NEURAL INFORMATION PROCESSING SYSTEMS

מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

Deconstructing Data Reconstruction:
Multiclass, Weight Decay and General Losses


Gon Buzaglo*


Niv Haim*


Gilad Yehudai


Gal Vardi

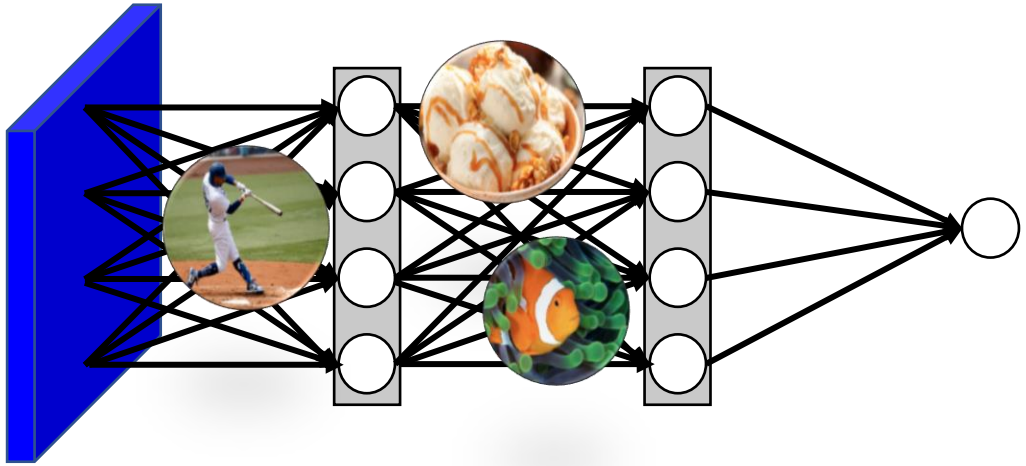

Yakir Oz


Yaniv Nikankin


Michal Irani

*Equal contribution

Understanding Memorization Through Reconstruction



- We reconstruct training data from trained classifiers.
- Motivation: Privacy

Main Technical Tool: Implicit Bias

- Neural Network $\Phi_\theta : \mathbb{R}^d \to \mathbb{R}$
- Weights $\theta \in \mathbb{R}^m$
- Samples $x_1, \ldots, x_n \in \mathbb{R}^d$

$$\forall j \in [m], \quad \theta_j - \sum_{i=1}^{n} \lambda_i y_i \nabla_{\theta_j} \Phi(\theta; x_i) = 0$$

$$\forall i \in [n], \quad \lambda_i \geq 0$$

$$\forall i \in [n], \quad y_i \Phi(\theta; x_i) \geq 1$$

$$\forall i \in [n], \quad \text{if } y_i \Phi(\theta; x_i) \neq 1 \;\; \lambda_i = 0$$

