



|              |                                                                               |
|--------------|-------------------------------------------------------------------------------|
| Title        | 汎化指標の網羅的評価の再考 : 現代ニューラルネットワークの課題への適応                                          |
| Author(s)    | 長沼, 大樹                                                                        |
| Citation     | サイバーメディアHPCジャーナル. 2026, 16, p. 52-55                                          |
| Version Type | VoR                                                                           |
| URL          | <a href="https://doi.org/10.18910/106234">https://doi.org/10.18910/106234</a> |
| rights       |                                                                               |
| Note         |                                                                               |

*The University of Osaka Institutional Knowledge Archive : OUKA*

<https://ir.library.osaka-u.ac.jp/>

The University of Osaka

# 汎化指標の網羅的評価の再考： 現代ニューラルネットワークの課題への適応

*Fantastic Generalization Measures Reloaded: Adapting to Modern Neural Network Challenges*

長沼 大樹

モントリオール大学 コンピュータ科学専攻 / Mila - Quebec AI Institute

## 1. はじめに

深層ニューラルネットワークは画像認識や自然言語処理など多岐にわたる分野で目覚ましい成果を上げているが、未知データに対する汎化性能を事前に予測する手段は依然として大きな課題である[1]。学習時に得られる情報のみからテスト時性能を高い信頼性で見積もる手法は、実応用の安全性・信頼性確保の観点から強く求められる。

Jiang らは 2020 年、CIFAR-10/SVHN を用いて 40 種類の汎化指標を網羅的にベンチマークし、PAC-Bayes 境界や Sharpness 系指標が良好な相関を示す一方、古典的なノルムベース指標は必ずしも汎化性能を反映しないことを報告した[2]。しかし、それから約 6 年が経過した現在、Transformer や基盤モデルの登場をはじめとする大きなパラダイムシフトを経験しているにも関わらず、現代的な要求に即した網羅的再評価は十分に行われてこなかった。

本研究では、大阪大学 D3 センターの大規模計算機システム「SQUID」を活用し、(i)分布外 (Out-of-Distribution、以下 OOD) 汎化、および(ii)信頼性校正 (Confidence Calibration) の二つの観点を中心に、43 種類の汎化指標の網羅的再評価を実施した。1 万通り以上のハイパーパラメータ設定でモデルを学習し、現代ニューラルネットワークに適した汎化指標群を体系的に明らかにすることを目指した。

## 2. 研究背景と問題設定

深層学習モデルの汎化を説明するために、VC 次元やパラメータ数に基づく古典的容量論[3]、PAC-Bayes 理論、ノルム・マージン基盤の複雑度評価、損失曲面の平坦性を捉える Sharpness 系指標など、

多くの理論的枠組みが提案されてきた。これらは原理的に IID (Independent and Identically Distributed) 設定下で導出されており、実応用で頻発する分布シフト下での挙動については体系的検証が乏しい。

近年では予測の信頼度と精度の整合性を測るキャリブレーション指標が安全性重視の応用場面で注目を集めており[4]、IID 環境下で良好にキャリブレートされたモデルが分布シフトに対する頑健性も高い傾向が指摘されている。しかし、これらを「汎化性能の予測子」として大規模に評価した研究は乏しい。本研究は、Jiang らの枠組み[2]を踏襲しつつ、(i)OOD 設定への評価範囲の拡張、(ii)キャリブレーション指標および情報量規準の追加、を二大改良点として位置づけ、現代の実応用要求を反映した汎化指標の網羅的再評価を行うものである。

## 3. 計算手法と汎化指標

### 3.1 評価対象の汎化指標

本研究では学習済みモデルと訓練データのみから算出可能な 43 種類の汎化指標を以下の 6 カテゴリに整理し、同一実装・同一評価条件で一括測定した。すなわち、(a)Baseline & Output-based (VC 次元、パラメータ数、出力エントロピー等)、(b)Norm & Margin-based ( $\ell_2/\ell_1$  ノルム、スペクトルノルム、Path-norm、Frobenius 距離等)、(c)Sharpness-based (SAM 型摂動損失、Hessian 上位固有値、PAC-Bayes 境界等)、(d)Optimization-based (勾配ノイズ分散、勾配ノルム、入力勾配ノルム等)、(e)Information Criteria (AIC、WAIC、TIC)、(f)Calibration & Confidence (ECE、MCE、ACE、温度スケール誤差等) の 6 種である。

