Changing the Training Data Distribution to Reduce Simplicity Bias Improves In-distribution Generalization

Dang Nguyen, Paymon Haddad, Eric Gan, and Baharan Mirzasoleiman

Department of Computer Science, UCLA

BigML

Not all minima are created equal

- In the in-distribution settings, when training and test come from the same distribution, minimizing the training loss generalizes well on the test data [1].

[1] Belkin, Mikhail, et al. "Reconciling modern machine-learning practice and the classical bias–variance trade-off." Proceedings of the National Academy of Sciences 116.32 (2019): 15849-15854.

Not all minima are created equal

- In the in-distribution settings, when training and test come from the same distribution, minimizing the training loss generalizes well on the test data [1].
- **However**, some global minima generalize better than others!

batch size	train accuracy	test accuracy	train loss
1	100.0 (100.0 - 100.0)	77.2 (77.7 - 76.4)	0.00 (0.00 - 0.00)
8	100.0 (100.0 - 100.0)	76.5 (76.7 - 75.9)	0.00 (0.00 - 0.00)
256	100.0 (100.0 - 100.0)	63.2 (63.4 - 61.3)	0.00 (0.00 - 0.00)
2048	100.0 (100.0 - 99.8)	60.2 (60.6 - 58.6)	0.00 (0.02 - 0.00)

Table 1. Train and test accuracy on CIFAR10, taken from [2].

[1] Belkin, Mikhail, et al. "Reconciling modern machine-learning practice and the classical bias–variance trade-off." Proceedings of the National Academy of Sciences 116.32 (2019): 15849-15854.
[2] https://www.dropbox.com/scl/fi/7lk8jkchj82oe7smh7b4w/Hossein_Mobahi_SAM_CSML_Talk.pdf?rlkey=1mc56v58cvcy480bflexfuioq&e=1&dl=0

Not all minima are created equal

- In the in-distribution settings, when training and test come from the same distribution, minimizing the training loss generalizes well on the test data [1]
- **However**, some global minima generalize better than others!

batch size	train accuracy	test accuracy	train loss
1	100.0 (100.0 - 100.0)	77.2 (77.7 - 76.4)	0.00 (0.00 - 0.00)
8	100.0 (100.0 - 100.0)	76.5 (76.7 - 75.9)	0.00 (0.00 - 0.00)
256	100.0 (100.0 - 100.0)	63.2 (63.4 - 61.3)	0.00 (0.00 - 0.00)
2048	100.0 (100.0 - 99.8)	60.2 (60.6 - 58.6)	0.00 (0.02 - 0.00)

Table 1. Train and test accuracy on CIFAR10, taken from [2].

[1] Belkin, Mikhail, et al. "Reconciling modern machine-learning practice and the classical bias–variance trade-off." Proceedings of the National Academy of Sciences 116.32 (2019): 15849-15854.
[2] https://www.dropbox.com/scl/fi/7lk8jkchj82oe7smh7b4w/Hossein_Mobahi_SAM_CSML_Talk.pdf?rlkey=1mc56v58cvcy480bflexfuioq&e=1&dl=0

UCLA

Improving ID generalization via data modification

Can we improve the ID performance by changing the data distribution of a clean dataset?

We do not assume any redundant, noisy, or harmful examples in the data.

Thus, we do not want to filter such examples!

The superior ID generalization of SAM

- Sharpness-aware minimization (SAM) [3] minimizes both loss and sharpness.

$$L_{\mathrm{SAM}}(w) = \max_{\|\epsilon\|_2 \leq p} L(w+\epsilon) = \underbrace{L(w)}_{\text{loss}} + \underbrace{[\max_{\|\epsilon\|_2 \leq p} L(w+\epsilon) - L(w)]}_{\text{sharpness}}$$

[3] Foret, Pierre, et al. "Sharpness-aware minimization for efficiently improving generalization." arXiv preprint arXiv:2010.01412 (2020).

The superior ID generalization of SAM

- Sharpness-aware minimization (SAM) [3] minimizes both loss and sharpness.

$$L_{\mathrm{SAM}}(w) = \max_{\|\epsilon\|_2 \le p} L(w+\epsilon) = \underbrace{L(w)}_{\text{loss}} + \underbrace{[\max_{\|\epsilon\|_2 \le p} L(w+\epsilon) - L(w)]}_{\text{sharpness}}$$

- SAM finds flatter local minima that generalize better than SGD!

