6S-03

深層学習のフィーチャに基づく学習モデル設計方法の提案と評価

太田 龍之介† 玉置 悠斗† 高井 直哉† 青山 幹雄†
南山大学 理工学部 ソフトウェア工学科†

1. 研究の背景と課題

近年，深層学習システムの利用が急増しており，要求する認識精度を達成する学習モデルを安定して生成する必要がある[2]. しかし，現状の開発方法において，データが学習に与える影響の分析が困難なため発見的な開発になっている.

本稿では以下の 2 点を研究課題とする.

(1) フィーチャ分析に基づき段階的に学習可能な深層学習モデル開発プロセスの提案
(2) 実際の画像データに提案方法を適用し，有効性と妥当性を評価する.

2. 関連研究

(1) 深層学習のモデル生成[1]

データのみの学習において，データのフィーチャを学習しモデル生成を行うが，フィーチャに基づいた開発方法は確立されていない.

(2) フィーチャ設計[3] [5]

機械学習モデルの精度を改善するために，適用対象となる問題の本質を表現したフィーチャに基づきデータを設計する技術体系. 特に，フィーチャの中から有用なフィーチャを選び出すことをフィーチャ選択と言う.

3. アプローチ

少量のデータを順次追加しながら学習を行い，フィーチャと学習曲線の関係を分析することで段階的に開発を行う. そこで，訓練誤差と汎化誤差の性質の違いを考慮し，訓練とテストを二重ループで実行する学習モデルの二重反復開発プロセスを提案する(図 1).

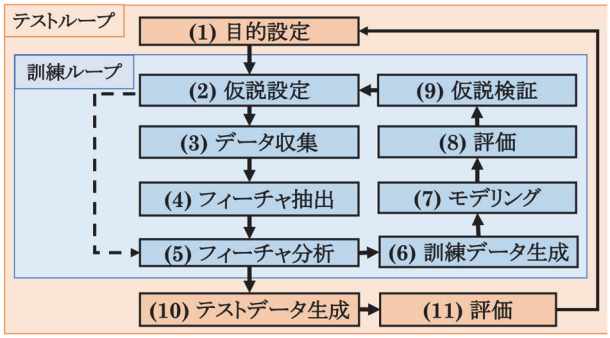

図 1 アプローチ

4. 学習モデルの二重反復開発プロセスの提案

4.1. 学習モデルの二重反復開発プロセス

本稿で提案する二重反復開発プロセスを図 2 に示す.

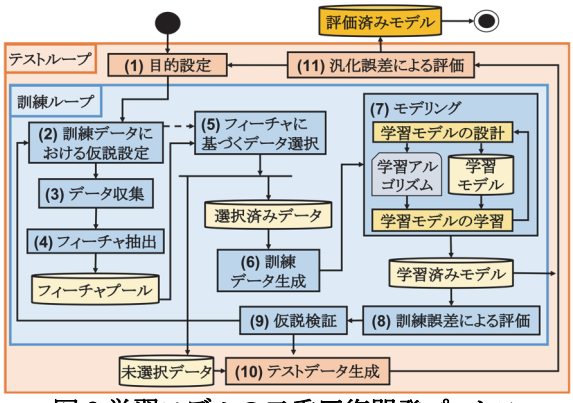

図 2 学習モデルの二重反復開発プロセス

開発プロセスの詳細は以下の通りである.

(1) 要求を満たすように目的を設定する.
(2) 目的達成に必要だと推定される訓練データについての仮説を設定する.
(3) 仮説設定に基づいたデータを収集する.
(4) データからフィーチャを抽出する.
(5) 分析プロセスをもとに訓練データを選択する.
(6) 選択済みデータから訓練データを生成する.
(7) 訓練データから学習モデルの生成を行う.
(8) 学習モデルの訓練誤差を評価する.
(9) (2)が満たされていればテストループに移る.
(10) 未選択データからテストデータを生成する.
(11) 学習モデルの汎化誤差を評価する.

