Tsinghua University

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Poster Tag: Thur-AM-358

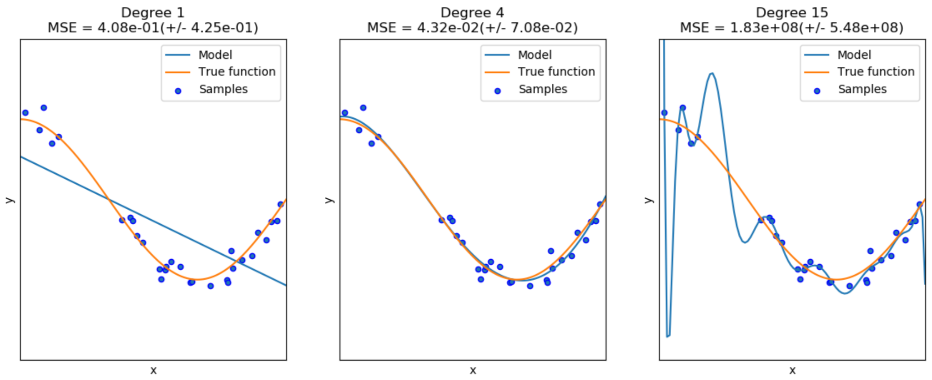
Gradient Norm Aware Minimization Seeks First-Order Flatness and Improves Generalization

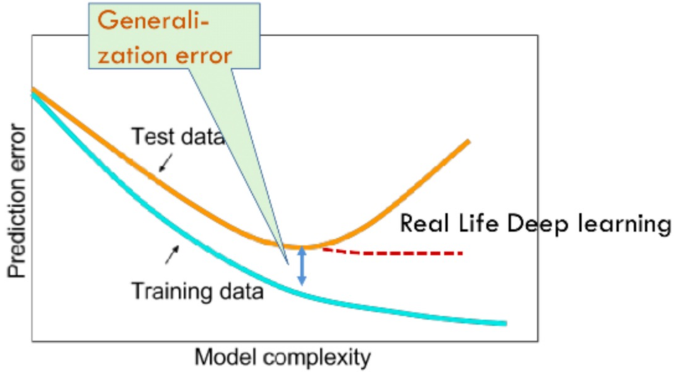
CVPR, 2023

***Xingxuan Zhang*[†], *Renzhe Xu*[†], *Han Yu*, *Hao zou*, *Peng Cui*[*]**

Generalization and Overfitting

- As the model overfits the training data, the generalization error increases.

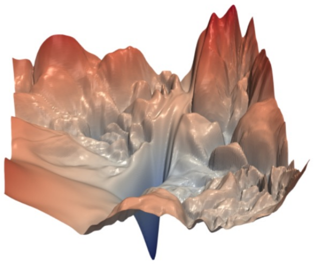




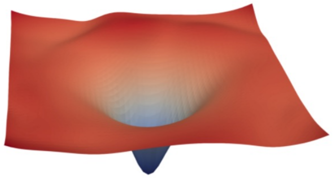
Ref: https://scikit-learn.org/stable/auto_examples/model_selection/plot_underfitting_overfitting.html

Generalization and Flatness

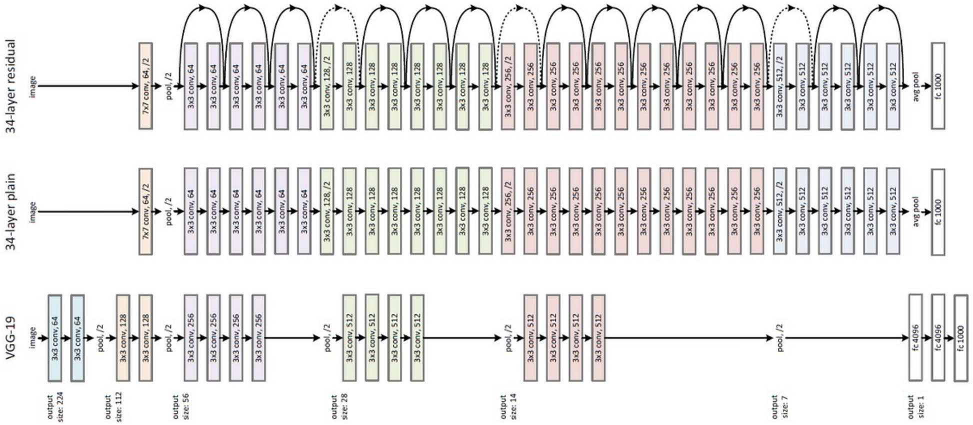
- Skip connections in ResNet lead to flat minima and better generalization