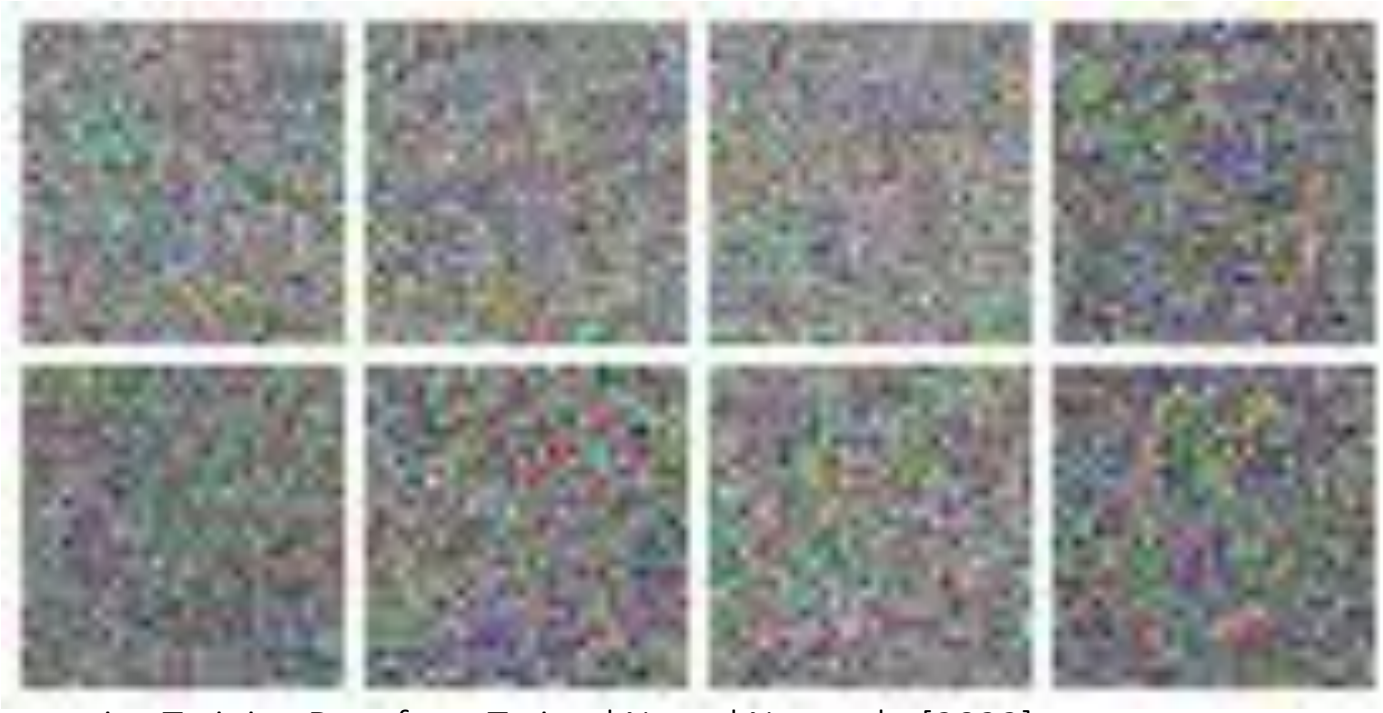
Soudry, Hoffer, Nacson, Gunasekar, and Srebro. The implicit bias of gradient descent on separable data [2018]
Lyu and Li. Gradient descent maximizes the margin of homogeneous neural networks [2019]
Ji and Telgarsky. Directional convergence and alignment in deep learning [2020]

Reconstructing Training Data from a Trained NN

$$L = \left\| \theta - \sum_i \lambda_i \, y_i \nabla_\theta \Phi(\theta; x_i) \right\|_2^2$$

Fixed

optimized



Haim, Vardi, Yehudai, Shamir, Irani. Reconstructing Training Data from Trained Neural Networks [2022]

Our Contributions

- Extension to multiclass classifiers
- Study the effects of weight decay
- Extension to regression losses (beyond Classification)
- Extension to CNN classifiers

Reconstruction from Multiclass Classifiers

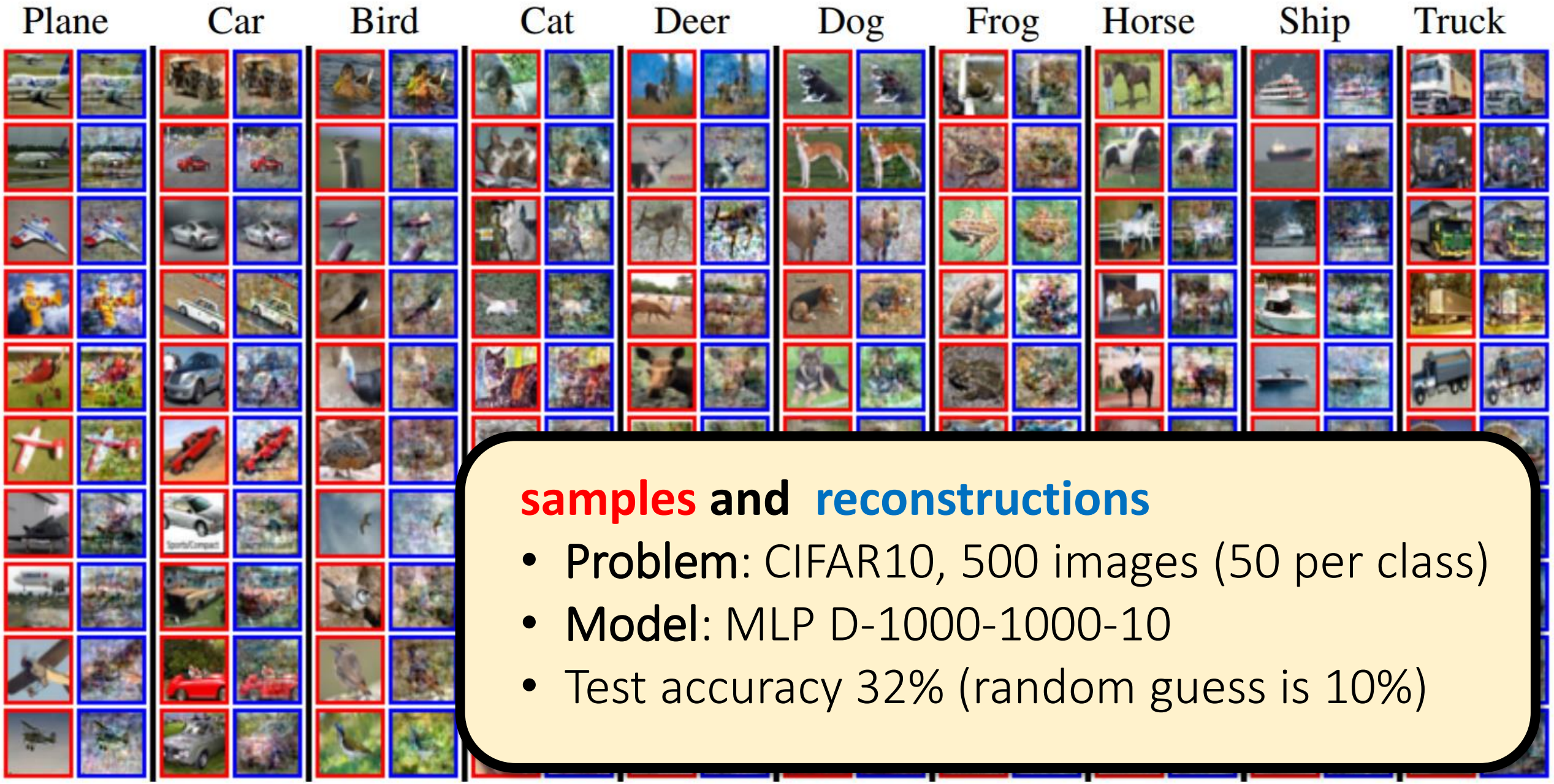
Given a trained Multiclass Classifier $\Phi(\boldsymbol{\theta}; \mathbf{x})$

