# 論文 / 著書情報
# Article / Book Information

| | |
|---|---|
| 題目(和文) | 深層学習による画像認識における推論根拠と未知不均衡ドメイン学習に関する研究 |
| Title(English) | |
| 著者(和文) | 桑島洋 |
| Author(English) | Hiroshi Kuwajima |
| 出典(和文) | 学位:博士(工学),<br>学位授与機関:東京工業大学,<br>報告番号:甲第12361号,<br>授与年月日:2023年3月26日,<br>学位の種別:課程博士,<br>審査員:田中 正行,奥富 正敏,蜂屋 弘之,中臺 一博,川上 玲 |
| Citation(English) | Degree:Doctor (Engineering),<br>Conferring organization: Tokyo Institute of Technology,<br>Report number:甲第12361号,<br>Conferred date:2023/3/26,<br>Degree Type:Course doctor,<br>Examiner:,,,, |
| 学位種別(和文) | 博士論文 |
| Type(English) | Doctoral Thesis |

# 深層学習による画像認識における
# 推論根拠と未知不均衡ドメイン学習に関する研究

東京工業大学工学院
システム制御系システム制御コース
桑島洋

2023 年 3 月 14 日

## 概要

大量のデータからデータ駆動で複雑なモデル（膨大なパラメータを持つ深いニューラルネットワーク）を学習する深層学習は、画像認識などで成果をあげ、自動車などのセーフティクリティカルな用途でも使われ始めている。深層学習モデルの開発では、第一に、学習データとテストデータを収集する。第二に、学習データから深層学習モデルをデータ駆動で学習する。第三に、深層学習モデルが推論した結果をテストデータで検証する。深層学習は、学習データとテストデータの分布が一致することを仮定しており、この仮定が成り立てば、高い性能を発揮する。

深層学習には「複雑なモデルの限界」と「データ駆動の限界」がある。深層学習モデルは、パラメータ数が膨大で構造も複雑であり、特徴量の活性値は観測できるが、実際の現象は把握できない。複雑なモデルの限界は、推論の過程や根拠を人間が理解できないため、問題を特定、修正、改善するというエンジニアリングが難しいことである。一方、深層学習を利用した高度運転支援システムや自動運転システムで事故が起きるなど、リスクが顕在化している。これらの事故は、道路を横断する大型トレーラーの側面や深夜に道路を横断する歩行者など、事故につながる希少な環境条件（ドメイン）に遭遇したときに起きている。学習データとテストデータの分布が一致するという深層学習の仮定に対して、実際の開発では、学習データとテストデータの分布の一致は不明である。現実的には、学習データには異なるドメインの標本が不均衡に混在し、標本の元ドメインはわからない。この状況を「未知不均衡ドメイン」と呼ぶ。未知不均衡ドメインの状況で、学習データに十分に含まれていない少数派ドメインにおいて深層学習モデルの性能が低下し、事故につながったと考えられる。データ駆動の限界は、未知不均衡ドメインに対して十分な性能を発揮できないことである。

ところで、深層学習では、学習データを所与のものとしてモデルを学習するだけではなく、現実の世界（母集団）から標本を抽出し学習データを作成することが重要である。データ作成に関連する技術として、アノテーションのコスト削減を主な目的とする能動学習がある。そのため、未知不均衡ドメインへのアプローチとして、「機械学習（モデル学習）」と「能動学習」がある。

そこで本研究では、深層学習による画像認識において、推論根拠と未知不均衡ドメイン学習（機械学習と能動学習）を基礎検討する。推論根拠に関する研究の目的は、特徴量に着目して深層学習の推論過程を分析し、人間に理解できる形式で提示する手法を構築することである。未知不均衡ドメイン学習に関する研究の目的は、未知不均衡ドメインの標本

を含む学習データに対して、多数派ドメインの性能を保持しつつ、少数派ドメインの性能を向上させる機械学習（モデル学習）と能動学習の手法を構築することである。

本論文は全 5 章で構成され、各章の概要は以下のとおりである。

第 1 章「緒言」では、本研究の背景や目的、本論文の構成などを述べる。

第 2 章「深層学習における推論根拠」では、深層学習モデルの特徴量に着目した推論根拠の分析方法と、クラウドソーシングを用いて分析結果の推論根拠の妥当性を評価する方法を提案する。提案手法で分析した深層学習モデルの推論根拠が人間の推論過程とおおむね一致し、人間に理解できる分析結果であることを実験で示す。また、推論根拠を理解することで、モデルの拡張や学習データの追加など、改善のための手掛かりが得られる可能性も示す。

第 3 章「深層学習における未知不均衡ドメイン機械学習」では、未知不均衡ドメインが混在する学習データを用いて深層学習モデルを学習し、ドメイン別のテストデータを用いて深層学習モデルをドメインごとにテストする「未知不均衡ドメイン機械学習」の問題設定を提案する。未知不均衡ドメイン機械学習の問題設定に対して、center loss（損失関数）と特徴量の空間における標本間の距離に基づいたミニバッチ抽出を組み合わせた手法を提案する。提案手法により、多数派ドメインの性能を維持しながら少数派ドメインの性能が向上することを実験で示す。

第 4 章「深層学習における未知不均衡ドメイン能動学習」では、第 3 章の内容を能動学習に拡張する。深層学習モデルの学習において能動学習が用いられる実際の状況を想定し、現実的な巨大データプールの生成と、現実的なモデル学習を考慮した実験設定を提案する。未知不均衡ドメイン機械学習を拡張し、能動学習とモデル学習を組み合わせることで未知不均衡ドメインに対処する「未知不均衡ドメイン能動学習」の問題設定を提案する。能動学習の現実的な実験設定のもとで、モデル学習と能動学習の手法の組み合わせを比較し、softmax margin による能動学習と第 3 章の提案手法（モデル学習）を組み合わせることで、多数派ドメインの性能を維持しながら少数派ドメインの性能が向上することを実験で示す。

最後に、第 5 章「結言」で本研究のまとめと今後の課題を述べる。

# 目次

# 図目次

# 表目次

# 第 1 章

# 緒言

2012 年に物体認識の競技会 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) で深層学習を用いた手法が優勝し [55]、物体認識で深層学習を用いることが主流になった。その後、数年の間に深層学習により画像認識タスクの拡張が起こり、物体認識（画像内の物体の種類）だけでなく、位置を矩形領域で推定する物体検出 [81] や物体領域を色分けするセマンティックセグメンテーション [25] が出現した。これらの深層学習による新しい問題設定が、画像認識の応用の幅を大きく広げた。

従来の機械学習は、人間が設計した特徴量に基づいて、分類器をデータ駆動で学習する。一方、深層学習は、大量のデータと、階層の深いニューラルネットワーク（これを深層学習モデルと呼ぶ）を用いることで、特徴量を人間が設計せず、特徴量と分類器の両方をデータ駆動で学習する [33]。深層学習モデルの開発では、最初に、学習データとテストデータを収集する。次に、学習データを用いて、深層学習モデルをデータ駆動で学習する。最後に、テストデータを用いて、深層学習モデルを検証する。データ駆動による深層学習モデルの学習は、学習データとテストデータの分布が一致することを仮定しており、この仮説が成り立てば、深層学習モデルは高い性能を発揮できる（図 1.1）。

深層学習技術は、自動運転システム（Automated Driving System, ADS）や先進運転支援システム（Advanced Driver-Assistance Systems, ADAS）などの自動車関係の応用や、画像認識による診断や AI によるゲノム解析などの医療関係の応用など、セーフティクリティカルな用途でも用いられ始めている。例えば、ADS は一般的に認知・判断・操作の 3 つの機能で構成されているが、特に認知機能と判断機能への深層学習の適用が進んでいる。認知機能では、人検出技術の導入 [110]、セマンティックセグメンテーションによる走行可能領域検の検出 [110]、判断機能では、深層強化学習を用いた自動運転 t と車線維持支援 [86] など、様々な方法が研究または実用化され、深層学習は ADS に欠かせな

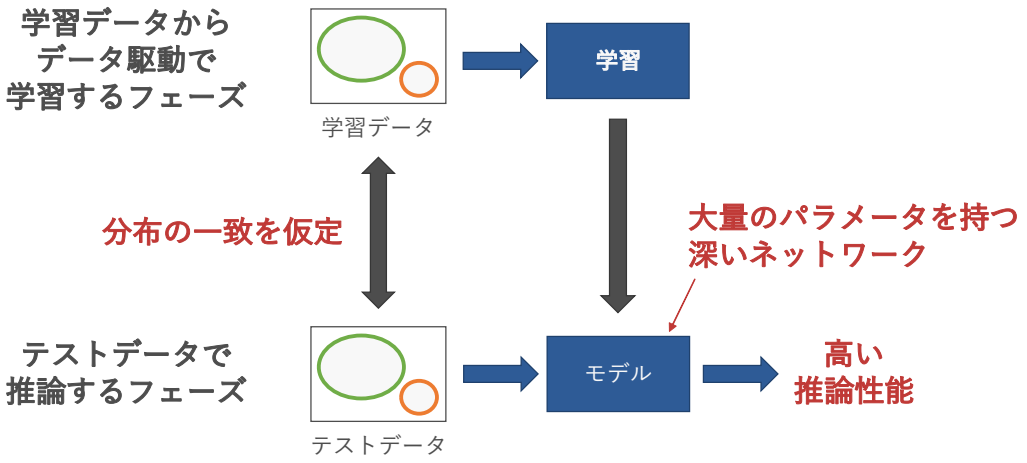



図 1.1: データ駆動による深層学習モデルの学習

い技術となっている。

ここで、従来ソフトウエアの開発と深層学習モデルの開発の相違点 [56] に着目する。従来のソフトウェア開発では、ソースコードを記述し、コンパイルして動作させる。人が設計する従来ソフトウェアの振る舞いは基本的には把握できる（並列処理など把握が動作理解が簡単ではない場合もある）。一方、深層学習モデルの開発では、前述の通り、データを作成し、モデルを学習し、学習済みモデルを動作させる。また、深いモデルのためパラメータ数は膨大である。このため、パラメータ数が膨大で構造も複雑、かつ、人間が直接設計せずデータから学習する深層学習モデルでは、特徴量の活性値は観測できるが実際の現象は理解できない（図 1.2）。

また、深層学習が社会に浸透し始める一方で、深層学習を搭載したシステムのリスクも顕在化している [112]。深層学習の自動車への応用では、2016 年に Tesla の Autopilot モード（ADAS）で発生した事故 [2]、2018 年に Uber が ADS の公道テスト走行中に発生した事故 [3] などが大きく報道された。2021 年には、運輸省道路交通安全局（NHTSA）が、Tesla の Autopilot と Traffic Aware Cruise Control による 12 件の緊急車両への追突事故の調査を開始した [4]。2016 年の Tesla 事故では、Autopilot 走行中に車線をまたぐトラックの側面に衝突した。国家運輸安全委員会 （NTSB）による原因調査が行われ、Autopilot 機能が、車の進路を横切るトラックを認識するように設計されていなかったためと結論付けられた。Uber の事故では、ADS の設計において横断歩道外に歩行者を考慮していなかったため、歩行者の発見が遅れたことなどが原因とされた。2021 年の調査対象の Tesla 事故では、緊急車両や関連する物体（点滅する警光灯、発煙筒、カラーコーン・

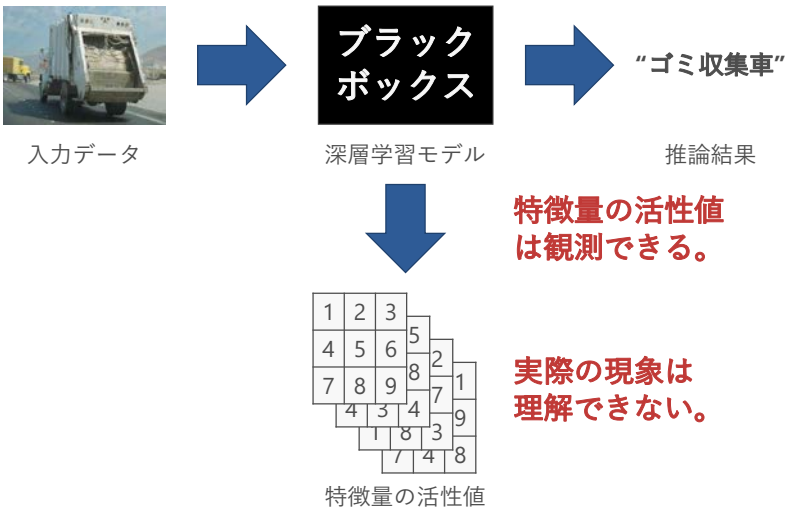



図 1.2: 深層学習のブラックボックス性


図 1.3: 未知不均衡ドメインの例

パイロン、安全反射ベスト、事故現場を守るために斜めに停められた緊急車両など）を学習データに含めていたことや、テストしていたことが問われている。いずれも、深層学習モデルが使われている ADAS/ADS による事故と推察され、重要だが学習データに十分に含まれていない希少な環境条件（ドメイン）に遭遇したときに発生している。深層学習では、学習データとテストデータの分布が一致することを仮定するが、現実的には、データには様々なドメインから生成された標本が、異なるバランスで混在し、標本の所属ドメインを知ることができない。本研究では、この状況を「未知不均衡ドメイン」と呼ぶ（図 1.3)。

ここまでに述べた深層学習のブラックボックス性やリスク顕在化は、2 つの深層学習の限界（データ駆動の限界と複雑なモデルの限界）を示している（図 1.4)。

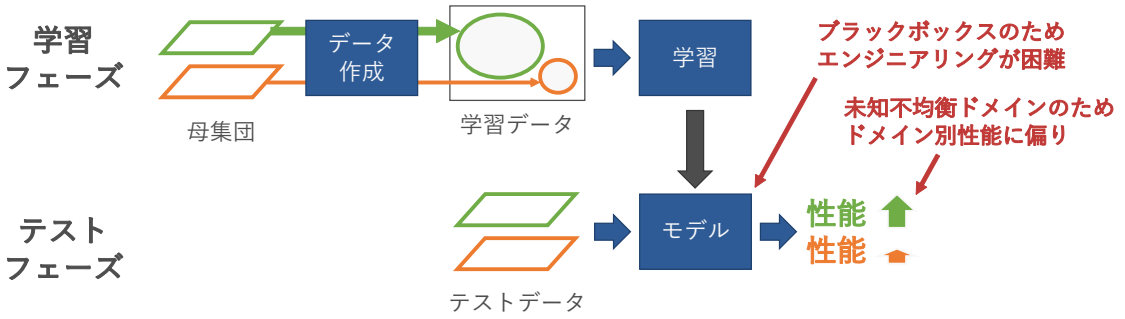



図 1.4: 深層学習の限界

1. **複雑なモデルの限界**

   モデルがブラックボックスで、問題の特定や修正などのエンジニアリングが難しい。

2. **データ駆動の限界**

   実際には学習データは未知不均衡ドメインであり、ドメインで性能が偏る。

第一の限界は、複雑なモデルの限界である。深層学習のモデルは、複雑な特徴抽出部を含む、階層の深いニューラルネットワークであり、複雑な構造で大量のパラメータを持つ。そのため、データ駆動で学習したモデルの内部処理はブラックボックスで、前述のようなリスクが顕在化しても、従来ソフトウェアのバグ修正のように、原因を分析し修正することはできない [56]。複雑なモデルの限界は、推論の過程や根拠を人間が理解できないため、問題を特定、修正、改善するというエンジニアリングが難しいことである。第二の限界は、データ駆動の限界である。深層学習では、学習データとテストデータの分布が一致することを仮定するが、実際の開発では、学習データとテストデータの分布の一致は不明で、学習データは未知不均衡ドメインである。未知不均衡ドメインの状況では、従来の学習方法は多数派ドメインに最適化してしまう。しかし、自動運転などの応用によっては、少数派ドメインの重要度が同等または高い場合がある。データ駆動の限界は、現実にはデータに混入することが避けられない未知不均衡ドメインに対して、ドメインにより性能が偏ってしまうことである。

ところで、データ駆動による深層学習では、モデル学習だけではなく、現実世界から標本を抽出するデータ作成も重要である（図 1.5）。モデル学習は、与えられた学習データから**モデルを学習**する。一方で、能動学習は、アノテーションが高コストの状況で、最小限の標本を母集団から抽出してアノテーションし、**学習データを作成**する。そのため、第二の限界である未知不均衡ドメインのアプローチとしては、モデル学習と能動学習がありうる。

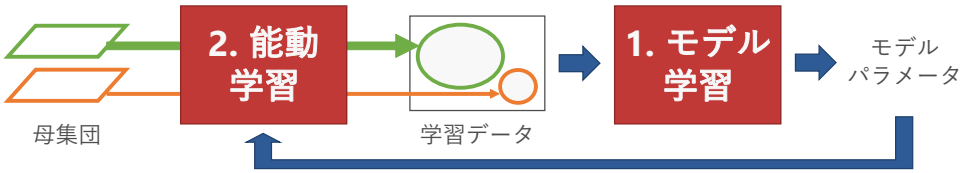



図 1.5: モデルの学習とデータの作成

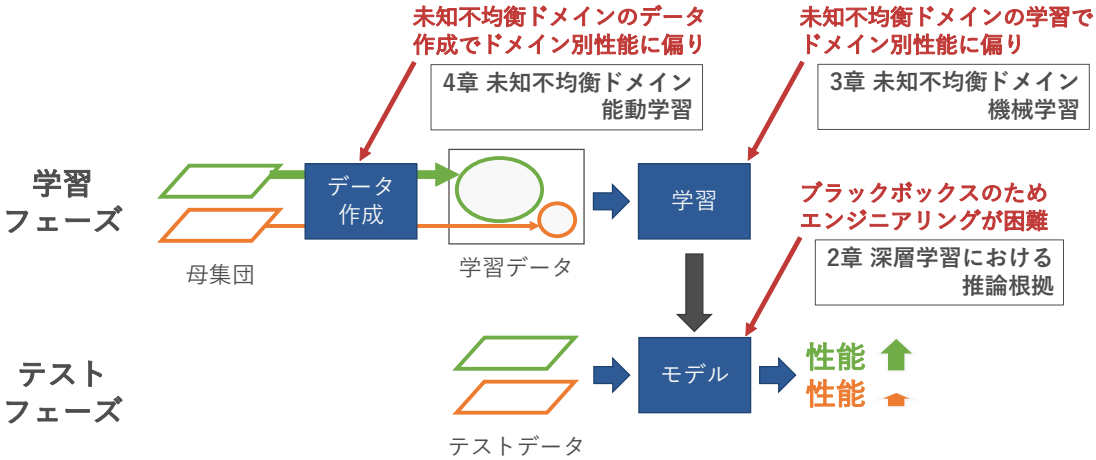



図 1.6: 本研究の目的

## 1.1 本研究の目的

本研究では、深層学習を用いた画像認識において、推論根拠と未知不均衡ドメイン学習の研究に取り組む（図 1.6）。

1. **深層学習における推論根拠**

   深層学習モデルがどのような過程を経て推論結果を生成しているかを分析し、人間に理解できる形式で提示する手法を構築する。

2. **未知不均衡ドメイン機械学習**

   訓練データに様々なドメインが様々なバランスで混在している状況（未知不均衡ドメイン）で、多数派ドメインの性能を保持しつつ、少数派ドメインの性能を向上する機械学習（モデル学習）の方法を構築する。

3. **未知不均衡ドメイン能動学習**

   深層学習システムにおける能動学習の現実的な実験設定のもと、未アノテーション

(a) ImageNet[23, 84, 80]


(b) 手書き文字認識 [61, 19, 47]

図 1.7: 実験に用いるデータセットの標本例

データ（データプール）に未知不均衡ドメインが含まれる状況で、多数派ドメインの性能を保持しつつ、少数派ドメインの性能を向上する能動学習（データプールから学習データを獲得）の方法を構築する。

なお、本論文で行った研究は基礎検討であり、それぞれの問題を表現できる異なるトイデータセットを用いる。

**2 章で用いるデータセット：**ImageNet[23, 84, 80]**（図** 1.7a**）**
特徴量と視覚属性の対応がわかりやすい自然画像のデータセット。人間が理解できる推論根拠の解析を生成する対象として適している。

**3 章と 4 章で用いるデータセット：**MNIST[23], EMNIST[84], USPS[80]**（図** 1.7b**）**
手書き文字認識のデータセット。3 種類の類似データセット（ドメイン）が存在し、未知不均衡ドメインのデータを模擬できる。

## 1.2 本論文の構成と概要

本論文の構成を図 1.8 に示す。

第 1 章「緒言」では、本研究の背景や目的、本論文の構成などを述べる。

第 2 章「深層学習の推論根拠の解析」[*1]では、深層学習ネットワークの推論過程を説明する方法を基礎検討する。特徴量の構造分析（活性化パターンの分析）と特徴量と自然言語を対応付ける（特徴量アノテーション）により、深層学習ネットワークの推論過程を説明する方法を提案する。特徴量の構造分析では、畳み込みニューラルネットワークの複数の特徴マップから推論に貢献した特徴量を取り出す手法を提案し、分類クラスによって貢献する特徴量（特徴マップ）が異なることを明らかにする。特徴量と自然言語の対応付けとして、特徴量の対応する入力画像の領域をアノテーションする手法を提案し、推論に貢

[*1] 第 2 章の研究成果は Kuwajima, Tanaka, and Okutomi (2019) [57] に基づく。ただし、(Kuwajima et al., 2019) [57] の著作権は Springer Nature Customer Service Centre GmbH に帰属する。

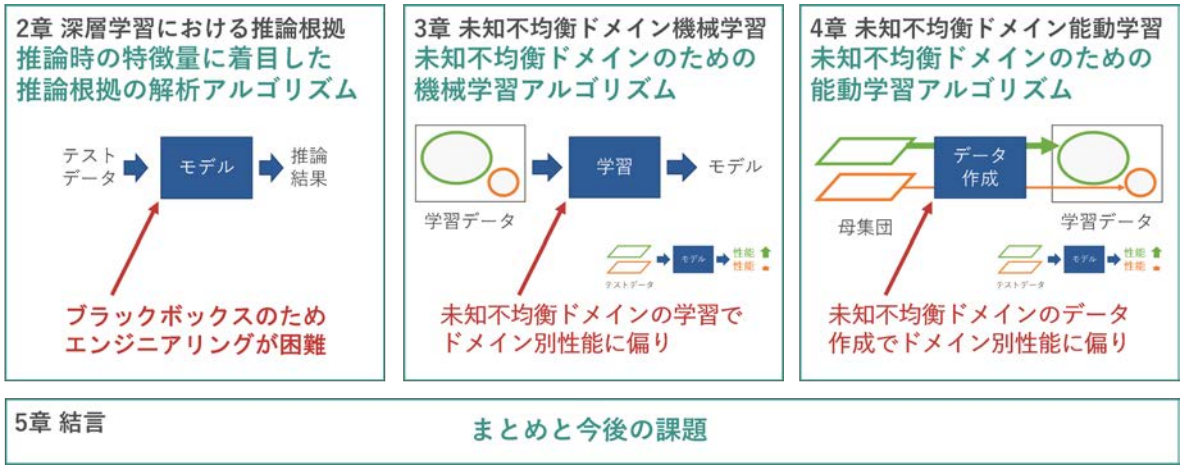



図 1.8: 本論文の構成

献する特徴量と組み合わせ、推論過程を説明する。また、説明の結果を評価するため、入力（画像）・出力（推論結果）・特徴量（中間表現）の間の整合性を見ることにより、（本論文などで提案する）深層学習ネットワークの推論過程の説明と、人間による推論過程との整合性を分析する方法を提案する。整合性分析では、画像特徴への注目（入力と特徴量の整合性、物理整合率）と、注目した画像特徴から推論結果を導き出す過程（特徴量と出力の整合性、論理整合率）の 2 つの観点で、深層学習ネットワークの推論過程と人間による推論過程の整合性を分析する。クラウドソーシングを使った実験により、深層学習ネットワークによる推論過程と人間による推論の過程が概ね一致することとともに、誤認識の場合の推論過程の分析により、ネットワークの拡張や学習データの追加など、改善のための次のアクションを示唆できる可能性を示す。

第 3 章「未知不均衡機械学習」[*2]では、様々な分布から発生したデータが含まれている学習データに対する機械学習（モデル学習）について基礎検討を行う。現実の応用では、学習データには複数のドメイン（例えば、手書き文字認識で、筆者の社会的属性で筆跡が変化する）が含まれ、あるドメインはより高い重要性やリスクを持つことがある。まず、学習データのドメイン構成が未知かつ不均衡で、テストデータではドメインごとに性能を評価する「未知不均衡ドメイン機械学習」の問題設定を提案する。この問題設定では、ドメイン別に性能を評価し、多数派ドメインと少数派ドメインの両方の性能を向上させることが必要である。次に、MNIST/EMNIST/USPS を混合して作成した不均衡データを用

[*2] 第 3 章の研究成果は Kuwajima, Tanaka, and Okutomi (2022) [58] に基づく。ただし、(Kuwajima et al., 2022) [58] の著作権は Society of Imaging Science and Technology (IS&T) に帰属する。

いて学習した特徴空間を観測し、標本数が少ないほど、少数派ドメインは多数派ドメインの遠方に分布するように特徴空間が構成されることを示す。この結果を受け、少数派ドメインの分布を多数派ドメインの分布に近づけることを狙い、未知不均衡ドメイン機械学習の手法として center loss（損失関数）と特徴空間でのセントロイド距離に基づいた重み付き標本抽出の組み合わせを提案し、MNIST/EMNIST/USPS を混合して作成した未知不均衡データを用いた実験で有効性を示す。

第 4 章「未知不均衡ドメイン能動学習」[*3]では、第 3 章を能動学習に拡張する。能動学習は、アノテーションのコストを削減するために、深層学習の登場以前から研究されてきた領域であり、前述の通り、深層学習には大量のデータが必要であるため、再び脚光を浴びている。そこで、まず、深層学習の文脈における能動学習の応用に即した現実的な実験設定を提案し、実験設定の選択により実験結果が大きく異なることを示す。次に、「未知不均衡ドメイン機械学習」を拡張し、複数のドメインが含まれるアノテーション前のデータプールから最適な学習データを獲得する能動学習と深層学習モデルを学習するモデル学習を組み合わせ、多数派ドメインと少数派ドメインの両方の性能を向上させる「未知不均衡ドメイン能動学習」の問題設定を行う。未知不均衡ドメイン能動学習の手法として、能動学習手法と機械学習（モデル学習）手法の組み合わせを検討し、softmax margin による単純な能動学習手法と第 3 章で構築した未知不均衡ドメイン機械学習（モデル学習）手法が有効であることを、MNIST/EMNIST/USPS を混合して作成した未知不均衡データプールを用いた実験で示す。

最後に、第 5 章「結言」で本論文のまとめと今後の課題を述べる。

[*3] 第 4 章の研究成果は Kuwajima, Tanaka, and Okutomi (2023) [59] に基づく。ただし、(Kuwajima et al., 2023) [59] の著作権は Society of Imaging Science and Technology (IS&T) に帰属する。

