On the Interplay of Curvature, Calibration and Out-of-Distribution Generalization: Insights from SAM and Focal Loss Analyses

ICCV23
Workshop on Uncertainty Quantification for Computer Vision

Hiroki Naganuma*[1,2], Masanari Kimura*[3] (equal contribution)
*: naganuma.hiroki@mila.quebec / masanari.kimura@zozo.com
[1] Mila [2] Université de Montréal [3] ZOZO RESEARCH

Introduction

Motivation and Background

- **Out-of-distribution (OOD) Generalization**
  - **In real-world applications, it is often the case that the test data obey a distribution different from the training data.**
  - Distributional shift violates the typical IID assumption for training.

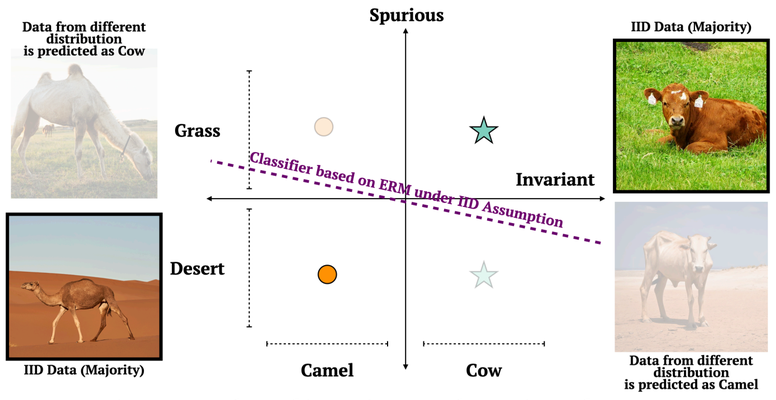

Fig: Examples of invariant and spurious features

- **Calibration (Measure of Uncertainty Reliablity)**
  - Modern DNNs (e.g., ResNet) **tend to be under/over-confidence, which is problematic because of the poor reliability of uncertainty** (which is measured by ECE(Expected Calibration Error)).
  - In applications such as medical imaging, DNNs cannot be used in decision making if the uncertainty reliability of the model is not high enough, such as in cases of overconfidence.

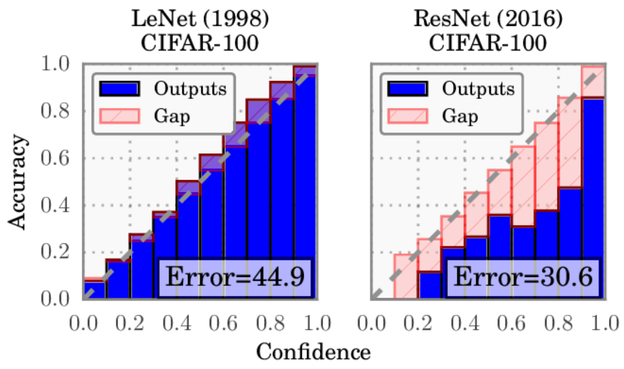

Fig: Reliability diagram (taken from [Chuan et al, 2017])

Definition 1 (Expected calibration error). For $(\boldsymbol{x}, y)$, let $\hat{y} = \arg\max p(y|\boldsymbol{x}; \boldsymbol{\theta})$. The expected calibration error is defined as

$$\mathrm{ECE}(\boldsymbol{\theta}) := \mathbb{E}\left[\left|\mathbb{P}\left(\hat{y} = y \middle| p(y|\boldsymbol{x}; \boldsymbol{\theta}) = s\right) - s\right|\right], \quad (1)$$

where s is the prediction probability.

Research Question

- What factors improve out-of-distribution generalization and calibration?

Hypothesis

- **Curvature is the factor that improves out-of-distribution generalization and calibration?**

Contribution

- We **generalize the behavior of algorithms** known to improve calibration and out-of-distribution generalization, respectively, in terms of curvature (Theorem 3).
- Our comparative experiments between Focal Loss and SAM investigated the presence or absence of curvature improvement and how this influenced the resolution of the tradeoff problem. **We demonstrated that curvature significantly contributes to improvements in OOD generalization and calibration.**

Curvature and Contractions

We introduce the concept of curvature and the contractions we use to test our hypothesis.

- On a Riemannian manifold, there's a unique torsion-free, metric connection ∇ known as the Levi-Civita connection.
- The curvature tensor related to the Levi-Civita connection is named the Riemannian curvature tensor, as a polynomial of Γ_{ij}^{l}

Hessian

Definition 4. The Hessian H^f of $f \in \mathcal{F}(\mathcal{M})$ with respect to the linear connection ∇ is the covariant derivative of the 1-form df as $H^f(X, Y) := (\nabla df)(X, Y)$.