Initialize:

- $\{\boldsymbol{x_1}, \ldots, \boldsymbol{x_m}\} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$
- $\{\boldsymbol{\lambda_1}, \ldots, \boldsymbol{\lambda_m}\} \sim \mathcal{U}[\mathbf{0}, \mathbf{1}]$

$$L_{\text{multiclass}} = \left\| \boldsymbol{\theta} - \sum_{i=1}^{m} \lambda_i \nabla_{\boldsymbol{\theta}} [\underbrace{y_i \Phi(\mathbf{x}_i; \theta)}_{\text{Distance from Boundary}}] \right\|_2^2$$

Reconstruction from Multiclass Classifiers



samples and reconstructions

- **Problem**: CIFAR10, 500 images (50 per class)
- **Model**: MLP D-1000-1000-10
- Test accuracy 32% (random guess is 10%)

Weight Decay allows Reconstruction from General Losses

$$\mathcal{L} = \sum_{i=1}^{n} \ell(\Phi(\mathbf{x}_i; \theta), y_i) + \lambda_{\text{WD}} \frac{1}{2} \|\theta\|^2$$

$$\nabla \mathcal{L} = 0$$

$$\theta - \sum_{i=1}^{n} \boxed{\lambda_i \frac{\partial \ell(\Phi(\mathbf{x}_i; \theta), y_i)}{\partial \Phi(\mathbf{x}_i; \theta)}} \nabla_\theta \Phi(\mathbf{x}_i; \theta)$$