# 第 2 章

# 深層学習における推論根拠

近年、深層学習技術は急速に発展しており、様々なシステムの実現に必須の技術となっている。しかし、深層学習モデルの推論過程はブラックボックスであり、高い透明性が求められるセーフティクリティカルなシステムにはあまり適していない。本章[*1]では、深層学習モデルの社会実装に向けて、ブラックボックス性という制約を解決するために、1）構造的特徴分析：推論に寄与する特徴量のリスト、2）言語的特徴分析：推論に寄与する各特徴量が対応する視覚属性を自然言語ラベルで表現したリスト、3）整合性分析：入力データ、推論（ラベル）、構造的特徴分析・言語的特徴分析の結果の間の整合性を測定する分析方法を開発した。提案手法は、深層学習モデルの内部で起こっている実際の推論過程を反映することで、深層学習モデルの透明性を高めた。説明文生成などの従来法はあるが、対象の深層学習モデルの推論過程を正確に反映することには主眼を置いていなかった。一方、提案手法は、可読性の高い結果を得るために従来法が用いていた、LSTM などのブラックボックスな手段は一切含んでいない。次に、実験により、定性的・定量的に分析結果を評価した結果、提案手法は深層学習モデルの透明性を向上させることを確認した。12,800 件の人間によるタスク（クラウドソーシング）で評価した結果、75% の作業者が入力データと特徴解析の結果が整合すると回答し、70% の作業者が推論（ラベル）と特徴解析の結果が整合すると回答した。これにより、推論根拠解析の提案手法は、人間が理解可能な推論根拠を出力しており、人間への説明として利用可能と類推される。また、提案手法である深層学習モデルの推論過程の解析方法の評価に加えて、分析結果の活用方法も検討した。深層学習モデルをどのように拡張するか（階層やパラメータ数）や、どの

[*1] 本章の研究成果は Kuwajima, Tanaka, and Okutomi (2019) [57] に基づく。ただし、(Kuwajima et al., 2019) [57] の著作権は Springer Nature Customer Service Centre GmbH に帰属する。

ような訓練データを追加収集するかなど、深層学習モデル改善に向けたエンジニアリングを見出すために、分析結果を活用できる可能性を示した。

本章の貢献は次の 3 点である。

1. 構造的特徴分析、言語的特徴分析、整合性分析から成る深層学習モデルの推論過程の分析方法を提案した。
2. 提案手法が深層学習モデルの透明性を向上させることを、実験と定性的・定量的な考察で示した。
3. 提案手法による深層学習モデルの推論過程の解析結果の活用方法を検討した。

本章の本節以降の構成は以下のとおりである。まず、2.1 で背景を述べる。2.2 節で DARPA の Explainable Artificial Intelligence プログラム（XAI）などで研究された深層学習の説明技術や可視化技術を紹介する。2.3 節では、深層学習モデルとして CaffeNet を、特徴量として conv5 を例に、CNN の推論を実行して特徴量の挙動（活性化の大きさ、特徴量 ID と視覚属性の関係）を把握する。ただし、注目する中間特徴量のインデックスを特徴量 ID と呼ぶ。2.4 節では、深層学習モデルの推論過程の分析方法（構造的特徴分析、言語的特徴分析、整合性分析）を提案する。2.5 節では、様々なデータに対して推論過程の分析を行う実験を行い、人間に解釈可能な分析結果を導出できることを定量的に評価し、2.6 節では、推論の正誤、物理整合率・論理整合率・推論整合率の高低の様々なバリエーションで分析結果例を詳細に議論する。分析結果から何が読み取れるか、深層学習モデルのエンジニアリングの示唆（データが足りない、ラベルが不足している、アノテーションが誤っている、など）を定性的に評価する。最後に、2.7 節で、深層学習モデルの推論根拠解析の研究をまとめる。

## 2.1 背景

深層学習モデルなどの深層学習技術により、人手で構築したルールベースのロジック [83, 66, 77, 106, 34, 98] ではなく、大量のデータから学習したモデルで、高度な環境認識や意思決定を行うシステムが普及している。深層学習の中で、特に深層学習は高い性能を達成しており、例えば物体認識では人間の正答率を超えている [1]。

自動運転システムのように、安全性が重視される環境認識や意思決定が必要なシステムでも、深層学習モデルは必須要素になりつつある [12]。深層学習モデルに高い信頼性を持たせるためには、高い性能と透明性の両方が重要である。特に、セーフティクリティカルなシステムには、透明性が求められる [60]。しかし、ニューラルネットワークなどの深層

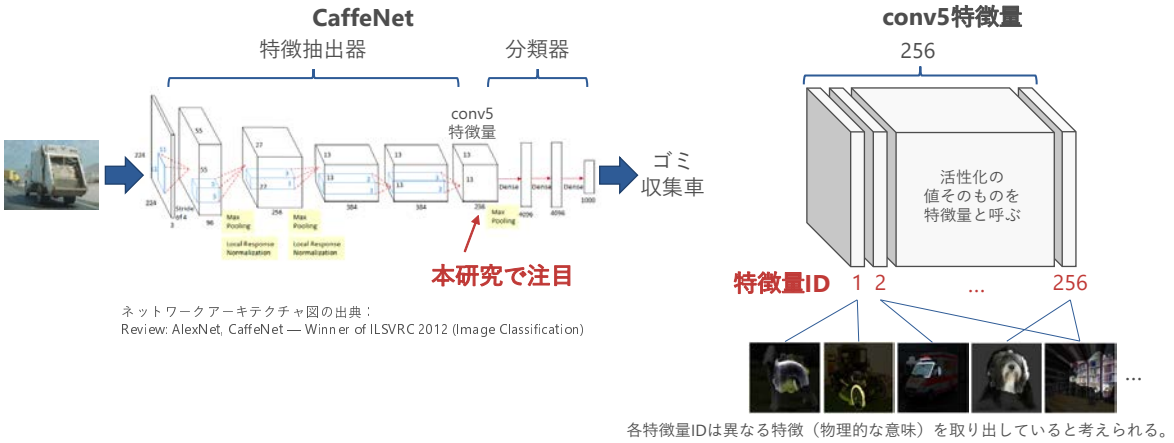



図 2.1: 特徴量 ID

学習モデルの推論過程は、ブラックボックスとみなされている。本章でいうブラックボックスとは、特徴の活性化は観測できるが、実際の現象が把握できない状態を指す。深層学習モデルは、性能は高いが透明性が低いため、自動運転のように深層学習モデルの結果が直接危険を引き起こす可能性があるセーフティクリティカルシステム [53] に、ブラックボックスである深層学習をそのまま適用することは難しい。

説明可能な AI（XAI）は、近年注目され急速に進展している関連研究分野である [69]。XAI 分野には、推論ネットワークが推論（ラベル）だけでなく、人間に理解しやすい説明を生成する研究がある。例えば、画像のキャプション生成や視覚説明（visual explanation）は、人間にとって理解しやすい自然言語による説明を行うための問題である。キャプション生成とは、入力画像に写っている物や状況を自然言語文によって記述する言語化手法である [105, 100]。視覚説明とは、LSTM などの説明対象の深層学習モデルとは別の、深層学習によるブラックボックスな説明モデルを用いて、深層学習モデルの推論の根拠を説明する [41]。説明モデルによる説明文生成と、説明対象の深層学習モデルによる分類は、異なるニューラルネットワークによって実行される。つまり、これらの先行研究では、説明対象の深層学習モデルの推論過程を必ずしも反映しない仕組み（説明モデル）を用いて、人間にとって可読性の高い説明文を生成している。また、説明文を生成する深層学習モデル（説明モデル）は、性能は高いが透明性が低い。そのため、結局、説明モデルの深層学習モデルにも新たなブラックボックス性の問題が存在してしまう。

そこで本章では、深層学習のブラックボックス性を解決するために、深層学習モデルの一例である畳み込みニューラルネットワーク（以下、CNN） [30, 55] の推論過程の透明性を向上させる分析手法を開発する。注目する中間特徴量（CNN の特徴マップ）のイン

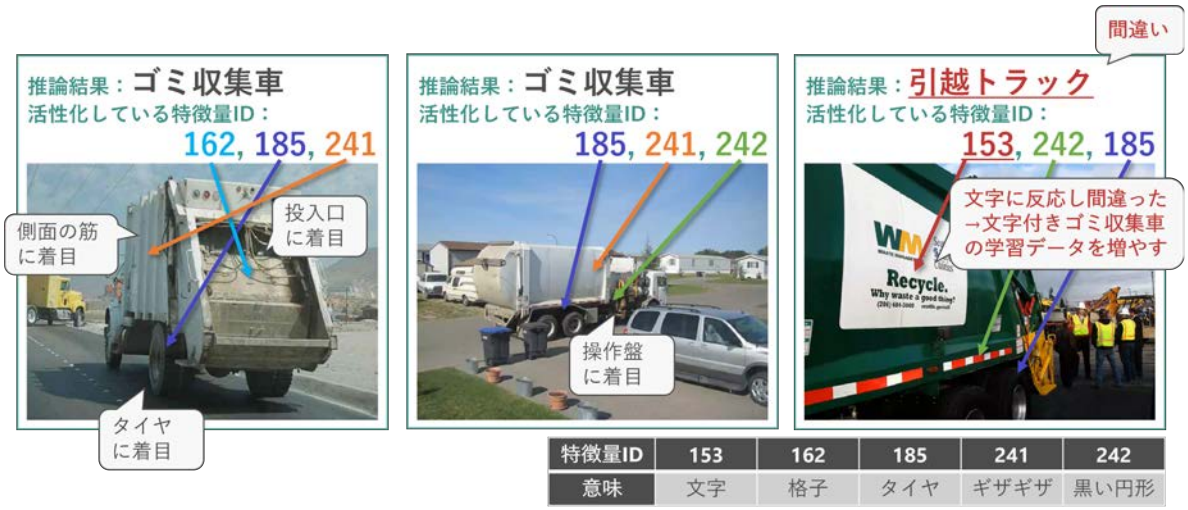



図 2.2: 深層学習における推論根拠の解析イメージ

デックスを特徴量 ID、活性化の値そのものを特徴量と呼ぶ（図 2.1）。各特徴量 ID は、異なる特徴（物理的な意味）を取り出していると考えられる。本章では、画像認識を例に、画像分類モデル CaffeNet とその特徴抽出器の最終段 conv5 特徴量に注目し、深層学習モデルで推論するにあたって最も貢献した特徴量の特徴量 ID を、推論根拠として提示する。図 2.2 に、深層学習における推論根拠の解析イメージを示す。同じ推論結果でも活性化している特徴量 ID が異なり、どこで間違ったかわかれば、モデル修正のヒントとなると考えられる。

本研究では、推論過程の分析として、1）構造的特徴分析、2）言語的特徴分析、3）整合性分析の 3 種類を検討する。構造的特徴分析とは、推論に寄与する特徴量 ID リストである。特徴量 ID は CNN の各フィルタが何を学習したかを意味するが、深層学習の過程で自動的に振られる ID（マジックナンバー）である。可読性は低いが、テスト時にシステムがプログラム的に推論過程を管理するのに有用である。一方、言語的特徴分析とは、構造的特徴分析で扱う各特徴量 ID について、視覚属性（意味）を記述した自然言語ラベルを付与したものである。この言語的特徴分析は、推論過程を人が理解するのに有効である。図 2.3 は、提案する特徴分析の例である。図 2.3a が入力画像、図 2.3b が推論（ラベル）を表し、図 2.3c は、図 2.3a から図 2.3b を推論するときの特徴量の分析である。図 2.3c の左列が構造的特徴分析、右列が言語的特徴分析である。最後に、整合性分析とは、入力画像、推論（ラベル）、前述の特徴分析の結果、の間の一貫性を測定するものである。整合性分析は、推論（ラベル）が間違っている場合の原因分析や、問題点を解決するための取りうるアクションの検討など、深層学習モデル改善の議論に役立てることができる。

(a) 入力画像

sorrel

(b) 推論（ラベル）

| Feature# | Visual attributes |
|---|---|
| 170 | animal legs |
| 132 | human legs, animal legs or beige |
| 218 | animals, furs or brown |

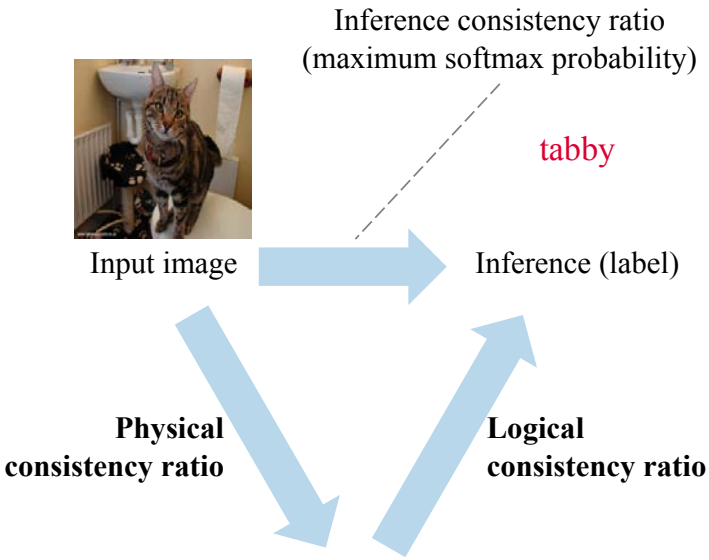
(c) 構造的特徴分析（左列）と言語的特徴分析（右列）

170 132 218

(d) 参考情報として提示する受容野（ネットワークが見ている領域）

図 2.3: 推論過程の透明性を向上させる特徴分析の例

図 2.4 に整合性分析の概念を示す。




| Feature# | Visual attributes |
|---|---|
| 155 | leopard patterns, faces of small animals, furs or two-tone brown/white |
| 56 | animals, furs or two-tone black/brown |
| 86 | furs or two-tone black/gray |

図 2.4: 入力画像、特徴量、推論（ラベル）の間の整合性分析

## 2.2 関連研究

深層学習モデルのブラックボックス性に対処するためには、人間が理解できる仕組みで、モデル内部から特徴量を抽出して推論根拠を解析し、推論結果と共に出力することが望ましい（図 2.5）。望ましい推論根拠解析の条件は以下にまとめられる。

1. 推論根拠の解析のために、モデルを変更しない。
2. 推論根拠の解析は、モデルの特徴量に基づく。
3. 推論根拠の解析方法は、ブラックボックスではない。
   （解析の過程を人間が理解できる。）
4. 推論根拠の解析方法は、学習データなどを用いて構築する。

（テストデータは用いない。）

5. 推論根拠の解析方法は、テストデータを用いて評価する。

しかし、入力データのみ着目する、新たなブラックボックス性を含むなど、関連研究は望ましい推論根拠を満足していない。

DARPA は 2017 年に、深層学習の説明技術を開発する Explainable Artificial Intelligence プログラム（XAI）を開始した [37]。XAI は 3 つの説明アプローチ、Deep Explanation、Interpretable Models、Model Induction を定義している。Deep Explanation と Interpretable Models は、説明対象の深層学習モデルの学習前に説明性の高い特徴や説明可能な因果関係モデルを設計する、事前アプローチである。Model Induction は、説明対象の深層学習モデルの学習後に新たな説明性の高いモデルを自動的に生成する、事後アプローチである。本章の研究も事後アプローチに分類されるが、XAI の Model Induction と異なる点は、説明のために新しいモデルを生成せず、説明対象の深層学習モデルの内部で実際に観測された活性（activation）を、直接分析する点である。

深層学習モデルの可視化は、近年活発に研究されている分野である [36, 71, 48]。先行研究では、基本的に入力画像の注目領域（attention）を、受容野やヒートマップで可視化している [90, 9, 70]。この可視化は、深層学習モデルが注目している入力画像中の領域を示している [11]。しかし、入力画像の注目領域とは、CNN 推論処理のごく初期段階の部分であり、可視化手法は CNN 推論処理の初期段階を明らかにしているに過ぎない。一方、本章では、入力画像だけでなく、深層学習モデルの推論過程についても分析を行うことを狙う。本章では、入力画像中の注目領域を後述の視覚属性（推論結果の材料になる視覚パターン）として示すことのみを目的に、参考情報として受容野を利用することとした。

入力空間に着目した深層学習モデルの可視化技術の他に、視覚属性と中間特徴（深層学習モデルの各ノードの活性化）の関係に着目した研究も存在する。先行研究では、黒、茶、毛皮といった視覚属性が、深層学習モデルのノードと関連していることが報告されている [26]。他の先行研究では、人間によって解釈可能な視覚属性と結びついている深層学習モデルのノード数をもって、解釈可能性を定量化することを試みている [8]。本章では、深層学習モデルの透明性向上の手段のひとつとして、視覚属性を用いた。

PJ-X（Pointing and Justification-based Explanation）は、DARPA XAI で取り組まれた説明手法のひとつである [75]。PJ-X は、入力空間の一部（つまり入力画像のある領域）に注目し、内省的説明（説明対象の深層学習モデルの推論過程に基づいた真の説明）と正当化説明（可読性の高い説明）を同時に行うことで、理解度の高い説明を行う。前者は入力空間の説明（前述の可視化と同様）を行うが、深層学習モデルの推論過程の分析は

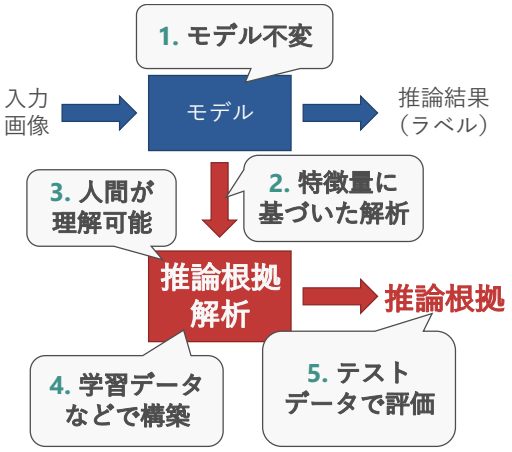



図 2.5: 望ましい推論根拠解析の条件

行わない。後者は、説明の生成に別のブラックボックスな深層学習手法である LSTM を用いる。PJ-X は、推論結果と深層学習モデルの特徴（活性、発火）との関係を分析しておらず、説明のために新たなブラックボックスな深層学習モデルを利用するため、本章の目的である透明性向上は行えない。

## 2.3 活性化パターンに着目した特徴量の観測

CaffeNet [24, 55] の conv5 特徴量を、ImageNet の学習データで観測し、特徴量の挙動を把握した。ImageNet には各クラス約 1,300 枚の学習画像がある。簡単のため、正解クラスの softmax 確率（softmax 層の値が最大であるもの、つまり softmax 確信度）が高い順で各クラス上位 100 枚の学習画像を選択し、実験に用いた。

まず、推論（ラベル）は活性の大きい特徴量に基づくという仮定を置いた。この仮定は、特に活性化関数 ReLU（CaffeNet で利用）に当てはまる。なぜなら、ReLU は正の単調関数であるから、活性の大きい特徴量が次の層に影響を与えるからである。

**仮定 1.** 推論過程で活性化している特徴量 ID は、推論（ラベル）に寄与している。

### 2.3.1 特徴量のスケール不定性

次に、conv5 の各特徴量 ID の活性化を調べると、特徴量 ID により活性化の大きさの範囲が異なることがわかった。したがって、特徴量 ID によって活性値のスケールは不定である。図 2.6 は、平均値が最も小さい特徴量 ID 94 と最も大きい特徴量 ID 22 の活性

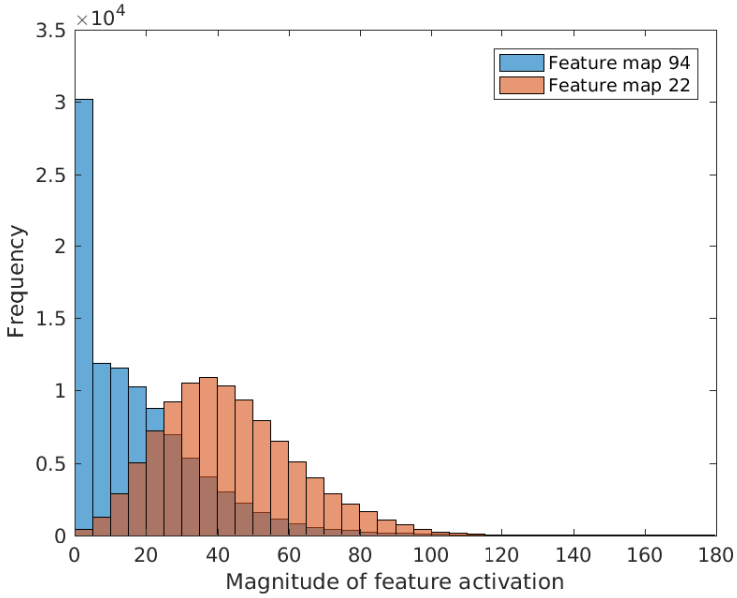



図 2.6: 特徴マップによって異なる活性化分布の観測

度の頻度を示したものである。特徴量 ID 94 の活性値の最頻値 0 に対して、特徴量 ID 22 の活性値の最頻値は 38 である。また、2 つの特徴量 ID で、特徴量の分布形状も異なる。ただし、活性化関数 ReLU によって負の値が無視されるため、活性値は正規分布していない。このように、特徴量 ID 94 と 22 では活性値の分布が異なることがわかる。以上の分析から、次の仮定を得る。

**仮定 2.** 異なる特徴量における活性値は、異なる値域を持つ。

### 2.3.2 特徴量 ID と視覚属性

図 2.7 では、3 つの視覚属性、毛皮状、ゴムタイヤ状、微細なセル状の模様と、2 つの特徴マップ 226 と 230 が対応していることを示す。左列は furly（毛皮状の模様）、右列は rubber tires（ゴムタイヤ状の模様）と fine cell patterns（微細な細胞状の模様）の視覚属性が、特徴マップ（特徴量 ID）226、230 に関連づいている。これら 2 つの特徴マップは、毛皮状という視覚属性を共有しているが、他の視覚属性にも対応している。このことから、次のような仮定を得る。

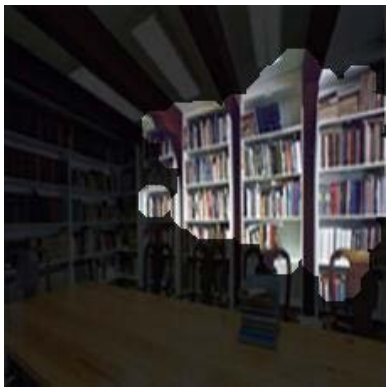
**仮定 3.** 特徴量 ID と視覚属性は多対多の関係にある。

furly

rubber tires

(a) 特徴マップ 226

furly



fine cell patterns

(b) 特徴マップ 230

図 2.7: 特徴マップと視覚属性の多対多関係

## 2.4 提案手法

本節では、前述の 3 つの仮定に基づいて、深層学習モデルの推論過程の透明性を向上させる特徴量の解析手法を提案する。2.4.1 節では、学習完了後と推論時に特徴解析を行い、3 種類の特徴量 ID を得る。2.4.2 節では、特徴量 ID と視覚属性を関連付けるため、手動で特徴量アノテーションを行う。2.4.3 節では、入力画像、特徴解析結果、推論結果（ラベル）の 3 点の整合性をクラウドソーシング（多数の被験者へのアンケート）により評価する。

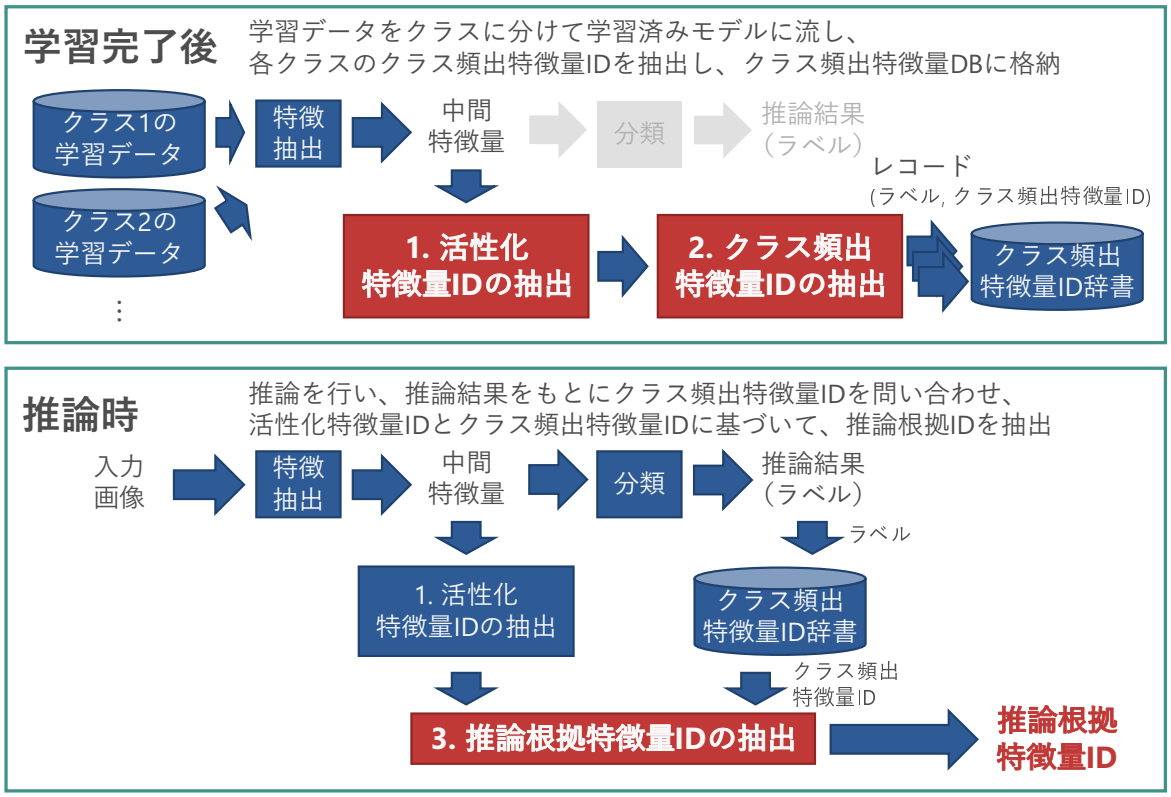



図 2.8: 推論根拠解析のアプローチ

### 2.4.1 特徴量の活性化に基づいた推論根拠の説明（構造的特徴分析）

構造的特徴分析では、以下の考え方に基づいて、入力画像から決まる活性化特徴量 ID と、推論結果から決まるクラス頻出特徴量 ID から、推論根拠特徴量 ID を抽出する。

**活性化特徴量 ID**

特徴量 ID の特徴量の値が大きければ、その特徴量 ID は活性化したと考える。

**クラス頻出特徴量 ID**

各クラスの標本を推論するとき、頻繁に活性化する特徴量 ID を、そのクラスを表す特徴量 ID と考える。

**推論根拠特徴量 ID**

クラス頻出特徴量 ID の中で、実際に活性化した特徴量 ID を、推論の根拠と考える。

図 2.8 に構造的特徴分析のアプローチ全体を示す。構造的特徴分析は、図 2.9、図 2.10、

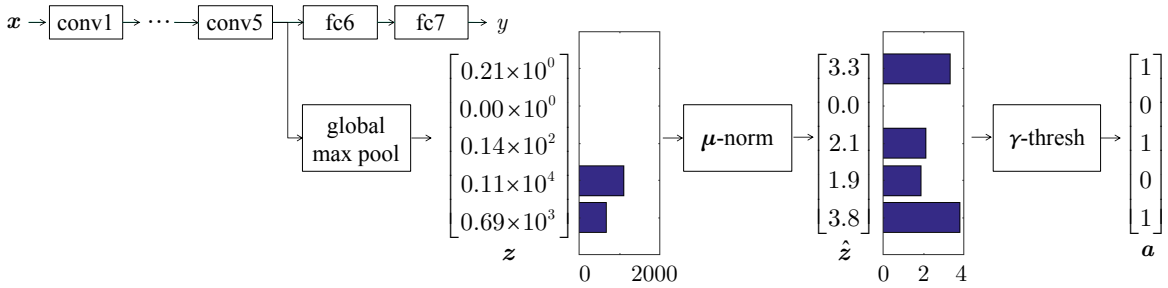



図 2.9: 活性化特徴量 ID

図 2.14 に示す 3 つの概念、1）活性化特徴量 ID、2）クラス頻出特徴量 ID、3）推論根拠特徴量 ID で構成される。ただし、活性化特徴量 ID と推論根拠特徴量 ID は推論ごとに抽出され、クラス頻出特徴量 ID はクラスごとに抽出される特徴量 ID である。