(a) without skip connections



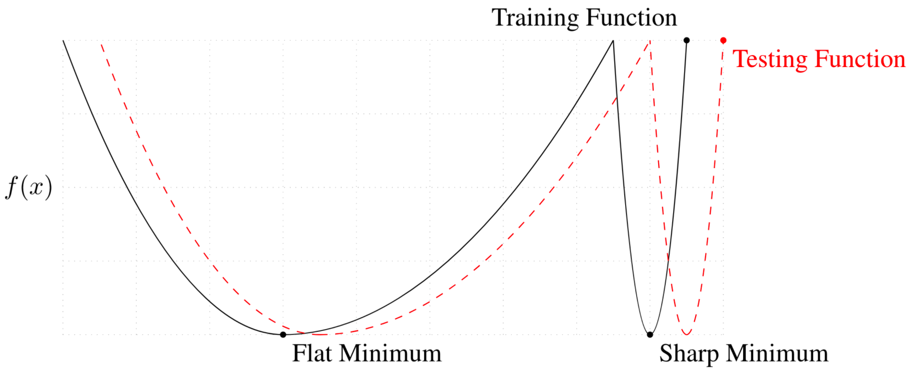
(b) with skip connections



Ref: He, Kaiming, et al. "Deep residual learning for image recognition." In *CVPR*. 2016
Li, Hao, et al. "Visualizing the loss landscape of neural nets." *Advances in neural information processing systems* 31 (2018).

Generalization and Flatness

- Recent works show that flat minima lead to better generalization



Theorem (stated informally) 1. *For any $\rho > 0$, with high probability over training set $\mathcal{S}$ generated from distribution $\mathscr{D}$,*

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

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

To make explicit our sharpness term, we can rewrite the right hand side of the inequality above as

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

Ref: Keskar, Nitish Shirish, et al. "On large-batch training for deep learning: Generalization gap and sharp minima." in ICLR 2017
Foret, Pierre, et al. "Sharpness-aware minimization for efficiently improving generalization." in ICLR 2021.

Zeroth-order flatness and first-order flatness

- Zeroth-order flatness – sharpness aware minimization (SAM)

Theorem (stated informally) 1. *For any $\rho > 0$, with high probability over training set $\mathcal{S}$ generated from distribution $\mathscr{D}$,*

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

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

To make explicit our sharpness term, we can rewrite the right hand side of the inequality above as

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

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

Input: Training set $\mathcal{S} \triangleq \cup_{i=1}^n \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$, Loss function $l : \mathcal{W} \times \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_+$, 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 $\mathcal{B} = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), ...(\boldsymbol{x}_b, \boldsymbol{y}_b)\}$;
  Compute gradient $\nabla_{\boldsymbol{w}} L_{\mathcal{B}}(\boldsymbol{w})$ of the batch's training loss;
  Compute $\hat{\boldsymbol{\epsilon}}(\boldsymbol{w})$ per equation 2;
  Compute gradient approximation for the SAM objective (equation 3): $\boldsymbol{g} = \nabla_w L_{\mathcal{B}}(\boldsymbol{w})|_{\boldsymbol{w}+\hat{\boldsymbol{\epsilon}}(\boldsymbol{w})}$;
  Update weights: $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \boldsymbol{g}$;
  $t = t + 1$;
end
return $\boldsymbol{w}_t$

Algorithm 1: SAM algorithm

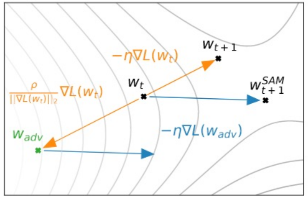


Figure 2: Schematic of the SAM parameter update.

Ref: Foret, Pierre, et al. "Sharpness-aware minimization for efficiently improving generalization." in ICLR 2021.

