

修士論文

意味的な画像概念の DNN 学習過程における汎化性能について

On generalization performance under DNN learning process of semantic image concepts

東京電機大学大学院 システムデザイン工学研究科

情報システム工学専攻 修士課程

23AMJ03 岩瀬 俊

研究指導教員 教授 前田 英作

要旨

深層ニューラルネットワーク（DNN）を用いたエンドツーエンド学習は、コンピュータビジョン（CV）および自然言語処理（NLP）の諸タスクにおいて高い性能を実証している。しかしながら、分散表現に依存する DNN の解釈可能性は依然として限定的であり、ブラックボックスとして扱われることが多い。この透明性の欠如により、DNN が何を、いつ、どのように学習するのかという深層学習のメカニズムに対する深い理解が妨げられている。複雑な画像認識タスクは一般的に、複数の意味的な画像概念を並行して学習することを伴い、異なる特徴が様々な時間スケールで学習される。従来研究では形状やテクスチャ特徴の学習が分析されてきたが、これらの概念は通常、与えられたデータセットに基づいて帰納的に定義されており、体系的な分析には制限があった。信頼性の高い分析を行うためには、適切かつ制御可能な難易度レベルと、それらを支持するための十分なデータを有する学習タスクの確立が不可欠である。これらの課題に対し、本研究では「数字」と「色」という2つの解釈可能な概念に着目し、サンプル間に固有のノイズを含む EMNIST Digits データセットに色情報を付加することで、100 クラスの分類タスクを構築した。その上で、標準的な条件下および追加的なラベルノイズ存在下における DNN 学習プロセスの分析を実施した。本研究の結果より、各概念の獲得難度に応じて学習のタイミングが異なること、また異なる概念の学習間に相互作用が存在することが明らかとなった。これらの知見は、深層学習における画像概念の学習プロセスに関する示唆を与えるものであり、画像ベースのタスクを超えた応用可能性を有するとともに、深層学習のダイナミクスの包括的な理解に寄与するものである。

Abstract

End-to-end learning using deep neural networks (DNNs) has demonstrated high performance across various computer vision (CV) and natural language processing (NLP) tasks. However, the interpretability of DNNs, which rely on distributed representations, remains limited, often rendering them as black boxes. This lack of transparency prevents a deeper understanding of the mechanics of deep learning, specifically regarding what, when, and how DNNs learn. In contrast, complex image recognition tasks generally involve learning multiple semantic image concepts in parallel, with different features learned at varying time scales. While previous studies have analyzed the learning of shape and texture features, these concepts are typically defined inductively based on the given dataset, limiting systematic analysis. For reliable analysis, it's crucial to establish learning tasks with appropriate, controllable difficulty levels and sufficient data to support them. To address these needs, we focused on two interpretable concepts—"numbers" and "colors"—and developed a classification task with 100 classes by adding color information to the EMNIST Digits dataset, which includes inherent noise across samples. We then analyzed the DNN learning process under standard conditions and with additional label noise. Our results reveal that the timing of learning differs depending on the difficulty of acquiring each concept and that there is an interaction between learning different concepts. These findings offer insights into the learning process of image concepts in deep learning, with potential applications beyond image-based tasks, contributing to a broader understanding of deep learning dynamics.

目次

第1章	序論	6
1.1	論文構成	7
第2章	先行研究	8
2.1	深層学習	8
2.2	二重降下現象	8
2.2.1	Model-wise double descent	9
2.2.2	Epoch-wise double descent	9
2.2.3	Sparse double descent	9
2.3	画像認識における形状・テクスチャ	10

图 目 次

表 目 次

第1章 序論

深層ニューラルネットワーク（DNN）は、画像認識や自然言語処理など、幅広い分野において著しい性能向上を遂げてきた。特に、ImageNet チャレンジにおける成功を契機に、DNN は視覚的タスクにおいて人間の能力を凌駕する成果を示している [1]。しかし、その性能の背後にある内部表現の形成メカニズムは未だ十分に解明されておらず、その「ブラックボックス」的な性質は重要な研究課題として残されている。モデルの信頼性や透明性を高めるためには、学習プロセスにおける視覚的特徴の獲得メカニズムを明らかにする必要がある。

本研究は、DNN の視覚的特徴の学習プロセスにおいて、特徴獲得の順序や相互作用のメカニズムを明らかにすることを目的とする。特に、学習過程における内部表現の動的変化を系統的に分析するため、新しい実験的枠組みを構築する。視覚的特徴として「数字」と「色」に注目し、データセットを制御することで、学習過程を観察可能な環境を整備する。