本研究では例として、CaffeNet の最も深い層にある最終畳み込み特徴量（空間次元を持つ特徴量）である conv5 中間特徴量の活性化に着目する。なお、AlexNet の最終畳み込み特徴量 conv5 は、物体や部品などの人間が解釈可能な高度な視覚概念を学習（獲得）していることが知られている [8]。AlexNet と類似の CaffeNet の最終畳み込み特徴量も、同様な視覚概念を学習していると考えられる。CaffeNet の入力画像を $\boldsymbol{x}$、出力を $\boldsymbol{y}$ とし、$\boldsymbol{x}_i^{\text{train}}$、$\boldsymbol{y}_i^{\text{train}}$ を学習データ、$\boldsymbol{x}^{\text{test}}$ をテストデータとする。

特徴量 ID の特徴量の値が大きければ、その特徴量 ID は活性化したと考え、活性化特徴量 ID を定義する。**活性化特徴量 ID a**（図 2.9）として、conv5 から生成する二値ベクトルを定義する。CNN の特徴マップは空間次元を持つが、今回の推論根拠の解析では特徴量の活性化の位置ではなく大きさに着目するため、活性化特徴量 ID $\mathbf{a}$ の抽出において空間次元を無視する。本研究では、テンソルからベクトルを得る最も単純な方法として、$13 \times 13 \times 256$ テンソルである conv5 を 256 次元の特徴ベクトル $\mathbf{z}$ に変換するために、global max pooling を用いた。つまり、空間次元を持つ特徴マップの一枚を、1 つの特徴量として捉え、空間次元を縮約する。次に、ベクトル $\mathbf{a}$ の要素は、関連する特徴量 ID、すなわち conv5 の特徴マップが活性化していれば 1 となるようにしたい。**仮定 2** によると、特徴量の活性値のスケールは不定であり、異なる特徴間の差異を捉えるためには、平均や分散などの統計情報を用いて正規化することが必要である。そこで本節では、平均値による正規化と閾値処理を用いることにした。つまり、$\mathbf{z}$ の各要素でスケールが異なるので、正規化して比較可能とするため、特徴ベクトル $\mathbf{z}$ から平均正規化特徴ベクトル $\hat{\mathbf{z}}$ を計算する。次に、$\hat{\mathbf{z}}$ を閾値 $\gamma$ で閾値処理すると、$\mathbf{x}$ に対応する活性化特徴量 ID（二値ベクトル）$\mathbf{a}$ を得る。

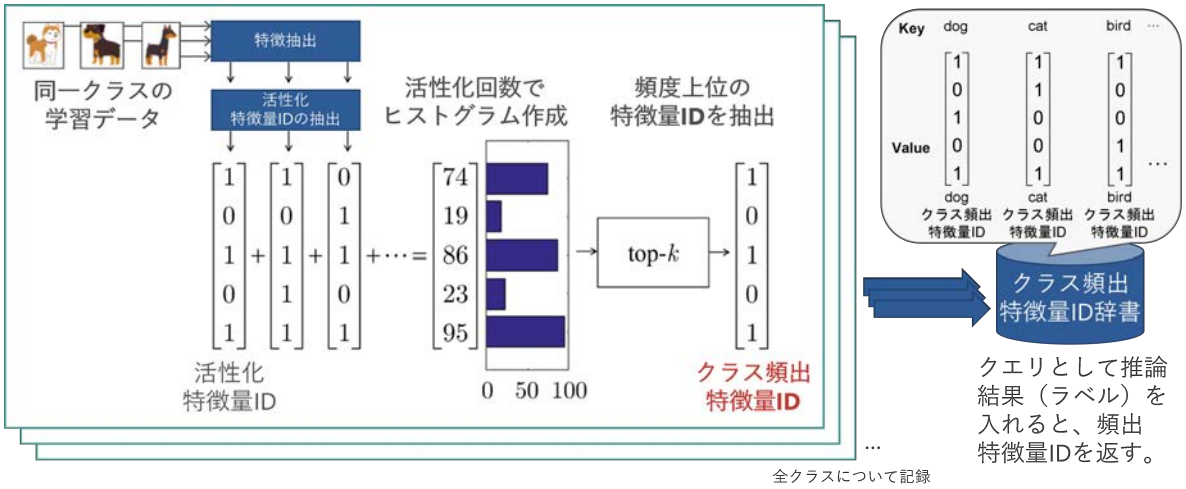



図 2.10: クラス頻出特徴量 ID

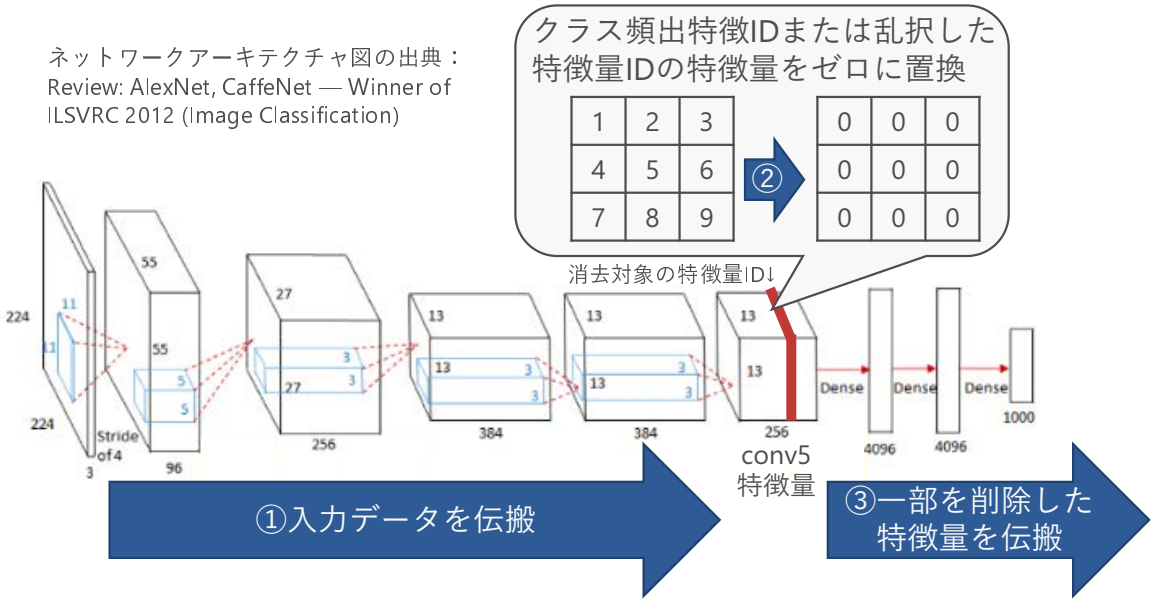



図 2.11: 特徴マップ削除モデル

各クラスの標本を推論するとき、頻繁に活性化する特徴量 ID を、そのクラスを表す特徴量 ID と考え、クラス頻出特徴量 ID を定義する。**クラス頻出特徴量 ID q**（図 2.10）は、各クラスの頻出特徴量 ID を示す二値ベクトルである。各クラスはそれぞれ異なる固有の頻出活性化パターン（あるクラスを推論結果として出力するときに高頻度で活性化する特徴量 ID）を持つと仮定し、以下の手順で算出する。図 2.10 は、クラス dog を例としたクラス頻出特徴量 ID の算出例である。学習完了後に、学習データ全体から取り出

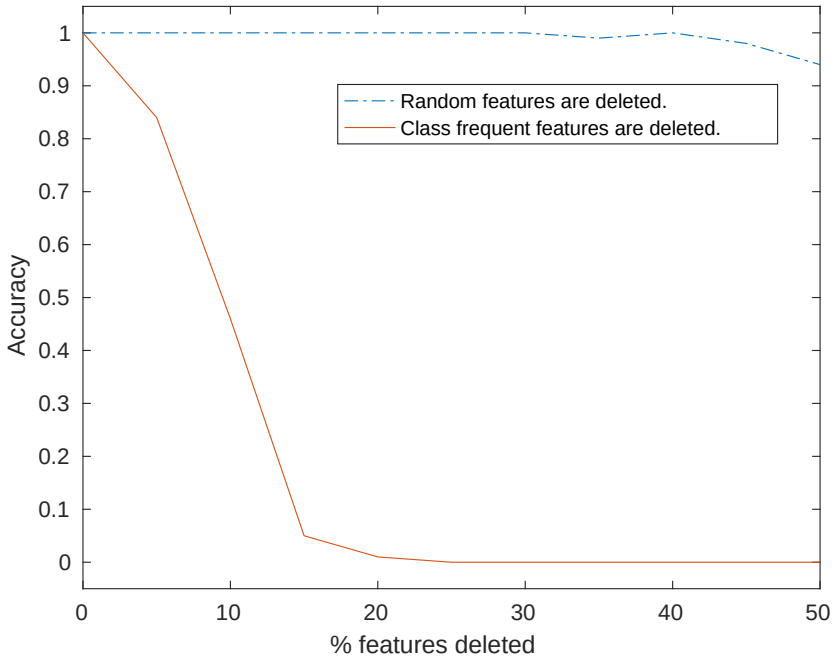



図 2.12: 特徴マップ削除による精度低下の観測

した dog クラスの学習データ $\boldsymbol{x}_i^{\text{train}}$ を二値化し、活性化特徴量 ID $\boldsymbol{a}_i^{\text{train}}$ を得る。$i$ について $\boldsymbol{a}_i^{\text{train}}$ の和を取ることで、dog クラスにおける各特徴量 ID の活性化回数（ヒストグラム）を得る。最後に、上位 $k$ 個の特徴量を抽出し、dog クラスのクラス頻出特徴量 ID とする（図 2.10 では $k = 3$）。学習完了後にすべてのクラスに対して上記の手続きを行い、クラス頻出特徴量 ID を抽出する。クラス頻出特徴量 ID を推論時に使用するため、$\boldsymbol{q}(\text{dog}) = [1, 0, 1, 0, 1], \boldsymbol{q}(\text{cat}) = [1, 1, 0, 0, 1], \boldsymbol{q}(\text{bird}) = [1, 0, 0, 1, 1]$ のようにクラス頻出特徴量 ID 辞書に保存する。ただし、クラス頻出特徴量 ID 辞書は、クエリとして推論結果（ラベル）を入力すると、頻出特徴量 ID を返す、ルックアップテーブル（参照テーブル）である。

クラス頻出特徴量 ID の有効性を確認するため、図 2.11 の要領で、CaffeNet において、乱択した特徴量 ID に対応する conv5 特徴量をゼロに置換した深層学習モデルと、クラス頻出特徴量 ID に対応する conv5 特徴量をゼロに置換した深層学習モデルを作成した。あるサンプルクラスに対して、それぞれの精度低下を図 2.12 に示す。クラス頻出特徴量 ID に対応する特徴量を消去すると、精度の急激な減衰が観測された。一方、乱択した特徴量 ID に対応する特徴量の消去に対しては、CNN の性能低下はわずかだった。この結果により、クラス頻出特徴量 ID に対応する特徴量の削除で性能が低下するため、クラス頻出特


図 2.13: 救急車クラスのクラス頻出特徴量 ID に対応する受容野（視覚属性）

徴量 ID は推論に貢献していると考えられる。一方、乱択した特徴量の削除では性能が低下しないため、CaffeNet など CNN には冗長な特徴量が含まれていると考えられる。

次に、クラス頻出特徴量 ID と推論（ラベル）の関係を理解することを目的に、クラス頻出特徴量 ID に対応する特徴マップについて、活性化しているノードが対応する入力画像の領域を可視化する。救急車クラスのクラス頻出特徴量 ID は 084、177、234、239、242 などである。そこで、救急車クラスのサンプル画像を入力に推論した際に、特徴量 ID 084、177、234、239、242 に対応する特徴マップの活性要素に対して、受容野を可視化し

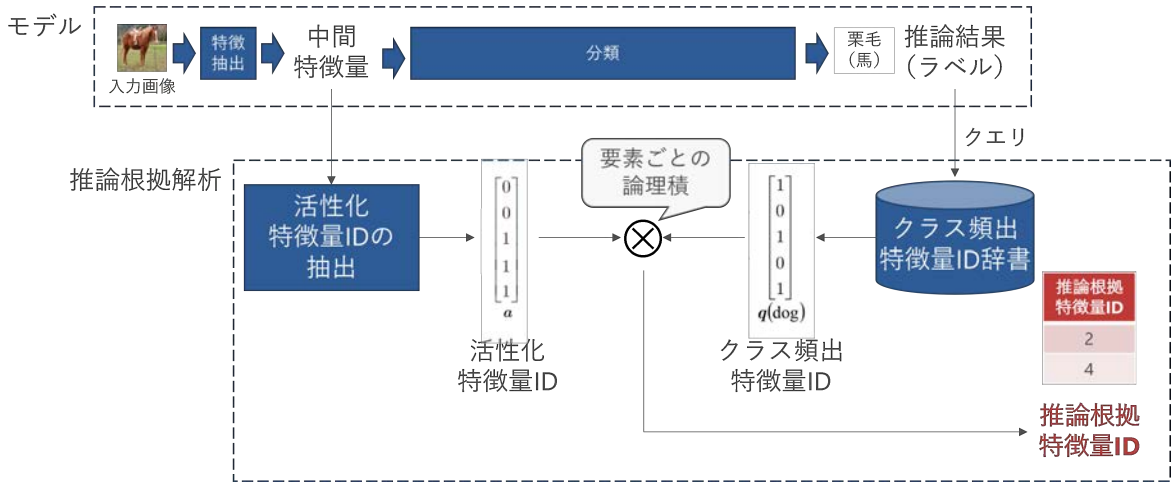



図 2.14: 推論根拠特徴量 ID

たものを図 2.13 に示す。左上から右上、左下から右下に、084、177、234、239、242 の特徴マップに対応している。ただし、受容野とは、各特徴マップの活性化した要素が対応する入力データの部分を可視化したものである。本研究では、簡単のために以下の手順で受容野を生成する。1）可視化対象の特徴マップについて、活性化特徴量 ID と同様の正規化と閾値処理により活性化特徴量要素を抽出して二値化する。2）その他の特徴マップ（特徴マップ 084 に対して受容野を生成する場合は、特徴マップ 084 以外のすべての特徴マップ）をゼロに置き換える。3）以上の変更を行った可視化対象の特徴マップとその他の特徴マップの両方を含む特徴量を、入力空間へ逆伝搬する。ただし、max pooling 層では、順伝搬の際に保存しておいた最大値位置を用いて、逆 pooling を行う。4）特徴量を入力空間へ逆伝搬した画像に対して、画像二値化と円形フィルタを用いた膨張処理などの後処理を行い、受容野を生成する。図 2.13 において、特徴量 ID 084 と 177 は白-赤（またはオレンジ）のツートンカラー、特徴量 ID 234 は窓、特徴量 ID 239 と 242 はタイヤに対応すると解釈でき、救急車の重要な視覚属性が受容野として観察された。深層学習モデルは、このような鍵になる視覚属性を基底にして推論を行っていると考えられる。この観察から、次の仮定を得る。

**仮定 4.** 推論（ラベル）のクラスに対して頻繁に活性化する特徴量 ID は、推論（ラベル）への貢献度が高い。

ある推論結果（ラベル）に対するクラス頻出特徴量 ID の中で、実際に活性化した特徴量 ID を、推論の根拠と考え、推論根拠特徴量 ID を定義する。**推論根拠特徴量 ID** $\boldsymbol{e}^{\text{test}}$

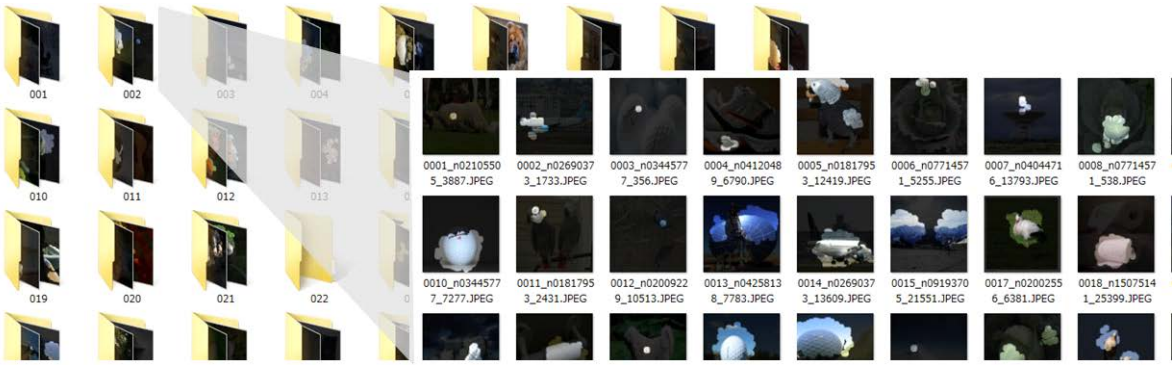



図 2.15: 作業者に示す特徴量アノテーション用データ（入力画像と受容野）

（図 2.14）は、テストデータ $\boldsymbol{x}^{\text{test}}$ に対する、活性化特徴量 ID $\boldsymbol{a}^{\text{test}}$ とクラス頻出特徴量 ID $\boldsymbol{q}^{\text{test}}$ の重なりと定義する。ただし、$\bigotimes$ は要素ごとの論理積を表す。**仮定 1** と**仮定 4** により、推論に寄与する特徴量 ID は、活性化特徴量 ID とクラス頻出特徴量 ID の両方に含まれる必要がある。図 2.14 の鎖線枠（モデル）は、特徴解析を行わない従来の推論である。図 2.14 の鎖線枠（推論根拠解析）のように、活性化特徴量 ID $\boldsymbol{a}^{\text{test}}$ は推論対象の標本 $\boldsymbol{x}^{\text{test}}$ を用いて計算する（ラベルは用いない）。クラス頻出特徴量 ID $\boldsymbol{q}^{\text{test}}$ の抽出にはラベルが必要だが、推論時には真値は得られない。そこで、CaffeNet の推論（ラベル）$y^{\text{test}}$ を採用して $\boldsymbol{q}(y^{\text{test}})$ を参照する。活性化する特徴量 ID の数は、入力画像によって変化するため、推論根拠特徴量 ID $\boldsymbol{e}^{\text{test}}$ の特徴ベクトルの要素数も、一般的には推論ごとに変化する。可読性を考え、推論根拠特徴量 ID $\boldsymbol{e}^{\text{test}}$ について、平均正規化活性度が大きい順に最大で上位 $\ell$ 個の特徴ベクトル要素のみを表示する。

### 2.4.2 特徴量 ID と視覚属性の関連付け（言語的特徴分析）

可読性のある分析結果を生成するため、それぞれの特徴マップが活性化した複数の入力画像を集め、それらの画像を説明対象の深層学習モデルに入力し、観測される視覚属性を人手でアノテーションする。つまり、各特徴マップが活性化したとき、深層学習モデルが何を見ているかを調べる。可読性のある視覚属性を実現する方法は様々あるが、最も単純な方法として、人間によるアノテーションを実施する。

**特徴量アノテーション用データ**は、学習データセットを用いて作成する。まず、学習データから適当な部分集合を選択する。次に、各特徴量 ID に対して、その特徴量 ID を含む推論根拠特徴量 ID を持つ画像を乱択する。ただし、学習データのラベルは既知であるため、特徴量アノテーションでは、真値ラベルを用いて推論根拠特徴量 ID を抽出でき

る。そのうえで、図 2.15 に示すような受容野を生成して人間のアノテーション作業者に提示し、アノテーションすべき視覚属性を指示する。各特徴量にフォルダを一つ作り、その特徴量が活性化した画像を配置する。ただし、画像には受容野を表示する。アノテーション作業者は、このフォルダ内の全画像の受容野を概観し、特徴量が意味する視覚属性をアノテーションする。

**特徴量アノテーションの工程**として、単一の視覚属性を表す複数の特徴量 ID の組み合わせと、複数の視覚属性の組み合わせを表す単一の特徴量 ID を、繰り返し作業で洗練する。**仮定 3** に基づいて、このような視覚属性と特徴量 ID の多対多関係をアノテーションするため、自由記述によるアノテーションから始めて繰り返し精緻化するため、1）オープンアノテーション、2）概念整理、3）クローズドアノテーションの 3 段階からなるアノテーション工程を定義する。図 2.16 に示すように、概念整理とクローズドアノテーションのステップを繰り返し、特徴量アノテーションを洗練させていく。オープンアノテーションでは、各特徴量 ID に対応する視覚属性を自由記述でアノテーションする。このステップは自由記述であるため、人間によるアノテーションにばらつきが生じる。次に、概念整理のステップでは、類似の視覚属性の統合、同じラベルをつけていたが実際は異なる視覚属性の分割、分割により必要となった視覚属性の新設（ラベル定義）などを行い、アノテーションのばらつきをおさえる。クローズドアノテーションでは、概念整理で定義した視覚属性（ラベル定義）に基づいて、各特徴量 ID に対応する視覚属性を選択式でアノテーションする。**仮定 3** に基づいて、一つの特徴量 ID に対して、複数の視覚属性のアノテーションを許容するため、視覚属性のラベルは可変長となる。

本研究の範囲では、特徴量アノテーションを単体で妥当性を検証することはしていない。推論根拠解析のロジックと特徴量アノテーションを併せて、2.4.3 節に示す通り、クラウドソーシングで妥当性を評価する。

### 2.4.3 入力画像・特徴量・推論の整合性分析

入力画像と推論（ラベル）、そして特徴解析（提案手法）の結果、すなわち推論根拠特徴量 ID と対応する視覚属性の整合性を測定する。この計測を行うことによって、提案手法が出力する推論根拠の妥当性を判断する。

推論根拠特徴量 ID と入力画像の整合性である**物理整合率**（physical consistency ratio, PCR）と推論根拠特徴量 ID と推論（ラベル）の整合性である**論理整合率**（logical consistency ratio, LCR）を提案する（図 2.4）。この 2 つの整合率は、人間による作業（クラウドソーシングによるアンケート）を通して評価される。一方、入力画像と推論（ラベル）


図 2.16: 特徴量アノテーションの繰り返し工程

の整合性については、推論（ラベル）のクラスに対応する softmax 確率、すなわち最大 softmax 確率を推論整合率（inference consistency ratio, ICR）として用いる。すべての比率は 0.0 から 1.0 の範囲の値をとる。

## 2.5 提案手法の妥当性

本節では、実験を通して、提案手法である特徴量解析方法を検証する。深層学習モデルの一例として、一般公開されている CaffeNet の推論過程を分析する。ただし、CaffeNet には ImageNet で事前学習したパラメータを搭載し、新たにモデルを学習しパラメータを更新しない。活性化特徴量 ID は前述の方法（閾値 $\gamma = 2$）で求める。クラス頻出特徴量 ID の特徴ベクトル要素数は $k = 5$、推論根拠特徴量 ID の特徴ベクトル要素数の最大値は $\ell = 3$ とする。推論過程の分析を補足する参考情報、および、人間による特徴アノテーションのデータとして、推論根拠特徴量 ID に対応する受容野を解析結果に付加する（図 2.3d）。

特徴量活性化の分析と同様に、1 クラスあたり 100 枚の学習画像を選び、conv5 の各特徴マップの平均値の計算、クラス頻出特徴量 ID の抽出、人間による視覚属性のアノテーションを行う。一方、人間（推論過程の解析結果を見るユーザーや特徴アノテーションの作業者）が ImageNet の 1,000 クラスを識別し、分析結果の理解や特徴量 ID の整合性判

表 2.1: 物理整合率と論理整合率を評価するための人手作業のタスク数

| 評価尺度 | クラス数 | サンプル数 | 作業者数 | 合計タスク数 |
|---|---|---|---|---|
| 物理整合率 | 32 | 10 | 20 | 6,400 |
| 論理整合率 | 32 | 10 | 20 | 6,400 |

断を正確に行うことは困難であるため、ImageNet の 1,000 クラスを 32 クラスに削減した。32 クラスは ImageNet の 1,000 クラスのサブセットで、WordNet の階層構造に従って、各クラスがほぼ同じ数の WordNet synsets を持つようにプログラムで自動選択した。

整合性分析のための人間による評価は、Amazon Mechanical Turk で行った。各入力画像の推論に対して行った特徴量解析の結果について、物理整合率と論理整合率を評価するための 2 つの質問を行った。

1. 推論根拠特徴量 ID に対応する視覚属性は、入力画像の全体と部分の、いずれかに関連する。
2. ある物体が推論根拠特徴量 ID に対する視覚属性を持つ場合、その物体は推論クラス（ラベル）の物体である。

1 問目は、推論（ラベル）を見せずに入力画像と推論根拠特徴量 ID（自然言語で記述した視覚属性と対応する受容野）だけを提示し、2 問目は、入力画像を見せずに推論根拠特徴量 ID（自然言語で記述した視覚属性）と推論（ラベル）だけを作業者に提示し、質問した。つまり、1 問目は画像と文章で構成される質問で、2 問目は文章だけの質問である。作業者が回答で選べる選択肢は、「強くそう思う」「そう思う」「そう思わない」「強くそう思わない」の 4 つで、回答取得後、前者 2 つと後者 2 つを統合し、それぞれ賛成と反対として扱った。この質問結果をもとに、物理整合率と論理整合率をそれぞれ評価する。各質問は、個人の偏りをなくすため、異なる作業者にアノテーションのタスクを重複して割り当てて、平均を求めて賛成の比率を得た。各比率は 0.0 (すべての作業者が反対) から 1.0 (すべての作業者が賛成) の範囲の値をとる。これらの比率を評価するために、合計 12,800 件の人間によるタスクを実施した（表 2.1)。また、推論クラス（ラベル）の softmax 確率を推論整合率として記録した。

クラウドソーシングによる評価の結果、実験データ全体について、物理整合率、論理整合率、推論整合率（softmax 確率の平均値）は、それぞれ 0.75、0.70、0.48 だった（図 2.17)。深層学習モデルによる推論の確信度（推論整合率、0.48）よりも、推論過程の分