Zeroth-order flatness and first-order flatness

- zeroth-order flatness and first-order flatness

Definition 3.1 (ρ-zeroth-order flatness). For any $\rho > 0$, the ρ-zeroth-order flatness $R_\rho^{(0)}(\boldsymbol{\theta})$ of function $\hat{L}(\boldsymbol{\theta})$ at a point $\boldsymbol{\theta}$ is defined as

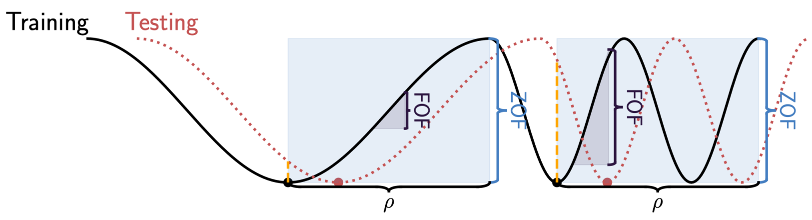
$$R_\rho^{(0)}(\boldsymbol{\theta}) \triangleq \max_{\boldsymbol{\theta}' \in B(\boldsymbol{\theta},\rho)} \left(\hat{L}(\boldsymbol{\theta}') - \hat{L}(\boldsymbol{\theta})\right), \quad \forall \boldsymbol{\theta} \in \Theta. \quad (2)$$

Here ρ is the perturbation radius that controls the magnitude of the neighborhood.

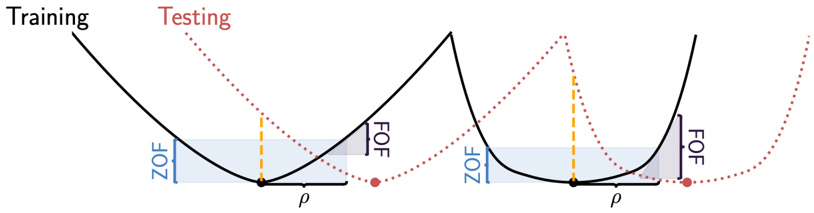
Definition 4.1 (ρ-first-order flatness). For any $\rho > 0$, the ρ-first-order flatness $R_\rho^{(1)}(\boldsymbol{\theta})$ of function $\hat{L}(\boldsymbol{\theta})$ at a point $\boldsymbol{\theta}$ is defined as

$$R_\rho^{(1)}(\boldsymbol{\theta}) \triangleq \rho \cdot \max_{\boldsymbol{\theta}' \in B(\boldsymbol{\theta},\rho)} \left\|\nabla \hat{L}(\boldsymbol{\theta}')\right\|, \quad \forall \boldsymbol{\theta} \in \Theta. \quad (3)$$

Here ρ is the perturbation radius that controls the magnitude of the neighbourhood.



(a)



(b)

GAM : gradient norm aware minimization

- The definition of GAM

Definition (Gradient norm Aware Minimization (GAM)). For any $\rho > 0$, GAM is defined as

$$R_{\rho}^{\mathrm{GNR}}(\boldsymbol{\theta}) \triangleq \rho \cdot \max_{\boldsymbol{\theta}' \in B(\boldsymbol{\theta}, \rho)} \left\| \nabla \hat{L}(\boldsymbol{\theta}') \right\|, \quad \forall \boldsymbol{\theta} \in \Theta. \tag{1}$$

Here ρ is the perturbation radius that controls the magnitude of the neighbourhood.

GAM : gradient norm aware minimization

- The approximation and optimization of GAM

$$\nabla R_\rho^{(1)}(\boldsymbol{\theta}) \approx \rho \cdot \nabla \left\| \nabla \hat{L}(\boldsymbol{\theta}^{\text{adv}}) \right\|, \quad \boldsymbol{\theta}^{\text{adv}} = \boldsymbol{\theta} + \rho \cdot \frac{\boldsymbol{f}}{\|\boldsymbol{f}\|},$$

$$\boldsymbol{f} = \nabla \left\| \nabla \hat{L}(\boldsymbol{\theta}) \right\|.$$

$$\forall \boldsymbol{\theta} \in \Theta, \quad \nabla \|\nabla \hat{L}(\boldsymbol{\theta})\| = \frac{\nabla^2 \hat{L}(\boldsymbol{\theta}) \cdot \nabla \hat{L}(\boldsymbol{\theta})}{\|\nabla \hat{L}(\boldsymbol{\theta})\|}.$$