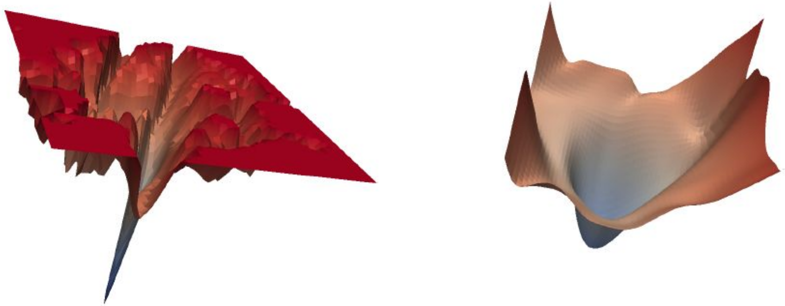

Figure 1. (left) Sharp minima of SGD (right) wide minima of SAM [3].

[3] Foret, Pierre, et al. "Sharpness-aware minimization for efficiently improving generalization." arXiv preprint arXiv:2010.01412 (2020).

UCLA

The superior ID generalization of SAM

- Sharpness-aware minimization (SAM) [3] minimizes both loss and sharpness

$$L_{\mathrm{SAM}}(w) = \max_{\|\epsilon\|_2 \le p} L(w+\epsilon) = \underbrace{L(w)}_{\text{loss}} + \underbrace{[\max_{\|\epsilon\|_2 \le p} L(w+\epsilon) - L(w)]}_{\text{sharpness}}$$

- SAM finds flatter local minima that generalize better than SGD!

Can we get insights from SAM to change the data distribution to improve ID generalization?

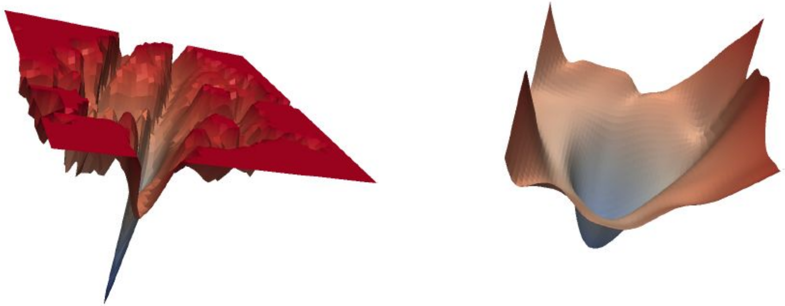

Figure 1. (left) Sharp minima of SGD (right) wide minima of SAM [3].

[3] Foret, Pierre, et al. "Sharpness-aware minimization for efficiently improving generalization." arXiv preprint arXiv:2010.01412 (2020).

SAM learns features more evenly than GD

We theoretically prove that SAM is less reliant on simplicity bias compared to GD.

[Informal] Consider a two-layer nonlinear CNNs, and a data with a fast-learnable and a slow-learnable feature. Then, starting from the same initialization, SAM learns the fast-learnable and slow-learnable features at a more uniform speed than GD, i.e., for every iteration $t \in [1, T_0]$:

$$\underbrace{\mathrm{SAM}_{fast}^{(t)} - \mathrm{SAM}_{slow}^{(t)}}_{\text{Feature learning gap in SAM}} < \underbrace{\mathrm{GD}_{fast}^{(t)} - \mathrm{GD}_{slow}^{(t)}}_{\text{Feature learning gap in GD}}$$

UpSample Early For Uniform Learning (USEFUL)

We propose a method to reduce the simplicity bias by changing the data distribution to mimic the training dynamic of SAM.

UpSample Early For Uniform Learning (USEFUL)

We propose a method to reduce the simplicity bias by changing the data distribution to mimic the training dynamic of SAM.

- Step 1: Identify examples with fast-learnable features via clustering model outputs in early training.
  - Such examples are provably separable, based on model's output early in training!

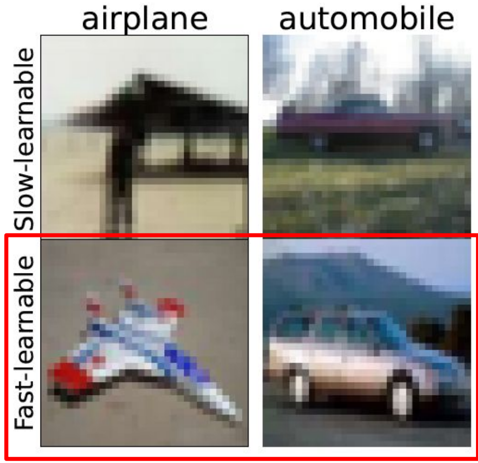

Figure 2. CIFAR10 images

UpSample Early For Uniform Learning (USEFUL)

We propose a method to reduce the simplicity bias by changing the data distribution to mimic the training dynamic of SAM.