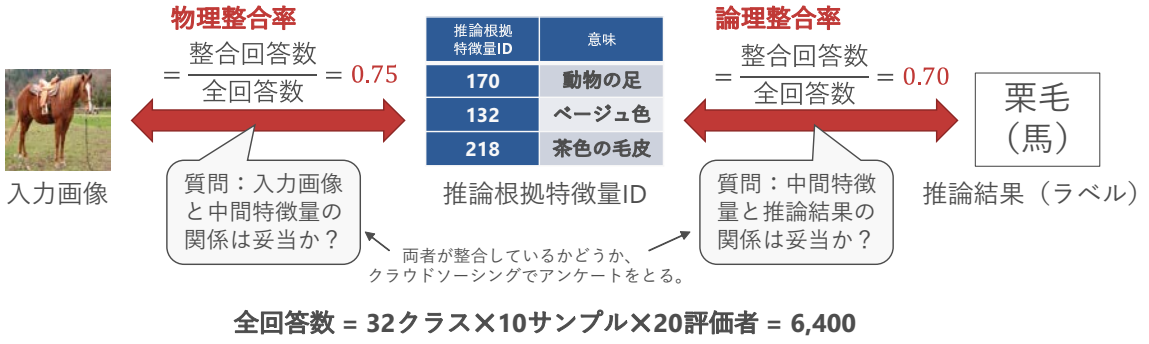
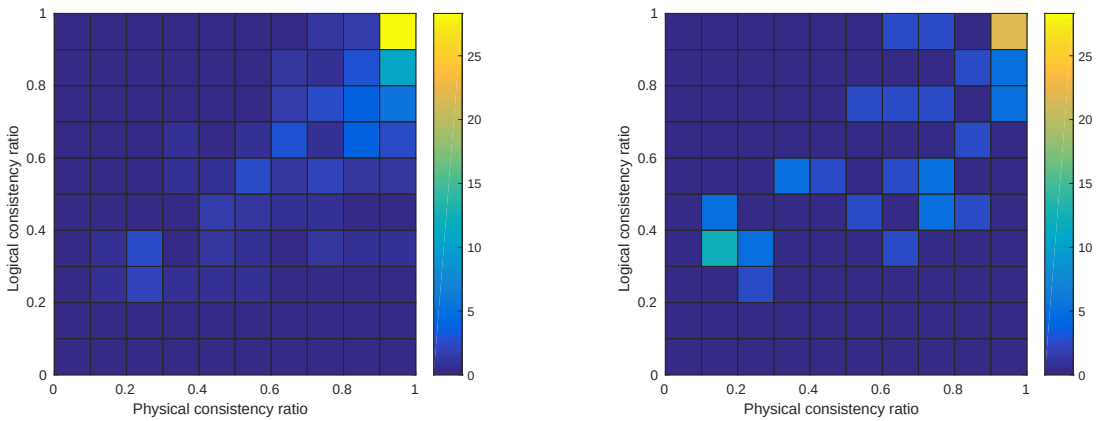



図 2.17: 整合性分析の結果




(a) 推論が正しいときの分布　　(b) 推論が間違っているときの分布

図 2.18: 整合性分析の結果得られた物理整合率と論理整合率の（離散）同時分布

析（物理整合率、論理整合率）が高いことがわかる。物理整合率と論理整合率は人間による評価であるため、本分析手法が生成した推論根拠について、人間によるコンセンサスが得られたと考えられる。提案手法の推論根拠の解析は、0.70 以上の人間が理解可能な推論根拠を出力しているため、提案手法は、人間への説明としてある程度利用可能と類推される。

図 2.18a と図 2.18b は、推論が正しい場合と間違っている場合の、物理整合率 (a) と論理整合率 (b) の同時確率分布（離散）を示している。異なる標本画像と異なる作業者による評価を行ったことにより、確率分布が得られた。推論が間違っている場合について、

ピークは低く、物理整合率と論理整合率はともに低い値から高い値まで全体に広がっていた。一方、推論が正しい場合には、両比率とも高くなる傾向があり、推論が間違っている場合の分布と比較してコントラストが高かった。したがって、推論が正しい場合の方が、推論が間違っている場合よりも、より良い（論理的にも物理的にも整合している場合が多い）分析結果が得られると言える。

## 2.6 特徴量 $\leftrightarrow$ 入力画像・推論の整合性に注目した考察

本節では、提案手法である深層学習モデルの推論過程の分析により、CNN の推論過程の透明性が向上することを示すとともに、深層学習モデルのさらなる精度向上のために、提案手法による分析結果により、どのようなデータ収集や深層学習モデル訓練の方策が可能になるか議論する。

現在、開発中の（学習とテストを繰り返して性能向上している）CNN があると想定する。現在学習中の CNN による推論（ラベル）と、提案手法による分析結果を用意する。図 2.19 と図 2.20 は正しい推論、図 2.21 と図 2.22 は誤った推論に対する分析結果である。ただし、図 2.19 および図 2.20 では、左から右に物理整合率（PCR）が増加し、下から上に論理整合率（LCR）が増加するように実験結果を配置した。図 2.19、図 2.20、図 2.21、図 2.22 の各画像は、CaffeNet への実際の入力画像サイズ $227 \times 227$ に変換されており、受容野は省略されている。

最初に、正しい推論結果に対する分析結果を考察する。

図 2.19 左列の画像で、物理整合率と推論整合率の両方が低い分析結果は、推論根拠特徴量 ID の特徴ベクトル要素数が最大値 $\ell = 3$ 未満になっていた。推論根拠特徴量 ID の数、つまり提示される根拠が少ないため、人間の作業者は、図 2.19 左列の画像の物理整合率を低く評価したと考えられる。推論根拠特徴量 ID に含まれる特徴ベクトル要素数が少ない場合は、softmax 確率の最大値である推論整合率も低いと考えられる。また、推論根拠特徴量 ID から得られる情報が少なければ、深層学習モデルの推論結果に対する革新も失われると解釈できる。このように、本章で仮説した推論過程が、実際の深層学習モデルの推論過程から大きく外れていないことが示唆された。

図 2.19 右列の画像は、物理整合率が高く、推論整合率が低い。一番下の画像は、視覚属性のラベルが適切でないため、論理整合率が低いと考えられる。カセットプレーヤーには左右に大きなスピーカーがあるはずで、特徴量 ID 196 と 171 はそれを表現している可能性がある。しかし、これらの特徴量 ID のラベル（視覚属性）は、 ゴムタイヤや丸みを帯びたもの（形状）であり、人間の作業者が特徴マップのラベル（視覚属性）とスピーカー

を結びつけるのは容易ではない。言語的特徴分析をより理解しやすくするため、視覚属性のラベル定義を改善する余地がある。

図 2.20 右上の画像は、物理整合率、論理整合率、推論整合率のすべてが高い。言語的特徴分析では、1）形状（細かい格子模様、細かい箱・ 丸の集積、ヒョウ柄)、2）色（赤・白のツートン)、3）具体物（黒い四 角窓、小動物の顔）の 3 種類の視覚属性が見られ、これらの視覚属性は人間にとっても救急車に関連するものであることがわかる。この例は、提案手法の分析が最もうまく動作した例である。

図 2.20 右下の画像は、物理整合率と推論整合性が高く、論理整合率が低い。物理整合に関して、推論根拠特徴量 ID は画像の物理的な特徴を捉えていない。論理積を取った結果、推論根拠特徴量 ID が一つしか残らないため、捉えられた物理特徴が少なかったと考えられる。論理整合に関して、推論根拠特徴量 ID は推論結果（ラベル）の論理的な特徴を捉えている。ただし、推論根拠特徴量 ID からは、競技としての野球ではなく、野球ボールが想定される。論理整合率が低いのは、言語的特徴分析における視覚属性と、推論クラス（野球）との関連付けが、人間には困難だったからと考えられる。「野球」はスポーツの一種であり、物理的な属性の特定は難しい。この例では、複数の硬式球が写った画像を野球クラスとしているが、野球の試合風景の画像も野球クラスとなるだろう。一方、推論整合率、すなわち最大 softmax 確率は 1.0 であるため、深層学習モデルはこの推論に非常に自信を持っている。この例から、学習されたニューラルネットワークは、本研究の仮説とは異なる推論過程で推論を行う可能性があることがわかる。

図 2.20 左上の画像は、論理整合率と推論整合性が高く、物理整合率が低い。言語的特徴分析の結果には視覚属性「毛皮」があるが、入力画像には見あたらない。この特徴量は、毛皮とは別の視覚属性に反応していることが示唆される。この例から、学習済み深層学習モデルには、人が理解できない特徴量があることがわかる。

図 2.20 左下の画像は、物理整合率と論理整合率が低く、推論整合性だけが高い。言語的特徴分析では、シャープな屋根・キャップや蓄積する細かい箱・円やゴムタイヤが示されているが、これらの視覚属性は抽象的・一般的であり、人間には見出せない（どのようなパターンなのか、どこに写っているのか、わからない）可能性がある。その上、これらの視覚属性が画像内にあったとしても、なぜそれが理髪店の椅子という推論（ラベル）と結びついているのか、人間には理解できない。この例は、上記 2 つの困難な状況が組み合わさったものである。これら 3 つの例は、透明性という点で、深層学習モデルの限界を示すものである。深層学習モデルの本質的な複雑さには、単純な推論根拠解析では透明化できないものがあるとが考えられる。

次に、誤った推論に対する分析結果について考察する。

図 2.21 の第 1 行第 1 列の例では、オシロスコープにアナログ時計が表示されている。CNN の推論（ラベル）は「カセットプレイヤー」であるが、正解ラベルは「アナログ時計」であり、推論を誤っている。物理整合に関して、推論根拠特徴量 ID に対応する視覚属性には「丸い形」（アナログ時計の円に対応)、「細かい箱・円」（オシロスコープのボタンやツマミ類に対応)、「四角い窓」（オシロスコープのディスプレイに対応）が現れており、画像の物理的な特徴を捉えている。論理整合に関して、推論根拠特徴量 ID に対応する視覚属性「丸い形」（スピーカー)、「細かい箱・円」（ツマミ類)、「四角い窓」（カセット挿入口）は、推論結果（ラベル）の論理的な特徴を捉えている。物理整合率は $\mathrm{PCR} = 0.70$ と高く、論理整合率も $\mathrm{LCR} = 0.50$ と低くはない。このため、CNN の推論は妥当であると考えられる一方で、推論整合率は非常に低い（$\mathrm{ICR} = 0.20 \times 10^{-4}$)。推論結果と正解ラベルが異なるのは、正解ラベルが妥当ではないと考えられる。このような場合には、正解ラベルの見直しが有効と考えられる。

図 2.21 の第 1 行第 2 列の例では、CNN の推論（ラベル）は「ヘビ」であるが、正解ラベルは鳥の一種である「アトリ（brambling)」であり、推論を誤っている。物理整合に関して、論根拠特徴量 ID に対応する視覚属性には「くねくね（squiggle)」という視覚属性が含まれている。論理整合に関して、推論根拠特徴量 ID は推論結果（ラベル）の論理的な特徴を捉えている。入力画像には確かに「くねくね」したパターン（屋根瓦による）が写っており、「くねくね」した物体はヘビであると論理的に推論できるため、深層学習モデルとしては論理が通っている。この状況を考察すると、鳥（および鳥に関係する視覚属性）の大きさは、「くねくね」視覚属性の大きさに比べて小さいため、CNN は視覚属性「くねくね」視覚属性を優先し、「ヘビ」と推論したと考えられる。このような場合は、解析結果を参考に、エンジニアリングの対応方針を意思決定ができる。意思決定の一つの方向性としては、屋根瓦などの特徴ある背景の画像を学習データに追加し、同様の状況への頑健性を高めることが考えられる。もう一つの方向性としては、画像分類では後景も考慮したラベルに変更する（前景ラベルと後景ラベルを分離し、この画像の場合は屋根瓦を後景ラベルに設定する）ことが考えられる。単純には、画像に対する正しいクラス定義として、屋根瓦が作る「くねくね」パターンが鳥よりも大きい場合、正解クラスを「アトリ（brambling)」（鳥）ではなく「屋根瓦」に見直す事も考えられる。アプリケーションによって、適切な対応方針は変化し、本研究で提案した解析では一意に特定できない。本研究の貢献は、このようなエンジニアリングの方針に関わる議論ができるような推論根拠の解析手法を提案したことである。

図 2.21 の第 2 行第 1 列の例は、CNN の推論（ラベル）は「そろばん」だが、正解ラベルは「樽」である。また、推論根拠特徴量 ID が一つしかなく、「細かい文字」と人間の直

感とは異なる（PCR = 0.30)。一方で、論理整合率は低くない（LCR = 0.55)。「細かい文字」という特徴から「そろばん」を想起することは可能だが、捉えた物理特徴が間違っていたと考えられる。この例には、樽の全体像は写っておらず、無数の樽が並べられている。このような画像の分類性能を向上するには、「樽」クラスの訓練データを見直す（無数の樽が写った画像を増やす）必要がある。

図 2.21 の第 2 行第 2 列の例は、CNN の推論（ラベル）は「カセットプレイヤー」だが、正解ラベルは「食器洗い機」である（画像には分解された食器洗い機と思われる物体が写っている)。推論根拠特徴量 ID が対応する視覚属性には、「細かい箱・丸」や「細かい箱・ 丸の集積」(スイッチやツマミ類に対応)、丸（中央下にあるモーターのように見えるものに対応）がある。物理整合率（PCR = 0.65）と論理整合率（LCR = 0.50）は低くない。この例では、物理的に特徴を把握しても、正解ラベルにたどり着くことが難しく、正解ラベルの見直しやラベルの追加（食器洗い機と分解された食器洗い機を分けるなど）を検討する必要がある。

図 2.22 の第 1 行第 1 列の例では、CNN の推論（ラベル）は「フラットコーテッド・レトリーバー」であるが、正しいラベルは「グローネンダール」であり、推論が誤っている。物理整合に関しては、推論根拠特徴量 ID が対応する視覚属性「毛皮、動物」「小動物の顔」などは、画像に写っている黒い犬の画像の物理的な特徴を捉えている。論理整合に関しては、推論根拠特徴量 ID は推論結果（ラベル）の論理的な特徴を捉えている（毛皮や小動物の顔があれば、それは犬であると論理的に推論できる)。しかし、現在学習できている特徴抽出だけでは、類似犬種までは区別できないため、推論が誤ったと考えられる。「フラットコーテッド・レトリーバー」と「グローネンダール」はどちらも黒い犬であり、CNN の誤った推論（ラベル）である「フラットコーテッド・レトリーバー」の推論根拠特徴量 ID に対応する視覚属性は、正しい推論（ラベル）である「グローネンダール」の推論根拠特徴量 ID に対応する視覚属性に類似している。これは、学習により獲得できている視覚属性だけでは、2 つのクラスを区別できないという例である。このような場合、より多くの学習データを集め、適切な視覚属性を獲得する必要がある。学習データ全体を増やすのではなく、その推論クラスの学習データを追加で収集するエンジニアリング方針が想定される。

図 2.22 の第 1 行第 2 列の例には、ハムが挟まったサンドイッチが写っている。CNN の推論（ラベル）は「フランスパン」であるが、正しいラベルは「Bakery（パン屋、製パン所、パン類)」である。推論根拠特徴量 ID が対応する視覚属性には「丸いパン」というわかりやすい特徴に加え、「黒い点」（ハムについた黒胡椒と思われる）などがある。物理整合率（PCR = 0.55）と論理整合率（LCR = 0.40）は低くないにもかかわらず、推論を

誤っている。この画像の物理特徴からは、パン屋ではなくパンそのものが想起される。ラベル「Bakery」はパン屋を表すのが自然だが、パン類も含むことができる曖昧さがある。このような正解ラベルを見直す余地がある。

図 2.22 の第 2 行第 1 列の例には、金属でできた水筒が写っているが、CNN の推論（ラベル）は「缶切り（can opener)」で正解ラベルは「樽」である。物理整合率が非常に高く（PCR = 0.90)、論理整合率も高い（LCR = 0.70）のが特徴である。このような場合は、画像に写っている特徴から正解ラベルに到達することが難しく、分類が困難な例である。様々な種類の「缶切り」訓練データを増やすことが考えられる。

図 2.22 の第 2 行第 2 列の例には、パン類、後りんご、サンドイッチ、シリアル、マカロニ、ピザ、クリームチーズを塗ったパンなど、様々な食物が写っているが、CNN の推論（ラベル）は「フランスパン」で正解ラベルは「グラニー・スミス・アップル（青りんごの品種)」のため、推論を誤っている。物理整合に関して、物理整合　推論根拠特徴量 ID が対応する視覚属性「黒い点（胡椒)、茶色、白、黒」「食べ物」などは、画像の物理的な特徴を捉えている。ただし、画像にはパン、青りんご、サンドイッチ、シリアル、マカロニ、ピザ、など、様々な物体が写っており、物理特徴は混在している。論理整合に関して、推論根拠特徴量 ID は推論結果（ラベル）の論理的な特徴を捉えられていない。このように様々な物体が写っている画像の正解ラベルが、それらの物体の一つであるケースは、分類が困難であり、正解ラベルの見直しが考えられる。例えば、様々な物体が混在していて何を正解とすべきか曖昧なので、テストデータで画像とクラスを分離することが考えられる。

深層学習モデルの精度が向上しない場合、次に取るべき対策を知ることは、応用において重要な課題である。今回の研究から、物理整合率が低い場合、深層学習モデルの特徴抽出部分の学習が不十分で、十分な視覚属性を捉えることができていないことを示唆する。一方、論理整合率が低い場合は、深層学習モデルの後段にある分類や回帰などの判定部分の訓練が不十分であることを示唆する。前者の場合、CaffeNet では conv5 以前の特徴抽出部のネットワーク層を増加させることが考えられる。後者の場合、conv5 より後の判定部のネットワーク層を増加させることが考えられる。

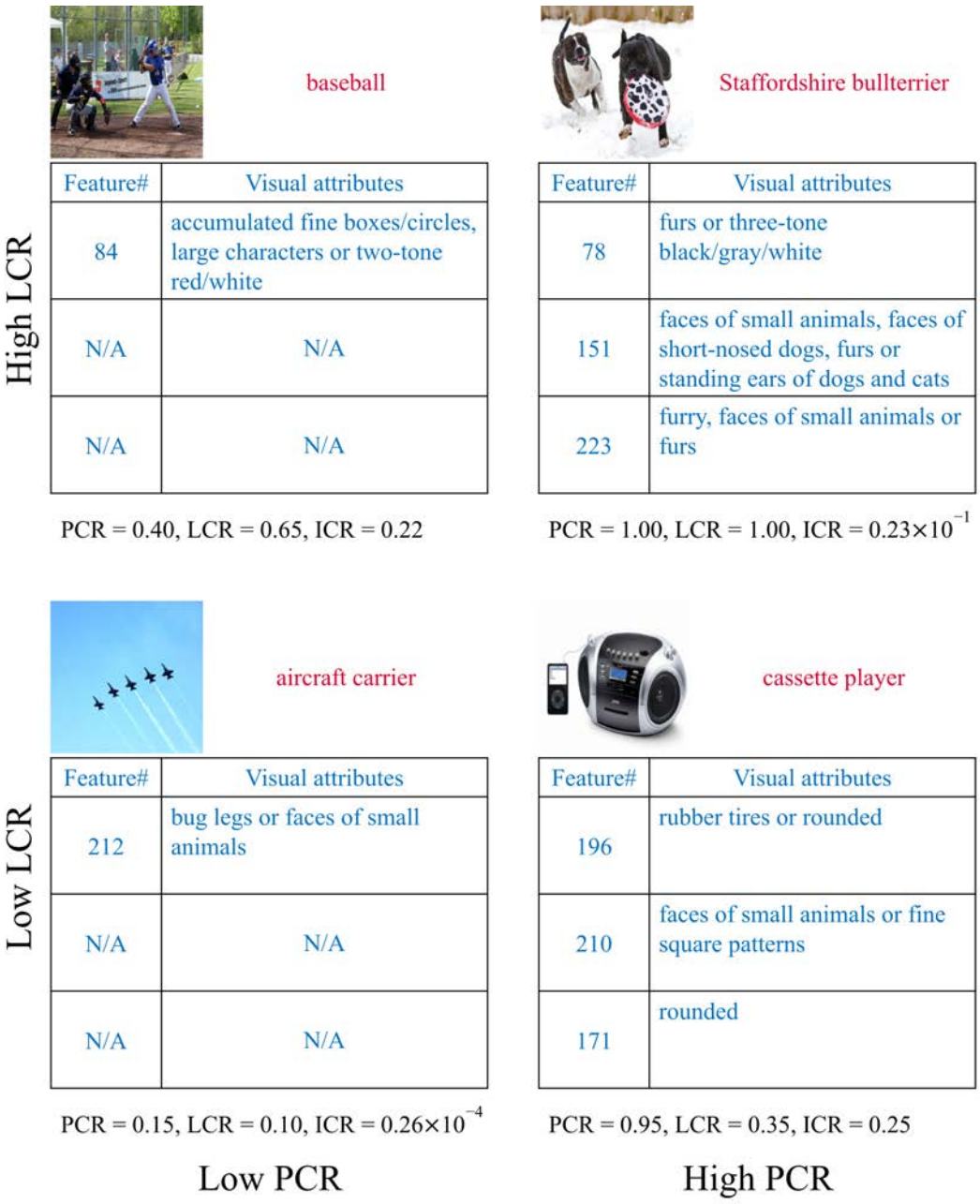



図 2.19: 正解時の特徴分析と整合性分析の結果 — 推論整合率（ICR）が低い場合

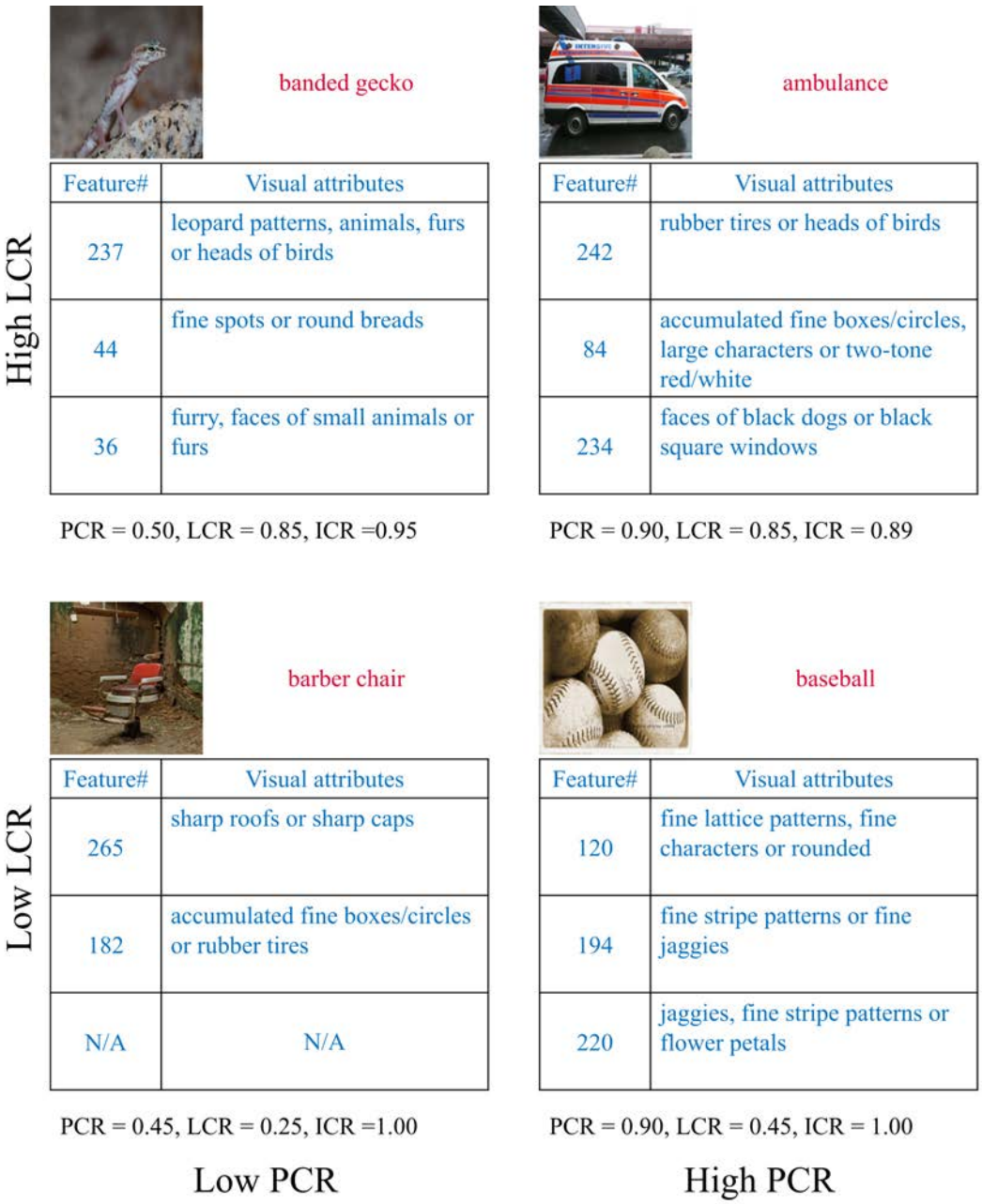



図 2.20: 正解時の特徴分析と整合性分析の結果 — 推論整合率（ICR）が高い場合

cassette player
(correct: analog clock)

| Feature# | Visual attributes |
|---|---|
| 171 | rounded |
| 182 | accumulated fine boxes/circles or rubber tires |
| 207 | rubber tires or square windows |

PCR = 0.70, LCR = 0.50, ICR = $0.20\times10^{-4}$

thunder snake
(correct: brambling)

| Feature# | Visual attributes |
|---|---|
| 191 | squiggle |
| 176 | squiggle or standing ears of dogs and cats |
| N/A | N/A |

PCR = 0.90, LCR = 0.90, ICR = $0.35\times10^{-1}$

abacus
(correct: barrel)

| Feature# | Visual attributes |
|---|---|
| 221 | fine characters |
| N/A | N/A |
| N/A | N/A |

PCR = 0.30, LCR = 0.55, ICR = 0.16

cassette player
(correct: dishwasher)

| Feature# | Visual attributes |
|---|---|
| 182 | accumulated fine boxes/circles or rubber tires |
| 207 | accumulated fine boxes/circles or rubber tires |
| 171 | rounded |

PCR = 0.65, LCR = 0.50, ICR = 0.11

図 2.21: 不正解時の特徴分析と整合性分析の結果

flat-coated retriever
(correct: groenendael)

| Feature# | Visual attributes |
|---|---|
| 136 | furry, animals, furs or black dots on brown |
| 184 | faces of small animals or furs |
| N/A | N/A |

PCR = 0.90, LCR = 0.95, ICR = 0.52

French loaf
(correct: bakery)

| Feature# | Visual attributes |
|---|---|
| 107 | furry, furs or black dots on white |
| 37 | round breads, gloss or fruits |
| N/A | N/A |

PCR = 0.55, LCR = 0.40, ICR = $0.41\times10^{-2}$

can opener
(correct: barrel)

| Feature# | Visual attributes |
|---|---|
| 171 | rounded |
| 195 | sharp caps |
| 222 | squiggle, rubber tires or rounded |

PCR = 0.90, LCR = 0.70, ICR = $0.66\times10^{-2}$

French loaf
(correct: Granny Smith)

| Feature# | Visual attributes |
|---|---|
| 6 | furry, black dots on brown, black dots, brown, two-tone brown/white or two-tone black/brown |
| 113 | foods or brown |
| N/A | N/A |

PCR = 0.55, LCR = 0.40, ICR = $0.41\times10^{-2}$

図 2.22: 不正解時の特徴分析と整合性分析の結果（続き）

## 2.7 本章のまとめ

本章では、セーフティクリティカルなアプリケーションにおける深層学習モデルのブラックボックス性に対処するため、推論過程の透明性を向上させる、1) 構造的特徴分析、2) 言語的特徴分析、3) 整合性分析の 3 種類の分析法を開発し、その結果を示した。そのうえで、分析手法と分析結果を定性的・定量的に考察し、データ収集や訓練などの深層学習モデルの開発工程で、提案手法による分析結果を活用する方針を議論することで、提案手法の有用性を示した。

深層学習モデルやアンサンブル学習などの複雑なアルゴリズムの透明性を定量的に評価することは困難であることが知られており [21]、本章の研究は問題の一部分を検討したに過ぎない。しかし、これまで深層学習モデルの推論過程はブラックボックスでしかなく、関連研究も可読性の向上に新たな深層学習モデルを利用するなど、ブラックボックス性そのものの解決を試みるものはなかった。本章の実験と考察が、透明性を前進させたことは確かである。深層学習モデルが誤った推論（ラベル）を生成した場合、改善する手がかりはなかったが、本手法により、ネットワークの拡張や学習データの収集など、次のアクションの可能性を示唆できる可能性があることがわかった。

