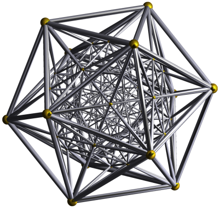
ACDL Summer Course 2023

Lecture 4: Robust Learning of Overparameterized Networks via Low-Dimensional Models

Qing Qu

EECS, University of Michigan

June 10th, 2023



The Trend of Large Models...

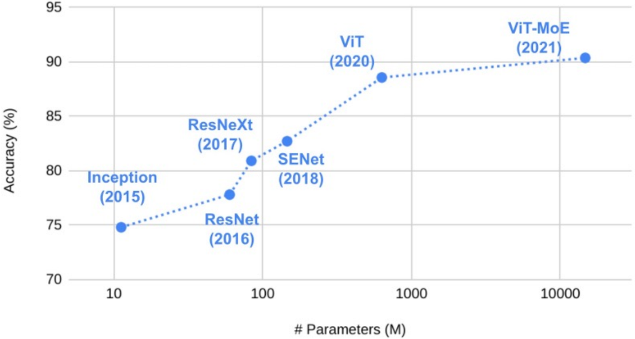


Figure: Accuracy vs. model size for image classification on ImageNet dataset

~23 million (# Parameters in ResNet-50) ≫ ~1 million (# Samples in ImageNet)

In principle, deep network can fit *any* training labels!
(*i.e.*, not only clean, but also corrupted labels)

The Curse of Overparameterization: Robustness

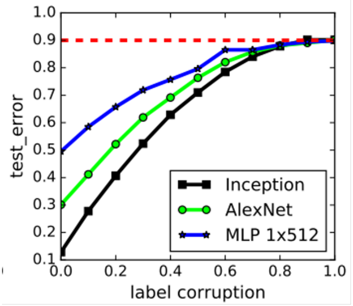


Figure: **Label memorization.**

The Curse of Overparameterization: Robustness

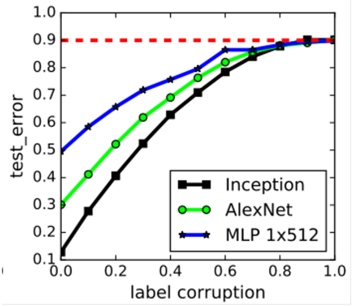


Figure: **Label memorization.**

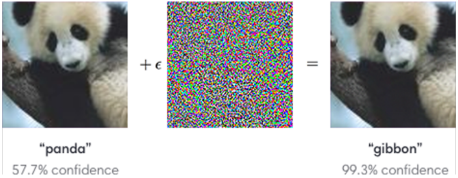


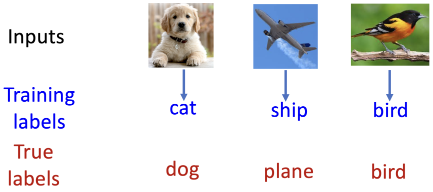
Figure: **Adversarial attack.**

Outline

Neural Collapse → Overfitting to Corruptions!

Label noise is common and often unavoidable

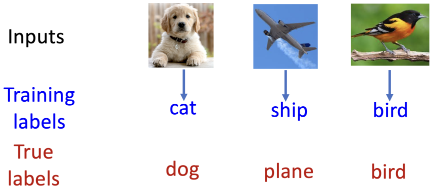
- Some proportion of the labels are incorrect (5-80%?)
- We don't know which labels are correct/incorrect



Neural Collapse → Overfitting to Corruptions!

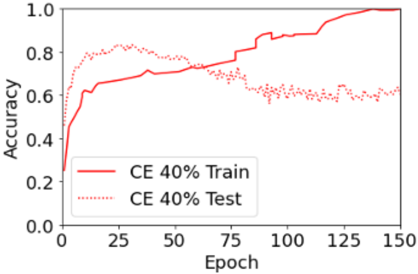
Label noise is common and often unavoidable

- Some proportion of the labels are incorrect (5-80%?)
- We don't know which labels are correct/incorrect



Neural Collapse always happens

- Perfectly fits noisy labels (ovefitting)
- Cannot predict well on new images



Neural Collapse → Overfitting to Corruptions!

Label noise is common and often unavoidable

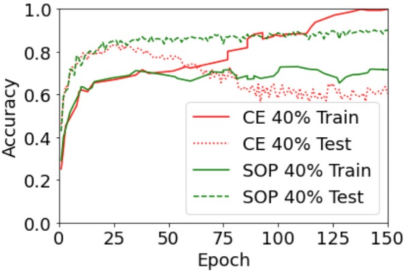
- Some proportion of the labels are incorrect (5-80%?)
- We don't know which labels are correct/incorrect