$$\frac{1}{\lambda_{WD}}$$

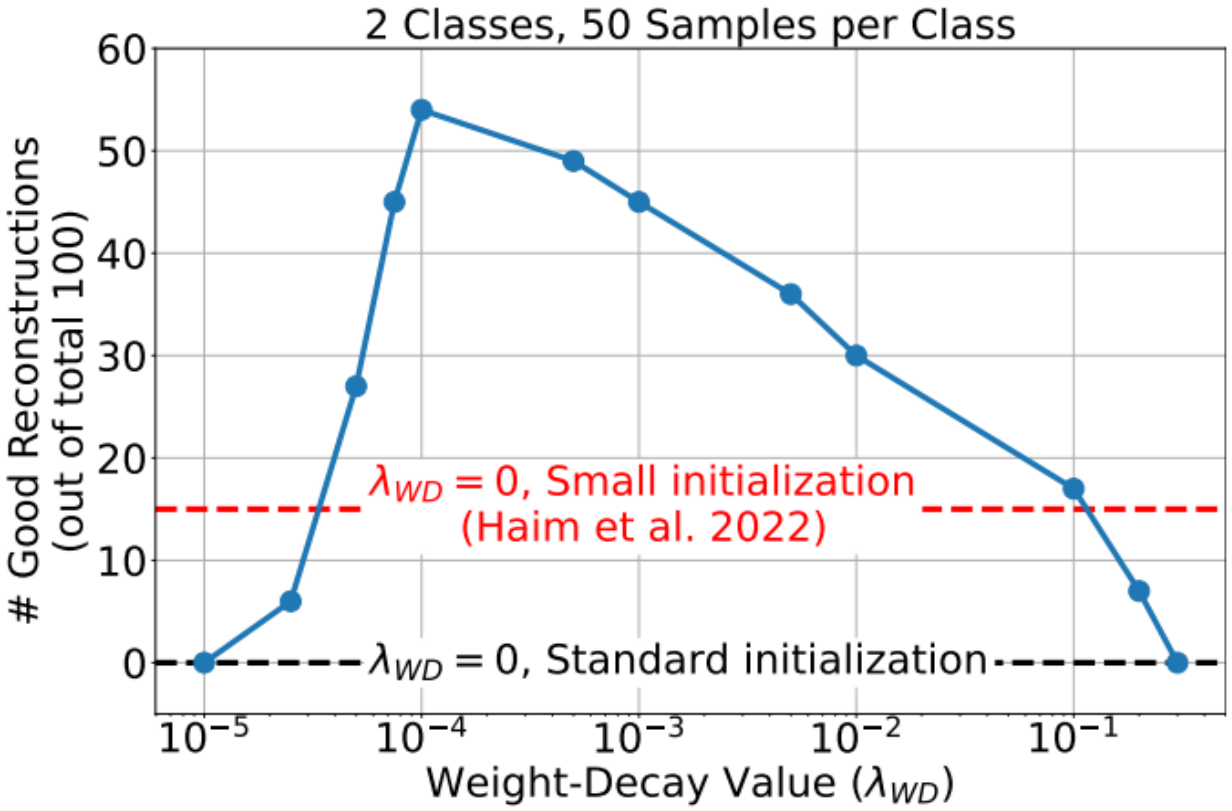
Role of Weight Decay and General Losses

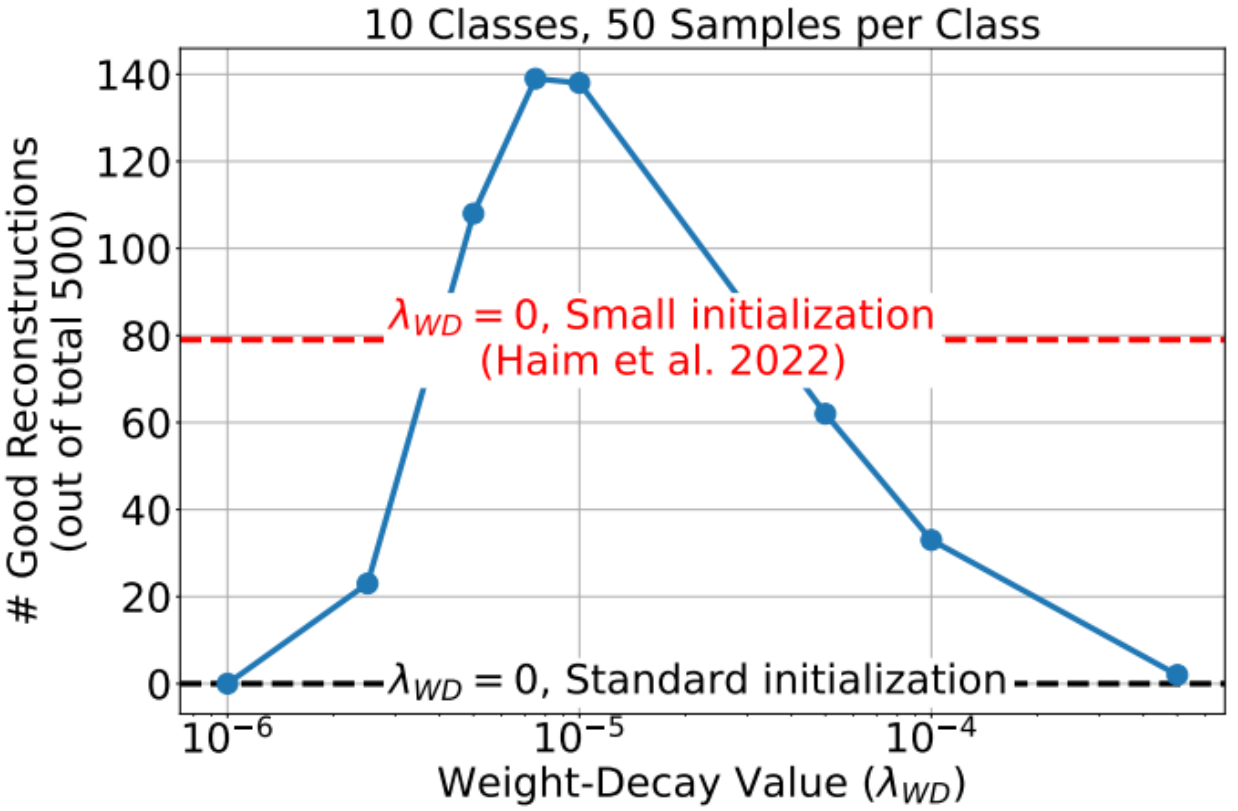
$$\ell(\Phi(\mathbf{x}_i;\theta), y_i) = (\Phi(\mathbf{x}_i;\theta) - y_i)^2$$