# 第 3 章

# 未知不均衡ドメイン機械学習

近年、深層学習は様々なアプリケーションに用いられ、成功を収めている。深層学習は、膨大な学習データの全体的な特徴を学習することを得意とする。しかし 実世界のアプリケーションでは、学習データは複数のドメイン（異なるデータ起源）を含むことが多く、ドメインによって重要度やリスクが異なる場合もある。本章[*1]では第一に、新しい問題設定として、未知不均衡ドメイン機械学習を提案する。提案する問題設定では、学習データにおいて標本のドメイン割り当て（どのドメインから生成されたデータか）が未知かつ不均衡であり、テスト時にはドメインごとにテストデータを用意して性能を評価する。ただし、第 4 章の未知不均衡ドメイン「能動学習」と対比させるため、本章では未知不均衡ドメイン「機械学習」と呼ぶ。第二に、分類タスクを例に、提案した問題設定に対する解法のアプローチを提案する。提案アプローチでは、深層学習学習モデルの特徴空間における各標本と全標本の重心（セントロイド）の距離に基づいた center loss と、同距離に基づいた重み付き標本抽出（重心から遠いデータを重点的に抽出）を組み合わせる。3 つの手書き文字認識データセット（MNIST, EMNIST, USPS）を組み合わせ、1 つの多数派ドメインに対して 1 つの少数派ドメインを加える実験設定と、1 つの多数派ドメインに対して 2 つの少数派ドメインを加える実験設定を行い、未知不均衡ドメイン機械学習の評価、つまり少数派ドメインを含むドメインごとの評価において、提案アプローチが関連手法より良い結果を達成することを示す。提案アプローチはマイナードメインにおいて平均 1% 以上精度を向上させることを示す。また、提案アプローチが、少数派ドメイン数が 1 と 2 の場合に想定通り動作することを示し、提案手法は任意の複数ドメインの混合に対して有

[*1] 本章の研究成果は Kuwajima, Tanaka, and Okutomi (2022) [58] に基づく。ただし、(Kuwajima et al., 2022) [58] の著作権は Society of Imaging Science and Technology (IS&T) に帰属する。

効であると推定する。

本章の貢献は次の 2 点である。

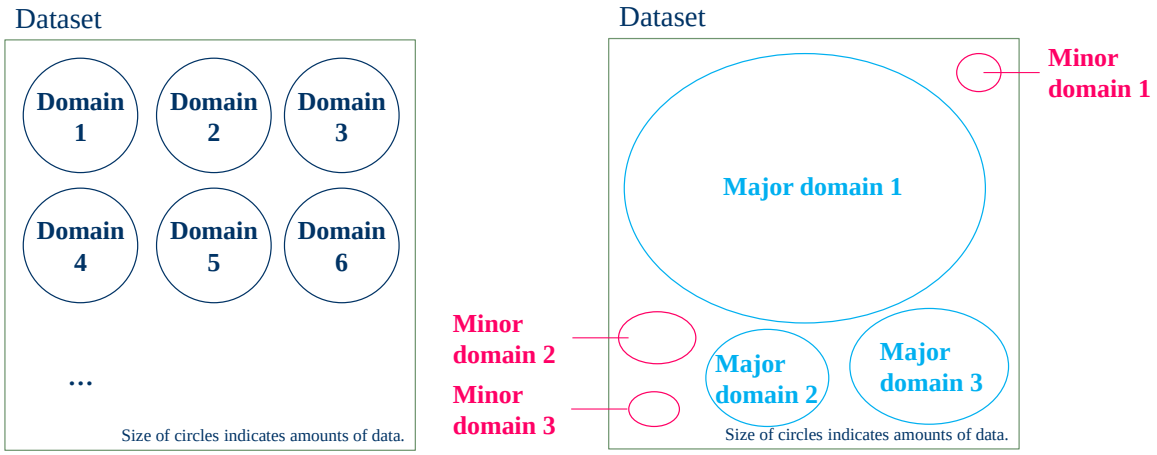
1. 新しい問題設定として、未知不均衡ドメイン機械学習を定式化する。
2. 未知不均衡ドメイン機械学習に有効なアプローチとして、center loss とセントロイド距離に基づいた重み付き標本抽出の組み合わせを特定し、有効性を実験で示す。

本章の本節以降の構成は以下のとおりである。まず、3.1 節で背景を述べる。3.2 節で center loss と関係する損失関数の先行研究、重み付き標本抽出、そしてドメイン適合の先行研究を紹介する。3.3 節で、未知不均衡ドメイン機械学習の問題設定を定式化し、3.4 節で、学習データのドメイン均衡と深層学習の結果（学習済み深層学習学習モデルの精度）の関係を観測する。3.5 節で、未知不均衡ドメイン機械学習に適した学習方法として、深層学習モデルの特徴空間に着目し、center loss とセントロイド距離に基づいた重み付き標本抽出（重心から遠いデータを重点的に抽出）の組み合わせを提案する。3.6 節で、既存手法を比較して提案法が少数派ドメインに対する深層学習モデルの性能を向上させることを実験で示し、性能向上の理由を考察する。最後に、3.7 節で、未知不均衡ドメイン機械学習の研究をまとめる。

## 3.1 背景

深層学習技術 [33] は近年急速に進歩し、様々なシステムで必要技術になりつつある。機械学習モデルは通常、標本損失（学習データ標本に対する推論結果と真値の乖離）の平均を最小化するように学習される。これは、機械学習が、多数派のドメインの標本を重視し、少数派のドメインの標本を重視しないことを意味する。しかし実際には、学習データには様々なドメインからの標本が含まれる。データセット中に対応する標本が少ないドメインを少数派ドメイン、データセット中に対応する標本が多いドメインを多数派ドメインとそれぞれ呼ぶ。図 3.1 に均衡ドメインと不均衡ドメインの模式図を示す。均衡ドメインの状況では、各ドメインから同程度の標本がデータセットに含まれる。一方、不均衡ドメインの状況では、多数派ドメインからは多数の標本を、少数派ドメインからは少数の標本が、データセットに含まれる。多数派ドメインと少数派ドメインは複数あっても良い。

ここで、ドメインとは、手書き文字認識では異なる個人、自動運転では異なる場所や環境条件、自動翻訳では異なる翻訳者、音声認識では方言やイントネーションの話し手など、異なるデータ起源である。データ収集時にドメインを洗い出し制御することは不可能であり、不均衡が生じることは避けられない（図 3.2）。一方で、深層学習の産業応用で




(a) 均衡ドメイン設定　　(b) 不均衡ドメイン設定

図 3.1: データセット中のドメイン均衡の模式図

は、少量の標本データが重要になることがある。例えば、自動運転やクレジットカード認証などにおける事故は、重要ではあるが稀なケースである。それらの事故標本数は、通常標本数に比べ、はるかに小さい。例えば、自動運転の場合、雨の日の夜中のデータは、晴れの日の昼間のデータより少ないのが普通である。一方、雨天時の事故リスクは晴天時よりもはるかに大きいと考えられる。図 1.3 に手書き文字認識と自動運転の場合の具体例を示す。このように、応用によっては少数派ドメインの重要度も高いが、未知不均衡ドメインに従来の学習を適用すると、多数派ドメインに最適化してしまう。未知不均衡ドメインの問題設定では、どのようなドメインでも高性能を導きたいため、ドメインごとにテストして性能を把握し向上させる。

図 3.3 は、機械学習モデルの特徴空間において、多数派ドメインと少数派ドメインの分布を模式的に示したものである。従来のランダムミニバッチ生成（stochastic gradient descent, SGD） [82, 50] では、少数派ドメインは多数派ドメインから離れた位置に分布している。そのため、深層学習モデルは多数派ドメインを適切に分離するための決定境界を学習し、多数派ドメインから遠く離れた少数派ドメインの標本に対する性能は低くなる可能性がある。しかし、セーフティクリティカルなシステムでは、少数派ドメインに対する性能も重要である。各標本のドメインが既知であれば、学習時にドメイン均衡抽出（少数派ドメインを重点的に抽出）できるが、現実的には多くの場合でドメイン情報は未知で

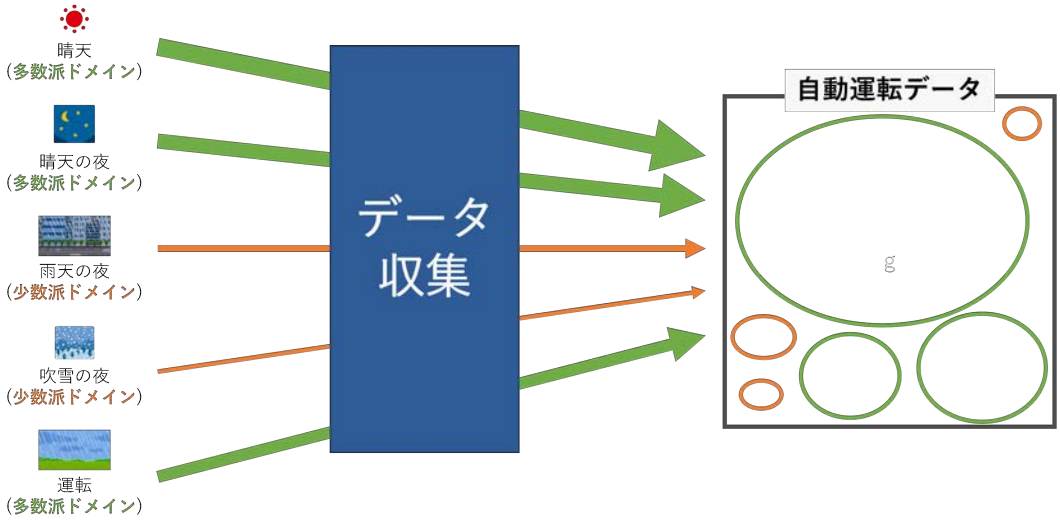



図 3.2: データ収集による不均衡ドメインの発生

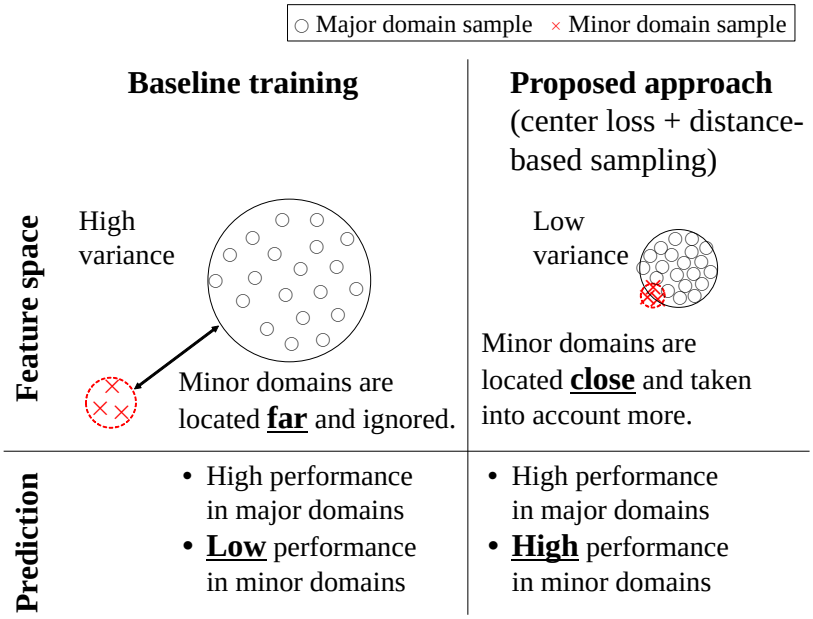



図 3.3: 不均衡ドメインに対する機械学習の想定

ある。

本章では、まず、未知不均衡ドメインに対する機械学習の問題設定を形式的に定義する。多くのドメイン適応（domain adaptation）技術 [72, 85] は、標本のドメイン情報（適合対象が学習時と異なるドメインである）が分かっている、既知ドメイン問題のみを対象としている。従って、ドメイン適応を未知ドメイン問題に適用することはできない。一方、

前述の通り、ドメイン情報が検出できれば、ドメイン均衡抽出 [14] を適用できる。しかし、本章で実験的に示す通り、異常検知を用いた少数派ドメイン標本検出はうまく機能しない [16, 15]。そこで本章では、図 3.3 の想定に基づいて、深層特徴空間ですべての標本の分散を小さくする center loss [101] と、深層特徴空間でのセントロイド距離に基づいた重み付き標本抽出を用いて、少数派ドメインの標本が多く含まれるミニバッチを生成することで、未知の少数派標本の性能を向上させることを目的とする。なお、一般的には、ドメインがオーバーラップしている場合や、階層的な場合も想定されるが、本研究では考慮しない。本研究では、ドメインが排他的で構造がフラットであることを仮定せず、一般の場合に有効な手法を検討する。

## 3.2 関連研究

まず、center loss [101] と関連する他の損失関数の先行研究を紹介する。Center loss とは、特徴空間における標本とセントロイド（クラスごとの標本平均）を近づけるための正則化である。Contrastive center loss [79] （対照的 center loss）は center loss を拡張した手法で、特徴空間において、center loss がクラス内分散を最小化するのに加えて、クラス間分散を最大化する。同様に、Contrastive loss（対照的損失）と triplet loss [42] は、特徴空間において positive sample 二つ組の距離を最小化し、negative sample 二つ組の距離を最大化する。Contrastive loss は、positive sample（例えば、同じクラスの標本）と negative sample（他クラスの標本）を選択する。Triplet loss は、1）アンカー標本、2）positive sample、3）negative sample の三つ組を選択し、positive sample 二つ組（1 と 2）と negative sample 二つ組（1 と 3）を構成する。Center loss、constrastive center loss、triplet loss のアルゴリズムより、特徴空間で、類似標本の分散を最小化し、非類似標本の分散を最大化することが、一般的なアプローチとわかる。本章では、少数派ドメインの精度回復のために、機械学習モデルの特徴量の分散最小化に関心があるため、center loss に着目する。

次に、重み付き標本抽出の先行研究を紹介する。各標本の特徴量に基づいて、標本数（重み付き標本抽出は）と損失の重みを制御する。Hard negative mining [91] は、正解するのが難しい標本のみを誤差逆伝播する手法である。SMOTE [18] は、不均衡なクラス [49] に対するデータ拡張 [89] の手法である。Hard negative mining は損失（正解と推論結果の差異）が大きい hard sample を積極的に選択するのに対し、SMOTE はクラス不均衡を補うためにデータ拡張を行う。しかし、高次元空間における hard sample の選択や少数クラス標本の生成は、球面集中現象 [92] やノイズ [103] の影響を受ける。セント

ロイド距離に基づいた重み付き標本抽出 [103] は、遠距離にある標本だけを選択するのではなく、標本確率の逆数に基づいて重み付き標本抽出を行うことで、様々な距離にある多様な negative sample を選択し、この高次元の問題に対処している。損失重み付けの手法である focul loss [63] は、物体検出器の学習における前景と背景の不均衡の解消を目的とし、すでに適切に分類できた標本を軽く重み付けする。

ドメイン適合 [72, 85] は，機械学習モデルをあるソースドメインで学習し、別のターゲットドメインでテストしたときの性能劣化に対処する研究領域である。ドメイン適合の問題設定では、標本のドメインラベルが既知と解釈できる。学習標本は常にソースドメインから、テスト標本は常にターゲットドメインから得られる。このように既知ドメイン設定であり、本研究の未知ドメイン設定とは異なる。深層教師ありドメイン適合 [72] は、教師あり（クラス既知）アプローチで、contrastive center loss と同様に、同じクラスについてはソースとターゲットの標本を近づけ、異なるクラスについてはソースとターゲットの標本を遠ざけるように、特徴抽出器を最適化する。一方、Maximum classifier discrepancy [85] は教師なし（クラス未知）アプローチで、特徴抽出器を共有し異なる分類器を持つ 2 つの分類器を作り、分類器の不一致を最小化と最大化を繰り返す。

## 3.3 未知不均衡ドメイン機械学習の問題設定

未知不均衡ドメイン機械学習を定式化する。学習標本、ラベル、標本のドメインラベルを $x$、$y$、$z$ とする。$N_z$ をドメイン数として、複数のドメインの学習標本とラベルの同時確率は、以下の混合分布になる。

$$p(x,y) = \sum_{z=0}^{N_z-1} p(z)p(x,y|z)\,, \tag{3.1}$$

$p(z)$ が既知ならば既知ドメイン問題、$p(z)$ が未知なら未知ドメイン問題と呼ぶ。$p(z)$ の分散が小さければ、つまりどのドメインの学習標本も同様の確率で生成するならば、ドメイン分布は均衡している。$p(z)$ の分散が大きければ、つまりドメインによって学習標本の発生確率が大きく異なれば、不均衡ドメインと呼ぶ。2 ドメインの場合を考えると、$p(z=0) \gg p(z=1)$ の状況が不均衡ドメイン問題である。ドメインが均衡しているならば、$p(z=0) \simeq p(z=1)$ となる。未知不均衡ドメイン機械学習の評価では、各ドメイン $z$ の性能である、ドメイン性能 $\mathrm{PERF}_z$ を導入する。

この未知不均衡機械学習の問題設定を図 3.4 に示す。従来の機械学習では多数派ドメイン Domain 1, 2, 3 などに機械学習モデルを最適化する。未知不均衡機械学習では、多数

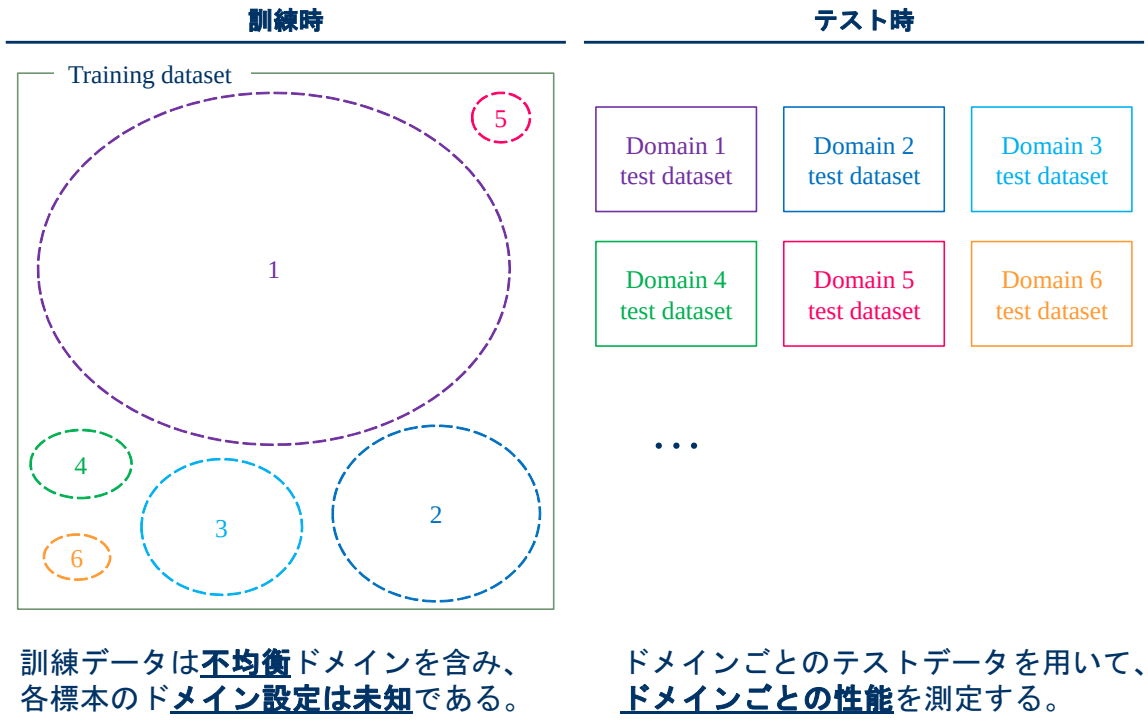



図 3.4: 未知不均衡ドメイン機械学習の問題設定

派ドメインと少数派ドメインの両方の性能を向上させる。

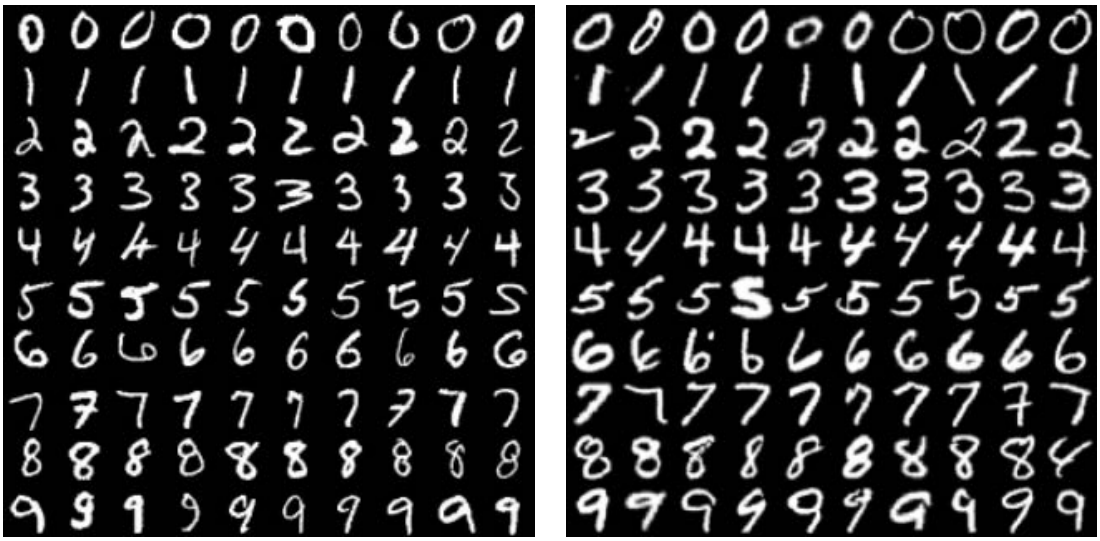
## 3.4 不均衡データから学習した特徴空間の観測

簡単のため、多数派ドメインと少数派ドメインの 2 ドメインの設定を考える。深層学習タスクの例として分類タスクに着目し、3 つの手書き文字認識データセット MNIST [61]、EMNIST [19]、USPS [47] を $32 \times 32$ にサイズを揃えて混合し、複数ドメインのデータを模倣する。3 種類のデータセットからは、多数派ドメインと少数派ドメインの二つ組を 6 組生成することができる。図 3.5 は、多数派ドメイン（MNIST）と少数派ドメイン（EMNIST）二つ組の例で、筆跡が異なっていることがわかる。各ドメイン二つ組に対して、手書き文字認識（分類タスク）を行う機械学習モデルとして、活性化関数 ReLU [29, 73] を用いた LeNet [62] を学習する。分類タスクにおける $\mathrm{PERF}_z$ の一例として、ドメイン精度 $\mathrm{ACC}_z$ を使用する。図 3.6a は、多数派ドメインの標本数を 500 と固定し、少数派ドメインの標本数に対して、6 組の $\mathrm{ACC}_z$ の平均を示したものである。

$f_{\theta,\phi}(x) = (h_\phi \circ g_\theta)(x)$ は学習済み機械学習モデルで、特徴抽出器 $g_\theta$ と分類器 $h_\phi$ に分解して表現している。本節では $f_{\theta,\phi}$ として LeNet を用い、LeNet の入力層から後ろから 2 番目の全結合層（F6）の前までのネットワークを $g_\theta$ とする。$g_\theta(x)$ を標本深層特徴量、$\mu_y = E_{x\sim p(x|y)}\left[g_\theta(x)\right]$ をクラス $y$ のセントロイド深層特徴量とする。標本深層特徴量とセントロイド深層特徴量の間の距離 $d = \|g_\theta(x) - \mu_y\|_2$ をセントロイド距離と定義する（図 3.7）。正解クラスが $y$ である学習標本 $x$ のセントロイド距離 $d$ は、クラス $y$ のセントロイド深層特徴量に基づいて計算される。図 3.8 は、多数派ドメインのクラスあたり標本数を 500 に固定し、少数派ドメインの標本数に対する 3 つの密度分布である。図 3.8 では、ドメイン二つ組の平均ではなく、多数派ドメインとして MNIST、少数派ドメインとして EMNIST を選択し、多数派ドメイン標本のセントロイド距離、少数派ドメイン標本のセントロイド距離、全標本のセントロイド距離の密度を示している。

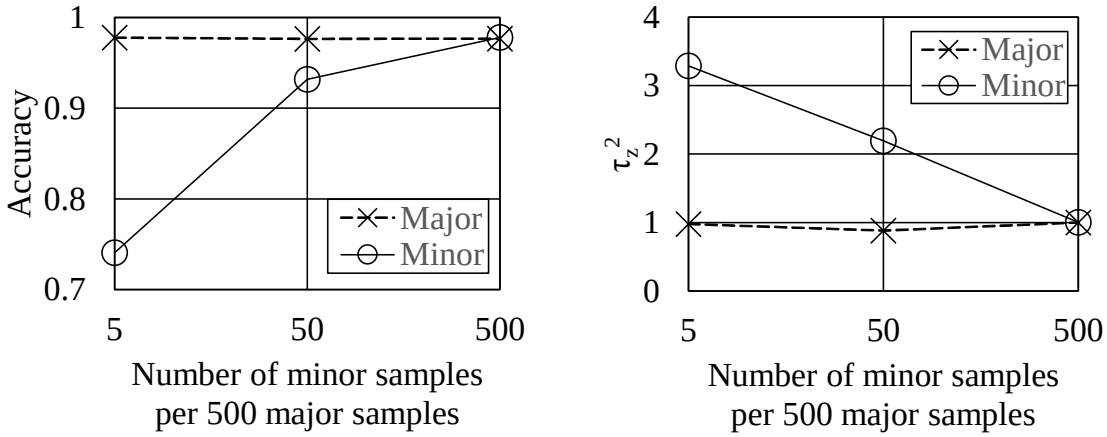
図 3.6a と図 3.8 より、少数派ドメインの標本数が多いと、少数派ドメインのテスト精度が向上する。したがって、仮に標本数を増やせるなら、少数派ドメイン性能を向上できると考えられる。一方、少数派ドメインの標本が少ないと、少数派ドメインのセントロイド距離は大きい。したがって、セントロイド距離が大きい標本が少数派ドメインである可能性が高い。そこで、未知不均衡ドメインの問題設定では少数派ドメイン標本が不明のため、距離が大きい標本を少数派ドメイン標本と仮定して増やす方針とする。

ここで、ドメイン標本間の近さを評価するために、ドメイン $z$ の 2 次の中心モーメン



(a) 多数派ドメインの標本　　(b) 少数派ドメインの標本

図 3.5: 画像処理における異なるドメインの例




(a) テスト精度の変化　　(b) ドメイン分離の変化

図 3.6: 少数派ドメインのサンプル数に対する不均衡ドメインの特性

ト [35] を正規化し、分類タスクにおける相対ドメイン分離 $\tau_z^2$ を定義する。

$$\tau_z^2 = \frac{E_{x,y\sim p(x,y|z)}\left[\|g_\theta(x) - \mu_y\|_2^2\right]}{E_{x,y\sim p(x,y)}\left[\|g_\theta(x) - \mu_y\|_2^2\right]} \tag{3.2}$$

図 3.6b では、多数派ドメインのクラスあたり標本数を 500 と固定し、少数派ドメインの標本数を変化させながら、多数派ドメインと少数派ドメインの相対ドメイン分離を計測し

