汎化改善のためのモデル診断

寺内 怜央[1], 佐藤 育郎[1,2], 川上 玲[1]
[1]東京工業大学, [2]デンソーITラボラトリ

Recognition and Learning Algorithm Laboratory

DENSO IT LAB

東京工業大学 Tokyo Institute of Technology

導入

背景 従来のモデル開発では，目標性能に到達するまでモデルの学習・評価を繰り返す必要があった.

目的
- 汎化に関する診断指標を導入することで，モデルの汎化改善の余地を明らかにする.
- 診断指標ならびに汎化性能を事後的に改善できる事後学習方法を開発する.

従来手法

SAM: Sharpness-Aware Minimization [P.Foret, et al., ICLR 2021]

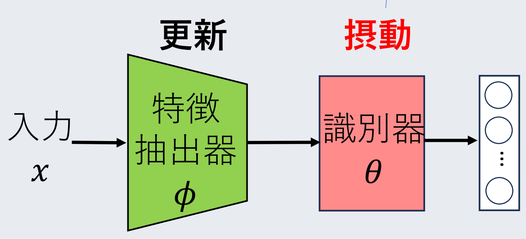
モデル内の全パラメタ空間における超球内部の最悪損失を最小化．局所解近傍の平坦性との関連が示されている.

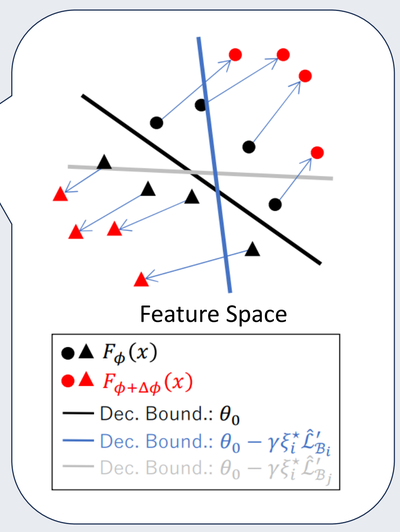
$$\Delta w = -\eta \left.\frac{\partial \mathcal{L}(w)}{\partial w}\right|_{w+\epsilon^*(w)}, \epsilon^*(w) = \arg\max_{\|\epsilon\|_p \le \rho} \mathcal{L}(w+\epsilon)$$

PoF: Post-training of Feature extractor [I. Sato, et al., ICML 2022]

バッチ勾配方向に摂動させた弱識別器に対して識別が行えるように特徴抽出器を学習する.

$$\Delta\phi = -\eta \frac{\partial \mathcal{L}_{\tilde{\mathcal{B}}}(\phi, \theta_0 - \gamma \xi^\star_{\mathcal{B}} \hat{\mathcal{L}}'_{\mathcal{B}})}{\partial \phi}$$





提案手法

モデル診断指標

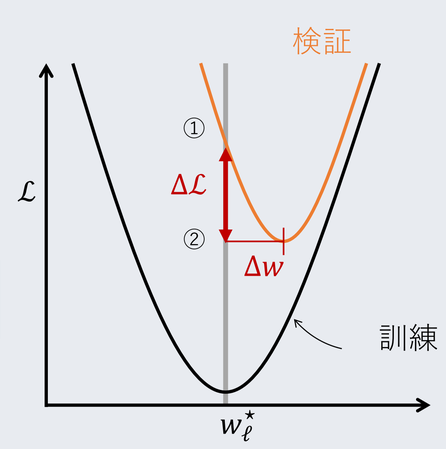
一般に，深層モデルは高層ほど過適合（学習・検証間の特徴分布のずれ）が生じやすい．ずれの一指標として$\Delta\mathcal{L}_\ell$を導入.

損失改善の余地：$\Delta\mathcal{L}_\ell$

学習・検証のデータ分布のずれから生じる**曲率に起因した検証損失の増分**：
①局所解における検証損失　と
②検証損失の最小値　　の差
を診断指標 $\Delta\mathcal{L}_\ell$ とする.

$$\Delta\mathcal{L}_\ell \equiv \mathcal{L}^V(w^*_\ell) - \min_{\Delta w} \mathcal{L}^V(w^*_\ell + \Delta w)$$

$\mathcal{L}^V$： 検証損失
w^*： 局所解



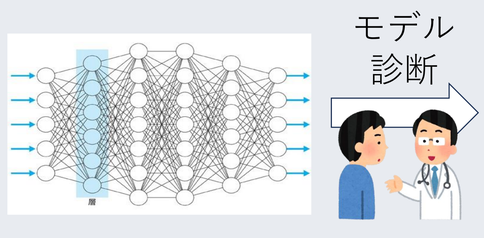
曲率指標：κ_ℓ

検証損失の変化分に対して，損失を最小化できるパラメタまでのパラメタの変動分との比をκ_ℓとする.