Algorithm 1 Gradient norm Aware Minimization (GAM)

Input: Batch size b, Learning rate η_t, Perturbation radius ρ_t, Trade-off coefficient α, Small constant ξ
$t \leftarrow 0$, $\boldsymbol{\theta}_0 \leftarrow$ initial parameters
while $\boldsymbol{\theta}_t$ not converged **do**
  Sample $W_t \leftarrow \{(x_1, y_1), (x_2, y_2), \ldots, (x_b, y_b)\}$
  $\boldsymbol{h}_t^{\text{loss}} \leftarrow \nabla L^{\text{oracle}}(\boldsymbol{\theta}_t)$ ▷ Calculate the oracle loss gradient $\nabla L^{\text{oracle}}(\boldsymbol{\theta}_t)$
  $\boldsymbol{f}_t \leftarrow \nabla^2 \hat{L}_{W_t}(\boldsymbol{\theta}_t) \cdot \frac{\nabla \hat{L}_{W_t}(\boldsymbol{\theta}_t)}{\|\nabla \hat{L}_{W_t}(\boldsymbol{\theta}_t)\| + \xi}$
  $\boldsymbol{\theta}_t^{\text{adv}} \leftarrow \boldsymbol{\theta}_t + \rho_t \cdot \frac{\boldsymbol{f}_t}{\|\boldsymbol{f}_t\| + \xi}$
  $\boldsymbol{h}_t^{\text{norm}} \leftarrow \rho_t \cdot \nabla^2 \hat{L}_{W_t}(\boldsymbol{\theta}_t^{\text{adv}}) \cdot \frac{\nabla \hat{L}_{W_t}(\boldsymbol{\theta}_t^{\text{adv}})}{\|\nabla \hat{L}_{W_t}(\boldsymbol{\theta}_t^{\text{adv}})\| + \xi}$ ▷ Calculate the norm gradient $\nabla R_\rho^{(1)}(\boldsymbol{\theta}_t)$
  $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \eta_t(\boldsymbol{h}_t^{\text{loss}} + \alpha \boldsymbol{h}_t^{\text{norm}})$
  $t \leftarrow t + 1$
end while
return $\boldsymbol{\theta}_t$

GAM : gradient norm aware minimization

- The properties of GAM

Hessian eigenvalue

Proposition 2.1. *Let $\boldsymbol{\theta}^*$ be a local minimum of $\hat{L}$. Suppose $\hat{L}$ can be second order Taylor approximated in the neighbourhood $B(\boldsymbol{\theta}^*, \rho)$, i.e., $\forall \boldsymbol{\theta} \in B(\boldsymbol{\theta}^*, \rho)$, $\hat{L}(\boldsymbol{\theta}) = \hat{L}(\boldsymbol{\theta}^*) + (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla^2 \hat{L}(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*)/2$. Then*

$$\lambda_{\max}\left(\nabla^2 \hat{L}(\boldsymbol{\theta}^*)\right) = \frac{R_\rho^{GNR}(\boldsymbol{\theta}^*)}{\rho^2}. \quad (2)$$

Generalization error bound

$$\mathbb{E}_{\epsilon_i \sim N(0, \rho^2/(\sqrt{d}+\sqrt{\log n})^2)}[L(\boldsymbol{\theta}+\boldsymbol{\epsilon})] \le \hat{L}(\boldsymbol{\theta}) + R_\rho^{GNR}(\boldsymbol{\theta}) + \sqrt{\frac{\frac{1}{4} d \log\left(1 + \frac{\|\boldsymbol{\theta}\|^2\left(\sqrt{d}+\sqrt{\log n}\right)^2}{d\rho^2}\right) + \frac{1}{4} + \log\frac{n}{\delta} + 2\log(6n+3d)}{n-1}} + \frac{M}{\sqrt{n}}.$$

Convergence

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\left\|\nabla L^{overall}(\boldsymbol{\theta}_t)\right\|^2\right] \le \frac{C_1 + C_2 \log T}{\sqrt{T}},$$

GAM : gradient norm aware minimization

- Experimental results on CIFAR-10 and CIFAR-100

Table 1. Results of GAM with state-of-the-art models on CIFAR-10 and CIFAR-100. The best results are highlighted in bold font.

