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

図 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[3]，VGGNet[4]，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-バリアンス 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 [**?**, **?**].

2.3　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 [5]．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の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 [6].

2.4　画像認識における二重降下現象と形状・テクスチャバイアスの関係

髙橋らの研究 [7] は，画像認識タスクにおける二重降下現象と，CNN の形状バイアスおよびテクスチャバイアスの変化の関連性を示唆した．この研究では，二重降下現象と形状・テクスチャバイアスの変化のタイミングに相関が見られ，テスト誤り率が上昇から下降に転じるタイミングと，形状バイアスが増加し始めるタイミングが一致することが示された．また，事前学習の有無による学習過程の違いも観測された．事前学習ありのモデルでは初期からテクスチャバイアスが高く，学習が進むと形状バイアスが増加する傾向が見られた．一方，事前学習なしのモデルでは学習初期は形状バイアスが高く，その後テクスチャバイアスが増加する傾向が確認された．

岩瀬らの研究 [8] では，さらに形状・テクスチャバイアスと二重降下現象に関する同期性を様々な角度から検証している．この研究では，CIFAR-10 と ResNet18 の組み合わせ以外の条件下で同期性が確認されたほか，どの層が要因となり，同期性が起こっているのかが明らかとなった．これらの知見は，画像認識モデルの学習ダイナミクスと二重降下現象の関係性に新たな洞察を提供している．この同期性を図 2.1 に示す．これまでの研究では，CNN の学習過程における形状とテクスチャという特徴の獲得と二重降下現象の関係について研究されてきた. そこで，本研究では形状とテクスチャではなく，色と数字という概念を設定し，２つの学習過程における概念獲得の過程の観測した．

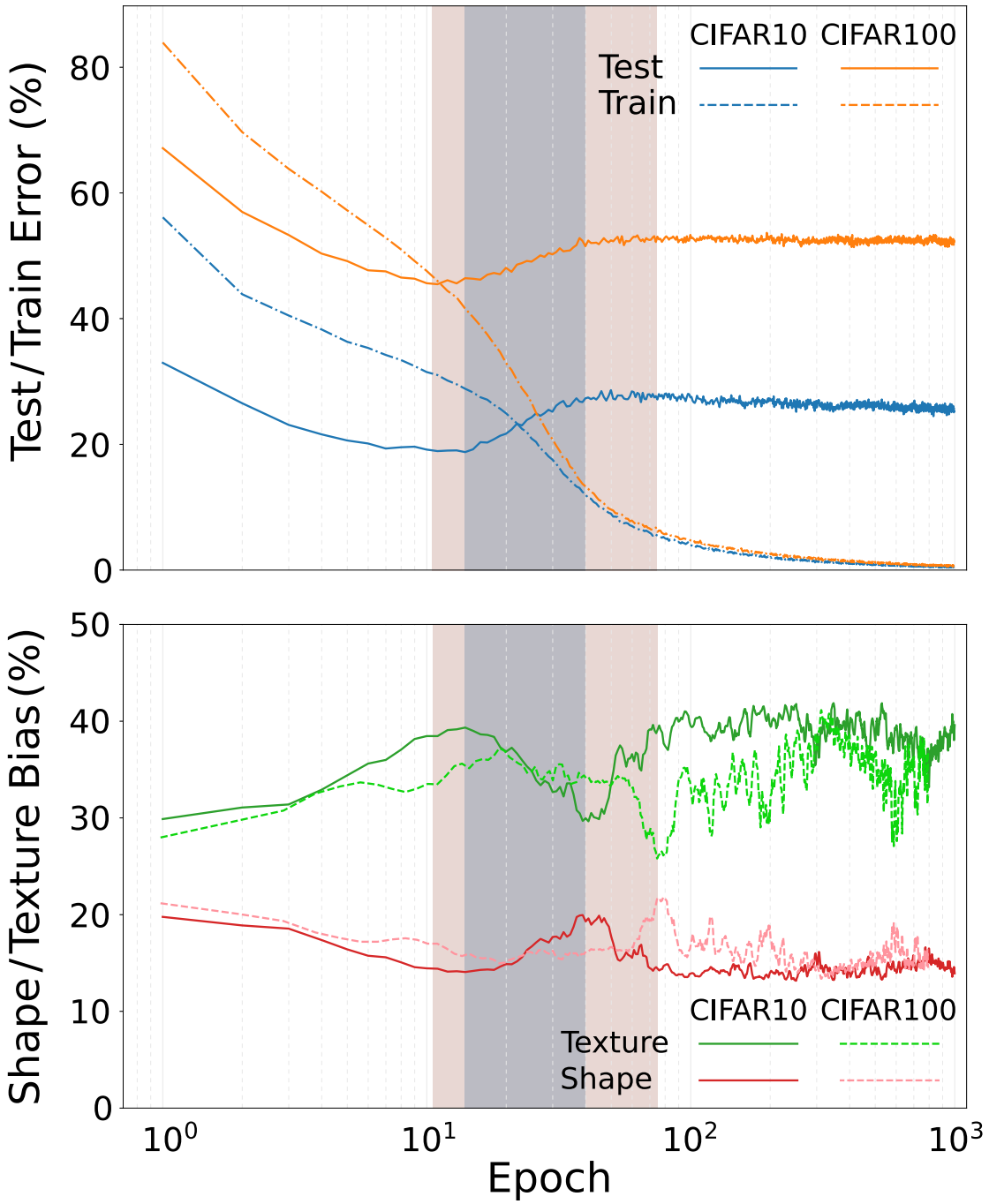

Fig. 2.1: ResNet18 の学習過程における二重降下現象と形状・テクスチャバイアスの変化の同期性．ImageNet で事前学習済みの ResNet18 を使用し，CIFAR-10 と CIFAR-100 の学習を行った際のテストエラー率．また，それぞれのエポックでのモデルにおける形状・テクスチャバイアスを計算した結果．二重降下現象を 3 つのフェーズに分けた際に，第 1 フェーズでは，テストエラー率が減少するときにはテキスチャバイアスが強くなる．次に過学習のタイミングで形状バイアスが強くなり，再度はエラー率が下がるときテクスチャバイアスを強くするように戻る．

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] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2012, pp. 1097–1105. [Online]. Available: https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. [Online]. Available: https://arxiv.org/abs/1409.1556

[5] 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.

[6] 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.

[7] 髙橋秀弥, 井上中順, 横田理央, 片岡裕雄, and 前田英作, "画像識別における形状・テクスチャ偏重度と二重降下現象の関係について," *IEICE Conferences Archives*, vol. IEICE-122, no. IEICE-PRMU-404,IEICE-IBISML-405, pp. IEICE–PRMU–13,IEICE–IBISML–13–IEICE–PRMU–16,IEICE–IBISML–16, Feb. 2023.

[8] S. Iwase, S. Takahashi, N. Inoue, R. Yokota, R. Nakamura, H. Kataoka, and E. Maeda, "On the relationship between double descent of cnns and shape/texture bias under learning process," in *"in press" International Conference on Pattern Recognition*, 2024.