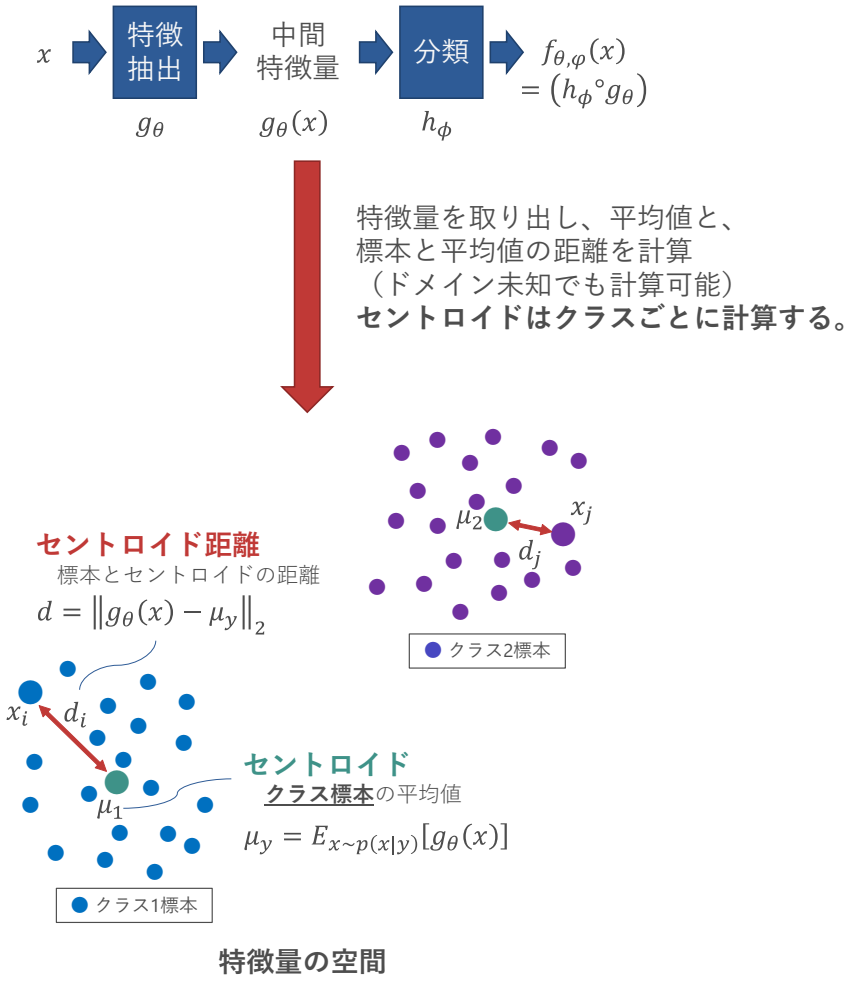
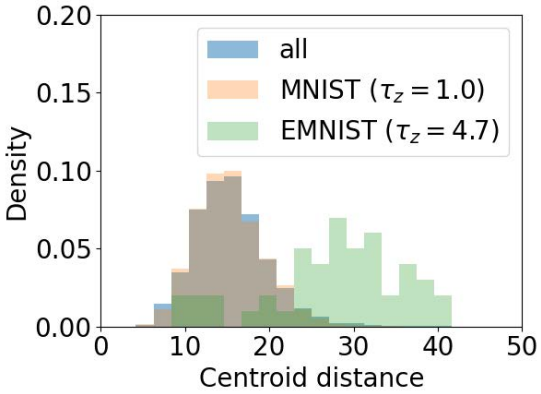


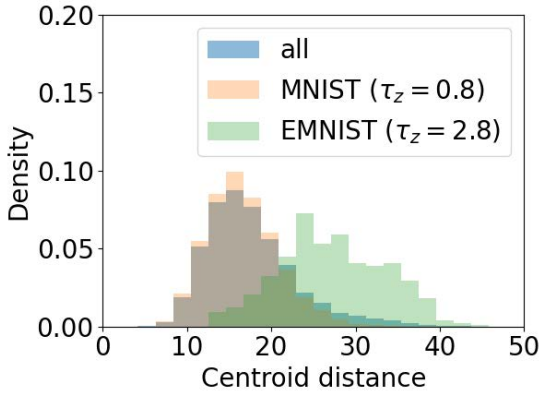

図 3.7: セントロイド距離の定義

た。ただし、相対ドメイン分離の値は、6 組のドメイン二つ組の平均である。図 3.6a と図 3.6b から、テスト精度と相対ドメイン分離の間に負の相関があることが読み取れる。この点からも、ドメイン分離が大きい（セントロイド距離が大きい）標本を少数派ドメイン標本と仮定できる。

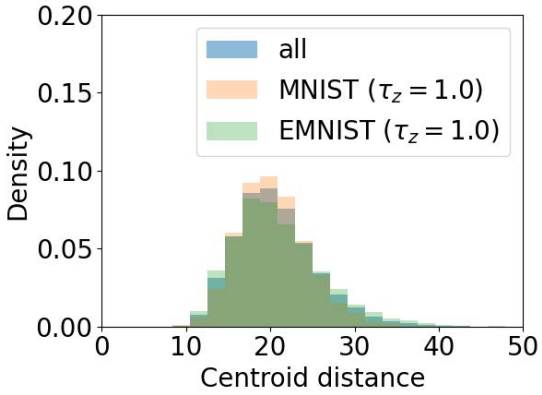
問題設定の締めくくりとして、3.2 節で述べた関連研究と未知不均衡ドメイン機械学習の違いを表 3.1 にまとめる。クラス不均衡分類は、学習データとテストデータともに不均衡クラスが含まれる問題設定である。対象ドメインは一つで、学習データとテストデータは同一分布である。ドメイン適合は、最初にソースドメインという単一ドメインの学習データで学習し、ドメイン適合の段階で、ラベルが付いていないターゲットドメインのデータを利用する。ここで、学習データのドメインとターゲットデータのドメインは異な




(a) 少数派ドメイン標本数 5 における結果




(b) 少数派ドメイン標本数 50 における結果

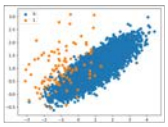
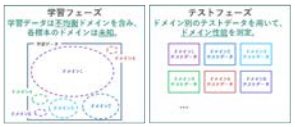



(c) 少数派ドメイン標本数 500 における結果

図 3.8: 少数派ドメインの標本数とセントロイド距離分布の関係

るため、複数ドメインとみなすこともできる。ただし、異なるドメインは混合されておらず、標本がどのドメインから生成されたかは既知である。また、（ドメイン適合としての）テストデータはターゲットドメインだけが含まれるため、単一ドメインである。本章で提案する未知不均衡ドメイン機械学習は、学習データに複数の多数派ドメインと少数派ドメインが混合され、ドメインは未知かつ不均衡である。一方、テストデータでは、すべてのドメインが分離されており、別々に評価されるため、複数ドメインである。複数のドメインにおける性能のバランスを取ることが、未知不均衡ドメイン機械学習の目的である。以上から、提案する未知不均衡ドメイン機械学習の問題設定は新規性があり、かつ、現実の機械学習応用を反映している。

表 3.1: 未知不均衡ドメイン機械学習と関連研究の問題設定の比較

| | クラス不均衡分類 | ドメイン適合 | 未知不均衡ドメイン |
|---|---|---|---|
| |  クラスは不均衡で、ドメインは均衡 | ソースドメイン<br>ラベル付き合成データなど<br>↓<br>ターゲットドメイン<br>ラベル無し実データなど<br>テスト対象はターゲットドメインのみ |  |
| 学習データ | （既知）単独ドメイン<br>• クラスが不均衡なドメイン<br>既知クラス | 既知複数ドメイン<br>• ソースドメイン<br>• ラベルなしターゲットドメイン<br>既知クラス | 未知複数ドメイン<br>• 多数派ドメインと少数派ドメインの混合<br>既知クラス |
| テストデータ | （既知）単独ドメイン<br>• クラスが均衡なドメイン<br>未知クラス | （既知）単独ドメイン<br>• ターゲットドメイン<br>未知クラス | 既知複数ドメイン<br>• 多数派ドメイン<br>• 少数派ドメイン<br>未知クラス |

## 3.5 Center loss（損失関数）と特徴量の分布に基づいたミニバッチ抽出法

3.4 節で示した通り、多数派ドメインと少数派ドメインの標本の均衡をとることは、ドメインが不均衡な場合に重要である。ドメインが既知ならば、重み付き標本抽出により、少数派ドメインの重みを大きくすることで、ドメインを均衡させることができる（これを均衡抽出と呼ぶ）。しかし、ドメインが未知ならば、各標本のドメインに基づいて均衡抽出することはできない。そこで、異常検知と均衡抽出を次のように組み合わせることができる。異常検知アルゴリズムによる異常スコアを用いて、多数派ドメインと少数派ドメインに分類する。分類結果を受け、予想ドメインに基づいた均衡抽出を行うことができる。しかし 3.6 節の実験で示すように、少数派ドメインの異常検知は難しいため、この単純なアプローチは機能しない。そこで本節では、分類タスクにおける未知不均衡ドメイン機械学習に対する実用的なアプローチを構築する。

3.3 節では、精度と相対ドメイン分離（正規化したドメイン内分散）の間に、負の相関があることを示した。そこで、ドメインを分類するのではなく、深層特徴量の分散を最小化することを試みる。そのために、center loss と標本間距離に基づいた重み付き標本抽出 [93] を用いる。ただし、標本間距離の代替としてセントロイド距離を利用する。Center loss と重み付きミニバッチ抽出を組み合わせることで、セントロイド距離を制御するアプローチを図 3.9 に示す。Center loss が分散の小さい特徴空間を作り出し、多くの標本を

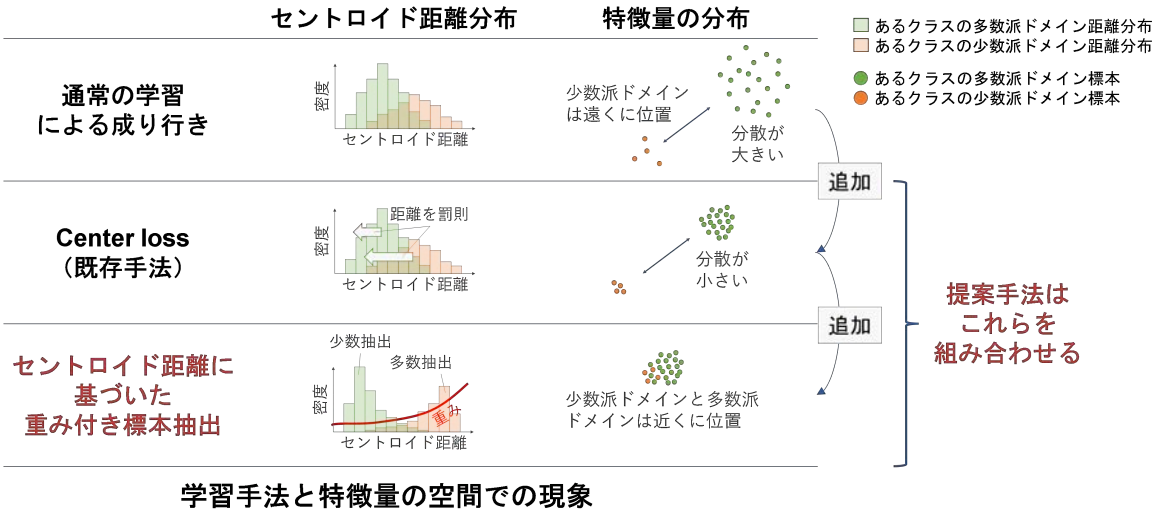



図 3.9: 未知不均衡ドメイン機械学習のアプローチ

正しく推論できるようにする。次に、そのような特徴空間でも依然、遠くに配置されてしまう標本を、重み付きミニバッチ抽出が重点的に学習する。

3.2 節で述べた通り、center loss は深層特徴量の分散を小さくする。一方、セントロイド距離に基づいた重み付き標本抽出の目的は、少数派ドメインの標本を重点的に選択することである。また、3.3 節では、少数派ドメインの標本はセントロイドから離れていることを示した。この観測を受け、深層特徴量の空間で、セントロイドから遠い標本は、少数派ドメインの標本であると仮定する。未知ドメインでは少数派ドメインだけを増やせないため、この仮定にもとづいて、セントロイド距離が遠い標本を多く抽出する。深層特徴量の空間でセントロイドから遠い標本の重みを高くすることで、少数派標本を多く含むであろうミニバッチを生成する [109]。まず、セントロイド距離 $d$ の関数として標本確率 $q(d)$ をモデル化する。次に、学習中にすべての標本のセントロイド距離 $d$ を保持し、全標本の $d$ に基づいて $q(d)$ のパラメータを推定する。$q(d)$ をガウス分布とする場合は、標本平均 $\bar{d}$ と標本分散 $s_d^2$ を、指数分布とする場合は母数 $\lambda_d$ を推定（フィッテイング）する。そして、未知ドメイン設定のもと、多数派ドメインと少数派ドメインを一様に選択するように、標本確率 $q(d)$ の逆数に比例する標本重み $q(d)^{-1}$ を用いて重み付き標本抽出を行い、ミニバッチ $\mathcal{B}$ を生成する（図 3.10）。最後に、全標本の $d$ を再計算するではなく $\mathcal{B}$ に含まれる標本の $d$ のみを更新する。

ここから、アルゴリズムの詳細を説明する。$i$、$j$、$c$ を全学習標本の添字、ミニバッチ $\mathcal{B}$ に含まれる学習標本の添字、クラスの添字とする。一度の学習ループで、機械学習モデル

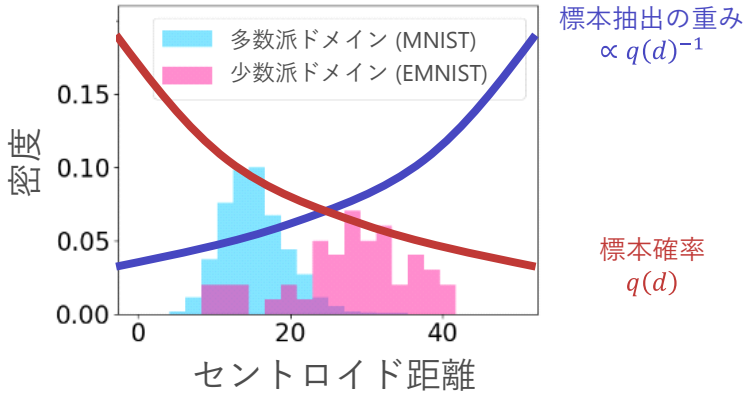



図 3.10: セントロイド距離に基づいた重み付き標本抽出

のパラメータ $\theta$ と $\phi$、全クラスのセントロイド深層特徴量 $\{\mu_c\}$、$\mathcal{B}$ に含まれる学習標本のセントロイド距離 $\{d_j\}$ を更新する。まず、分類の損失（例えば softmax [10] 交差エントロピー誤差）と center loss $\frac{1}{2}\sum_{j\in\mathcal{B}}\|g_\theta(x_j)-\mu_{y_j}\|_2^2$ を誤差逆伝播し、パラメータ $\theta$ と $\phi$ を更新する。次に、center loss の誤差逆伝播により全クラスのセントロイド深層特徴量 $\{\mu_c\}$ を更新する。最後に、ミニバッチ $\mathcal{B}$ に含まれる学習標本に対して $\|g_\theta(x_j)-\mu_{y_j}\|_2^2$ を計算し、慣性項付きでセントロイド距離 $\{d_j\}$ を更新する。振動を避けるため、セントロイド距離の更新に係数 $\alpha$ の慣性項を適用する。本手法の擬似コードを Algorithm 1 に示す。

通常の SGD (stochastic gradient descent) アルゴリズムは、一度選択した標本は二度と選択しない非復元抽出（sampling without replacement） [44] である。また、SGD は 1 エポックですべての学習標本をちょうど一度だけ選択する。一方、提案アプローチは復元抽出（sampling with replacement） [95] であり、重み付き標本抽出で少数派ドメインの学習標本を何度も選択する。したがって、Algorithm 1 では、ミニバッチ数（1 エポックあたり）とエポック数を考慮しない。そこで、学習データセットのサイズをバッチサイズで割ったものを 1 エポックあたりの抽出回数として定義し、学習の進捗を追跡する。

**Algorithm 1:** Center loss と距離に基づいた重み付きミニバッチ標本を組み合わせた未知不均衡ドメイン機械学習（分類タスク）

**Input:** 学習データ $\{(x_i, y_i)\}$, 深層学習ネットワーク $f_{\theta,\phi} = h_\phi \circ g_\theta$, 慣性項 $\alpha$
**Output:** パラメータ $\theta$（特徴抽出部）, $\phi$（分類部）
$\theta$ と $\phi$ を初期化
セントロイド深層特徴量 $\{\mu_c\}$ を初期化
セントロイド距離 $\{d_i\}$ を初期化
**repeat**
  $\{d_i\}$ に基づいて分布 $q(d)$ のパラメータを推定
  重み $q(d_j)^{-1}$ を用いてミニバッチ $\mathcal{B} = \{(x_j, y_j)\}$ を標本化
  $\mathcal{B}$ の分類損失と center loss に基づいて $\theta$ と $\phi$ を更新
  $\mathcal{B}$ の center loss に基づいて $\mu_c$ を更新
  各 $(x_j, y_j) \in \mathcal{B}$ について、$d_j \leftarrow \alpha d_j + (1-\alpha)\|g_\theta(x_j) - \mu_{y_j}\|_2$ を更新
**until 学習終了**;

## 3.6 未知不均衡ドメイン機械学習における損失関数と抽出手法の比較

3.5 節の通り、本手法は center loss（損失関数の一種）とセントロイド距離に基づいた重み付き標本抽出（標本抽出手法の一種）で構成される。本節では、損失関数と標本抽出の二軸で、提案手法を既存手法と比較する（図 3.11）。損失関数として focal loss [63] と center loss [101] を、抽出法として入力データの空間と深層特徴量の空間で実行した局所外れ値因子法（local outlier factor, LOF） [16]、交差エントロピー誤差、セントロイド距離に基づいた重み付き標本抽出を比較する。実験により、提案法全体（center loss とセントロイド距離に基づいた重み付き標本抽出の組み合わせ）の性能と、center loss とセントロイド距離に基づいた重み付き標本抽出のそれぞれ単体の効果を確認する。

実験設定について述べる。ドメインを既知から未知に変更したほかは、3.3 節と同様の実験設定を用いる。活性化関数を ReLU に置き換えた LeNet で手書き文字認識（分類タスク）を学習し、F6 層で深層特徴量 $g_\theta(x)$ に着目する。バッチサイズは 128、すなわち 1 エポックあたりの反復回数が約 $5,000/128$ 回とし、100 エポック後に精度を測定した。

実験設定のうち、特にデータ設定について述べる。MNIST、EMNIST、USPS データ

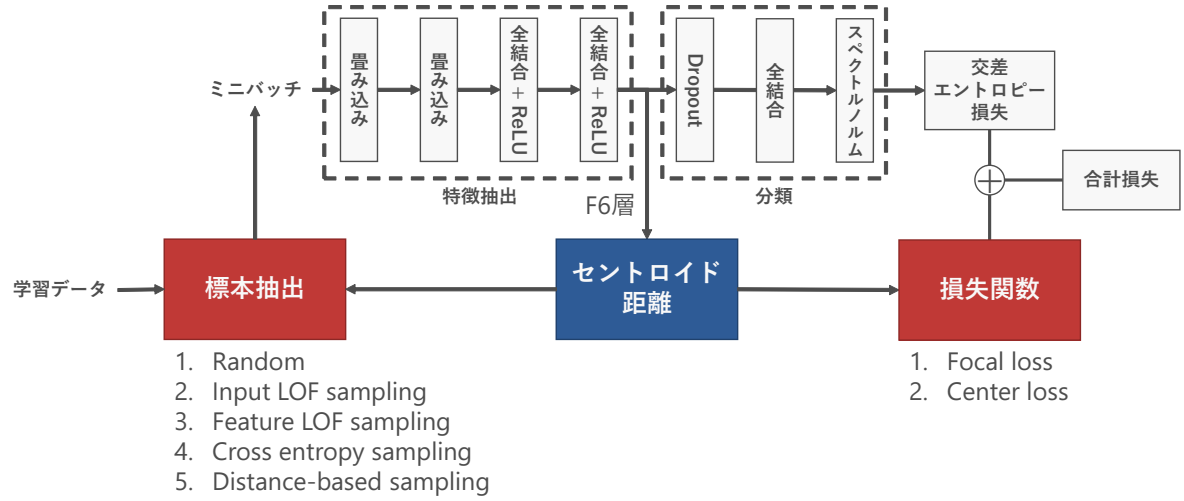
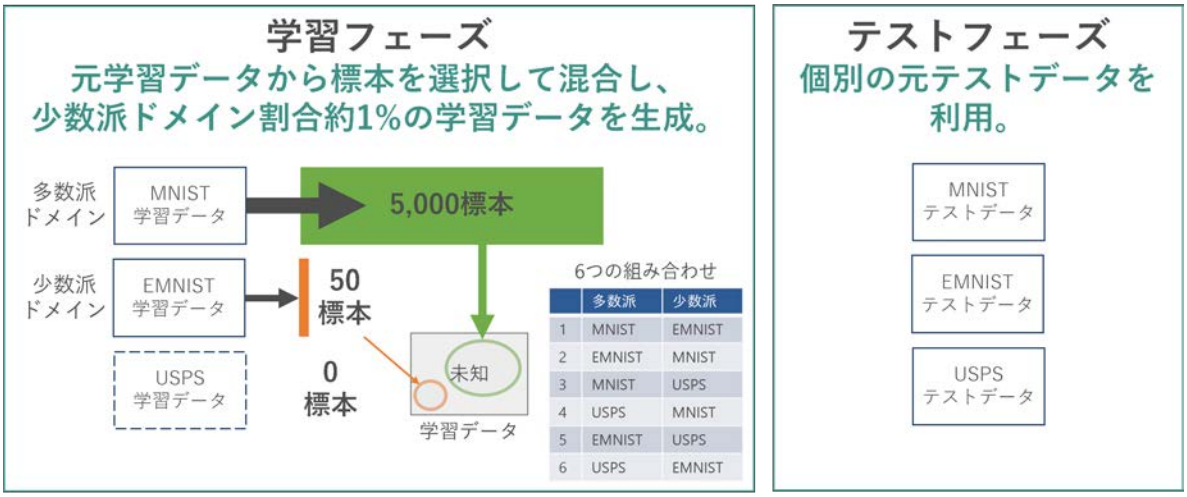



図 3.11: 未知不均衡ドメイン機械学習の比較実験対象

（M、E、U と略す）による二つ組ドメインと三つ組ドメインについて未知不均衡ドメインデータを作成し、実験を行う。例えば、二つ組ドメイン M/E は多数派ドメインが MNIST で少数派ドメインが EMNIST、三つ組ドメイン M/E,U は多数派ドメイン MNIST と二つの少数派ドメイン EMNIST と USPS を表す。一つのドメイン組み合わせに対して、異なるシード値で 4 回実験を行った平均値を計算する。分布の表示など、平均値を計算することができない場合は、代表ドメイン設定として、M/E 二つ組を用いる。多数派ドメインはクラスあたり 500 標本、少数派ドメインはクラスあたり 5 標本とし、学習データとして合計約 $5,000$ 枚の画像を用意した。データ作成の要領を図 3.12 と図 3.13 に図示する。

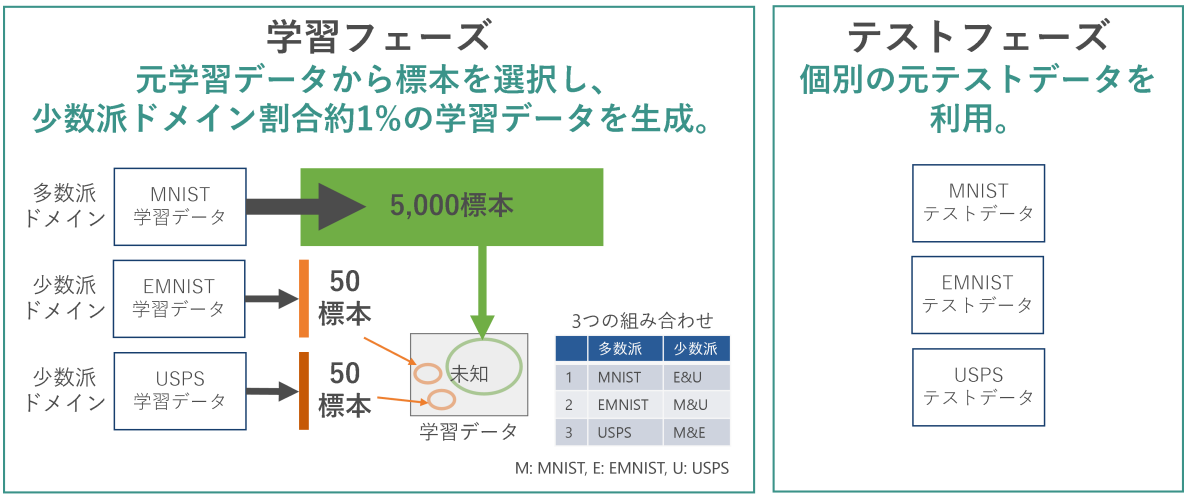
比較手法のハイパーパラメータと、設計選択について述べる。セントロイド距離に基づいた重み付き標本抽出で用いる距離分布 $q(d)$ として、指数分布を用いる。LOF スコアと交差エントロピー誤差も同様に指数分布でモデル化する。$d(x)$ 更新時には係数 $\alpha = 0.9$ の慣性項を用いる。入力 LOF スコアは、入力（標本画像）$x$ に対して学習開始時に一度だけ計算する一方、パラメータ $\theta$ は学習とともに更新されるため、深層 LOF スコアは、深層特徴量 $g_\theta(x)$ に対して各エポック毎に計算する。Focal loss では、焦点パラメータ（focusing parameter）$\gamma = 2$ を用いる。

表 3.2 と表 3.3 は、2 ドメイン（1 つの多数派ドメインと 1 つの少数派ドメイン）の二つ組と 3 ドメイン（1 つの多数派ドメインと 2 つの少数派ドメイン）の三つ組に対する、ドメイン精度 $\mathrm{ACC}_z$ である。まず少数派ドメインについて、2 ドメインの二つ組と 3 ドメインの三つ組の組み合わせの半数以上で、提案手法によりドメイン精度が大幅に改善した（$2.0\% \rightarrow 3.5\%$）。残りの組み合わせでも、従来法と同等の少数派ドメイン精度が得られ




※図は多数派：MNIST、少数派：EMNISTの組み合わせの場合

図 3.12: 2 ドメイン設定における未知不均衡ドメイン学習の実験データ作成




※図は多数派：MNIST、少数派：EMNIST,USPSの組み合わせの場合

図 3.13: 3 ドメイン設定における未知不均衡ドメイン学習の実験データ作成

た（最良の従来法に対して ±0.6% 以内）。一方、多数派ドメインについても、2 ドメイン設定の 2 組を除いて、提案手法が最高のドメイン精度を達成した。さらに、この例外 2 組の精度劣化は 0.01% と 0.08% にとどまった。セントロイド距離に基づいた重み付き標本抽出は、セントロイド距離が小さい多数派ドメインの標本を相対的に軽視するため、直感的には多数派ドメイン精度に悪影響を及ぼすと考えられる。しかし、この実験結果より、多数派ドメインでは大きな性能低下がないことがわかる。また、すべてのドメイン組み合わせを平均しても、少数派ドメイン精度で提案法は他手法に大きく差をつけ（2 ドメイン