		CIFAR-10				CIFAR-100			
Model	Aug	SGD	SGD + GAM	SAM	SAM + GAM	SGD	SGD + GAM	SAM	SAM + GAM
ResNet18	Basic	$95.32_{\pm0.13}$	**96.17**$_{\pm0.21}$	$96.10_{\pm0.20}$	**96.75**$_{\pm0.18}$	$78.32_{\pm0.32}$	**79.53**$_{\pm0.30}$	$79.27_{\pm0.16}$	**80.45**$_{\pm0.25}$
ResNet18	Cutout	$95.99_{\pm0.13}$	**96.46**$_{\pm0.20}$	$96.64_{\pm0.13}$	**96.99**$_{\pm0.23}$	$78.73_{\pm0.13}$	**79.89**$_{\pm0.31}$	$79.43_{\pm0.15}$	**80.80**$_{\pm0.14}$
ResNet18	RA	$96.07_{\pm0.07}$	**96.52**$_{\pm0.09}$	$96.64_{\pm0.17}$	**97.06**$_{\pm0.13}$	$78.62_{\pm0.32}$	**79.82**$_{\pm0.24}$	$79.71_{\pm0.15}$	**80.97**$_{\pm0.29}$
ResNet18	AA	$96.13_{\pm0.05}$	**96.71**$_{\pm0.07}$	$96.75_{\pm0.08}$	**97.17**$_{\pm0.08}$	$78.88_{\pm0.15}$	**80.56**$_{\pm0.21}$	$80.58_{\pm0.25}$	**81.59**$_{\pm0.24}$
ResNet101	Basic	$96.35_{\pm0.08}$	**96.98**$_{\pm0.11}$	$96.82_{\pm0.16}$	**97.20**$_{\pm0.15}$	$80.47_{\pm0.13}$	**82.21**$_{\pm0.40}$	$82.03_{\pm0.12}$	**83.13**$_{\pm0.07}$
ResNet101	Cutout	$96.56_{\pm0.18}$	**97.22**$_{\pm0.05}$	$97.07_{\pm0.08}$	**97.36**$_{\pm0.24}$	$80.53_{\pm0.30}$	**82.36**$_{\pm0.24}$	$81.60_{\pm0.35}$	**83.40**$_{\pm0.13}$
ResNet101	RA	$96.68_{\pm0.25}$	**97.33**$_{\pm0.30}$	$97.12_{\pm0.18}$	**97.40**$_{\pm0.23}$	$80.60_{\pm0.28}$	**82.40**$_{\pm0.31}$	$82.19_{\pm0.34}$	**83.28**$_{\pm0.20}$
ResNet101	AA	$96.78_{\pm0.14}$	**97.39**$_{\pm0.18}$	$97.18_{\pm0.11}$	**97.42**$_{\pm0.1}$	$81.83_{\pm0.37}$	**83.19**$_{\pm0.15}$	$82.44_{\pm0.47}$	**83.94**$_{\pm0.23}$
WRN28_2	Basic	$94.82_{\pm0.07}$	**95.69**$_{\pm0.13}$	$95.47_{\pm0.08}$	**95.85**$_{\pm0.08}$	$75.45_{\pm0.25}$	**77.21**$_{\pm0.31}$	$77.04_{\pm0.18}$	**77.69**$_{\pm0.20}$
WRN28_2	Cutout	$95.70_{\pm0.20}$	**96.41**$_{\pm0.18}$	$96.22_{\pm0.13}$	**96.39**$_{\pm0.22}$	$76.80_{\pm0.45}$	**78.58**$_{\pm0.24}$	$78.04_{\pm0.43}$	**79.33**$_{\pm0.12}$
WRN28_2	RA	$95.75_{\pm0.16}$	**96.35**$_{\pm0.13}$	$96.22_{\pm0.08}$	**96.49**$_{\pm0.20}$	$76.73_{\pm0.27}$	**78.66**$_{\pm0.03}$	$77.88_{\pm0.29}$	**78.96**$_{\pm0.13}$
WRN28_2	AA	$95.44_{\pm0.06}$	**95.98**$_{\pm0.09}$	$96.07_{\pm0.08}$	**96.44**$_{\pm0.09}$	$77.35_{\pm0.02}$	**79.05**$_{\pm0.10}$	$78.64_{\pm0.23}$	**79.50**$_{\pm0.21}$
WRN28_10	Basic	$95.73_{\pm0.10}$	**96.61**$_{\pm0.15}$	$96.78_{\pm0.80}$	**97.29**$_{\pm0.11}$	$81.40_{\pm0.13}$	**83.45**$_{\pm0.09}$	$83.41_{\pm0.04}$	**84.31**$_{\pm0.06}$
WRN28_10	Cutout	$96.74_{\pm0.03}$	**96.97**$_{\pm0.05}$	$97.35_{\pm0.16}$	**97.56**$_{\pm0.12}$	$81.53_{\pm0.40}$	**83.69**$_{\pm0.08}$	$82.38_{\pm0.15}$	**84.43**$_{\pm0.13}$
WRN28_10	RA	**97.14**$_{\pm0.04}$	$96.83_{\pm0.03}$	**97.58**$_{\pm0.07}$	$97.49_{\pm0.03}$	$81.65_{\pm0.18}$	**83.84**$_{\pm0.09}$	$82.79_{\pm0.06}$	**84.68**$_{\pm0.13}$
WRN28_10	AA	$96.93_{\pm0.12}$	**97.05**$_{\pm0.04}$	$97.48_{\pm0.06}$	**97.67**$_{\pm0.08}$	$81.99_{\pm0.11}$	**84.02**$_{\pm0.18}$	$83.84_{\pm0.30}$	**84.81**$_{\pm0.21}$
PyramidNet110	Basic	$96.19_{\pm0.11}$	**97.11**$_{\pm0.14}$	$97.26_{\pm0.05}$	**97.51**$_{\pm0.09}$	$82.74_{\pm0.12}$	**84.91**$_{\pm0.09}$	$85.01_{\pm0.09}$	**85.25**$_{\pm0.06}$
PyramidNet110	Cutout	$96.82_{\pm0.09}$	**97.32**$_{\pm0.21}$	$97.49_{\pm0.06}$	**97.91**$_{\pm0.14}$	$83.31_{\pm0.21}$	**85.20**$_{\pm0.19}$	$84.90_{\pm0.03}$	**85.46**$_{\pm0.10}$
PyramidNet110	RA	$97.15_{\pm0.21}$	**97.80**$_{\pm0.22}$	$97.60_{\pm0.09}$	**98.01**$_{\pm0.10}$	$84.04_{\pm0.19}$	**86.47**$_{\pm0.14}$	$85.33_{\pm0.27}$	**85.64**$_{\pm0.20}$
PyramidNet110	AA	$97.11_{\pm0.01}$	**97.85**$_{\pm0.02}$	$97.61_{\pm0.14}$	**97.95**$_{\pm0.10}$	$84.48_{\pm0.03}$	**85.92**$_{\pm0.03}$	$85.69_{\pm0.17}$	**86.35**$_{\pm0.18}$