本研究では、EMNIST Digits データセットに色情報を付加し、視覚的特徴の学習ダイナミクスを解析するための制御可能なデータセットを構築する。これにより、視覚的特徴の学習順序、獲得速度、および特徴間の相互作用を詳細に分析する。学習過程の評価には、モデルの精度を色や数字のみのエラー率に分解し、概念ごとの精度を観察する。

本研究の貢献は以下の3点に要約される。第一に、DNN の学習プロセスを解析するための制御可能な実験環境を構築した。第二に、異なる視覚的特徴の学習順序および相互作用を定量的に評価し、特徴獲得メカニズムを明らかにした。第三に、学習環境におけるノイズの影響を詳細に調査し、モデルのロバスト性と学習効率に関する重要な知見を得た [2]。

これらの成果は、深層学習における内部表現の理解を深めるとともに、効率的かつ解釈可能な学習アルゴリズムの設計に向けた新たな指針を提供するものである。

1.1 論文構成

第1章「序論」では、深層学習の急速な発展とその多岐にわたる応用分野について概観し、特に画像認識分野における性能向上とその背景にある主要な技術的進展について述べた。さらに、深層ニューラルネットワーク（DNN）の高い性能を支える内部表現の獲得メカニズムに関する既存の理論的枠組みと、経験的に観測される挙動との間に見られるギャップを指摘し、本研究の対象とする課題の明確化を行った。最後に、DNNの学習プロセスにおける視覚的特徴の獲得順序や相互作用の解明が、モデルの透明性向上や学習効率の改善に寄与することを示し、本研究の目的と貢献点を述べた。

第2章「先行研究」では、深層学習における視覚的特徴の獲得の研究がどのように進展してきたかを概観し、特に、画像認識における形状とテクスチャの重要性に関する先行研究を紹介する。さらに、深層学習における重要な経験的に知られる二重降下現象に関する最近の研究成果を紹介し、深層学習の学習プロセスにおける特徴獲得のメカニズムに関する理論的知見と実験的結果を紹介する。

第3章「視覚的特徴の獲得を分析するための実験設定」では、深層学習における、視覚的特徴の獲得メカニズムを紐解くための実験設定を提案し、その設定の理由と目的と述べる。

第4章「実験」では、第4章に基づいて行った検証結果、各種パラメータが与える実験結果への影響について検証する。

第5章「考察」では、実験結果についての考察を行う。

第6章「結論」では、本研究を総括する。

第7章「今後の展望」では、本研究で得られた知見から今後の方向性を示す。

第2章 先行研究

2.1 深層学習

一般に、機械学習で使用されるモデルは決定木、サポートベクターマシン (SVM)、ニューラルネットワークなどが存在する。決定木は得られた予測に対して、どの説明変数が影響したのかの判断が容易であり、説明可能性が高いことで知られている。一方で、ニューラルネットワークは、パーセプトロンを筆頭に、層の増加やネットワークの複雑化が図られてきた。黎明期においては、非線形な問題をとけるように知見が盛り込まれた SVM や、生物が持つ視覚野の知見から提案されたネオコグニトロンなどの画期的な手法が提案されてきた。その中でも、ネオコグニトロンに端を発する、畳み込みニューラルネットワーク (CNN) は、LeNet[?] により、誤差逆伝播法が導入され、2010 年代以降には、AlexNet[?], VGGNet[?], ResNet[?], と急速に進化を遂げてきた。このような深層化されたニューラルネットワークは興味深い性質や振る舞いを示す。しかし、そのような性質がどのような機序によって引き起こされるかについての完全な合意はとられていない。

2.2 二重降下現象

機械学習において、モデルの性能はモデルの複雑性（例えば、パラメータ数）と深い関係があり、モデルのパラメータ数が不足することによるアンダーフィッティング (Underfitting) [?] や、過剰なパラメータによるオーバーフィッティング (Overfitting) [?] などの現象が知られている。モデルの複雑性が増すにつれて、初めは性能が向上し（アンダーフィッティングを克服）、その後過剰な複雑性により性能が低下するとされていた。これは U 字型のカーブ、いわゆるバイアス-バリエンス トレードオフ [?] として知られている。

ところが近年発見された Double Descent[?] と呼ばれている現象は、モデルの複雑性がさらに増すと、性能が再び向上する。つまり、最初の U 字型のカーブ（アンダーフィッティングからオーバーフィッティングへの移行）の後、さらに複雑性が増加すると、新たな性能向上のフェーズが現れるのである。過剰パラメータを持つディープニューラルネットワークが、理論的にはオーバーフィッティングを起こすべきなのに、実際には優れた汎化性能を示す場合がある [?, ?]。

この Double Descent は、Belkin ら [?] によって決定木や二層のニューラルネットワークで確認され、その後、Nakkiran ら [?] が、ディープニューラルネットワーク (DNN) においても観察されること、学習エポック数の増加に対しても Double Descent が起こることを示した。さらに、パラメータの枝刈りによるスパース性の増加に対しても Double Descent が起こ

ることが報告されている [?]. パラメータ数, 学習エポック数, スパース性の増加に伴って観察される Double Descent は, それぞれ, Model-wise Double Descent, Epoch-wise Double Descent, Sparse Double Descent と呼ばれている [?, ?].