表 3.2: 2 ドメイン設定の未知不均衡ドメイン機械学習におけるドメイン別精度

| | | M/E | E/M | M/U | U/M | E/U | U/E | Average |
|---|---|---|---|---|---|---|---|---|
| Random | Major | 0.9835 | 0.9839 | 0.9832 | 0.9647 | 0.9849 | 0.9655 | 0.9776 |
| | Minor | 0.6153 | 0.6148 | 0.9342 | 0.9186 | 0.7141 | 0.6466 | 0.7406 |
| Input LOF sampling | Major | 0.9822 | 0.9837 | 0.9842 | 0.9639 | 0.9851 | 0.9636 | 0.9771 |
| | Minor | 0.6276 | 0.6420 | 0.9332 | 0.9223 | 0.7115 | 0.5911 | 0.7380 |
| Feature LOF sampling | Major | 0.9838 | 0.9844 | 0.9844 | 0.9657 | 0.9845 | 0.9644 | 0.9779 |
| | Minor | 0.6219 | 0.6393 | 0.9321 | 0.9089 | 0.7136 | 0.5998 | 0.7359 |
| Cross entropy sampling | Major | 0.9817 | 0.9850 | 0.9852 | 0.9635 | 0.9853 | 0.9630 | 0.9773 |
| | Minor | 0.6319 | 0.6292 | 0.9331 | 0.9135 | 0.7228 | 0.6079 | 0.7397 |
| Distance-based sampling | Major | 0.9834 | 0.9834 | 0.9828 | 0.9639 | 0.9846 | 0.9604 | 0.9764 |
| | Minor | 0.6287 | 0.6188 | 0.9231 | 0.9050 | 0.6908 | 0.6065 | 0.7288 |
| Focal loss | Major | 0.9834 | 0.9818 | 0.9820 | 0.9641 | 0.9848 | 0.9644 | 0.9768 |
| | Minor | 0.6186 | 0.6225 | 0.9321 | 0.9162 | 0.7204 | 0.6324 | 0.7404 |
| Center loss | Major | 0.9903 | 0.9909 | **0.9911** | 0.9692 | **0.9907** | 0.9706 | 0.9838 |
| | Minor | 0.6976 | 0.7046 | **0.9377** | 0.9424 | 0.8002 | 0.7621 | 0.8074 |
| Center loss + dist. sampling (proposed) | Major | **0.9910** | **0.9911** | 0.9903 | **0.9711** | 0.9906 | **0.9709** | **0.9842** |
| | Minor | **0.7330** | **0.7372** | 0.9363 | **0.9493** | **0.8242** | **0.7632** | **0.8239** |

設定で 1.65%、3 ドメイン設定で 1.57%)、多数派精度で他手法を上回った。これらの実験結果により、提案アプローチが未知不均衡ドメイン機械学習に有効であることが示された。また、2 ドメインと 3 ドメインの実験結果から、提案アプローチが任意の複数ドメインに対しても有効であることが帰納的に推定できる。以下、実験結果の詳細について説明する。

表 3.2 より、提案手法は M/E、E/M、E/U の組み合わせで、それぞれ 3.5%、3.2%、2.4% の精度向上を達成し、他の手法を凌駕した。その他の M/U、U/M、U/E の組み合わせでは、提案手法は他手法と同等（精度差は −0.2% から 0.7%）だった。表 3.3 では、提案法は E/M,U の組み合わせで、3.1% と 1.6% という大幅な少数派ドメイン精度向上を達成し、他手法より優れていた。その他の M/E,U、U/M,E の組み合わせでは、片方の少数派ドメインのドメイン精度が大幅に向上したが（2.7%、2.9%)、もう一方の少数派ド

表 3.3: 3 ドメイン設定の未知不均衡ドメイン機械学習におけるドメイン別精度

| | | M/E,U | E/M,U | U/M,E | Average |
|---|---|---|---|---|---|
| Random | Major | 0.9827 | 0.9841 | 0.9646 | 0.9771 |
| | Minor | 0.6294 0.9250 | 0.6810 0.7332 | 0.8939 0.6556 | 0.7725 |
| Input LOF sampling | Major | 0.9831 | 0.9848 | 0.9644 | 0.9774 |
| | Minor | 0.6109 0.9180 | 0.6871 0.7393 | 0.8768 0.6149 | 0.7664 |
| Feature LOF sampling | Major | 0.9818 | 0.9835 | 0.9625 | 0.9759 |
| | Minor | 0.6150 0.9230 | 0.6616 0.7242 | 0.8855 0.6390 | 0.7619 |
| Cross entropy sampling | Major | 0.9822 | 0.9849 | 0.9621 | 0.9764 |
| | Minor | 0.6283 0.9219 | 0.6960 0.7415 | 0.8868 0.6340 | 0.7749 |
| Distance-based sampling | Major | 0.9821 | 0.9842 | 0.9623 | 0.9762 |
| | Minor | 0.6233 0.9250 | 0.6648 0.7373 | 0.8900 0.6246 | 0.7681 |
| Focal loss | Major | 0.9831 | 0.9838 | 0.9634 | 0.9768 |
| | Minor | 0.6063 0.9198 | 0.7122 0.7514 | 0.8934 0.6507 | 0.7766 |
| Center loss | Major | 0.9904 | 0.9906 | 0.9704 | 0.9838 |
| | Minor | 0.7070 **0.9331** | 0.7493 0.8077 | **0.9336** 0.7526 | 0.8261 |
| Center loss + dist. sampling (proposed) | Major | **0.9910** | **0.9907** | **0.9705** | **0.9840** |
| | Minor | **0.7340** 0.9307 | **0.7796 0.8236** | 0.9279 **0.7818** | **0.8392** |

メインのドメイン精度はわずかに低下した（−0.24%、−0.57%）。

最後に、center loss とセントロイド距離に基づいた重み付き標本抽出の効果を別々に確認する。表 3.2、表 3.3 の右端列にある平均結果によると、損失関数に関して、ベースライン（損失関数：分類損失のみ、抽出法：乱択）、focal loss、center loss の中で、center loss が最も精度が高かった。一方、抽出法に関しては、セントロイド距離に基づいた重み付き標本抽出は他の方法（入力 LOF 抽出法、深層 LOF 抽出法、交差エントロピー誤差抽出法）に対して、ほとんどの条件で精度が低かった。しかし、提案手法で center loss とセントロイド距離に基づいた重み付き標本抽出を組み合わせることで、多数派ドメインと少数派ドメイン両方のドメイン精度において全ての手法を平均的に上回った。特に、center loss に追加でセントロイド距離に基づいた重み付き標本抽出を適用すると、少数派ドメインの平均ドメイン精度は 1.5% 以上も向上した。この結果から、セントロイド距離に基づいた重み付き標本抽出は、center loss によって標本をセントロイドに十分近づけ、セン

トロイド距離が大きい標本のコントラストを際立たせることで機能すると考えられる。一方、提案手法により、多数派ドメインのドメイン精度も向上している。提案手法は、少数派標本だけを狙うのではなく、セントロイド距離が大きい標本を重点的に学習するので、セントロイドから遠い多数派ドメイン標本も重視され、多数派ドメインのドメイン精度も向上したと考えられる。

## 3.7 本章のまとめ

本章では、新たな問題設定として未知不均衡ドメイン機械学習を提案し定式化した。提案問題設定では、学習データは、異なるドメインから不均衡に発生した標本の混合であると仮定する。セーフティクリティカルなシステムなど特定の産業応用では、リスクの高い少数派ドメインが重要であり、多数派ドメインと同様に少数派ドメインでの性能を向上させることが課題である。そこで、特徴空間における各標本と全標本の重心の距離（セントロイド距離）を近づける center loss と、セントロイド距離の遠い標本を重点的に選択するミニバッチ抽出を組み合わせ、効果的な手法を提案した。提案手法は、多数派ドメインのドメイン精度を損なうことなく、少数派ドメインで大幅なドメイン精度向上を達成した。

# 第 4 章

# 未知不均衡ドメイン能動学習

近年成功を収めている深層学習は、大量のデータを必要とする。大量データを入手するには、データ収集とアノテーション（教師値の付与）が必要である。能動学習は、深層学習の登場以前から、アノテーションのコストを削減するために広く研究・利用されてきたが、深層学習の出現により改めて注目を集めている。多くの深層学習応用では、能動学習でアノテーションすべき情報量の多いデータを絞り込んで、学習時間を短縮することを狙っている。本章[*1]では、まず、能動学習の実験設定を調査する。実験設定の違いにより、能動学習の実験結果が大きく変化することを示し、能動学習を応用する現実的なユースケースを反映した、実用的な実験設定を導出する。次に、能動学習における未知不均衡ドメインの問題設定を提案する。任意のデータセットには複数のドメインが含まれる。例えば、手書き文字認識における個人は異なる社会的属性を持つ。未知不均衡ドメインを考慮しなければ、収集した大量データの中から多数派ドメインの標本が多く取り出されて学習データに入り、少数派ドメインの標本は学習データにはあまり含まれない。しかし、テスト段階では、多数派ドメインと少数派ドメインの両方を正確に推論する必要がある。そのため、能動学習でも未知不均衡ドメインの問題を考慮する必要がある。前章では未知不均衡ドメイン機械学習を提案したが、本章では未知不均衡ドメインの問題設定を能動学習にも拡張する。図 1.5 では、未知不均衡ドメインのアプローチとして、モデル学習と能動学習があることを述べた。未知不均衡ドメイン機械学習では、モデル学習のみで未知不均衡ドメインに対処したが、未知不均衡ドメイン能動学習では、能動学習とモデル学習を組み合わせ、多数派ドメインの性能を維持しつつ、少数派ドメインの性能を向上する。前

[*1] 本章の研究成果は Kuwajima, Tanaka, and Okutomi (2023) [59] に基づく。ただし、(Kuwajima et al., 2023) [59] の著作権は Society of Imaging Science and Technology (IS&T) に帰属する。

述の実用的な実験設定のもとで、未知不均衡ドメイン能動学習の手法を比較し、softmax マージンを用いた単純な能動学習法と、前章で提案した center loss と距離に基づいた抽出法（モデル学習法）の組み合わせが、最も高い性能を達成できることを示す。

本章の貢献は次の 2 点である。

1. 能動学習の実験設定を調査し、実験設定が実験結果に大きな影響を与えることを実験的に示す。
2. 現実的な実験設定で、未知不均衡ドメイン能動学習の様々な方法を評価する。

本章の本節以降の構成は以下の通りである。まず、4.1 で背景を述べる。次に、4.2 節で能動学習と未知不均衡ドメイン機械学習の先行研究を紹介する。4.3 節では、能動学習の実験設定を調査し、実験設定の 2 つのポイントとしてプールデータ拡張と学習データ拡張を示し、現実的な実験設定を提案する。4.4 節では、未知不均衡ドメインがデータプールに含まれる場合の能動学習である、未知不均衡ドメイン能動学習の問題設定を述べる。そして、4.5 節にて、実験設定による能動学習実験結果の変化と、提案した現実的な実験設定のもとで、未知不均衡能動学習の実験結果を示す。最後に 4.6 節で、未知不均衡ドメイン能動学習の研究をまとめる。

## 4.1 背景

近年、急速に進歩した深層学習技術は、様々なシステムに必要な要素技術となっている。深層学習技術の性能は、膨大なパラメータを持つ複雑な深層学習モデルの構造と、そのような膨大なパラメータを最適化できる計算機資源によって実現可能となる。一方で、学習（最適化）のためには大量のデータが必要であり、データの効率的な収集とアノテーションは、深層学習における重要課題である。人間の作業が必要なアノテーションは、作業時間とコスト（人件費）がかかる。実際の産業応用では常に予算は限られるため、収集したすべての標本をアノテーションすることはできない。このような現実の状況を想定し、能動学習とは、アノテーションを効率的に行う技術で [87, 88, 96, 97, 20, 32]、深層学習が登場する以前から研究されてきた。能動学習では、収集したデータから最も情報量の多い標本を選択し、人間がアノテーションを行う。そして、累積的にアノテーションしたデータを用いて深層学習モデルを学習することで、最小限のアノテーションコストで性能を向上させる。能動学習は、すでに自動運転 [7, 40]、医療画像 [43]、医療診断、微生物学、製造業 [96] などの幅広い分野で研究されている。しかし、そのような学術研究における能動学習の実験設定と、実応用での能動学習のユースケースを、完全に一致させること

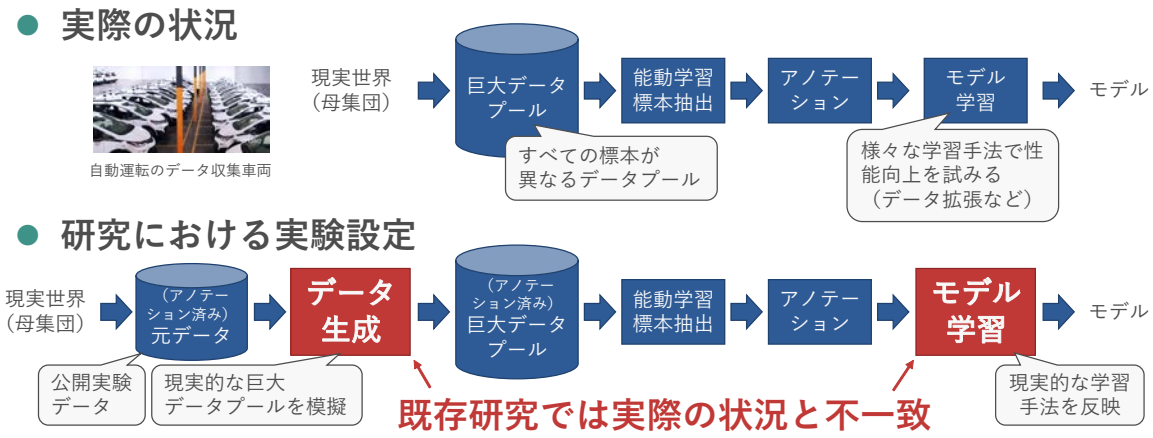



図 4.1: 能動学習の実際の活用状況と研究における実験設定

はできない。そのため、実応用での能動学習のユースケースを反映した実験を慎重に設計しなければならない。本章では、1）実用的な能動学習の実験設定と、2）未知不均衡ドメインの能動学習に適した手法を調査する。

図 4.1 は、アノテーションコストを抑えなければならない実応用における実際の状況と、研究における実験設定を示す。実応用の現場では、図 4.1 上段のように真のデータ分布（収集環境）から膨大なデータ標本を収集する。一方、研究実験では、図 4.1 下段のようにデータ生成を行った上で、能動学習を実行する。研究における実験設定では、能動学習が活用される現実の状況における大量の未アノテーションデータを模倣するため、現実を反映した巨大データプールの作成が重要である。現実の状況で収集されるデータプールの標本数は、深層学習の研究で用いられる標準的な実験データセット [61, 19, 47, 107] よりもはるかに巨大であり [46, 78]、深層学習の実験データセットをそのまま能動学習の実験に用いることはできない。しかし、現実の状況を再現するために、能動学習の実験で新しい標本を大量に収集することも現実的ではない。このように、実際の環境で能動学習手法を評価することはできないため、適切な実験設定を慎重に設計する必要がある。実験設定の設計は、意味のある実験結果を得るために不可欠である。

また、どのようなデータも、複数のドメインを含んでいる。3 章では、未知不均衡ドメイン学習として、学習データに複数ドメインが含まれる状況を考えたが、アノテーション前のデータも同様である。例えば、手書き文字認識では、コミュニティ、年齢、性別などの社会的属性が異なる個人を含む。3 章と同様、標本数の多いドメインを多数派ドメイ

ン、標本数の少ないドメインを少数派ドメインとし、多数派ドメインと少数派ドメインが混在する状況をドメイン不均衡、標本のドメインが不明の状況を未知と呼ぶ。多くの実世界の応用では、データに複数ドメインが含まれ、ドメインは不均衡かつ未知である。データプールの大部分は多数派ドメインの標本であるため、通常、学習した深層学習モデルは多数派ドメインで最も良い性能を発揮する。しかし、特にセーフティクリティカルなシステムでは、少数派ドメインの性能も重要である。例えば、自動運転システムやクレジット決済システムにおける事故が少数派ドメインであり、標本数は少ないものの重大な結果をもたらす。そのため、深層学習モデルの多数派ドメインにおける性能を維持したまま、少数派ドメインにおける性能を向上させることが必要である。

## 4.2 関連研究

能動学習は、データプールからアノテーションすべき標本を選択する技法である。能動学習の研究では、データプールから標本を選択することを獲得と呼ぶ。より良い深層学習モデルを得るため、能動学習と深層学習を繰り返す。能動学習では、深層学習モデルの推論結果の不確実性が高い標本や、分類問題での決定境界に近い標本など、学習済みの深層学習モデルにとって情報が多い標本を獲得する。獲得した標本をアノテーションし、これまでにアノテーションした標本と合わせて学習データとし、深層学習を行う。能動学習による繰り返し回数により、深層学習モデルを得るための時間が決まる。そのため、能動学習の獲得モード（アノテーションすべき標本を一つずつ獲得するか、バッチで獲得するか）は、特に学習時間が長くなる深層学習において重要である。標本獲得の基準と獲得モードの二つの観点で、能動学習の手法を整理する。

最初に、標本獲得の基準について述べる。分類問題では、深層学習モデルの推論の不確実性を評価する最も単純な方法として、softmax 値を用いることが考えられる [17]。Softmax 値は、分類問題で入力標本が各分類に属する確率を表す。ただし softmax 値は、softmax 層を最終層に持つ深層学習ネットワークで、最終層への入力に softmax 関数を適用することで得られる。Softmax 値を用いて、推論結果の不確実性の代替となる値を計算することができる。例えば、softmax 値の最大値、最大と二位の softmax 値の差、などが考えられる。検出タスク（矩形回帰）やセグメンテーションなどの他のタスクでは、softmax 値とは異なる形式の不確実性の表現を検討する必要がある。Sofmax 値を使う単純な方法の他に、ベイズ推論 [13] を用いて推論結果の不確実性を定量化するアプローチである、ベイジアンニューラルネットワーク [74, 64, 65] がある。ベイジアンニューラルネットワークを能動学習に利用する手法として、BALD [45] がある。

次に、獲得モードについて述べる。複雑な深層学習モデルの学習に要する時間は長いため、能動学習と深層学習の繰り返し回数を削減するニーズがある。そこで、能動学習の標本獲得としてバッチ方式と逐次方式が研究されている [38]。逐次方式は、能動学習で標本を一つずつ獲得するため、最適な標本を獲得することができるが、標本が一つ増えるごとに深層学習を行う必要がある。バッチ方式は、能動学習で一度に複数の標本を獲得する。その結果、一定の標本数の学習データを得るまでの深層学習の試行回数は、逐次能動学習では多く、バッチ能動学習では少なくなる。バッチ能動学習は最適な標本集合を獲得できない可能性があるが、時間効率の面で深層学習に適用しやすい [38, 6]。BatchBALD [52] は、ベイジアンニューラルネットワークを用いた能動学習である BALD [45]（逐次方式）のバッチ方式版である。BatchBALD は、バッチ方式により時間効率を高めながら、一回の獲得バッチに含まれる複数標本の独立性を確保することで、最適な学習データが集めることを目的とする。

前述のとおり能動学習は複雑な特徴空間を持つ深層学習と併用することで再び注目されているが、近年の能動学習の研究では、深層学習の特徴量の活用や、深層学習により出現した問題設定（セマンティックセグメンテーション）固有の拡張などが研究されている。深層学習の複雑な特徴空間を用いることで、単純な標本間の距離が意味をなさない画像や動画などの高次元データに対して、能動学習の効果を向上する手法 [76] や、セマンティックセグメンテーションなどラベル自体が高次元で一つの標本をアノテーションするだけでも高コストな状況で、一つの標本の中で限定領域だけをアノテーション対象として獲得することで、高次元ラベル全体にアノテーションすることを回避する [104] など手法が研究されている。

本章に関係する能動学習の最新研究として、能動学習へのデータ拡張の適用がある。データ拡張とは、学習データにある標本を変換して異なる標本を生成すし、学習データ量を増やす（能動学習ではなく、モデル学習のための）技術である [28, 102, 89, 99]。また、ほとんどの最新の能動学習研究 [104, 76] は、まだ能動学習技術と学習技術の統合を考慮していない。しかし、LADA (Look-Ahead Data Acquisition) は、データプール中の未アノテーション標本と、その未アノテーション標本をデータ拡張した標本の両方を能動学習に利用することで、能動学習とデータ拡張を統合することを試みた [51]。

一般的に、深層学習は学習データ全体の特性を統計的に捉えることを得意とする。しかし、現実の深層学習の応用では、学習データに複数のドメインが含まれることが一般的であり、ドメインによっては重要度やリスクが異なる。3 章の未知不均衡ドメイン機械学習では、各学習データ標本のドメイン起源が未知かつ各ドメインの標本数が不均衡（多数派ドメインと少数派ドメインが混在している）という条件で、ドメインごとにテストデータ

を用意して各ドメインの性能を評価することで、現実の深層学習の応用に近い問題への対処を試みた [58]。3 章の未知不均衡ドメイン学習は深層学習モデルの学習のみに焦点を当てたが、未知不均衡ドメインの問題は能動学習のデータプールにも存在する。

## 4.3 能動学習の実験設定

能動学習の実験設定では、図 4.1 に示すように、アノテーション済みデータのラベルを隠して能動学習手法の性能を確認する。能動学習の活用状況に即し、現実的な巨大なアノテーション済みデータプールを作成するかが重要なポイントの一つである。もう一つのポイントは、能動学習においても、学習データのデータ拡張やバリデーションデータのサイズなど、現実の実務に即した深層学習の設定を考慮することである。一般的な能動学習の実験設定を概観したうえで、現実的な実験設定について述べる。

近年の能動学習研究 [52, 51, 76, 104] の実験設定を検討する。第一に、能動学習の実験で用いるデータプールの大きさが重要である。能動学習を実務で利用する場合は、アノテーションなしデータの量は膨大であることが前提である。しかし、能動学習の研究では、実験に用いるアノテーション付きデータセット（元データ）の標本数が十分でない場合が多い。そこで、元データの標本を複製する。例えば、ある研究では [52]、元データ標本のデータ要素（画像データの画素値）にガウス雑音を加えることで、仮想的に巨大なデータプールを生成する。その結果、生成されたデータプールの中に類似した標本が含まれる。このようなプールデータ生成により、データプールの分布は元データの分布から変化してしまう。

膨大なアノテーション済みデータプールから、能動学習手法が情報の多い標本を獲得し、学習データに組み込んだ上で深層学習モデルを学習する。このとき、データ拡張などの深層学習モデルの学習技術によっては能動学習手法の優劣を逆転させる可能性もある。そこで第二に、能動学習の研究であっても、深層学習モデルを学習する戦略が重要である。4.2 節で述べた通り、能動学習の研究では、最近まで学習時にデータ拡張が組み合わせられることはなかった。しかし、LADA は能動学習においてデータ拡張が重要であることを指摘し、能動学習とデータ拡張を組み合わせる手法を開発した [51]。

さらに、現実的な設定とするためにバリデーションデータのサイズを考慮する。能動学習の既存研究 [52] では、学習データの数十倍のバリデーションデータを用いるものがある。しかし、現実の深層学習の応用では、通常、学習データよりもバリデーションデータが小さい。巨大なバリデーションデータにアクセスできる状況では、バリデーションデータを学習データとして利用することが自然である。

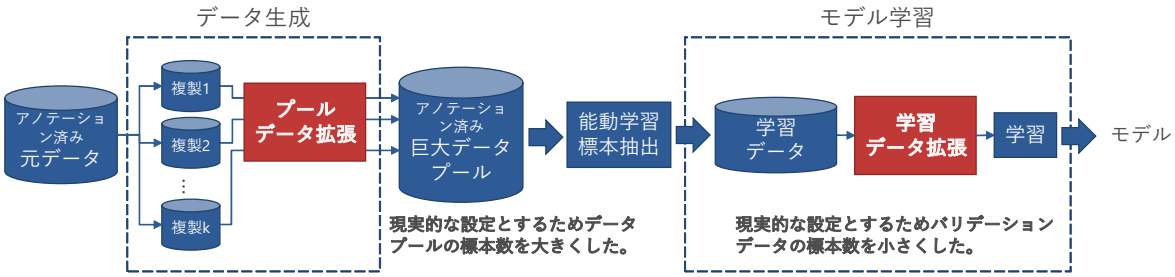



図 4.2: データ生成とモデル学習に着目した能動学習の現実的な実験設定

次に、図 4.1 に示した実際の状況に近い実験設定として、図 4.2 に示す現実的な実験設定を検討する。本章で提案する現実的な実験設定のポイントは以下の 2 点である。

1. データの多様性を確保するため、プールデータ拡張をデータ生成に適用する。
2. 実開発で多用され性能向上に寄与する学習データ拡張をモデル学習に適用する。

第一に、アノテーション済みの元データの標本を複製したあとに、データ拡張を適用することで巨大データプールを生成する。その結果、データプールには多種多様な標本が存在する、実世界の応用に近い条件が得られる。第二に、データプールから能動学習により標本を獲得して学習データに統合し、深層学習モデルを学習する。深層学習モデルの学習では、推論性能を向上させるための基本的な深層学習の実践として、データ拡張を行う。また、現実的な設定のために、深層学習モデルの開発におけるデータ利用の実践として、バリデーションデータの標本数を学習データの標本数と同等以下とする。ここから、図 4.2 の 2 つのポイントである、プールデータ拡張と学習データ拡張について説明する。

### 4.3.1 プールデータ拡張

能動学習の実験で用いる巨大なアノテーション済みデータプールが多様な標本を含むことが望ましい。そのため、元データを単純に複製するだけでは不十分である。そこで、現実的な能動学習の実験設定（図 4.2）では、アノテーション済みの元データを $k$ 回複製した後、重複する標本にデータ拡張 [28, 102, 89, 99] を適用し、巨大アノテーション済みデータプールを作成する。プールデータ拡張の手法として、ランダムアフィン変換 [27] やランダム切り抜き [94, 111] など、より現実の画像のバリエーションに近い変換を用い

る。このようなデータ生成プロセスで、データプール内の多様性を確保することができる。データプールの生成過程におけるデータ拡張を、プールデータ拡張と呼ぶ。

### 4.3.2 学習データ拡張

能動学習においても、深層学習モデルの学習時にデータ拡張を行うことができる [51]。プールデータ拡張と区別するために、本研究では深層学習モデルの学習時のデータ拡張を学習データ拡張と呼ぶ。学習データ拡張は、深層学習モデルの学習結果（学習済みモデルの精度）を向上させるため、現実の深層学習の開発でデータ拡張を用いることが多い。例えば、多くの深層学習フレームワークにはデータ拡張が基本機能として実装され、すぐに使えるようになっている。そこで、現実的な能動学習の実験設定（図 4.2）では、学習データ拡張を行うこととする。

### 4.3.3 その他の現実的な設定

その他の設定を述べる。バリデーションデータは、深層学習モデルの学習の一部として、モデル選択（深層学習ニューラルネットワークの設計など）、ハイパーパラメータのチューニング、学習の早期停止などのために用いられる。深層学習では、学習データを深層学習モデルのパラメータ更新に用い、バリデーションデータをハイパーパラメータの選択などに用いる。そのため、深層学習モデルのパラメータ更新に、バリデーションデータを直接用いることはない。そのため、通常は学習データがバリデーションデータより大きくなるように、入手したアノテーション済みデータを学習データとバリデーションデータに分割する。そこで、現実的な能動学習の実験設定（図 4.2）でも、学習データと同程度かそれ以下のサイズのバリデーションデータを用意することとする。