- Step 1: Identify examples with fast-learnable features via clustering model outputs in early training.
  - Such examples are provably separable, based on model's output early in training!
- Step 2: Upsample examples that are not in the cluster of points containing fast-learnable features.

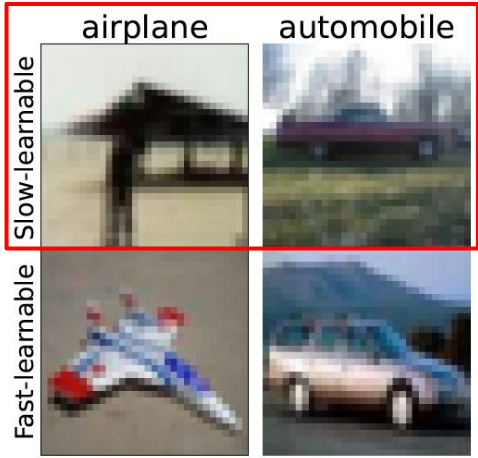

Figure 2. CIFAR10 images

UpSample Early For Uniform Learning (USEFUL)

We propose a method to reduce the simplicity bias by changing the data distribution to mimic the training dynamic of SAM.

- Step 1: Identify examples with fast-learnable features via clustering model outputs in early training.
  - Such examples are provably separable, based on model's output early in training!
- Step 2: Upsample examples that are not in the cluster of points containing fast-learnable features.
- Step 3: Restart training on the modified data distribution.

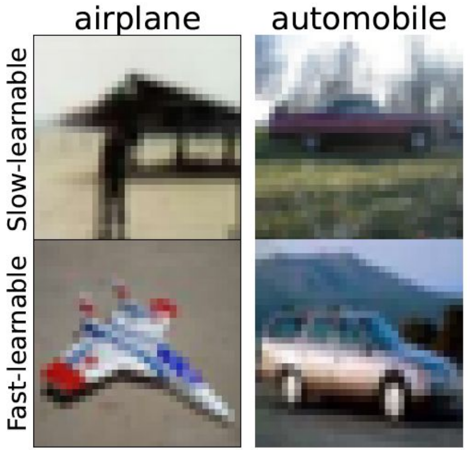

Figure 2. CIFAR10 images

Experimental results

Our method improves the performance of both SGD and SAM, achieving SOTA results in a variety of settings.

