Sharpness-Aware Minimization

Farid Davletshin

Modern Optimization Methods

December 26, 2022

SAM in a few words

SAM is an optimization algorithm that:

- Minimizes loss value **AND** sharpness
- Is efficient and easy to implement
- Strongly improves generalization (SOTA on Imagenet, CIFAR, SVHN, and others)
- Robust to label noise

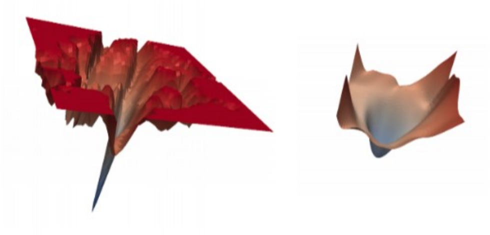


Figure: (left) A sharp minimum to which a ResNet trained with SGD converged. (right) A wide minimum to which the same ResNet trained with SAM converged.

SAM in a few words

- Is more interpretable

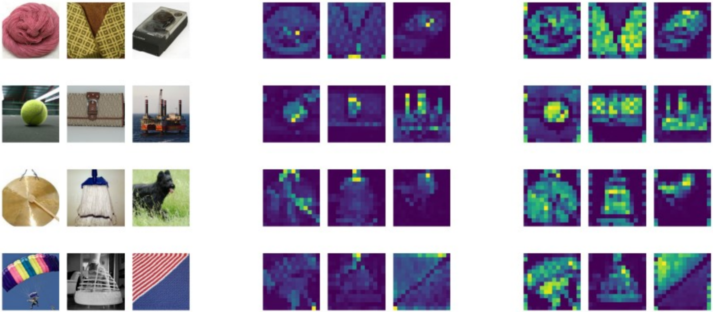


Figure 3: Raw images (**Left**) and attention maps of ViT-S/16 with (**Right**) and without (**Middle**) sharpness-aware optimization.

SOTA for today

Task	Dataset	Model	Metric Name	Metric Value	Global Rank	Uses Extra Training Data	Result	Benchmark
Fine-Grained Image Classification	Birdsnap	EffNet-L2 (SAM)	Accuracy	90.07%	# 1	✓		Compare
Image Classification	CIFAR-10	PyramidNet (SAM)	Percentage correct	98.6	# 28	✓		Compare
Image Classification	CIFAR-100	PyramidNet (SAM)	Percentage correct	89.7	# 28	✓		Compare
Image Classification	CIFAR-100	EffNet-L2 (SAM)	Percentage correct	96.08	# 1	✓		Compare
Image Classification	Fashion-MNIST	Shake-Shake (SAM)	Percentage error	3.59	# 2	×		Compare
			Accuracy	96.41	# 3	×		Compare
Fine-Grained Image Classification	FGVC Aircraft	EffNet-L2 (SAM)	Top-1 Error Rate	4.82	# 1	✓		Compare
Image Classification	Flowers-102	EffNet-L2 (SAM)	Accuracy	99.65%	# 4	✓		Compare
Fine-Grained Image Classification	Food-101	EffNet-L2 (SAM)	Accuracy	96.18	# 1	✓		Compare

SOTA for today

Image Classification	ImageNet	ResNet-152 (SAM)	Top 1 Accuracy	81.6%	# 440	×		Compare
			Top 5 Accuracy	95.65	# 107	×		Compare
Image Classification	ImageNet	EfficientNet-L2-475 (SAM)	Top 1 Accuracy	88.61%	# 32	✓		Compare
			Number of params	480M	# 775	✓		Compare
			Hardware Burden	None	# 1	✓		Compare
			Operations per network pass	None	# 1	✓		Compare
Fine-Grained Image Classification	Oxford-IIIT Pet Dataset	EffNet-L2 (SAM)	Top-1 Error Rate	2.90%	# 1	✓		Compare
			Accuracy	97.10%	# 1	✓		Compare
Fine-Grained Image Classification	Stanford Cars	EffNet-L2 (SAM)	Accuracy	95.96%	# 4	✓		Compare
Image Classification	SVHN	WRN28-10 (SAM)	Percentage error	0.99	# 1	×		Compare

Neural Network training

- Training dataset $\mathcal{S} \triangleq \cup_{i=1}^{n}\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$ drawn i.i.d. from distribution $\mathscr{D}$
- Neural network with weights $\boldsymbol{w} \in \mathcal{W} \subseteq \mathbb{R}^d$;
- Per-data-point loss function $l : \mathcal{W} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$,
- Training loss $L_S(\boldsymbol{w}) \triangleq \frac{1}{n}\sum_{i=1}^{n} l(\boldsymbol{w}, \boldsymbol{x}_i, \boldsymbol{y}_i)$ which approximates $L_{\mathscr{D}}(\boldsymbol{w}) \triangleq \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim D}[l(\boldsymbol{w}, \boldsymbol{x}, \boldsymbol{y})]$.

- Train the network to get $\boldsymbol{w}$ having low population loss $L_{\mathscr{D}}(\boldsymbol{w})$.