We also have the explicit formula as $H^f(X, Y) = Xdf(Y) - df(\nabla_X Y) = XY(f) - (\nabla_X Y)f$, and the local coordinates representation is

$$H_{ij}^f = \frac{\partial^2 f}{\partial\theta^i \partial\theta^j} - \Gamma_{ij}^k \frac{\partial f}{\partial\theta^k}. \quad (5)$$

Spectral Radius

Definition 6 (Spectral radius)**.** Let $\lambda_1, \ldots, \lambda_m$ be the eigenvalues of Hessian H^f for some function $f \in \mathcal{F}(\mathcal{M})$. The spectral radius of H^f is defined as $A(H^f) := \max_{i\in[0,m]}\{|\lambda_1|, \ldots, |\lambda_m|\}$, and $A(H^f) = \lim_{k\to\infty} \|H^{fk}\|^{1/k}$.

Laplacian

Definition 7 (Laplacian)**.** For every linear connection ∇, any metric $\mathfrak{g}$ and C^2-function f, Laplacian operator Δf is defined as $\Delta f = div(grad\ f)$, where div is the divergence of vector field and $grad$ is the gradient vector field of f.

Algorithms Subjected in Our Analysis

We target two popular and standard algorithms that have been used for calibration and out-of-distribution generalization respectively.

Focal Loss

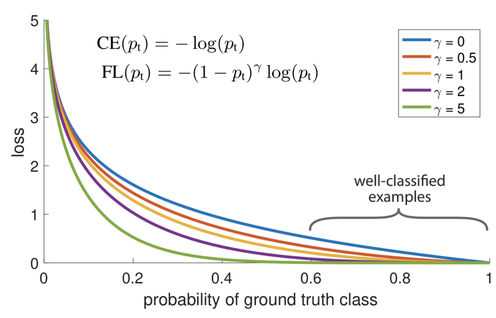

Fig: Focal Loss and Cross Entropy Loss (taken from [TY Lin et al, 2017])

Sharpness Aware Mnimization

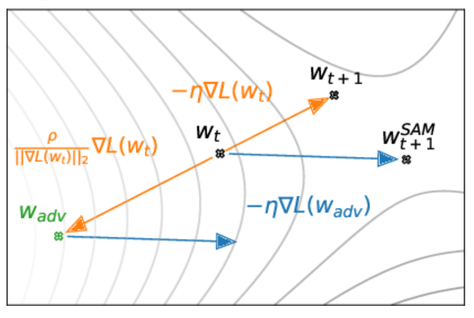

Fig: Schematic of the SAM parameter update (taken from [P Foret et al, 2017])

Theoretical Results

Theorem 3 (Curvature contractions)**.** For some loss function $f \in \mathcal{F}(\mathcal{M})$, we have the following.

i) Full-batch SAM behaves as the regularizer of spectral radius of H^f.

ii) 1-batch SAM and Focal Loss behave as the regularizers of the Laplacian of f.

Corollary 1. If loss landscape is locally convex around some point p, that is, H^f is positive semidefinite, then the regularization term of full-batch SAM is upper bounded by the regularization term of 1-batch SAM with some scaling factor $c \in \mathbb{R}$, as $A(H^f) \leq c\Delta f$.

Corollary 2. If the loss landscape has locally concave directions at some point p, full-batch SAM is more robust than 1-batch SAM or Focal Loss regularization.

Experiments

Tasks (Datasets)

- CIFAR10 -> CIFAR10.1
- ImageNet1K->ImageNetV2
- DomainBed
  - PACS
  - VLCS
  - OfficeHome
  - DomainNet

→ Each data point in the scatter plot represents an experiment with different ϱ and γ values. The trend suggests that **while Focal Loss seems to sacrifice ECE to improve ACC, the application of SAM appears to address this trade-off.**

Comparison Between Focal Loss and Sharpness Aware Minimization

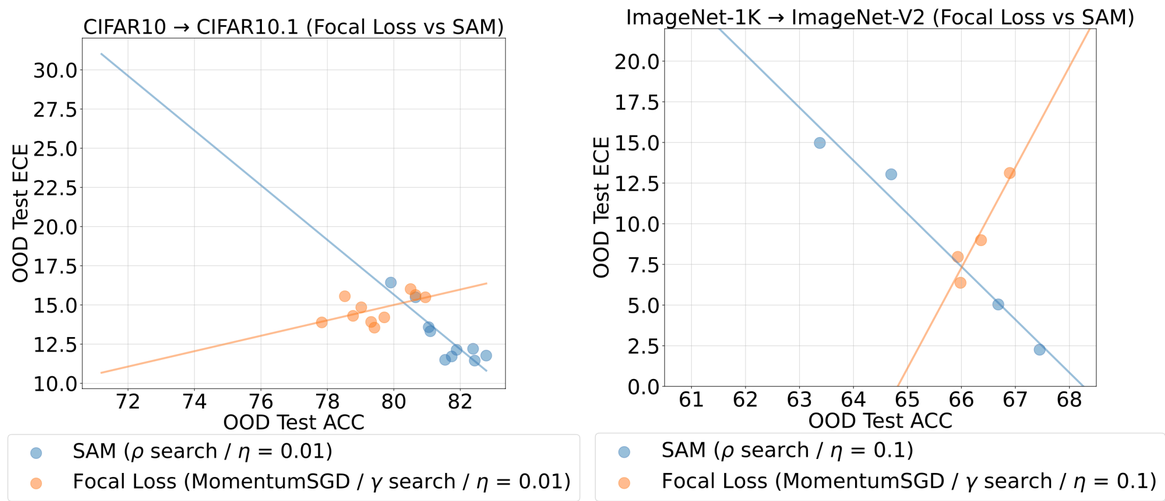

Fig: Trade-off between ECE and Accuracy (ACC) observed with Focal Loss and its mitigation using SAM in CIFAR10.1 and ImageNet-V2.