Inputs

Training labels	cat	ship	bird
True labels	dog	plane	bird

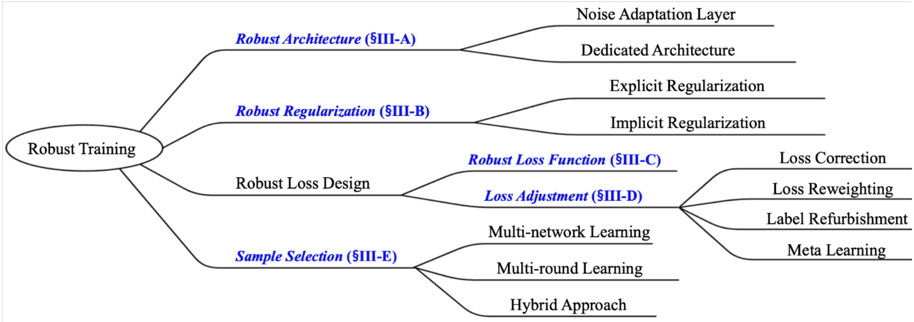
Neural Collapse always happens

- Perfectly fits noisy labels (ovefitting)
- Cannot predict well on new images



Prior Work on Robust Deep Learning for Noisy Labels

Various (heuristic or principled) methods have been proposed[1]



[1] Song et al., Learning from noisy labels with deep neural networks: A survey, IEEE TNNLS, 2022.

A Sparse Over-Parameterization (SOP) Method

Observation: Only a small fraction of the labels are corrupted, so that the label noise is **sparse**.

$$\begin{array}{r}\text{“cat”}\\ \text{“dog”}\\ \\ \\ \end{array}\underbrace{\overset{\boldsymbol{y}}{\begin{bmatrix}0\\1\\\vdots\\0\end{bmatrix}}}_{\text{Corrupted label}} = \underbrace{\overset{f(\boldsymbol{x};\boldsymbol{\theta})}{\begin{bmatrix}1\\0\\\vdots\\0\end{bmatrix}}}_{\text{True label}} + \underbrace{\overset{\boldsymbol{s}}{\begin{bmatrix}-1\\1\\\vdots\\0\end{bmatrix}}}_{\text{Sparse noise}}$$

[2] Wright et al., Robust face recognition via sparse representation, TPAMI, 2008. Candes et al., Robust principal component analysis? JACM, 2011.

A Sparse Over-Parameterization (SOP) Method

Observation: Only a small fraction of the labels are corrupted, so that the label noise is **sparse**.

$$\begin{array}{r}\text{“cat”}\\ \text{“dog”}\\ \\ \\ \end{array}\overset{\boldsymbol{y}}{\begin{bmatrix}0\\ {\color{blue}1}\\ \vdots\\ 0\end{bmatrix}} = \overset{f(\boldsymbol{x};\boldsymbol{\theta})}{\begin{bmatrix}{\color{red}1}\\ 0\\ \vdots\\ 0\end{bmatrix}} + \overset{\boldsymbol{s}}{\begin{bmatrix}{\color{red}-1}\\ {\color{blue}1}\\ \vdots\\ 0\end{bmatrix}}$$

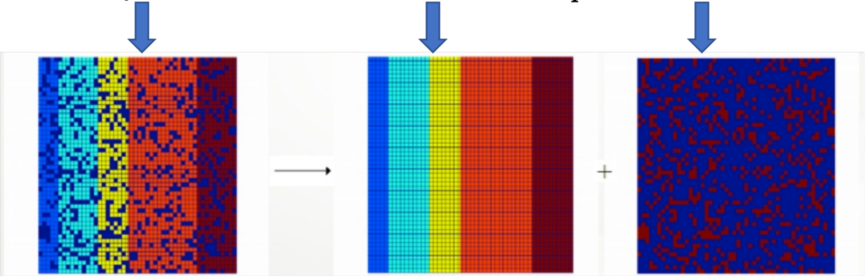
Corrupted label | True label | Sparse noise

Idea from the past: we developed principled methods for dealing with sparse corruption in Compressive Sensing: Robust PCA[2]

[2] Wright et al., Robust face recognition via sparse representation, TPAMI, 2008. Candes et al., Robust principal component analysis? JACM, 2011.

A Sparse Over-Parameterization (SOP) Method

$$\underbrace{\boldsymbol{y}_i}_{\text{noisy label}} = f(\underbrace{\boldsymbol{x}_i}_{\text{input}}; \underbrace{\boldsymbol{\Theta}^\star}_{\text{params.}}) + \underbrace{\boldsymbol{s}_i^\star}_{\textbf{sparse}\text{ label noise}}$$



Exact Separation of Sparse Corruption with Incoherence between Data and Noise

A Sparse Over-Parameterization (SOP) Method

Our approach:[3] minimize the distance between $\boldsymbol{y}$ and $f(\boldsymbol{\theta};\boldsymbol{x}) + \boldsymbol{s}$

$$\min_{\boldsymbol{\theta},\boldsymbol{u}_i,\boldsymbol{v}_i} \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{\mathrm{CE}}\Big(f(\boldsymbol{x}_i;\boldsymbol{\theta}) + \underbrace{\boldsymbol{u}_i \odot \boldsymbol{u}_i - \boldsymbol{v}_i \odot \boldsymbol{v}_i}_{\textbf{over-parameterize } \boldsymbol{s}_i \text{ to promote sparsity}}, \boldsymbol{y}_i\Big)$$

[3] Liu, Zhu, Qu, You, Robust Training under Label Noise by Over-parameterization, ICML'22.

