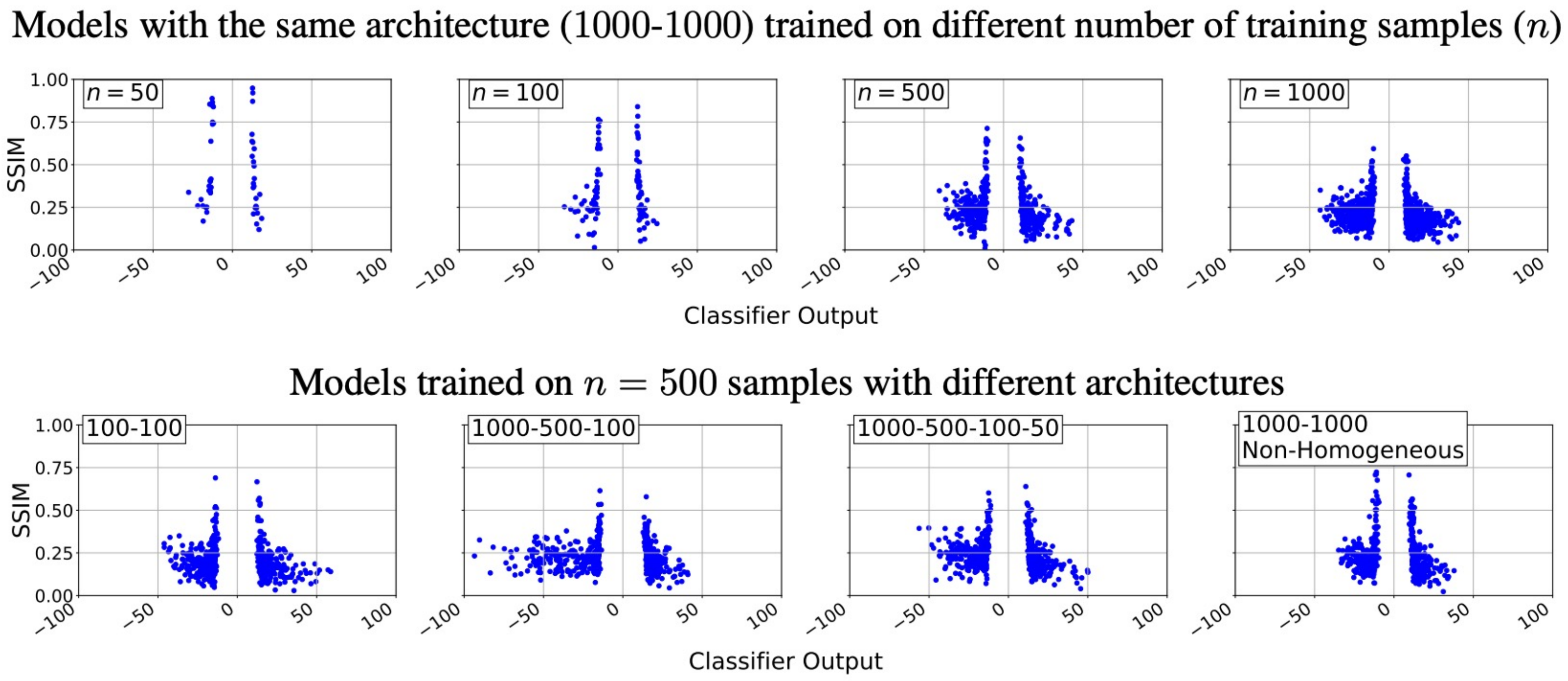


Figure 4: Each point represents a training sample. The y-axis is the highest SSIM score achieved by a reconstruction of this sample, the x-axis is the output of the model. **Top:** The effect of training the same model on different number of training samples (n). **Bottom:** The effect of training models with different architectures (on $n = 500$ training samples). The right-most plot shows a 3-layer non-homogeneous MLP (with bias terms in all hidden layers). See discussion in Section 5.3.

1. https://www.youtube.com/watch?v=sEsHNEEPegM&t=1s
2. Haim, Niv, et al. "Reconstructing training data from trained neural networks." *Advances in Neural Information Processing Systems* 35 (2022): 22911-22924.

Summary

- Strength:
  - Exact dataset reconstruction technique
  - Based on the theory of implicit bias of gradient descent for homogenous models.

- Weaknesses:
  - Multiple assumptions:
    - Model trained to achieve 100% accuracy on training set.
    - Homogenous model
  - Experiments on MNIST and CIFAR10 but difficult to extend to large datasets and large models.
  - Mainly extracts training points close to the margin (support).