このうち(e)と(f)は、Jiang らの先行研究には含まれておらず、本研究の独自拡張である。情報量規準は対数尤度と複雑度補正を組み合わせる古典的枠組みを深層学習向けに適応した実装を採用し、キャリブレーション指標は予測確率分布と経験的精度の整合度を多様な統計量で評価した。

### 3.2 学習設定とデータセット

学習には、CIFAR-10[5]に加え、OOD 評価用として CIFAR-10-C および CIFAR-10-P[6]を採用し、さらに DomainBed の PACS、VLCS ベンチマークを用いてドメインシフト環境を構成した。モデルアーキテクチャは、CIFAR-10 系で 3 層 CNN (SimpleCNN)、ResNetV2-32、Network in Network (NiN) の 3 種を、DomainBed 系で ImageNet-1k 事前学習済み ResNet-50+線形分類器を用いた。

ハイパーパラメータは、最適化手法 (SGD/RMSProp/Adam)、学習率、weight decay、バッチサイズ、Dropout 率、ランダムシードの各軸に対してグリッド探索を行い、合計で 1 万通り以上の学習条件を構成した。本規模の網羅的探索を実現できたのは、まさに SQUID の大規模並列計算リソースの恩恵である。

### 3.3 評価指標と相関分析手法

汎化ギャップの予測精度は、Kendall の順位相関係数  $\tau$  を用いて評価した。各汎化指標  $\mu$  と評価対象の汎化ギャップ  $g$  について、ハイパーパラメータ軸  $a$  以外を固定した固定条件グループ  $S$  内で  $\tau$  を計算し、有効シード集合での平均をとってグループ代表値  $\tau^-(S)$  とした。さらに、軸  $a$  に対する固定条件グループ集合  $S_a$  にわたる平均

$$\tau_{\text{mean}}(a) = (1/|S_a|) \sum \tau^-(S)$$

を算出することで、各軸変動に対する汎化指標の順位整合性を体系的に可視化した。なお、IID 汎化ギャップ (GenGap\_CIFAR10) に加え、OOD 評価指標として訓練精度-CIFAR-10-C/P テスト精度の差 (OODGenGap\_C/P)、ならびに IID テスト精度-OOD テスト精度の差 (OODTestGap\_C/P) の双方を採用し、IID と OOD の両側面からの予測力を同時に検証できる枠組みとした。

## 4. SQUID 上での大規模並列実験

本研究の中核をなす 1 万通り以上の学習試行は、大阪大学 D3 センターのスーパーコンピューター SQUID (汎用 CPU ノード群および GPU ノード群) の公募型利用枠を用いて実施した。1 学習試行あたり、CIFAR-10 の 100 エポック学習および 43 指標の同一バッチでの一括計測を行うため、典型的に数 GB 規模のチェックポイント・勾配統計を保持しながら計算する必要がある。これを単一マシンで網羅的に実行することは現実的でなく、SQUID の数百ノード規模の並列実行能力が不可欠であった。

実験のスケジューリング面では、ハイパーパラメータ集合をジョブ配列化し、ノード割り当てを動的に管理する独自パイプラインを構築した。これにより、計算資源を最大限効率的に利用しつつ、再現性を担保するためのシード固定・チェックポイントの世代管理・指標計算の冪等化を実現した。最終的に得られた実験データはモデル毎におよそ 43 指標×6 個以上の評価対象 (GenGap\_CIFAR10、OODGenGap\_C、OODGenGap\_P、OODTestGap\_C、OODTestGap\_P、ドメインシフト指標等) の組み合わせをもち、解析に供する規模となった。

また、Sharpness 系や Hessian ベースの指標は二階微分情報を要するため計算コストが大きい。Hutchinson 推定や少数 Lanczos 反復、ノード内 GPU 並列とジョブ並列の二段並列化により、現実的な所要時間で網羅測定を可能にした。SQUID の大容量メモリと高速インターコネクトが、計算量の重い指標群をスコープに含められた決定的要因である。

## 5. 結果と考察

### 5.1 カテゴリ別の IID/OOD 相関の俯瞰

図 1 に、各汎化指標の IID 汎化ギャップに対する Granulated Kendall  $\tau$  (横軸) と OOD 汎化ギャップに対する  $\tau$  (縦軸) の散布を、汎化指標 6 カテゴリ別に示す。各点は単一の汎化指標とハイパーパラメータ軸の組であり、原点を中心とした象限が IID/OOD の順位整合性の符号関係を表す。