4.2. フィーチャ分析プロセス

フィーチャ分析プロセスを図 3 に示す.

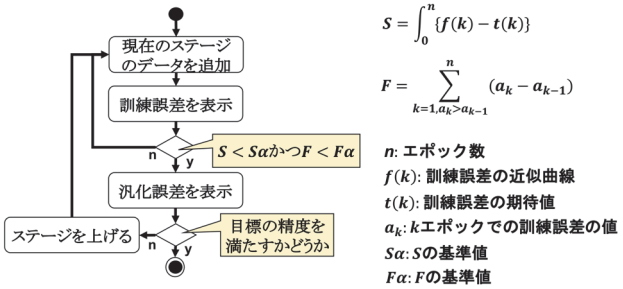

図 3 フィーチャ分析プロセス

分析を段階的に行うために，訓練誤差と汎化誤差の推移から追加するデータを選択する. データのフィーチャ数に応じて学習曲線が変化するため，フィーチャ数をもとにデータをグループ化する. そして，フィーチャ数が最小のグループ(ステージ 1)から学習を行うことで，学習プロセスを段階的に実行可能とする. さらに，訓練誤差が一定の基準値を満たし目標の精度に達していない場合はフィーチャ数が必要な数に達していないと判断し，ステージを上げてデータを追加する. これによって，学習に効果のあるフィーチャを特定し，学習を効率化する.

A Feature-Based Design Method of Learning Model for Deep Learning and its Evaluation
†Ryunosuke Ota, Yuto Tamaki, Naoya Takai, Mikio Aoyama, Department of Software Engineering, Nanzan University.

5. プロトタイプの実装

提案方法を実行するプロトタイプのアーキテクチャを図 4 に示す. 各プロセスは Python で実装し, モデルの学習は深層学習ライブラリの Chainer[4]を使用した.

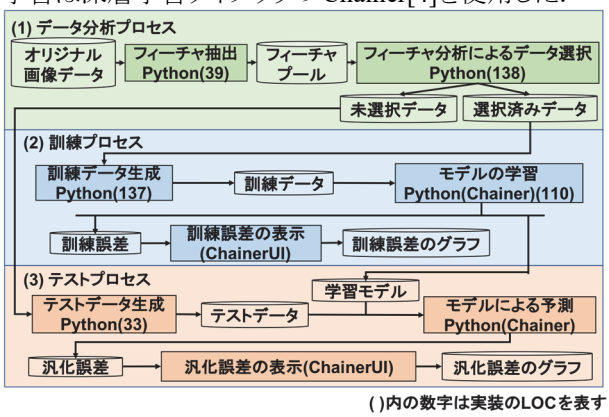

図 4 プロトタイプのアーキテクチャ

6. 実データへの適用

提案方法の有効性と妥当性を評価するために 3 種類のペットボトルの画像データ 12,000 枚に適用した. 適用条件を表 1 に示す. 適用 1 は基準値ごとに 3 回, 適用 2 は 5 回実行した. また, ステージはフィーチャ数の範囲ごとにランク付けしたもの, データ数は n ループ目のデータ数を表す.

表 1 提案方法の適用条件

	基準値	ステージ	データ数	達成条件
適用 1	1:$S\alpha$=30,$F\alpha$=1.5 2:$S\alpha$=20,$F\alpha$=1.0 3:$S\alpha$=10,$F\alpha$=0.5	1:600-700 2:700-800 3:800-900	$30(n^2-n+5)$	精度 94%以上
適用 2	1:$S\alpha$=30,$F\alpha$=1.5	1:600-700 2:700-800 3:800-900	$150n$	6 ループ終了

7. 評価

例題をペットボトル画像の 3 クラス分類とし, 訓練データとテストデータに対して 300 エポック学習, テストした. 提案方法と従来の学習方法(ランダムにデータを学習させた場合)による学習モデルの精度を比較した. 学習モデルの評価関数は平均二乗誤差を採用する.