A Sparse Over-Parameterization (SOP) Method

Our approach:[3] minimize the distance between $\boldsymbol{y}$ and $f(\boldsymbol{\theta};\boldsymbol{x})+\boldsymbol{s}$

$$\min_{\boldsymbol{\theta},\boldsymbol{u}_i,\boldsymbol{v}_i} \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{\mathrm{CE}}\Big(f(\boldsymbol{x}_i;\boldsymbol{\theta}) + \underbrace{\boldsymbol{u}_i \odot \boldsymbol{u}_i - \boldsymbol{v}_i \odot \boldsymbol{v}_i}_{\textbf{over-parameterize } \boldsymbol{s}_i \text{ to promote sparsity}}, \boldsymbol{y}_i\Big)$$

- Here the over-parameterization $\boldsymbol{u}_i \odot \boldsymbol{u}_i - \boldsymbol{v}_i \odot \boldsymbol{v}_i$ introduces implicit algorithmic regularization [Vaskevicius et al.'19, Zhao et al.'19]

$$\text{variational form } \|\boldsymbol{s}\|_1 = \min_{\boldsymbol{s}=\boldsymbol{u}\odot\boldsymbol{u}-\boldsymbol{v}\odot\boldsymbol{v}} \frac{1}{2}(\|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|^2)$$

[3] Liu, Zhu, Qu, You, Robust Training under Label Noise by Over-parameterization, ICML'22.

A Sparse Over-Parameterization (SOP) Method

Our approach:[3] minimize the distance between $\boldsymbol{y}$ and $f(\boldsymbol{\theta}; \boldsymbol{x}) + \boldsymbol{s}$

$$\min_{\boldsymbol{\theta}, \boldsymbol{u}_i, \boldsymbol{v}_i} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{\mathrm{CE}}\Big(f(\boldsymbol{x}_i; \boldsymbol{\theta}) + \underbrace{\boldsymbol{u}_i \odot \boldsymbol{u}_i - \boldsymbol{v}_i \odot \boldsymbol{v}_i}_{\text{over-parameterize } \boldsymbol{s}_i \text{ to promote sparsity}}, \boldsymbol{y}_i\Big)$$

- Here the over-parameterization $\boldsymbol{u}_i \odot \boldsymbol{u}_i - \boldsymbol{v}_i \odot \boldsymbol{v}_i$ introduces implicit algorithmic regularization [Vaskevicius et al.'19, Zhao et al.'19]

$$\text{variational form } \|\boldsymbol{s}\|_1 = \min_{\boldsymbol{s} = \boldsymbol{u} \odot \boldsymbol{u} - \boldsymbol{v} \odot \boldsymbol{v}} \frac{1}{2}\left(\|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|^2\right)$$

- Why not use explicit regularization?

$$\min_{\boldsymbol{\theta}, \{\boldsymbol{s}_i\}} \frac{1}{N} \sum_{i=1}^{N} \underbrace{\mathcal{L}_{\mathrm{CE}}\big(f(\boldsymbol{x}_i; \boldsymbol{\Theta}) + \boldsymbol{s}_i, \boldsymbol{y}_i\big)}_{\to 0} + \underbrace{\lambda \|\boldsymbol{s}_i\|_1}_{\to 0}$$

[3] Liu, Zhu, Qu, You, Robust Training under Label Noise by Over-parameterization, ICML'22.

A Sparse Over-Parameterization (SOP) Method

Our approach:[4] minimize the distance between $\boldsymbol{y}$ and $f(\boldsymbol{\theta}; \boldsymbol{x}) + \boldsymbol{s}$

$$\min_{\boldsymbol{\theta}, \boldsymbol{u}_i, \boldsymbol{v}_i} \; \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{\mathrm{CE}}\Big(f(\boldsymbol{x}_i; \boldsymbol{\theta}) + \underbrace{\boldsymbol{u}_i \odot \boldsymbol{u}_i - \boldsymbol{v}_i \odot \boldsymbol{v}_i}_{\textbf{over-parameterize } \boldsymbol{s}_i \text{ to promote sparsity}}, \boldsymbol{y}_i\Big)$$

Training: gradient descent with a discrepant learning rate:

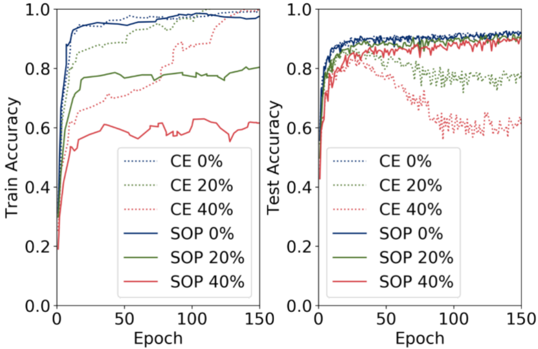
$$\begin{aligned}
\boldsymbol{\theta} \;&\leftarrow\; \boldsymbol{\theta} - \tau \frac{\partial}{\partial \boldsymbol{\theta}} L(\{\boldsymbol{u}_i, \boldsymbol{v}_i\}; \boldsymbol{\theta}) \\
\boldsymbol{u}_i \;&\leftarrow\; \boldsymbol{u}_i - \alpha\tau \frac{\partial}{\partial \boldsymbol{\theta}} L(\{\boldsymbol{u}_i, \boldsymbol{v}_i\}; \boldsymbol{u}_i) \\
\boldsymbol{v}_i \;&\leftarrow\; \boldsymbol{v}_i - \alpha\tau \frac{\partial}{\partial \boldsymbol{\theta}} L(\{\boldsymbol{u}_i, \boldsymbol{v}_i\}; \boldsymbol{v}_i)
\end{aligned}$$

Ideally, the implicit regularization drives the GD dynamics to the desired solution.

