Robust Learning via Over-parameterization

Sheng Liu
NYU

Zhihui Zhu
University of Denver

Qing Qu
UMich

Chong You
Google NYC

Blessing of Over-Parameterization

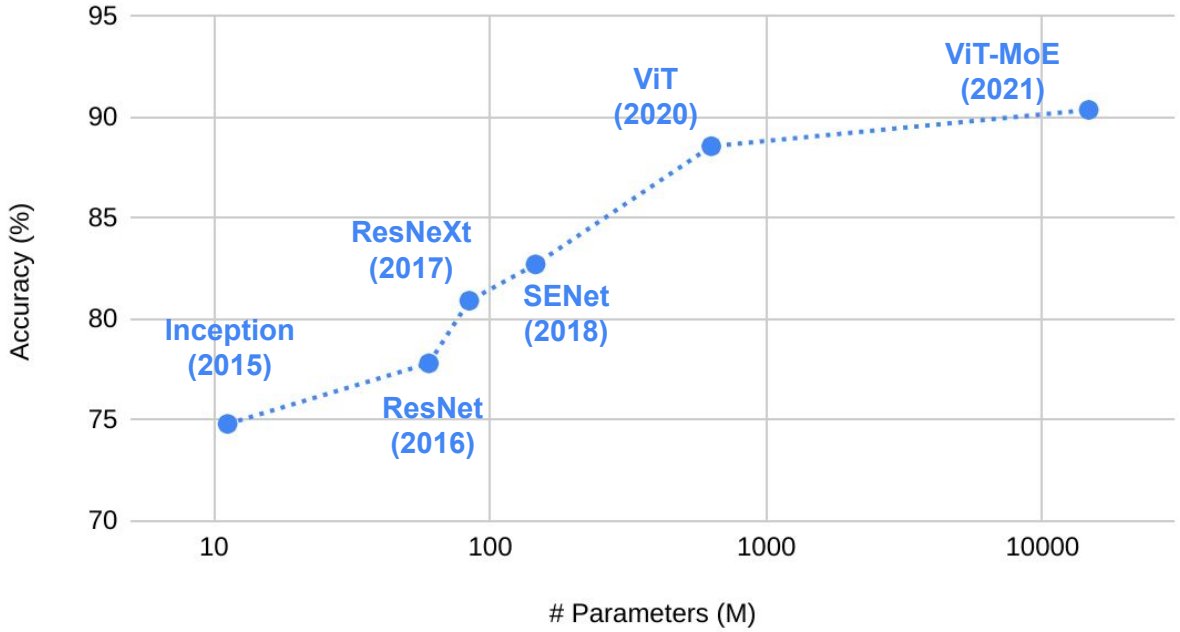


Figure: Accuracy vs. model size for image classification on ImageNet dataset

Over-Parameterization ➪➪➪ Overfitting!

Baseline (CE, dotted curves):

- Train Acc. = 100% (overfitting!)
- Test Acc. = ↗↗↗ ↘↘↘

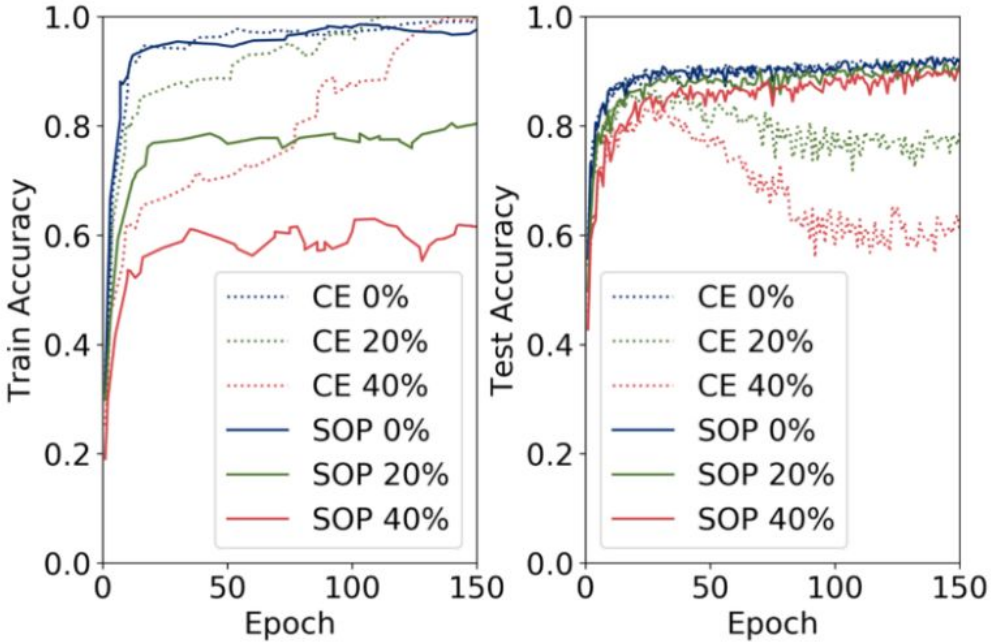


Figure: Learning curves for CIFAR-10 classification *under {0, 20%, 40%} label noise*.