Curvature Analysis for Focal Loss and Sharpness Aware Minimization

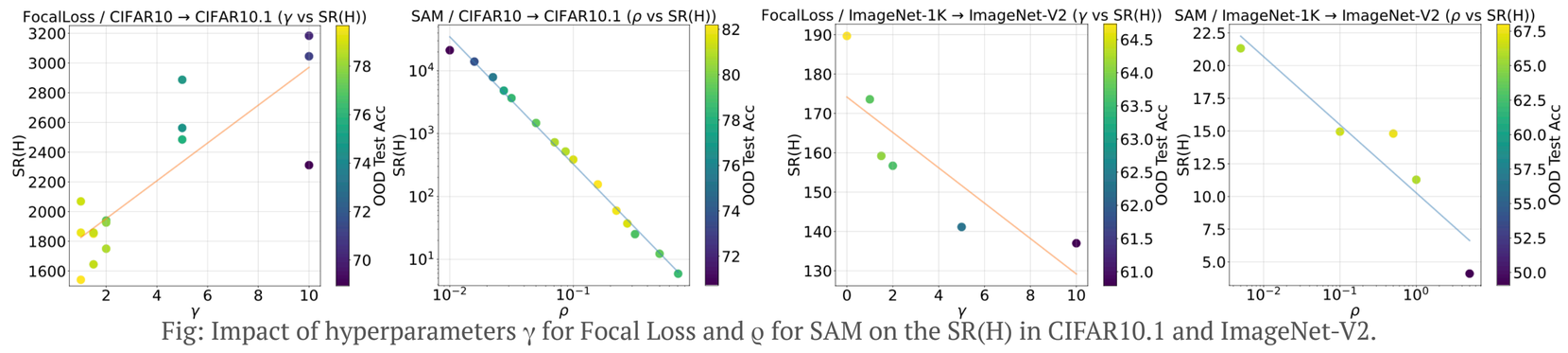

Fig: Impact of hyperparameters γ for Focal Loss and ϱ for SAM on the SR(H) in CIFAR10.1 and ImageNet-V2.

↑ As γ and ϱ increase, SR(H) is generally reduced, with **SAM typically achieving lower SR(H) values than Focal Loss**. However, while a larger γ in Focal Loss results in a lower SR(H) in ImageNet-V2, it may lead to accuracy deterioration if increased excessively. With SAM, an optimal ϱ is observed to yield the best out-of-distribution generalization.

Batch Size Effect for Sharpness Aware Minimization

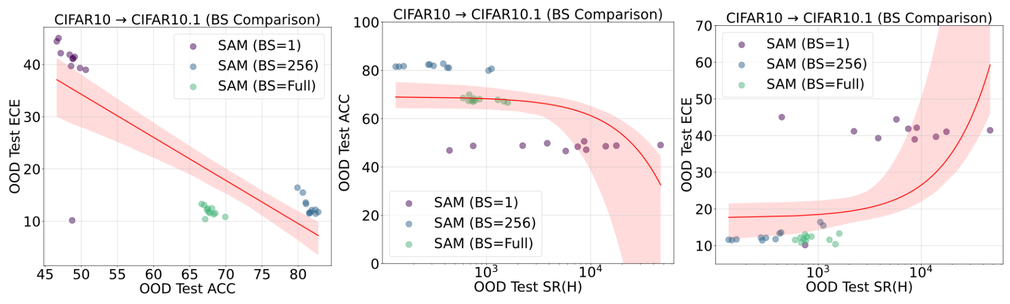

Fig: Effects of batch size variations on the behavior of SAM in the CIFAR10.1 task. Each color represents an experiment with a different batch size.

← With an appropriate batch size (BS=256), both ACC and ECE performance is good. BS=Full performs well only in terms of ECE, while BS=1 results in suboptimal performance for both indicators. This results implies that **there can be instances of suboptimal learning resulting in low validation accuracy yet possessing small curvature.**

References

[Chuan et al, 2017] Chuan Guo,Geoff Pleiss,Yu Sun,and Kilian Q Weinberger. On calibration of modern neural networks. In International conference on machine learning, pages 1321–1330. PMLR, 2017.
[TY Lin et al, 2017] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Doll´ar. Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision, pages 2980–2988, 2017
[P Foret et al, 2017] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020