GAM : gradient norm aware minimization

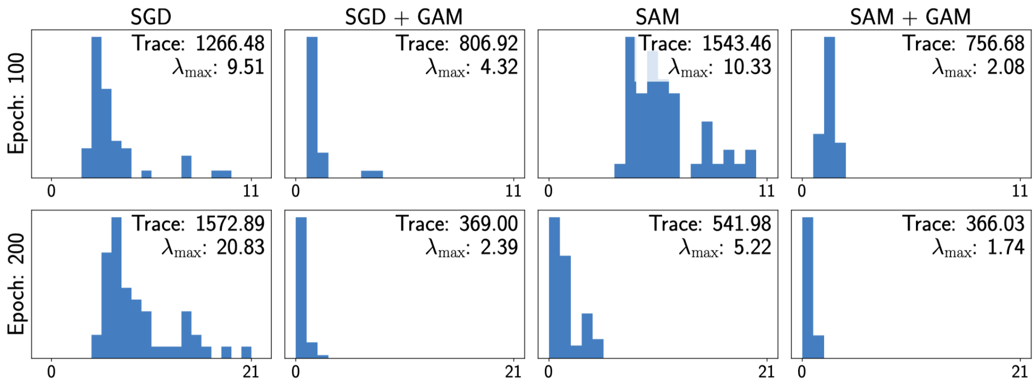
- Experimental results on ImageNet and transfer learning