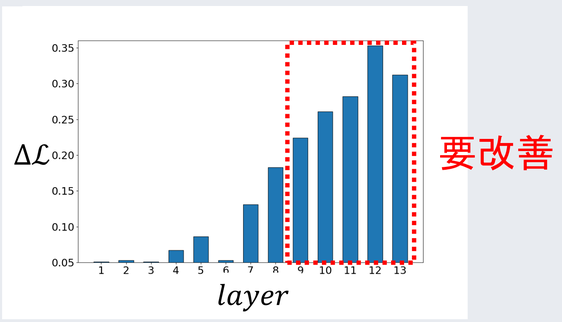
$$\Delta\kappa_\ell \equiv \Delta\mathcal{L}_\ell \,/\, |\Delta w|^2$$

診断方法

指標の計算により，損失改善の余地を層ごとに把握.



モデル診断



要改善

事後学習手法: PoL (Post-training of ℓ-th Layer)

曲率κが最大となる層のパラメタにPoFと同様の摂動を与え，曲率に起因した損失の増分 $\Delta\mathcal{L}$ が最大となる層のパラメタを更新する.

評価

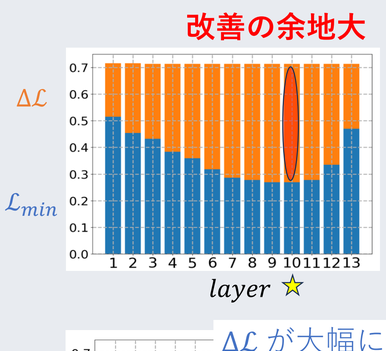
目的 学習済みモデルの診断指標を計測し事後学習による汎化性能の改善度合いを調べる.

条件
- ネットワーク：ViT-B/16
- データセット：CIFAR 100
- Fine Tuning：SGD(60ep)
- 事後学習：PoF, PoL
- 指標：テスト損失，$\Delta\mathcal{L}$，κ^2，計算時間，メモリ使用量

結果
- SGDによる局所解は高層で過適合が生じやすい.
- たった一層のみの更新でPoFに匹敵する損失改善を達成.

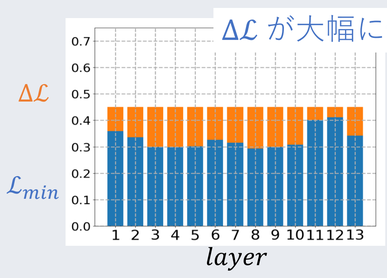
展望
- 分類誤差は下がらず，今後は別の代理損失の適用も試みる.
- 最適な摂動層の追求およびその理論的説明について研究する.

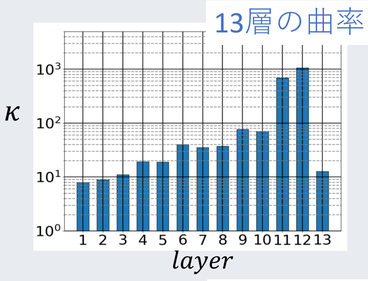
ImageNet 事前学習 モデル → 60-ep SGD fine-tuning モデル



60-ep SGD fine-tuning モデル → PoF 事後学習 モデル

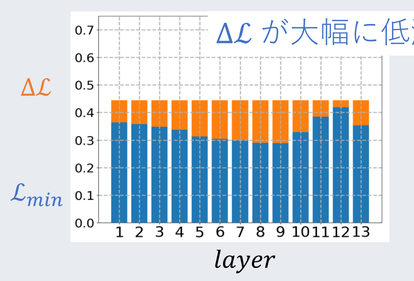
摂動13層, 更新0-12層

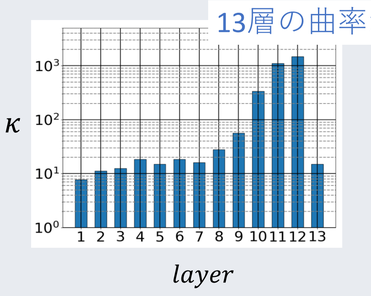




60-ep SGD fine-tuning モデル → PoL 事後学習 モデル

★摂動13層, 更新10層





モデル	テスト損失	計算時間* (s/ep)	メモリ使用量* (MiB)
SGD	0.7155	-	-
+PoF	0.4507	817	21,614
+PoL(ours)	**0.4446**	**443**	**6274**

計算量の削減

*1GPU,バッチサイズ128での学習

本研究は東京工業大学「DENSO IT LAB 認識・学習アルゴリズム共同研究講座」における研究プロジェクト「機械学習の実応用に向けた品質基盤開発」の成果である.