samples and reconstructions

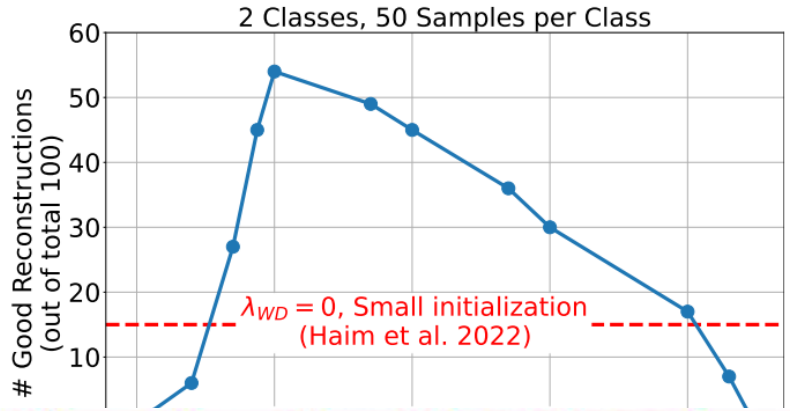
- **Problem**: CIFAR10 with $\{-1,1\}$ labels, 300 images
- **Model**: MLP D-1000-1000-1
- **Training loss**: MSE

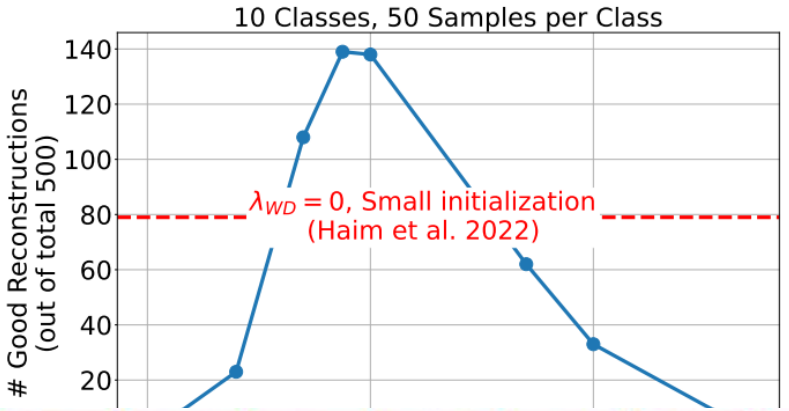
Weight Decay Increases Reconstructability





Weight Decay Increases Reconstructability

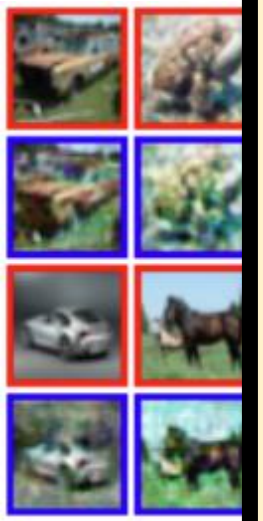
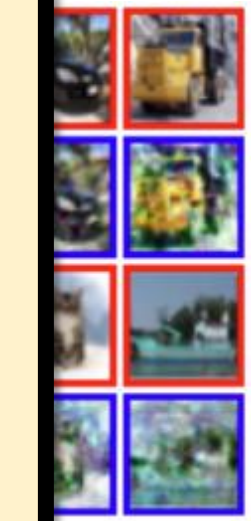




samples and reconstructions

Model: CNN!

Weight Decay Increases Reconstructability

Thank you!

arXiv: 2307.01827