### 5.3 IID と OOD で異なる有効性

weight decay 軸の検討では、Information Criteria、Calibration、Optimization-based の複指数指標が OOD 汎化ギャップに対して正の  $\tau_{\text{mean}}$  を示す一方、IID 側に対しては負の  $\tau_{\text{mean}}$  を示す領域が観測された。同一の指標であっても IID/OOD のいずれを評価対象とするかで順位整合性の符号が反転しうることが明確となった。これは、IID 環境で良好なモデル選択指標が必ずしも OOD 汎化を予測するとは限らないという実用上の警鐘であると同時に、IID 単独では予測力に乏しいと見なされてきた Information Criteria や Calibration 系指標が OOD 設定下では有効な信号源となりうることを示している。

### 5.4 普遍的に予測力を持つ単一指標は存在しない

全体として、Sharpness 系指標は比較的安定した予測力を保持する一方、IID と OOD のいずれの評価対象、いずれのアーキテクチャにおいても一貫して高い順位整合性を示す「普遍的な単一指標」は確認されなかった。本結果は、特定の汎化指標のみに依拠したモデル選択は本質的にリスクを伴うこと、および評価対象の分布特性・モデル系列に応じて指標群を組み合わせる必要があることを強く示唆する。

DomainBed 系の PACS/VLCS でも、IID 試験ドメインでの有効指標と未観測ターゲットドメインへの汎化の有効指標が必ずしも一致せず、IID モデル選択の限界[7]と整合する結果が得られた。

## 6. おわりに

本研究では、SQUID の大規模計算リソースを活用し、CIFAR-10/CIFAR-10-C/CIFAR-10-P および DomainBed 系ベンチマーク上で、43 種類の汎化指標を 1 万通り以上のハイパーパラメータ条件で網羅評価した。その結果、(1) IID と OOD で有効性が反転する指標群が体系的に存在すること、(2) Information Criteria および Calibration 系指標が OOD 設定下で新たな信号源となりうること、

(3) 単一指標による普遍的なモデル選択は困難であることを明らかにした。

本成果の一部は情報処理学会全国大会 (IPJSJ 2026) [9]においてすでに口頭発表を行い、包括版を論文[8]として国際会議 ICML 2026 に投稿中である。「OOD 設定下で初めて有効性を示す Calibration・Information Criteria 系指標」という知見は、安全性が求められる医療・自動運転等の応用における信頼性評価に直接資するものと考ええる。今後は分布シフトの種類および severity の拡張、Transformer 系・基盤モデル系への適用、複合的モデル選択基準の確立を進めていきたい。

末筆ながら、大規模計算機システム SQUID の利用機会を提供してくださった大阪大学 D3 センターおよび関係各位、ならびに共同研究者の皆様に深く感謝申し上げる。

## 参考文献

- (1) C. Zhang, et al., Understanding deep learning requires rethinking generalization, ICLR, 2017.
- (2) Y. Jiang, et al., Fantastic Generalization Measures and Where to Find Them, ICLR, 2020.
- (3) V. N. Vapnik, The Nature of Statistical Learning Theory, Springer, 1995.
- (4) C. Guo, et al., On Calibration of Modern Neural Networks, ICML, 2017.
- (5) A. Krizhevsky, Learning Multiple Layers of Features from Tiny Images, Tech. Rep., Univ. of Toronto, 2009.
- (6) D. Hendrycks, T. Dietterich, Benchmarking Neural Network Robustness to Common Corruptions and Perturbations, ICLR, 2019.
- (7) I. Gulrajani, D. Lopez-Paz, In Search of Lost Domain Generalization, ICLR, 2021.
- (8) S. Nakai, Y. Fadhoun, K. Mathlouthi, K. Yoshida, G. Talluri, I. Mitliagkas, H. Naganuma, Revisiting Generalization Measures Beyond IID: An Empirical Study under Distributional Shift, ICML, 2026 (投稿中).
- (9) 中井 蒼頼, 吉田 晃太朗, 長沼 大樹, 深層学習における広義の汎化指標の特性分析, 情報処理学会全国大会 (IPJSJ 2026), 2026.