[4] Liu, Zhu, Qu, You, Robust Training under Label Noise by Over-parameterization, ICML'22.

A Sparse Over-Parameterization (SOP) Method

$\{0\%, 20\%, 40\%\}$ percent of labels for CIFAR-10 training data are randomly flipped uniformly to another class. Use ResNet34.



Observation: Compared to vanilla training, SOP does not overfit to wrong labels and obtain better generalization performance.

Theoretical Justification on SOP

A simple model: assume $f(\boldsymbol{x};\boldsymbol{\theta})$ is a scalar function and can be approximated by first-order Taylor expansion

$$f(\boldsymbol{x};\boldsymbol{\theta}) \approx f(\boldsymbol{x};\boldsymbol{\theta}_0) + \langle \nabla f(\boldsymbol{x};\boldsymbol{\theta}_0), \boldsymbol{\theta} - \boldsymbol{\theta}_0 \rangle$$

Theoretical Justification on SOP

A simple model: assume $f(\boldsymbol{x};\boldsymbol{\theta})$ is a scalar function and can be approximated by first-order Taylor expansion

$$f(\boldsymbol{x};\boldsymbol{\theta}) \approx f(\boldsymbol{x};\boldsymbol{\theta}_0) + \langle \nabla f(\boldsymbol{x};\boldsymbol{\theta}_0), \boldsymbol{\theta} - \boldsymbol{\theta}_0 \rangle$$

WLOG, assume $f(\boldsymbol{x};\boldsymbol{\theta}_0) + \langle \nabla f(\boldsymbol{x};\boldsymbol{\theta}_0), \boldsymbol{\theta}_0 \rangle = 0$. For N training samples,

$$\begin{bmatrix} f(\boldsymbol{x}_1;\boldsymbol{\theta}) \\ \vdots \\ f(\boldsymbol{x}_N;\boldsymbol{\theta}) \end{bmatrix} \approx \begin{bmatrix} \nabla f(\boldsymbol{x}_1;\boldsymbol{\theta}_0)^\top \\ \vdots \\ \nabla f(\boldsymbol{x}_N;\boldsymbol{\theta}_0)^\top \end{bmatrix} \boldsymbol{\theta} = \boldsymbol{J} \cdot \boldsymbol{\theta}$$

Theoretical Justification on SOP

A simple model: assume $f(\boldsymbol{x};\boldsymbol{\theta})$ is a scalar function and can be approximated by first-order Taylor expansion

$$f(\boldsymbol{x};\boldsymbol{\theta}) \approx f(\boldsymbol{x};\boldsymbol{\theta}_0) + \langle \nabla f(\boldsymbol{x};\boldsymbol{\theta}_0), \boldsymbol{\theta} - \boldsymbol{\theta}_0 \rangle$$

WLOG, assume $f(\boldsymbol{x};\boldsymbol{\theta}_0) + \langle \nabla f(\boldsymbol{x};\boldsymbol{\theta}_0), \boldsymbol{\theta}_0 \rangle = 0$. For N training samples,

$$\begin{bmatrix} f(\boldsymbol{x}_1;\boldsymbol{\theta}) \\ \vdots \\ f(\boldsymbol{x}_N;\boldsymbol{\theta}) \end{bmatrix} \approx \begin{bmatrix} \nabla f(\boldsymbol{x}_1;\boldsymbol{\theta}_0)^\top \\ \vdots \\ \nabla f(\boldsymbol{x}_N;\boldsymbol{\theta}_0)^\top \end{bmatrix} \boldsymbol{\theta} = \boldsymbol{J} \cdot \boldsymbol{\theta}$$

This leads to the following corrupted observation problem

$$\boldsymbol{y} = \boldsymbol{J} \cdot \boldsymbol{\theta}_\star + \boldsymbol{s}_\star$$

where $\boldsymbol{\theta}_\star$ is the underlying groundtruth parameter, and $\boldsymbol{s}_\star$ is sparse.