(1) 適用 1 における精度の比較

画像 750, 1050, 1410 枚での提案方法と従来の学習方法の精度を比較したグラフを図 5 に示す.

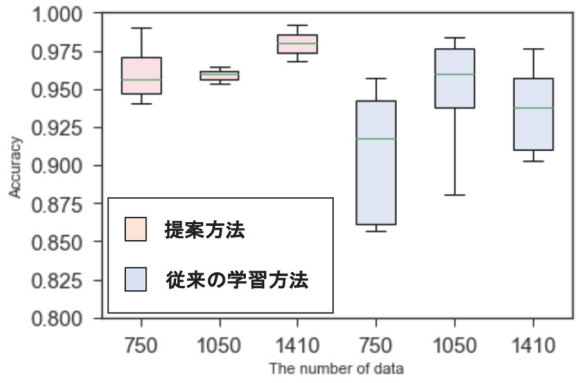

図 5 提案方法と従来の学習方法の精度

従来の学習方法での精度はデータ数ごとの標準偏差は全て 0.020 以上に対し, 提案方法では全て 0.020 未満となった. さらに, 同じデータ数での精度の平均値においては, 提案方法が従来の学習方法よりも上回っていることが確認でき, 最大改善率はデータ数 750 で 5.4%だった.

(2) 適用 2 におけるデータ数の増加に伴う精度の推移

データ数の増加に伴う精度の推移(図 6)と各データ数における精度の平均値(表 2)を示す.

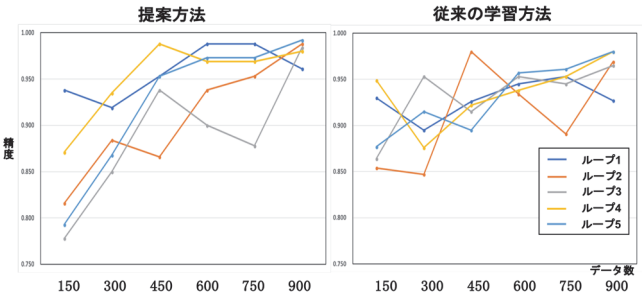

図 6 データ数の増加に伴う精度の推移

提案方法では, 従来方法と比較してデータ数と精度が比例関係にあることが確認できる. さらに, データ数 450, 600, 750, 900 での精度が全体的に上回っているため, 比較的高い精度で学習が早く収束していることがわかる.

表 2 各データ数における精度の平均値

	150	300	450	600	750	900
提案方法	0.839	0.891	0.940	0.954	0.952	0.981
従来方法	0.895	0.897	0.928	0.945	0.941	0.964

8. 考察

提案方法では, フィーチャと学習曲線の関係に着目した段階的な開発により, 適切なフィーチャ数を操作できるので同じデータ数でも平均して高い精度を実現できていると言える. さらに, フィーチャに基づいたデータ選択により訓練データのランダム性が軽減されるため, 一度に学習させる場合と比較して安定した精度を保つことができた.

9. 今後の課題

今後, 他のデータ, 深層学習モデルへ提案方法を適用し有効性と妥当性を評価する. また, 二重反復開発プロセスにおいてフィーチャ分析プロセスの検討を行う.

10. まとめ

フィーチャに基づくデータ選択によって, 段階的に学習可能な深層学習モデル開発プロセスを提案した. 提案方法のプロトタイプを実装し, 実際の画像データに適用し, その有効性と妥当性を示した.

11. 参考文献

[1] I. Goodfellow, et al., Deep Learning, MIT Press, 2016.
[2] 丸山 宏, 城戸 隆, 機械学習工学へのいざない, 人工知能, Vol. 33, No. 2, 2018 年 3 月, pp. 124-131.
[3] S. Ozdemir, et al., Feature Engineering Made Easy, Packt, 2018.
[4] Preferred Networks, Chainer, https://chainer.org/.
[5] J. Li, et al., Feature Selection: A Data Perspective, ACM Computing Surveys, Vol. 50, No. 6, Dec. 2017, pp. 1-45.