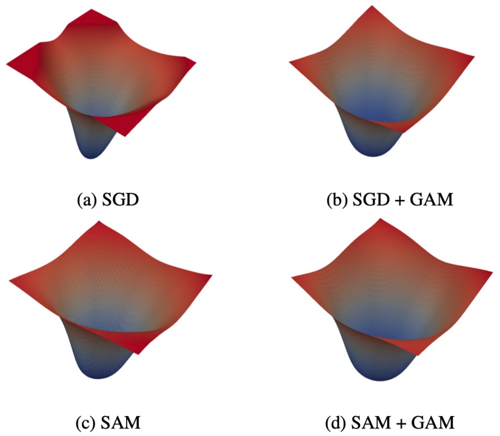
Model	Dataset	Base Opt	Base + GAM	SAM	SAM + GAM
ResNet50	Top-1	$76.01_{\pm 0.19}$	$\mathbf{76.59}_{\pm 0.15}$	$76.47_{\pm 0.11}$	$\mathbf{76.86}_{\pm 0.15}$
ResNet50	Top-5	$92.75_{\pm 0.08}$	$\mathbf{93.10}_{\pm 0.08}$	$93.07_{\pm 0.05}$	$\mathbf{93.22}_{\pm 0.06}$
ResNet101	Top-1	$77.69_{\pm 0.08}$	$\mathbf{78.45}_{\pm 0.10}$	$78.35_{\pm 0.12}$	$\mathbf{78.70}_{\pm 0.12}$
ResNet101	Top-5	$93.76_{\pm 0.09}$	$\mathbf{94.09}_{\pm 0.12}$	$94.02_{\pm 0.06}$	$\mathbf{94.15}_{\pm 0.12}$
ViT-S/32	Top-1	$68.26_{\pm 0.22}$	$\mathbf{69.95}_{\pm 0.16}$	$69.73_{\pm 0.05}$	$\mathbf{70.15}_{\pm 0.18}$
ViT-S/32	Top-5	$87.39_{\pm 0.19}$	$\mathbf{88.11}_{\pm 0.26}$	$87.91_{\pm 0.30}$	$\mathbf{88.23}_{\pm 0.18}$
ViT-B/32	Top-1	$71.15_{\pm 0.14}$	$\mathbf{73.58}_{\pm 0.06}$	$73.10_{\pm 0.18}$	$\mathbf{73.70}_{\pm 0.10}$
ViT-B/32	Top-5	$90.12_{\pm 0.07}$	$\mathbf{91.15}_{\pm 0.19}$	$91.03_{\pm 0.06}$	$\mathbf{91.50}_{\pm 0.16}$

	EfficientNet-b0				Swin-t			
Dataset	SGD	SGD + GAM	SAM	SAM + GAM	AdamW	AdamW + GAM	SAM	SAM + GAM
Stanford Cars	82.14	**83.50**	83.21	**83.98**	83.50	**84.90**	83.55	**85.29**
CIFAR-10	86.26	**87.37**	86.95	**87.97**	91.32	**92.06**	91.77	**92.55**
CIFAR-100	63.75	**64.85**	64.29	**65.03**	72.88	**73.78**	73.99	**74.30**
Oxford_IIIT_Pets	91.03	**91.80**	91.65	**91.96**	93.49	**93.87**	93.59	**94.03**
Food101	82.54	**82.69**	82.57	**83.01**	86.38	**86.89**	86.64	**87.03**

GAM : gradient norm aware minimization

- GAM Hessian spectrum and visualization





Thanks！

Xingxuan Zhang[†], Renzhe Xu[†], Han Yu, Hao zou, Peng Cui[*]. Gradient Norm Aware Minimization Seeks First-Order Flatness and Improves Generalization. *CVPR, 2023, Highlight.*

Github: https://github.com/xxgege/GAM