Theoretical Justification on SOP

We over-parameterize the sparse noise by $\boldsymbol{u} \odot \boldsymbol{u} - \boldsymbol{v} \odot \boldsymbol{v}$ and solve

$$\min_{\boldsymbol{\theta},\boldsymbol{u},\boldsymbol{v}} g(\boldsymbol{\theta},\boldsymbol{u},\boldsymbol{v}) = \frac{1}{2}\|\boldsymbol{J} \cdot \boldsymbol{\theta} + \boldsymbol{u} \odot \boldsymbol{u} - \boldsymbol{v} \odot \boldsymbol{v} - \boldsymbol{y}\|_2^2$$

using gradient descent with *discrepant learning rates*

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \mu \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_t, \boldsymbol{u}_t, \boldsymbol{v}_t), \ \begin{bmatrix} \boldsymbol{u}_{t+1} \\ \boldsymbol{v}_{t+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{u}_t \\ \boldsymbol{v}_t \end{bmatrix} - \alpha\mu \begin{bmatrix} \nabla_{\boldsymbol{u}} g(\boldsymbol{\theta}_t, \boldsymbol{u}_t, \boldsymbol{v}_t) \\ \nabla_{\boldsymbol{v}} g(\boldsymbol{\theta}_t, \boldsymbol{u}_t, \boldsymbol{v}_t) \end{bmatrix}$$

Theoretical Justification on SOP

We over-parameterize the sparse noise by $\boldsymbol{u} \odot \boldsymbol{u} - \boldsymbol{v} \odot \boldsymbol{v}$ and solve

$$\min_{\boldsymbol{\theta},\boldsymbol{u},\boldsymbol{v}} g(\boldsymbol{\theta},\boldsymbol{u},\boldsymbol{v}) = \frac{1}{2}\|\boldsymbol{J}\cdot\boldsymbol{\theta} + \boldsymbol{u}\odot\boldsymbol{u} - \boldsymbol{v}\odot\boldsymbol{v} - \boldsymbol{y}\|_2^2$$

using gradient descent with *discrepant learning rates*

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \mu\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}_t,\boldsymbol{u}_t,\boldsymbol{v}_t),\ \begin{bmatrix}\boldsymbol{u}_{t+1}\\ \boldsymbol{v}_{t+1}\end{bmatrix} = \begin{bmatrix}\boldsymbol{u}_t\\ \boldsymbol{v}_t\end{bmatrix} - \alpha\mu\begin{bmatrix}\nabla_{\boldsymbol{u}} g(\boldsymbol{\theta}_t,\boldsymbol{u}_t,\boldsymbol{v}_t)\\ \nabla_{\boldsymbol{v}} g(\boldsymbol{\theta}_t,\boldsymbol{u}_t,\boldsymbol{v}_t)\end{bmatrix}$$

Theorem (informal) If gradient descent with infinitesimally small initialization and step size μ converges to $(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{u}},\widehat{\boldsymbol{v}})$, then $(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{u}}\odot\widehat{\boldsymbol{u}} - \widehat{\boldsymbol{v}}\odot\widehat{\boldsymbol{v}})$ is an optimal solution to the following convex problem

$$\min_{\boldsymbol{\theta},\boldsymbol{s}} \frac{1}{2}\|\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{s}\|_1,\ \text{s.t. } \boldsymbol{y} = \boldsymbol{J}\cdot\boldsymbol{\theta} + \boldsymbol{s},$$

solving which exactly recovers $(\boldsymbol{\theta}_\star,\boldsymbol{s}_\star)$ when $\boldsymbol{J}$ is *incoherent* [Candes & Tao'05].

Theoretical Justification on SOP

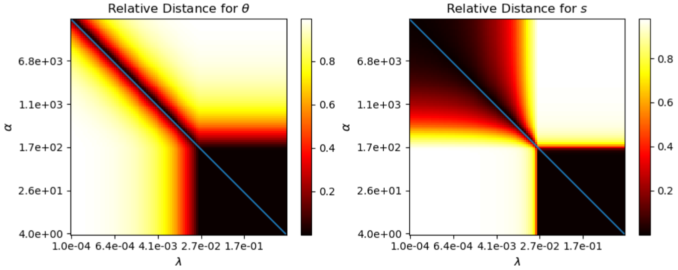


Figure: **The SOP and the convex problem produce the same solutions with $\alpha = -\frac{\log \gamma}{2\lambda}$.**

Superior Performance with Training Efficiency

	CIFAR-10				CIFAR-100			
	Symmetric			Asym	Symmetric			Asym
Methods	20%	50%	80%	40%	20%	50%	80%	40%
CE	87.2	80.7	65.8	82.2	58.1	47.1	23.8	43.3
MixUp	93.5	87.9	72.3	-	69.9	57.3	33.6	-
DivideMix	96.1	94.6	93.2	93.4	77.1	74.6	60.2	72.1
ELR+	95.8	94.8	93.3	93.0	77.7	73.8	60.8	77.5
SOP+	**96.3**	**95.5**	**94.0**	**93.8**	**78.8**	**75.9**	**63.3**	**78.0**

Ours

Superior Performance with Training Efficiency

	CIFAR-10				CIFAR-100			
	Symmetric			Asym	Symmetric			Asym
Methods	20%	50%	80%	40%	20%	50%	80%	40%
CE	87.2	80.7	65.8	82.2	58.1	47.1	23.8	43.3
MixUp	93.5	87.9	72.3	-	69.9	57.3	33.6	-
DivideMix	96.1	94.6	93.2	93.4	77.1	74.6	60.2	72.1
ELR+	95.8	94.8	93.3	93.0	77.7	73.8	60.8	77.5
SOP+	**96.3**	**95.5**	**94.0**	**93.8**	**78.8**	**75.9**	**63.3**	**78.0**

Ours

CE	Co-teaching+	DivideMix	ELR+	SOP (**ours**)	SOP+ (**ours**)
0.9h	4.4h	5.4h	2.3h	**1.0h**	**2.1h**

Table: **Comparison of total training time** in hours on CIFAR-10 with 50% symmetric label noise.

