4R-06

半教師付き学習における一貫性損失減衰の影響解析

Relationship between Consistency Loss Decay and Generalization in Semi-Supervised Learning

楢木 悠士 *1 Yuji Naraki　本川 哲哉 *2 Tetsuya Motokawa　長沼 大樹 *3 Hiroki Naganuma

*1早稲田大学 基幹理工学部
School of Fundamental Science and Engineering, Waseda University

*2筑波大学大学院 図書館情報メディア研究科
Graduate School of Library Information and Media Studies, University of Tsukuba

*3モントリオール大学, Mila
Université de Montréal, Mila - Quebec Artificial Intelligence Institute

半教師付き学習では, 教師ラベルなしのデータを学習データとして用いるため, 入力に二つの異なる摂動を与え, 出力を比較する一貫性損失を用いることで高い汎化性能を達成している. 分類問題において最小化を行う代理損失は, 一貫性損失を分類損失に加算する形で定式化される. 一般的に分類損失は学習終盤において序盤から大きく減少するが, 一貫性損失は終盤においても比較的高い値を保つ. 本研究では汎化能力と相関のある指標として損失関数のヘッセ行列の固有値に着目し, 一貫性損失を減衰させた場合のモデルの汎化能力への影響を調査した.

1. はじめに

深層学習を用いた手法において一部教師ラベルのないデータを学習に取り組む取り組みは半教師付き学習と呼ばれる. 半教師付き学習では, 一貫性損失を制約として加えることで教師ラベルのないデータの利用を可能にし, 同じラベル数用いた教師付き学習よりも性能が向上することが知られている. しかしながら, 一貫性損失の挙動が汎化性能にどのように影響を与えているのかは明らかではない. 本研究では, 半教師付き学習手法における, 一貫性損失の汎化性能への影響理解を目的とし, 損失の学習過程における挙動を分類損失・一貫性損失に切り分けて観察を行った. さらに, 一貫性損失を用いない・もしくは減衰させた場合の Loss Landscape を観察することで, 汎化性能向上の解釈を与えた.

2. 背景

2.1 半教師付き学習手法

画像認識分野においての半教師付き学習手法のうち, 本研究では, 入力やモデルにノイズを加えた場合もノイズなしの場合と同じ出力であることを期待する一貫性に基づく手法を扱う. この手法では Teacher-Student と呼ばれる設定で学習を行う. ラベルなしデータに対して, Teacher モデルと Student モデルの出力の一貫性損失を代理損失に加えることで半教師付き学習を実現している. 一貫性損失は出力の二乗平均誤差や KL ダイバージェンスで計算される.

Π モデル [1]: Π モデルでは Teacher モデルと Student モデルのアーキテクチャは同一であり, その重みを共有する. ラベルなしデータに対して, 入力とモデルに 2 種類の異なる摂動を加え, その二つの出力の一貫性損失を代理損失に加える. ラベル付きデータに対しては分類損失のみを計算する.

Temporal Ensembling[1] Temporal Ensembling では一貫性損失として Student 出力と収束安定化のため Teacher 出力の指数移動平均（Exponential Moving Average; EMA）を用いる. ただし, Temporal Ensembling では EMA を用いるため Teacher 出力をすべて保持する必要がある.

Mean Teacher[2]: Mean Teacher は Temporal Ensembling を擬似的に再現する手法であり, Teacher モデルの重みを Student モデルの重みの EMA で定義する. これにより,Temporal Ensembling では Teacher 出力をすべて保持する必要があった問題が解決された. Teacher モデルの重みを Student の EMA で定義することで, 摂動を実現するのと同時に収束を安定させ, 性能の向上をもたらした.

本研究では Π モデルと Mean Teacher を用いる.

2.2 汎化性能の指標

本研究では, 汎化性能を測る指標としてヘッセ行列の固有値を用いる. 損失関数の解周辺が平坦さが, 汎化性能と相関があると知られている. 収束解付近の勾配が 0 になると仮定すると, 損失関数の曲率は式 1 のようにヘッセ行列で近似することができる. ヘッセ行列は式 2 で計算される.

$$L\left(\boldsymbol{w}^{*}+d\boldsymbol{w}\right)-L\left(\boldsymbol{w}^{*}\right)\sim\frac{1}{2}d\boldsymbol{w}^{T}\boldsymbol{H}d\boldsymbol{w} \tag{1}$$

$$\boldsymbol{H}(\boldsymbol{w})=\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^{2}}{\partial\boldsymbol{w}\partial\boldsymbol{w}^{T}}L(\boldsymbol{w};\boldsymbol{x}_{i},\boldsymbol{y}_{i}) \tag{2}$$

ヘッセ行列のある固有値が小さい時, その固有値に対応する固有ベクトルの次元方向で, 損失関数の曲率が小さいと言える. 我々はこの性質からヘッセ行列の固有値の和である Trace と固有値分布を調査した. ただし, ReLU 関数を使用した時, Batch Normalization などのリパラメトリゼーション操作を行うと, Trace で汎化性能を測れないことが示されている [3]. 本研究ではこの点を考慮したモデルを用いた.

3. 一貫性損失と汎化性能

一貫性損失を用いた半教師付き学習では同じラベル数を用いた教師付き学習よりもヘッセ行列の Trace が低い値を示し, 汎化性能が高い [4]. また, 図 1 に示すとおり, Π モデルと Mean Teacher では分類損失に比べて, 一貫性損失が大きく, 特に学習終盤でも一貫性損失は比較的大きな値を持つ. ただし, 一貫性損失には学習の発散を防ぐため学習初期の数エポックに対して Warmup を施しているため, 学習初期に一貫性損失に小さいな係数がかかる.