This Talk: Over-parameterization *without* Overfitting

Baseline (CE, dotted curves):

- Train Acc. = 100% (overfitting!)
- Test Acc. = ↗↗↗ ↘↘↘

Our work (SOP):

- Train Acc. ≅ correct_labels%
- Test Acc. = ↗↗↗↗↗

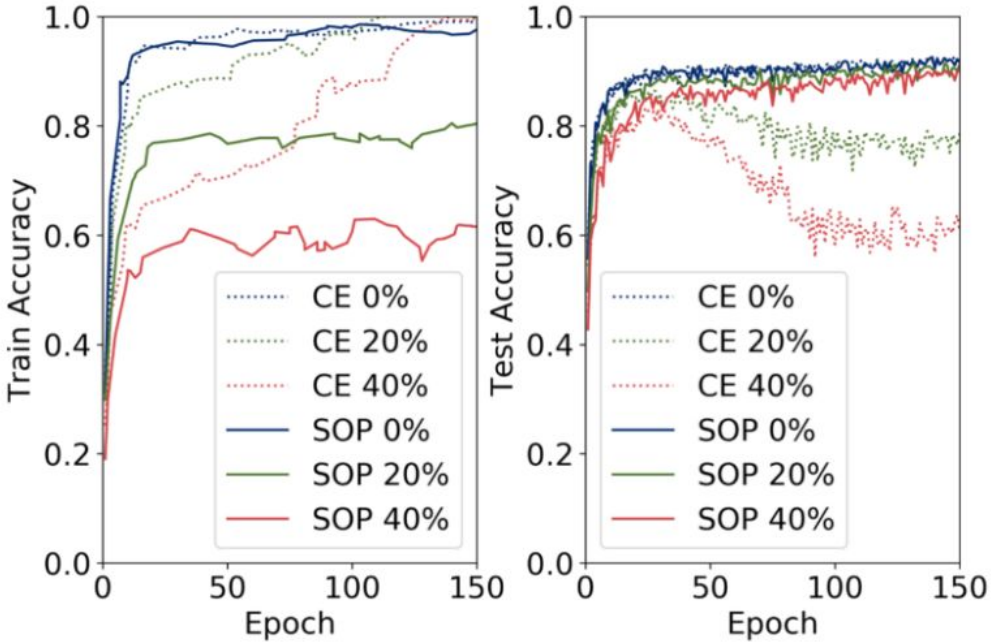


Figure: Learning curves for CIFAR-10 classification *under {0, 20%, 40%} label noise*.

A Sparse Over-parameterization (SOP) Approach

Our Strategy: Sparse modeling of the label noise s^*_i

$$\min_{\{u_i, v_i\},\, \Theta} \mathscr{L}(\{u_i, v_i\}, \Theta) := \sum_i \ell_{CE}\big(f(x_i; \Theta) + \underbrace{u_i \odot u_i - v_i \odot v_i}_{\text{Sparse Over-parameterization (SOP)}}, y_i \big)$$

Hadamard product

The idea is to have

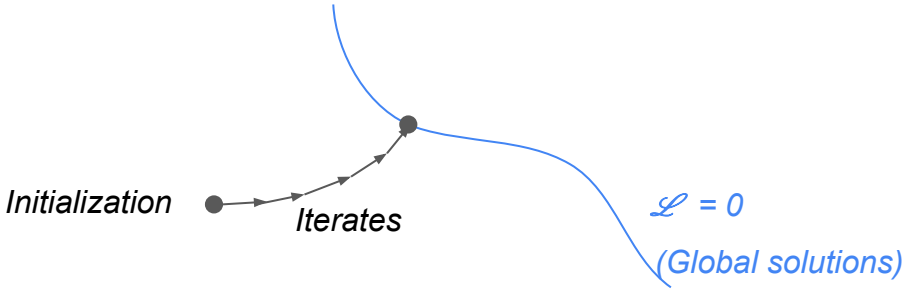
- $f(x_i; \Theta)$ converges to $f(x_i; \Theta^*)$,
- $u_i \odot u_i - v_i \odot v_i$ converges to s^*_i,

and such a (correct) solution is a global optimal solution!

However, "incorrect" global optimal solutions exist! (e.g., $u_i = v_i = 0$, $f(x_i; \Theta) = y_i$)

Algorithm Matters!

We rely on a particularly designed algorithm to find a particular solution



- Initialization: A small value for $\{u_i, v_i\}$ (e.g., *i.i.d. N(0, 1e-16)*)
- Iterates: Gradient descent with a discrepant learning rate

$$\Theta \leftarrow \Theta - \quad \boldsymbol{\tau} \cdot \partial \mathscr{L}(\{u_i, v_i\}, \Theta) \,/\, \partial \Theta$$

$$u_i \leftarrow u_i - \boldsymbol{\tau} \cdot \boldsymbol{\alpha} \cdot \partial \mathscr{L}(\{u_i, v_i\}, \Theta) \,/\, \partial u_i, \quad i = 1, \ldots, N$$

$$v_i \leftarrow v_i - \boldsymbol{\tau} \cdot \boldsymbol{\alpha} \cdot \partial \mathscr{L}(\{u_i, v_i\}, \Theta) \,/\, \partial v_i, \quad i = 1, \ldots, N$$

Results: High Accuracy

	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**

… and Fast!

CE	Co-teaching+	DivideMix	ELR+	SOP	SOP+
0.9h	4.4h	5.4h	2.3h	1.0h	2.1h

NEW SOTA on CIFAR-10**N***

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
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**
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**
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**

Two-network based

Two-network based

Semi-supervised

Ours

* LEARNING WITH NOISY LABELS REVISITED: A STUDY USING REAL-WORLD HUMAN ANNOTATIONS, ICLR ‘22

Thank you!