SOP on CIFAR-10 with human annotated noisy labels

CIFAR-10N: provide CIFAR-10 with human annotated noisy labels[5]

Label Set	CIFAR-10N Aggregate	CIFAR-10N Random 1	CIFAR-10N Random 2	CIFAR-10N Random 3	CIFAR-10N Worst
Noise Rate	9.03%	17.23%	18.12%	17.64%	40.21%

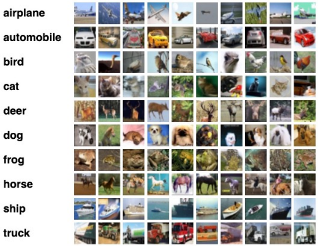
[5] Wei et al., Learning with noisy labels revisited: A study using real-world human annotations, ICLR 2022.

SOP on CIFAR-10 with human annotated noisy labels

CIFAR-10N: provide CIFAR-10 with human annotated noisy labels[5]

Label Set	CIFAR-10N Aggregate	CIFAR-10N Random 1	CIFAR-10N Random 2	CIFAR-10N Random 3	CIFAR-10N Worst
Noise Rate	9.03%	17.23%	18.12%	17.64%	40.21%

- Annotated by 747 independent workers
- Provide 5 noisy label sets for CIFAR-10 train images:
- **Random** $i = 1, 2, 3$: the i-th submitted label for each image;
- **Aggregate**: aggregation of three noisy labels by majority voting
- **Worst**: label set with the highest noise rate



[5] Wei et al., Learning with noisy labels revisited: A study using real-world human annotations, ICLR 2022.

New SOTA on CIFAR-10N

Method	CIFAR-10N Clean	Aggregate	Random 1	Random 2	Random 3	Worst	
CE (Standard)	92.92 ± 0.11	87.77 ± 0.38	85.02 ± 0.65	86.46 ± 1.79	85.16 ± 0.61	77.69 ± 1.55	
Forward T (Patrini et al., 2017)	93.02 ± 0.12	88.24 ± 0.22	86.88 ± 0.50	86.14 ± 0.24	87.04 ± 0.35	79.79 ± 0.46	
Backward T (Patrini et al., 2017)	93.10 ± 0.05	88.13 ± 0.29	87.14 ± 0.34	86.28 ± 0.80	86.86 ± 0.41	77.61 ± 1.05	
GCE (Zhang & Sabuncu, 2018)	92.83 ± 0.16	87.85 ± 0.70	87.61 ± 0.28	87.70 ± 0.56	87.58 ± 0.29	80.66 ± 0.35	
Co-teaching (Han et al., 2018)	93.35 ± 0.14	91.20 ± 0.13	90.33 ± 0.13	90.30 ± 0.17	90.15 ± 0.18	83.83 ± 0.13	
Co-teaching+ (Yu et al., 2019)	92.41 ± 0.20	90.61 ± 0.22	89.70 ± 0.27	89.47 ± 0.18	89.54 ± 0.22	83.26 ± 0.17	
T-Revision (Xia et al., 2019)	93.35 ± 0.23	88.52 ± 0.17	88.33 ± 0.32	87.71 ± 1.02	87.79 ± 0.67	80.48 ± 1.20	
Peer Loss (Liu & Guo, 2020)	93.99 ± 0.13	90.75 ± 0.25	89.06 ± 0.11	88.76 ± 0.19	88.57 ± 0.09	82.00 ± 0.60	
ELR (Liu et al., 2020)	93.45 ± 0.65	92.38 ± 0.64	91.46 ± 0.38	91.61 ± 0.16	91.41 ± 0.44	83.58 ± 1.13	
ELR+ (Liu et al., 2020)	**95.39 ± 0.05**	94.83 ± 0.10	94.43 ± 0.41	94.20 ± 0.24	94.34 ± 0.22	91.09 ± 1.60	Two-network based
Positive-LS (Lukasik et al., 2020)	94.77 ± 0.17	91.57 ± 0.07	89.80 ± 0.28	89.35 ± 0.33	89.82 ± 0.14	82.76 ± 0.53	
F-Div (Wei & Liu, 2020)	94.88 ± 0.12	91.64 ± 0.34	89.70 ± 0.40	89.79 ± 0.12	89.55 ± 0.49	82.53 ± 0.52	
Divide-Mix (Li et al., 2020)	**95.37 ± 0.14**	**95.01 ± 0.71**	**95.16 ± 0.19**	**95.23 ± 0.07**	**95.21 ± 0.14**	**92.56 ± 0.42**	Two-network based
Negative-LS (Wei et al., 2021)	**94.92 ± 0.25**	91.97 ± 0.46	90.29 ± 0.32	90.37 ± 0.12	90.13 ± 0.19	82.99 ± 0.36	
JoCoR (Wei et al., 2020)	93.40 ± 0.24	91.44 ± 0.05	90.30 ± 0.20	90.21 ± 0.19	90.11 ± 0.21	83.37 ± 0.30	
CORES^2 (Cheng et al., 2021)	93.43 ± 0.24	91.23 ± 0.11	89.66 ± 0.32	89.91 ± 0.45	89.79 ± 0.50	83.60 ± 0.53	
CORES* (Cheng et al., 2021)	94.16 ± 0.11	**95.25 ± 0.09**	94.45 ± 0.14	94.88 ± 0.31	94.74 ± 0.03	91.66 ± 0.09	
VolMinNet (Li et al., 2021)	92.14 ± 0.30	89.70 ± 0.21	88.30 ± 0.12	88.27 ± 0.09	88.19 ± 0.41	80.53 ± 0.20	
CAL (Zhu et al., 2021a)	94.50 ± 0.31	91.97 ± 0.32	90.93 ± 0.31	90.75 ± 0.30	90.74 ± 0.24	85.36 ± 0.16	
PES (Semi) (Bai et al., 2021)	94.76 ± 0.2	94.66 ± 0.18	**95.06 ± 0.15**	**95.19 ± 0.23**	**95.22 ± 0.13**	**92.68 ± 0.22**	Semi-supervised
SOP (Liu et al., 2022)	N/A	**95.61 ± 0.13**	**95.28 ± 0.13**	**95.31 ± 0.10**	**95.39 ± 0.11**	**93.24 ± 0.21**	Ours