Not all minima created are equal

Training a neural network with different optimization strategies (for example, change batch size), we get:

- Perfect fit on the training set.
- Training loss approaching zero.
- But very different test accuracy.

Table: Train and test accuracy for a convolutional network trained on CIFAR10, for different batch sizes (reproducing an experiment from [SMN+16])

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)

Conclusion: Some global minima generalize better than others

Main theorem

Theorem

For any $\rho > 0$, with high probability over training set $\mathcal{S}$ generated from distribution $\mathscr{D}$,

$$L_{\mathscr{D}}(\boldsymbol{w}) \leq \max_{\|\epsilon\|_2 \leq \rho} L_{\mathcal{S}}(\boldsymbol{w} + \epsilon) + h(\|\boldsymbol{w}\|_2^2 / \rho^2),$$

where $h : \mathbb{R}^+ \to \mathbb{R}^+$ is a strictly increasing function (under some technical conditions on $L_{\mathscr{D}}(\boldsymbol{w})$).

Simplifying lower bound

Re-arranging the terms to make the sharpness term more explicit:

$$[\max_{\|\epsilon\|_2 \leq \rho} L_{\mathcal{S}}(\boldsymbol{w}+\epsilon) - L_{\mathcal{S}}(\boldsymbol{w})] + L_{\mathcal{S}}(\boldsymbol{w}) + h(\|\boldsymbol{w}\|_2^2/\rho^2).$$

The expression of h heavily depends on the proof method, we substitute the second term with $\lambda\|\boldsymbol{w}\|_2^2$ for standard L2 regularization.

This gives us the SAM objective:

$$\min_{\boldsymbol{w}} L_{\mathcal{S}}^{SAM}(\boldsymbol{w}) + \lambda||\boldsymbol{w}||_2^2 \qquad \text{where} \qquad L_{\mathcal{S}}^{SAM}(\boldsymbol{w}) \triangleq \max_{||\epsilon||_p \leq \rho} L_S(\boldsymbol{w}+\epsilon),$$

Solving min-max problem

Do a first order approximation of the objective:

$$\epsilon^*(\boldsymbol{w}) \triangleq \arg\max_{\|\epsilon\|_p \le \rho} L_{\mathcal{S}}(\boldsymbol{w}+\epsilon) \approx \arg\max_{\|\epsilon\|_p \le \rho} L_{\mathcal{S}}(\boldsymbol{w}) + \epsilon^T \nabla_{\boldsymbol{w}} L_{\mathcal{S}}(\boldsymbol{w})$$
$$= \arg\max_{\|\epsilon\|_p \le \rho} \epsilon^T \nabla_{\boldsymbol{w}} L_{\mathcal{S}}(\boldsymbol{w}).$$

Well known solution to the dual norm problem:

$$\hat{\epsilon}(\boldsymbol{w}) = \rho \,\text{sign}\left(\nabla_{\boldsymbol{w}} L_{\mathcal{S}}(\boldsymbol{w})\right) |\nabla_{\boldsymbol{w}} L_{\mathcal{S}}(\boldsymbol{w})|^{q-1} / \left(\|\nabla_{\boldsymbol{w}} L_{\mathcal{S}}(\boldsymbol{w})\|_q^q \right)^{1/p} \quad (2)$$

Computing the SAM gradient

$$\nabla_{\boldsymbol{w}} L_{\mathcal{S}}^{SAM}(\boldsymbol{w}) \approx \nabla_{\boldsymbol{w}} L_{\mathcal{S}}(\boldsymbol{w}+\hat{\epsilon}(\boldsymbol{w})) = \frac{d(\boldsymbol{w}+\hat{\epsilon}(\boldsymbol{w}))}{d\boldsymbol{w}} \nabla_{\boldsymbol{w}} L_{\mathcal{S}}(\boldsymbol{w})|_{\boldsymbol{w}+\hat{\epsilon}(w)}$$
$$= \nabla_{\boldsymbol{w}} L_{\mathcal{S}}(\boldsymbol{w})|_{\boldsymbol{w}+\hat{\epsilon}(\boldsymbol{w})} + \frac{d\hat{\epsilon}(\boldsymbol{w})}{d\boldsymbol{w}} \nabla_{\boldsymbol{w}} L_{\mathcal{S}}(\boldsymbol{w})|_{\boldsymbol{w}+\hat{\epsilon}(\boldsymbol{w})}.$$

The algorithm

Input: Training set $\mathcal{S} \triangleq \cup_{i=1}^{n}\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$, Loss function $l : \mathcal{W} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, Batch size b, Step size $\eta > 0$, Neighborhood size $\rho > 0$.