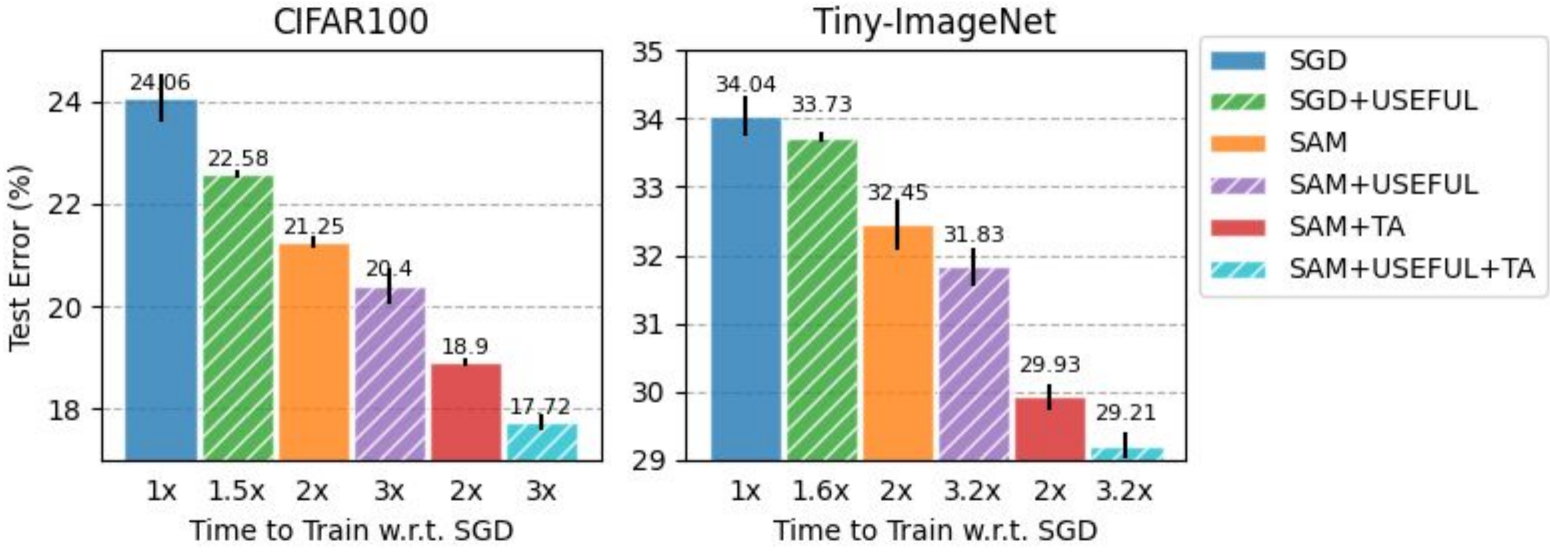

Figure 3. Test classification errors of ResNet18 on different datasets.

Experimental results

Our method improves the performance of both SGD and SAM, achieving SOTA results in a variety of settings.

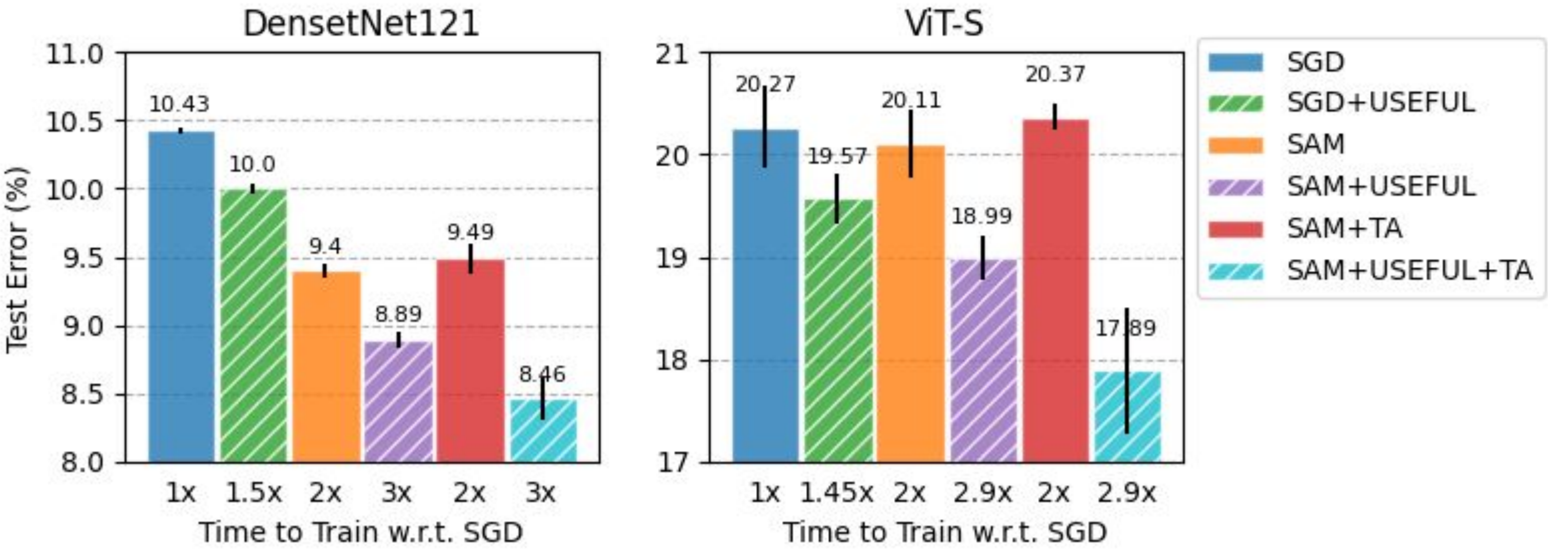

Figure 4. Test classification errors of different architectures on CIFAR10.

Thank you!
Please come visit our poster at
Session 5: Fri 13 Dec 11 AM - 2 PM PST

Dang Nguyen   nguyentuanhaidang@gmail.com   @dangnth97   https://hsgser.github.io/