Sparse modeling gives super performance again label noise[6]

[6] Wei et al., Learning with noisy labels revisited: A study using real-world human annotations, ICLR 2022.
Liu, Zhu, Qu, You, Robust Training under Label Noise by Over-parameterization, ICML'22.

Outline

Deep Image Prior[7]

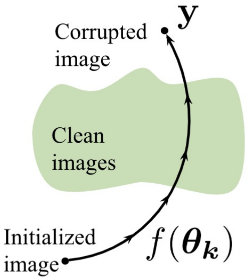
- **Goal:** given a corrupted image $\boldsymbol{y} = \boldsymbol{x}_{\star} + \boldsymbol{s}$, recover the clean image $\boldsymbol{x}_{\star}$ from the noisy observation

[7] Ulyanov D, Vedaldi A, Lempitsky V. Deep image prior[J]. International Journal of Computer Vision, 2020, 128(7).

Deep Image Prior[7]

- **Goal:** given a corrupted image $\boldsymbol{y} = \boldsymbol{x}_\star + \boldsymbol{s}$, recover the clean image $\boldsymbol{x}_\star$ from the noisy observation
- **Idea:** using a deep network $f(\boldsymbol{\theta})$ to fit the observation $\boldsymbol{y}$:

$$\min_{\boldsymbol{\theta}} \ell(\underbrace{\boldsymbol{y}}_{\textbf{corrupted image}}, \underbrace{f(\boldsymbol{\theta})}_{\textbf{recovered image}})$$

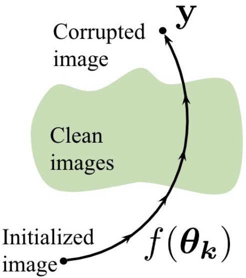


[7] Ulyanov D, Vedaldi A, Lempitsky V. Deep image prior[J]. International Journal of Computer Vision, 2020, 128(7).

Deep Image Prior[7]

- **Goal:** given a corrupted image $\boldsymbol{y} = \boldsymbol{x}_\star + \boldsymbol{s}$, recover the clean image $\boldsymbol{x}_\star$ from the noisy observation
- **Idea:** using a deep network $f(\boldsymbol{\theta})$ to fit the observation $\boldsymbol{y}$:

$$\min_{\boldsymbol{\theta}} \ell(\underbrace{\boldsymbol{y}}_{\text{corrupted image}}, \underbrace{f(\boldsymbol{\theta})}_{\text{recovered image}})$$



- **Early stopping:** As the network is highly **overparameterized**, early stopping is needed.

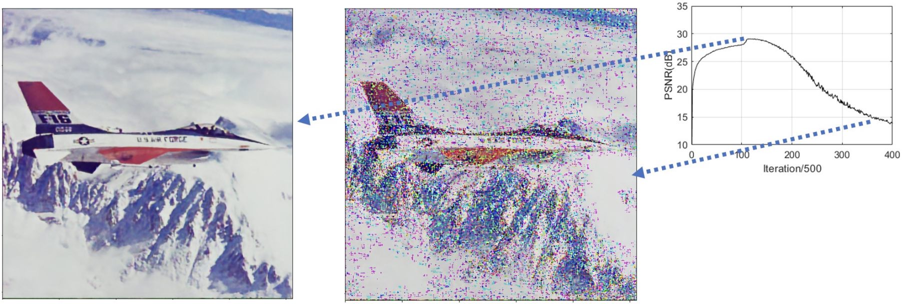
[7] Ulyanov D, Vedaldi A, Lempitsky V. Deep image prior[J]. International Journal of Computer Vision, 2020, 128(7).