Output: Model trained with SAM

Initialize weights $\boldsymbol{w}_0$, $t = 0$;

while *not converged* **do**

  Sample batch $\mathcal{B} = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), ...(\boldsymbol{x}_b, \boldsymbol{y}_b)\}$;

  $\delta(\boldsymbol{w}) = \nabla_{\boldsymbol{w}} L_{\mathcal{B}}(\boldsymbol{w})$;

  $\hat{\epsilon} = \frac{\delta(\boldsymbol{w}_t)}{\|\delta(\boldsymbol{w}_t)\|}$;

  $\boldsymbol{w}_{\mathrm{adv}} = \boldsymbol{w}_t + \hat{\epsilon}$;

  $\boldsymbol{g} = \delta(\boldsymbol{w}_{\mathrm{adv}})$;

  $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \boldsymbol{g}$;

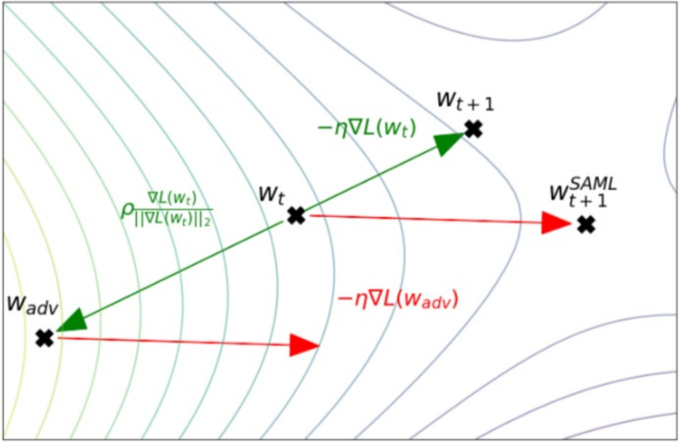
  $t = t + 1$;

end

return $\boldsymbol{w}_t$

The algorithm

Figure: One update of SAM against one update of plain gradient descent.



Robustness to corrupted labels

Method	Noise rate (%)			
	20	40	60	80
[SOA+19]	94.0	92.8	90.3	74.1
[ZS18]	89.7	87.6	82.7	67.9
[LYL+19]	87.1	81.8	75.4	-
[CLCZ19]	89.7	-	-	52.3
[HQJZ19]	92.6	90.3	43.4	-
MentorNet [JZL+17]	92.0	91.2	74.2	60.0
Mixup [ZCDLP17]	94.0	91.5	86.8	76.9
MentorMix [JHLY19]	**95.6**	**94.2**	91.3	**81.0**
SGD	84.8	68.8	48.2	26.2
Mixup	93.0	90.0	83.8	70.2
Bootstrap + Mixup	93.3	92.0	87.6	72.0
SAM	95.1	93.4	90.5	77.9
Bootstrap + SAM	95.4	**94.2**	**91.8**	79.9

Table: Test accuracy on the clean test set for models trained on CIFAR-10 with noisy labels. Lower block is our implementation, upper block gives scores

Multiple Columns

Heading

1. Statement
2. Explanation
3. Example

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Integer lectus nisl, ultricies in feugiat rutrum, porttitor sit amet augue. Aliquam ut tortor mauris. Sed volutpat ante purus, quis accumsan dolor.

Obtained results

Table: Test error rates for ResNet trained on CIFAR-10, with and without SAM.

CIFAR-10	Epoch	top-1	top-k
No SAM	100	15.34	0.91
No SAM	200	12.94	0.89
No SAM	400	11.24	**0.8**
SAM	100	**14.5**	**0.7**
SAM	200	**12.48**	**0.55**
SAM	400	**11.07**	1.08

Obtained results

Table: Test error rates for ResNet trained on CIFAR-100, with and without SAM.

CIFAR-100	Epoch	top-1	top-k
No SAM	100	89.93	66.33
No SAM	200	90.19	66.19
No SAM	400	**48.98**	**20.6**
SAM	100	**57.6**	**25.5**
SAM	200	**55.25**	**24.93**
SAM	400	78.09	43.27

FGSM (Fast Gradient Sign Method)

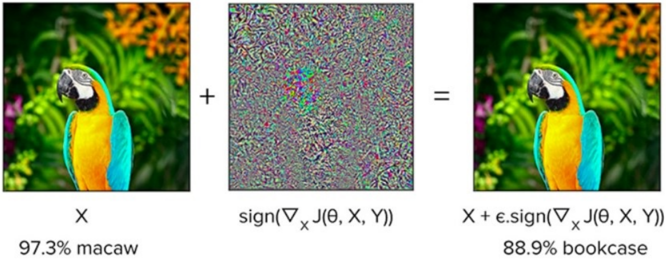


Figure 1: The Fast Gradient Sign Method (FGSM) for adversarial image generation

$$\hat{\epsilon}(\boldsymbol{w}) = \rho \ \mathrm{sign}\left(\nabla_{\boldsymbol{w}} L_{\mathcal{S}}(\boldsymbol{w})\right)$$

References

Pierre Foret, Ariel Kleiner, Hossein Mobahi, 2021
Sharpness-aware Minimization for Efficiently Improving Generalization
ICLR Spotlight 12(3), 656 – 678.

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, 2022
Adaptive Sharpness-Aware Minimization for Scale-Invariant Learning of Deep Neural Networks
Proceedings of Machine Learning Research (ICML) 9(5),538 – 567.

The End