ここで, 一貫性損失が学習後半においても比較的高い値を保つ事による汎化性能への影響を調査すべく, 本研究では, 一貫性損失を減衰させた時のヘッセ行列の Trace の比較, Loss Landscape の観察を行う.

連絡先: yuji.1277@akane.waseda.jp

Copyright ©2021 Information Processing Society of Japan.
All Rights Reserved.

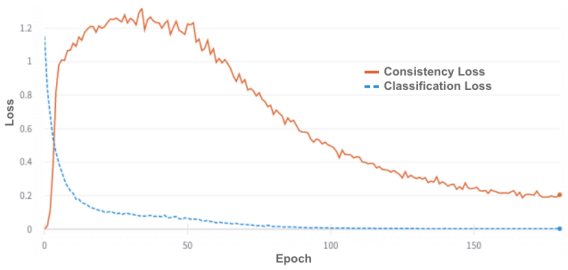

図 1: 一貫性損失と分類損失の推移

4. 実験

機械学習フレームワークとしては PyTorch[5] を, 画像分類タスクのデータセットとして CIFAR-10*1 を使用した. DNN モデルは C3 モデル [6] を用いて学習を行った. 半教師付き学習手法としては Π モデルと Mean Teacher を用いる. 一貫性損失の減衰の有無で場合分けを行い実験を行う. 学習はすべての実験に共通して 180 エポック実施する. また, ヘッセ行列の固有値の計算には PyHessian[7] を用いた. すべての実験は 10 種の異なるシード値を設定し, それらの平均の値をもとに比較を行った.

4.1 一貫性損失の減衰

表 2 に一貫性制約を用いた半教師付き学習における一貫性損失を減衰させた時のヘッセ行列の Trace の変化を示す. 一貫性損失の減衰は Consine 関数を用いたスケジューリングで実現する. 式 3 の通りに一貫性損失のスケール更新を行った.

$$\eta_{cur} = \min\left\{0.01, \frac{1}{2}\left(1 + \cos\left(\frac{N_{cur}}{N_{\max}}\pi\right)\right)\right\} \tag{3}$$

η_{cur} は一貫性損失のスケール, N_{cur} は学習が終了したエポック数, $N_{\max}$ は学習全体のエポック数を表す.

5. 結果・考察

Π モデルと Mean Teacher において一貫性損失の減衰の有無による最終更新時の勾配ノルムとヘッセ行列の Trace の変化を表 1, 2 に示す.

表 1: 一貫性損失減衰の有無と最終更新時の勾配ノルム

	減衰なし	減衰あり
Π モデル	2.75	0.0526
Mean Teacher	2.64	0.0512

表 2: 一貫性損失減衰の有無とヘッセ行列の Trace

	減衰なし	減衰あり
Π モデル	115.2	180.2
Mean Teacher	100.7	157.9

表 1 から, 4.1 で示した減衰を行うことで勾配ノルムが小さくなることが確認された. また, 表 2 から一貫性損失を減衰させるとヘッセ行列の Trace が大きくなることが示された. さらに, 図 2 に示すとおりヘッセ行列の固有値分布の観察の結果, 減衰させてしまった場合, 大きな固有値が複数現れており, 複数の次元で曲率が急峻な Loss Landscape になることが確認できた. これらのことから, 一貫性損失が学習終盤においても大きい値を持つことがモデルの汎化性能に寄与していると結論付けられる.

*1 https://www.cs.toronto.edu/~kriz/cifar.html

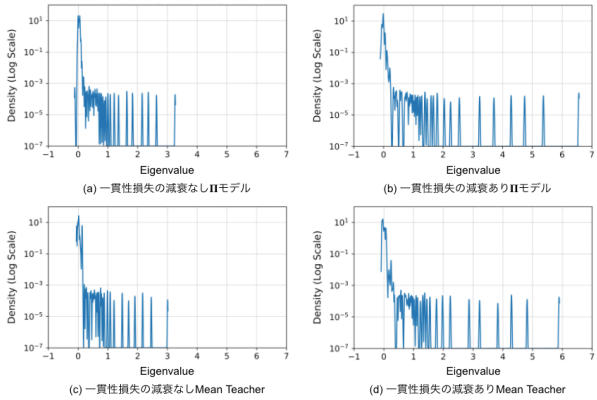

図 2: 一貫性損失減衰のスケジューリング

6. おわりに

本研究では, 一貫性の制約を用いる半教師付き学習では一貫性損失が学習後半まで大きい値を持つことで汎化性能を向上させていることが確認できた. 今後の課題として, C3 モデルでの検証だけではなく ResNet[8] などのモデルでの検証が必要である.

参考文献

[1] Samuli Laine and Timo Aila. *Temporal Ensembling for Semi-Supervised Learning*. ICLR 2017, 2017

[2] Antti Tarvainen and Harri Valpola. *Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results*. NIPS 2017, 2017

[3] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. *Sharp minima can generalize for deep nets*. In Proceedings of Machine Learning Research, volume 70, 2017.

[4] Yuji Naraki, Tetsuya Motokawa, and Hiroki Naganuma. *Towards Understanding the Effect of Model Averaging in Semi-Supervised Learning from the Perspective of Hessian Spectrum Analysis*. The 23rd Information-Based Induction Sciences Workshop 2020, 2020.

[5] Adam Paszke et al. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. NeurIPS 2019, 2019

[6] Zhewei Yao, Amir Gholami, Qi Lei, Kurt Keutzer, and Michael W. Mahoney. *Hessian-based analysis of large batch training and robustness to adversaries*. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, pages 4954–4964, 2018.

[7] Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W. Mahoney. *Pyhessian: Neural networks through the lens of the hessian*. 2020.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. *Deep Residual Learning for Image Recognition*. CVPR 2016, 2016

Copyright ©2021 Information Processing Society of Japan.
All Rights Reserved.