A Case Study: Robust Image Recovery with Sparse Noise

$$\min_{\boldsymbol{\theta}} \| \underbrace{\boldsymbol{y}}_{\text{corrupted input}} - \underbrace{f(\boldsymbol{\theta})}_{\text{recovered image}} \|_1$$

sparse corruption

A Case Study: Robust Image Recovery with Sparse Noise



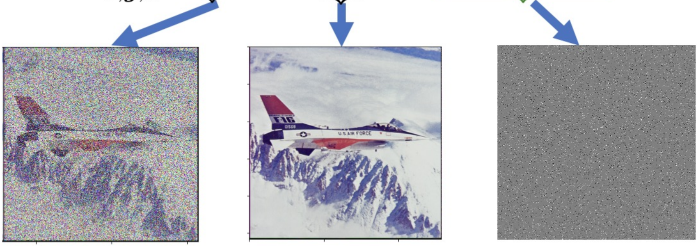
Early termination solution
(impractical!)

Global solution: $f(\boldsymbol{\theta}) \approx \boldsymbol{y}$
(overfitting!)

Robust Recovery without Overfitting?

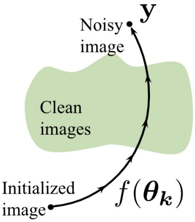
Method: sparse (double) overparameterization:[8]

$$\min_{\boldsymbol{\theta},\boldsymbol{g},\boldsymbol{h}} \| \underbrace{\mathbf{y}} - (\underbrace{f(\boldsymbol{\theta})} + \underbrace{\boldsymbol{g}\odot\boldsymbol{g}-\boldsymbol{h}\odot\boldsymbol{h}})\|_2^2$$

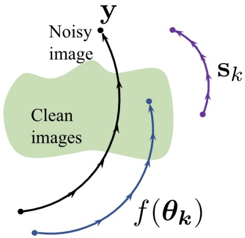


[8]You C, Zhu Z, Qu Q, Ma Y. Robust recovery via implicit bias of discrepant learning rates for double over-parameterization. NeurIPS'20.

Sparse Overparameterization Method



DIP

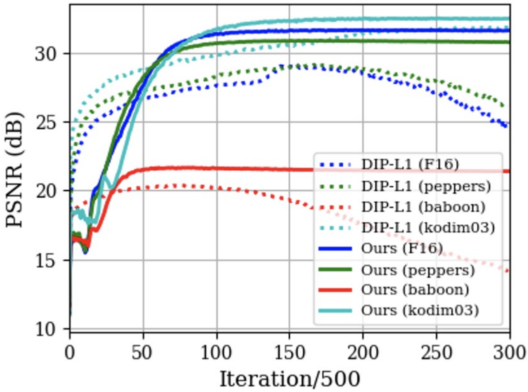


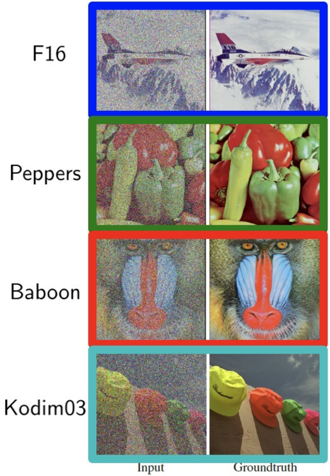
Ours

- **Optimization:** gradient descent with discrepant learning rate:

$$
\begin{aligned}
\boldsymbol{\theta} &\leftarrow \boldsymbol{\theta} - \tau \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\{\boldsymbol{u}, \boldsymbol{v}\}; \boldsymbol{\theta}) \\
\boldsymbol{u} &\leftarrow \boldsymbol{u} - \alpha\tau \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\{\boldsymbol{u}, \boldsymbol{v}\}; \boldsymbol{u}) \\
\boldsymbol{v} &\leftarrow \boldsymbol{v} - \alpha\tau \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\{\boldsymbol{u}, \boldsymbol{v}\}; \boldsymbol{v})
\end{aligned}
$$

Experiments on Real Images





No early stop, no parameter tuning!

Outline

References

SCAN ME

SCAN ME

1 Liu S, Zhu Z, Qu Q, You C, Robust Training under Label Noise by Over-parameterization, ICML'22.

2 You C, Zhu Z, Qu Q, Ma Y. Robust recovery via implicit bias of discrepant learning rates for double over-parameterization. NeurIPS'20.

3 Li X, Yaras C, An Y, Ravishankar S, and Qu Q. Data-Free Deep Image Recovery from Sparsely Corrupted Incomplete Measurements, In preparation, 2023.

Conclusion and Coming Attractions

Take-home Message: We can achieve better robustness in learning our overparameterized deep models by exploiting the low-dimensional structures in the data and network.

Thank You! Questions?

Call for Papers

- IEEE JSTSP Special Issue on Seeking Low-dimensionality in Deep Neural Networks (SLowDNN) Manuscript Due: **Nov. 30, 2023**.
- Conference on Parsimony and Learning (CPAL) January 2024, Hongkong, Manuscript Due: **Aug. 28, 2023**.

SCAN ME

SCAN ME