2.2.1 Model-wise double descent

Yang らは, バイアス-分散のトレードオフに関する古典的な理論を, 広範な実験を通して再検討した [?]. 彼らは, 分類理論が予測するようにバイアスが単調減少する一方で, 分散は単峰性の挙動を示すことを発見した. このバイアスと分散の組み合わせは, 3つの典型的なリスクカーブパターンを示唆しており, すでに報告されている多くの実験結果と一致している. また, Somepalli らは, 新たな決定境界可視化手法を提案し, 二重降下におけるエラーの悪化する領域において, 決定境界が断片化していることを報告している [?]. さらに, Curth らは決定木などの Belkin らが二重降下を観察した条件において, 2つの次元の軸による U 字のカーブの重ね合わせによって二重降下が起きると報告し, 深層学習における二重降下においても, この観点は良い指針になることを示唆している [?].

2.2.2 Epoch-wise double descent

統計的シミュレーションの結果から, 学習過程における二重降下に関するいくつかの仮説が浮かび上がってきた. これらの仮説はデータの特徴に焦点を当てている. 例えば, Stephenson らは, 二重降下は遅いが有益な特徴によって起こると仮定し, 理想的な線形モデルにおいてデータの主成分を除去することで二重降下の挙動を除去できることを示している [3]. 一方, Pezeshki らは実験により, 異なる時間スケールで学習された特徴が二重降下を引き起こすことを発見している [?]. さらに, Heckel らは, モデルの異なる部分が異なるエポックで学習することによる, 複数のバイアスと分散の重複トレードオフが二重降下を引き起こすとしている [?]. そのうえで, 層間で学習率を変えることで二重降下を緩和できることを実証している.

2.2.3 Sparse double descent

モデルのスパース性が高まるにつれて, つまり多くのパラメータがゼロまたは非常に小さくなるにつれ, まず性能の向上が見られる [?, ?]. しかし, ある点を境に性能は低下する. さらにスパース性を高めると, 性能は再び向上する. このことは, ネットワークの刈り込みのような方法で達成可能な適度なスパース性が, モデルのオーバーフィッティングを抑制し, 汎化性能を向上させることを示唆している. また, 枝刈り前のモデルの大きさに関わらず, 枝刈り後の精度は一定である可能性が示唆されている [?].

2.3 画像認識における形状・テクスチャ

Geirhos らは、ImageNet で学習した CNN が、分類のために特に画像のテクスチャを重視することを示した [4]。彼らは、相反する形状とテクスチャ情報を持つ画像を CNN に入力し、出力が形状ベースのラベルとテクスチャベースのラベルのどちらに一致するかをチェックした。この結果に基づいて、CNN が認識において形状とテクスチャのどちらを優先するかを分析した。一方、Islam らは、ニューロンの潜在表現に基づくモデルにおいて、形状とテクスチャのどちらを重視するかを定量的に判断する方法を提案した [?]。この方法によって、CNN がどの特徴に偏重するかを定量的に分析することができる。さらに、Ge らは人間の視覚系のモデル化を試み、Human Vision System (HVS) を開発した。HVS は、画像分類時にどの特徴（形状、テクスチャ、色など）が最も重要な役割を果たすかを定量的に評価可能である [5]。

本研究は、画像理解と二重降下における CNN に関する先行研究を基礎としている。Islam らの手法を使用して、CNN 学習中のテクスチャと形状情報に関する知識の獲得と二重降下現象との関係を明らかにすることを試みた。Islam らの手法を利用し、最終畳み込み層における形状・テクスチャそれぞれをエンコードするニューロン数を推定、推定した割合の値を形状・テクスチャ偏重度としている。

関連図書

- [1] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “Imagenet large scale visual recognition challenge,” *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
- [2] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015. [Online]. Available: <http://arxiv.org/abs/1409.0473>
- [3] C. Stephenson and T. Lee, “When and how epochwise double descent happens,” *CoRR*, vol. abs/2108.12006, 2021.
- [4] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, “Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness,” in *ICLR*, 2019.
- [5] Y. Ge, Y. Xiao, Z. Xu, X. Wang, and L. Itti, “Contributions of shape, texture, and color in visual recognition,” in *ECCV*, S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., 2022, pp. 369–386.