## 4.4 未知不均衡ドメイン能動学習の問題設定

能動学習の実験で用いるデータプールは、様々なドメインの標本を含む。一方、各ドメインは均衡、つまり各ドメインの標本が同程度データプールに含まれているわけではない。ここで、標本数が多いドメインを多数派ドメイン、少ないドメインを少数派ドメインと呼ぶ。しかし、実際の産業応用では少数派ドメインが重要な状況がある。例えば、自動運転の事故やクレジットカードの不正利用は重要であるが、標準的な標本よりはるかに少ない。そのため、多数派ドメインの性能を維持したまま、少数派ドメインの性能を向上させることが重要である。未知不均衡ドメイン機械学習の問題設定は 3.3 節で述べた。未知

不均衡ドメイン能動学習では、能動学習とモデル学習を組み合わせ、多数派ドメインの性能を維持しつつ、少数派ドメインの性能を向上する。

## 4.5 実験

本節では、1. 能動学習の実験設定と 2. 未知不均衡ドメイン能動学習の実験を行う。

能動学習とモデル学習の共通実験設定を述べる。すべての実験で、能動学習の獲得標本数は 10 で、獲得モードはバッチ方式とする。獲得標本数とは、能動学習手法が 1 回の反復で獲得する標本の数である。実験において、逐次方式とバッチ方式の能動学習手法を比較する場合は、1 回の能動学習の反復の中で、データプールから獲得スコアが大きい順に複数の標本を取りだすことで、逐次方式の能動学習手法を簡易的にバッチ方式として動作させることができる。データプールの初期標本数は $300,000$ である。乱択した 20 標本を含む初期学習データから開始し、学習データの最大標本数が 320、すなわち、能動学習の反復回数が 30 回、獲得標本数の合計が 300 個で終了する。なお、学習データの最大標本数は先行研究 [52] に整合させた。深層学習モデルの学習では、活性化関数を ReLU [29, 73] に置き換えた LeNet [62] のベイジアンニューラルネットワークに、分類タスクである手書き数字認識を学習させる。推論時には、MC dropout [31, 5, 22] を実行し、得られた複数のロジット（softmax 関数 [17] の入力）の平均を計算した上で softmax 関数に入力し、softmax 確率を得る。Dropout 確率は 0.5 とし、MC dropout の推論標本数は 10 とした。深層学習モデルの学習には復元抽出 [95] を用い、バッチサイズ 128 で 40 エポック学習後にテスト精度を測定した。

能動学習では学習データを増やしながら繰り返し精度を計測する。能動学習が選択した累積標本数を獲得標本数と呼ぶ。能動学習を実行すると、横軸を獲得標本数、縦軸をテスト精度とする学習曲線が結果として得られる。本章の実験では、能動学習手法の評価に指標 ALC (Area under Learning Curve) を用いる [39]。ALC は、能動学習における学習曲線の正規化下部面積で、0.0 から 1.0 の値をとる。図 4.3 に ALC の概要を示す。能動学習が最初に標本を獲得してすぐに精度 100% を達成し、能動学習の終了まで維持した場合、ALC が 1.0 となる。能動学習やモデル学習の結果にはばらつきが生じるため、全ての実験において 0 から 3 の乱数シードで 4 回の試行を行う。

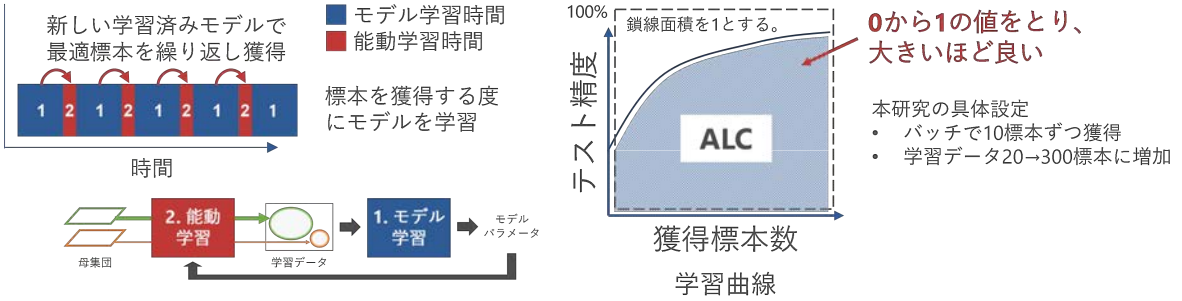



図 4.3: Area under Learning Curve (ALC) の概要

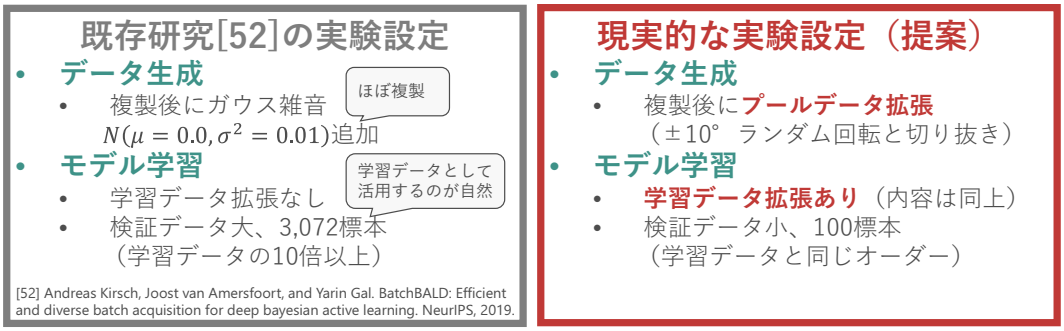


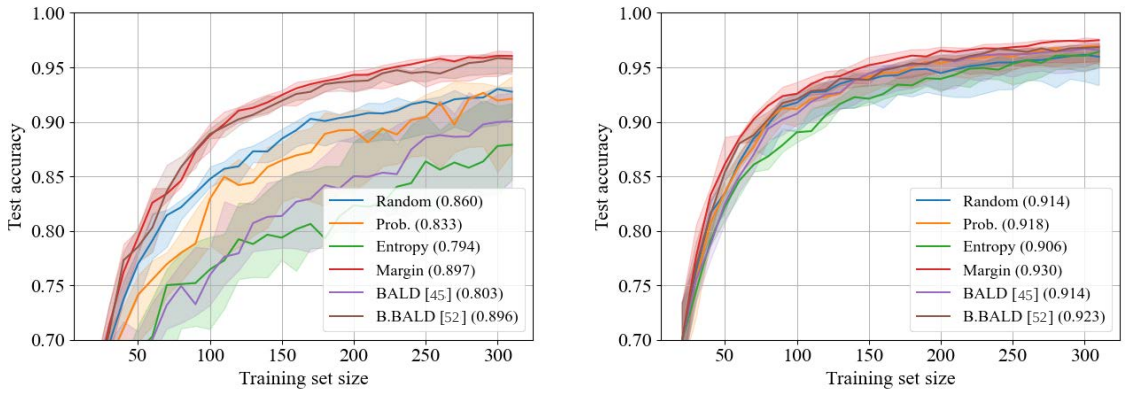

図 4.4: 比較実験における具体的な実験設定

### 4.5.1 従来の能動学習実験設定と現実的な能動学習実験設定の比較

本節では、既存の能動学習研究 [52] の実験設定例と、本章が提案する現実的な実験設定を比較する（図 4.4）。

既存の能動学習の実験設定例では、元データを複製して巨大なアノテーション済みデータプールを作成する。プールデータ拡張は行わず、$\mu = 0.0, \sigma = 0.1$ のガウス雑音を画像データの各画素値に付加することで、完全に同一な標本は含まれることを回避する。深層学習モデルは学習データ拡張なしで学習する。バリデーションデータの標本数は $3,072$ で、学習データの最大標本数 320 の約 10 倍である。

一方、本章で提案する現実的な実験設定では、アノテーション済みの元データを複製した後にプールデータ拡張を行い、巨大なアノテーション済みデータプールを生成する。深層学習モデルは、学習データ拡張を用いて学習する。ただし、プールデータ拡張と学習データ拡張には、同じデータ拡張を用いる。データ拡張の方法は、角度 $-10°$ から $10°$ のランダム回転とランダム切り抜きである。バリデーションデータの標本数は 100 で、学習




(a) 既存の能動学習研究 [52] の実験設定例での結果　　(b) 現実的な実験設定（提案）での結果

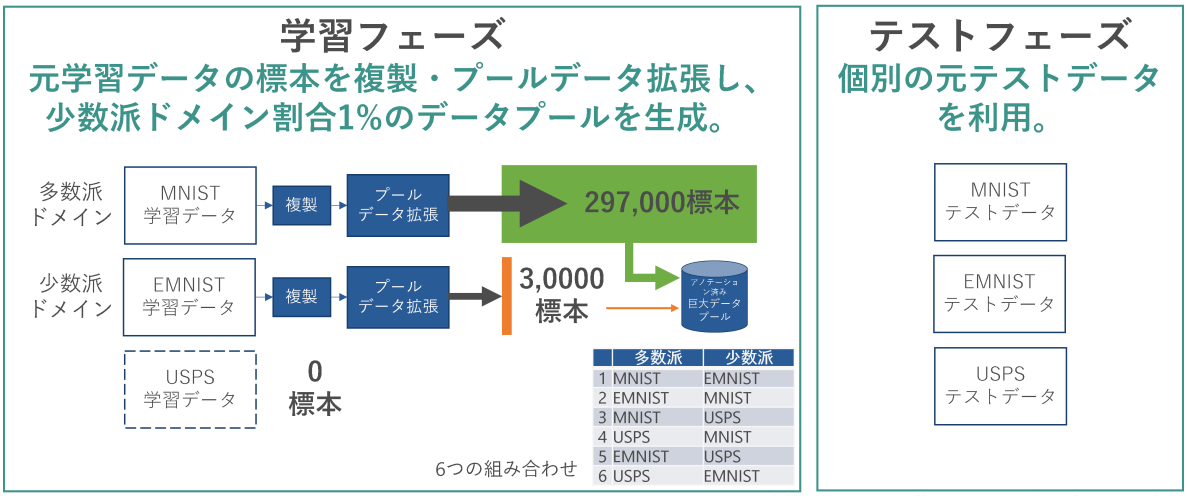
図 4.5: 異なる能動学習実験設定における能動学習アルゴリズムのテスト精度

データの最大標本数 320 と同オーダーである。

図 4.5 は、既存研究 [52] の実験設定例（図 4.5a）と提案の現実的な実験設定（図 4.5b）における、様々な能動学習手法のテスト精度である。太線は、3 つのデータセット MNIST [61], EMNIST [19], USPS [47] に対する 4 回試行した全 12 試行の中央値で、影は四分位を表す。ALC スコアを凡例の能動学習手法名の右横に示す。

図 4.5a では、逐次方式である BALD [45] は、乱択による能動学習のベースラインより、性能が低かった。一方、図 4.5b では、すべての能動学習手法が同様のテスト精度を達成した。また、テスト精度の分散も小さかった。実験設定を変更することで、BALD [45] の性能は改善した。BALD [45] は図 4.5a と図 4.5b ともに、学習データの標本数が増えれば性能は向上するが、図 4.5b では特に学習曲線が急峻である。BatchBALD [52] でも、図 4.5b の能動学習開始時点では、乱択による能動学習のベースライン同等の性能だった。図 4.5b を拡大すると、最も単純な能動学習の一つである softmax マージンが、能動学習の最初から最後まで最も高い精度を達成した。テスト精度の分散についても、図 4.5b より図 4.5a が小さかった。

以上のように、異なる実験設定で得られた実験結果には、大きな隔たりが観測された。そのため、能動学習の手法を現実的な設定で評価することが重要である。




※図は多数派：MNIST、少数派：EMNISTの組み合わせの場合

図 4.6: 2 ドメイン設定における未知不均衡ドメイン能動学習の実験データ作成

### 4.5.2 未知不均衡ドメイン能動学習における能動学習とモデル学習手法の比較

前節では、能動学習手法の比較において、実験設定の重要性が明らかになった。本節では、図 4.2 の現実的な実験設定を用いて、未知不均衡ドメイン能動学習の手法を検討する。

まず、実験設定の詳細を述べる。データプールの中に未知不均衡ドメインを再現する。MNIST [61], EMNIST [19], USPS [47] の元データ（それぞれ M, E, U と略す）のうち、2 つの元データを選択した 6 つのドメインペアを作り、各ドメインペアにデータプールを作る。例えばドメインペア M/E は、多数派ドメイン MNIST と少数派ドメイン EMNIST の組み合わせを表す。そして、ドメインごとのテスト精度を計測し、ドメインごとの学習曲線と ALC スコアを得る。ここで、平均 ALC スコアとは、ドメインごとの ALC スコアの平均（6 ドメインペアの平均）とする。本実験では、データプールの 99% が多数派ドメインの標本で、残りを少数派ドメインの標本とする。つまり、少数派率は 1% である（図 4.6）。各標本のドメインは最初のデータプール生成時にのみ使用され、能動学習アルゴリズムに対してはドメインの割り当てを未知とする。

能動学習手法のベースラインとして乱択（乱数による獲得）、softmax 確率、softmax マージン、softmax エントロピーの 3 つの sotmax 法、BALD [45] と BatchBALD [52] の 2 つのベイズ法を比較する。Softmax 法は、1. softmax 確率は、推論クラスの確率、すなわち最大 softmax、2. softmax マージンは、第一位と第二位の最大 softmax 値の差、3.

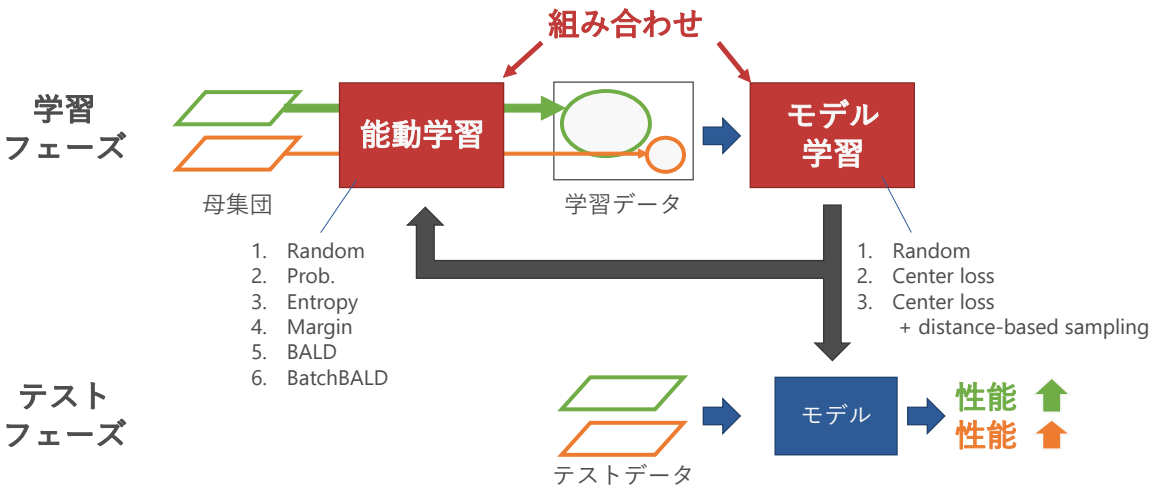



図 4.7: 未知不均衡ドメイン能動学習の能動学習手法とモデル学習手法の組み合わせ

softmax エントロピーは、softmax 値から計算したエントロピーをそれぞれ用いて、データプールから標本を獲得する。一方、モデル学習手法のベースラインとして乱択（乱数によるミニバッチ標本抽出）、center loss と乱択の組み合わせ、center loss とセントロイド距離（図 3.7）に基づいたミニバッチ標本抽出法の組み合わせの 3 学習手法を評価する。これらのモデル学習手法は、3 章で比較した未知不均衡ドメイン機械学習の手法の一部である。なお、center loss とセントロイド距離に基づいたミニバッチ標本抽出法は、LeNet の F6 層における深層特徴空間について実行する。能動学習手法とモデル学習手法の組み合わせを図 4.7 に示す。

表 4.1 に実験結果を示す。Center loss 付き乱択、center loss 付きの距離に基づいたミニバッチ標本抽出をそれぞれ R、R+C、D+C と略す。MNIST [61], EMNIST [19], USPS [47] を組み合わせた 6 つの多数派/少数派ドメインペアそれぞれに対して、個別 ALC スコアと平均 ALC スコアを示す。各ドメインペアの個別 ALC スコアは、4 つの乱数シードを用いた 4 回の試行の平均値である。太字は、各ドメインペアの個別 ALC スコア（多数派ドメイン ALC スコア、少数派ドメイン ALC スコア）と平均 ALC で、最良の手法を表す。

第一に、表 4.1 の最下段の平均 ALC スコアに注目すると、未知不均衡ドメインに対する能動学習（ドメイン別の結果）は、単純な能動学習（多数派ドメインだけの結果）と異なることがわかる。例えば、能動学習法：softmax エントロピーとモデル学習法：center loss の組み合わせは、多数派ドメインの ALC スコアでは、乱択ベースライン（0.911）よりも良い（0.913）結果を示している。しかし、この組み合わせは、少数派ドメインを考慮

している平均 ALC スコアでは、乱択ベースライン（0.872）よりも悪い（0.869 と 0.871）結果となる。このことから、少数派ドメインも考慮して能動学習アルゴリズムを評価するためには、未知不均衡ドメインの考え方を取り入れることが必要とわかる。多数派ドメインも少数派ドメインも考慮した平均 ALC スコアにより、softmax マージンによる能動学習と center loss と距離に基づいたミニバッチ標本抽出法による学習が、他の手法を凌駕していることがわかる。

ここで、表 4.1 の平均 ALC スコアと、多数派ドメインと少数派ドメインの個別 ALC スコアに注目する。第二に、平均 ALC スコアの観点では、能動学習は softmax マージン・学習は center loss と距離に基づいたミニバッチ標本抽出の組み合わせは、6 組のうち 3 組のドメインペアで他を上回り、残り 3 組のうち 2 組で最良の手法に匹敵する性能を達成した。最後に残ったドメインペア U/M では、最良手法より平均 ALC スコアが 0.01 低かった。第三に、少数派ドメイン ALC スコアでは、6 ドメインペアのうち 2 ドメインペアで、能動学習手法は softmax マージン、モデル学習手法は center loss とセントロイド距離に基づいたミニバッチ標本抽出の組み合わせが最高の性能を達成した。最新の能動学習法の一つである BatchBALD [52] は明示的に不均衡ドメインに対処する手法ではないが、他の 2 ドメインペアで少数派ドメイン ALC スコアで最良の結果を達成した。BatchBALD [52] は、U/M と U/E の 2 ドメインペアで、能動学習手法 softmax マージンとモデル学習手法 center loss と距離に基づいたミニバッチ標本抽出の組み合わせより 0.01 高い少数派ドメイン ALC スコアを達成したが、M/U と E/U という残りの 2 ドメインペアでは同等だった。第四に、多数派ドメイン ALC スコアでは、U/E を除くドメインペアで、能動学習は softmax マージン、モデル学習は center loss とセントロイド距離に基づいたミニバッチ標本抽出の組み合わせが常に他の手法を上回っていた。U/E でも 2 番手であり、同手法と最良手法との差は ALC スコアでわずか 0.003 だった。

以上の実験結果から、softmax マージンによる能動学習と、center loss とセントロイド距離に基づいたミニバッチ標本抽出の組み合わせは、多数派ドメインでの性能を維持したまま、少数派ドメインでの性能を向上させ、ほとんどの領域において平均 ALC スコアの観点でも最も優れていることがわかった。

表 4.1: 未知不均衡ドメイン能動学習におけるドメイン別 ALC スコア [39]

| Active learning | | Random acquisition | | | Softmax probability | | | Softmax entropy | | | Softmax margin | | | BALD [45] | | | BatchBALD [52] | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model training | | R | R+C | D+C | R | R+C | D+C | R | R+C | D+C | R | R+C | D+C | R | R+C | D+C | R | R+C | D+C |
| M/E | Major | 0.922 | 0.930 | 0.925 | 0.926 | 0.932 | 0.928 | 0.911 | 0.921 | 0.913 | 0.933 | 0.941 | **0.943** | 0.917 | 0.924 | 0.928 | 0.930 | 0.934 | 0.927 |
| | Minor | 0.890 | 0.897 | 0.880 | 0.883 | 0.886 | 0.876 | 0.841 | 0.847 | 0.853 | 0.897 | 0.909 | **0.914** | 0.884 | 0.889 | 0.889 | 0.892 | 0.898 | 0.897 |
| | Mean | 0.906 | 0.914 | 0.903 | 0.904 | 0.909 | 0.902 | 0.876 | 0.884 | 0.883 | 0.915 | 0.925 | **0.928** | 0.900 | 0.906 | 0.909 | 0.911 | 0.916 | 0.912 |
| E/M | Major | 0.929 | 0.937 | 0.935 | 0.928 | 0.932 | 0.937 | 0.916 | 0.927 | 0.925 | 0.943 | 0.948 | **0.949** | 0.922 | 0.928 | 0.920 | 0.936 | 0.928 | 0.935 |
| | Minor | 0.888 | 0.900 | 0.900 | 0.896 | 0.908 | 0.906 | 0.888 | 0.899 | 0.895 | 0.908 | 0.919 | **0.920** | 0.886 | 0.906 | 0.890 | 0.906 | 0.898 | 0.912 |
| | Mean | 0.908 | 0.919 | 0.918 | 0.912 | 0.920 | 0.921 | 0.902 | 0.913 | 0.910 | 0.926 | **0.934** | **0.934** | 0.904 | 0.917 | 0.905 | 0.921 | 0.913 | 0.923 |
| M/U | Major | 0.915 | 0.923 | 0.927 | 0.929 | 0.931 | 0.935 | 0.917 | 0.919 | 0.922 | 0.935 | 0.941 | **0.944** | 0.924 | 0.928 | 0.928 | 0.929 | 0.937 | 0.935 |
| | Minor | 0.787 | 0.761 | 0.755 | **0.792** | 0.761 | 0.754 | 0.759 | 0.730 | 0.740 | 0.772 | 0.771 | 0.780 | 0.776 | 0.750 | 0.750 | 0.785 | 0.758 | 0.758 |
| | Mean | 0.851 | 0.842 | 0.841 | 0.860 | 0.846 | 0.844 | 0.838 | 0.825 | 0.831 | 0.854 | 0.856 | **0.862** | 0.850 | 0.839 | 0.839 | 0.857 | 0.847 | 0.847 |
| U/M | Major | 0.895 | 0.903 | 0.903 | 0.910 | 0.910 | 0.913 | 0.895 | 0.897 | 0.899 | 0.916 | **0.925** | **0.925** | 0.902 | 0.897 | 0.896 | 0.903 | 0.903 | 0.905 |
| | Minor | 0.708 | 0.726 | 0.721 | 0.801 | 0.805 | 0.808 | 0.775 | 0.789 | 0.785 | 0.792 | 0.784 | 0.785 | 0.834 | 0.840 | 0.838 | 0.842 | 0.850 | **0.862** |
| | Mean | 0.802 | 0.815 | 0.812 | 0.855 | 0.858 | 0.861 | 0.835 | 0.843 | 0.842 | 0.854 | 0.854 | 0.855 | 0.868 | 0.868 | 0.867 | 0.872 | 0.876 | **0.884** |
| E/U | Major | 0.923 | 0.929 | 0.933 | 0.930 | 0.933 | 0.938 | 0.914 | 0.915 | 0.920 | 0.941 | **0.945** | **0.945** | 0.922 | 0.927 | 0.923 | 0.930 | 0.936 | 0.939 |
| | Minor | 0.874 | 0.861 | 0.850 | 0.872 | 0.855 | 0.857 | 0.834 | 0.826 | 0.837 | **0.882** | 0.881 | 0.873 | 0.852 | 0.848 | 0.833 | 0.880 | 0.860 | 0.843 |
| | Mean | 0.898 | 0.895 | 0.892 | 0.901 | 0.894 | 0.898 | 0.874 | 0.870 | 0.879 | 0.911 | **0.913** | 0.909 | 0.887 | 0.888 | 0.878 | 0.905 | 0.898 | 0.891 |
| U/E | Major | 0.885 | 0.895 | 0.901 | 0.905 | 0.911 | 0.906 | 0.892 | 0.898 | 0.898 | 0.916 | **0.926** | 0.923 | 0.894 | 0.903 | 0.898 | 0.905 | 0.915 | 0.914 |
| | Minor | 0.846 | 0.856 | 0.855 | 0.865 | 0.868 | 0.878 | 0.859 | 0.863 | 0.866 | 0.881 | 0.887 | 0.887 | 0.879 | 0.884 | 0.878 | 0.888 | **0.903** | 0.895 |
| | Mean | 0.865 | 0.876 | 0.878 | 0.885 | 0.889 | 0.892 | 0.875 | 0.881 | 0.882 | 0.898 | 0.906 | 0.905 | 0.887 | 0.894 | 0.888 | 0.896 | **0.909** | 0.904 |
| Avg. | Major | 0.911 | 0.920 | 0.921 | 0.921 | 0.925 | 0.926 | 0.908 | 0.913 | 0.913 | 0.931 | **0.938** | **0.938** | 0.914 | 0.918 | 0.916 | 0.922 | 0.925 | 0.926 |
| | Minor | 0.832 | 0.833 | 0.827 | 0.852 | 0.847 | 0.846 | 0.826 | 0.826 | 0.829 | 0.855 | 0.859 | 0.860 | 0.852 | 0.853 | 0.846 | **0.865** | 0.861 | 0.861 |
| | Mean | 0.872 | 0.877 | 0.874 | 0.886 | 0.886 | 0.886 | 0.867 | 0.869 | 0.871 | 0.893 | 0.898 | **0.899** | 0.883 | 0.885 | 0.881 | 0.894 | 0.893 | 0.894 |

## 4.6 本章のまとめ

本章では、能動学習手法が異なる実験設定で異なる振る舞いをすることを示し、実験設定が適切な能動学習の手法を選択するために重要であることを示した。現実的な実験設定として、1. 大規模データプールの生成時のプールデータ拡張と、2. 現実的な深層学習モデルの学習方法を模擬する学習データ拡張の実験への導入を提案した。現実的な設定とするためバリデーションデータのサイズも変更した。次に、3 章を拡張して、実応用で重要な未知不均衡ドメインを能動学習に導入した。最後に、提案した現実的な実験設定において、未知不均衡ドメインに対して最適な能動学習の手法とモデル学習の手法の組み合わせを検討した。その結果、softmax マージンを用いた能動学習と center loss とセントロイド距離に基づいたミニバッチ標本抽出を用いた学習の組み合わせが、多数派ドメインと少数派ドメインの両方に対して有効であることがわかった。

# 第 5 章

# 結言

深層学習には、データ駆動の限界と複雑なモデルの限界がある。複雑なモデルの限界とは、推論の過程や根拠を人間が理解できないため、問題を特定、修正、改善するというエンジニアリングが難しいことである。データ駆動の限界とは、現実の未知不均衡ドメインに対して、ドメインで性能が偏ることである。これらの深層学習の限界に対処するため、本研究では、深層学習を用いた画像認識において、推論根拠と未知不均衡ドメイン学習の研究に取り組んだ。ただし、未知不均衡ドメインのデータとは、様々なドメイン（収集条件）の標本が異なる均衡（バランス）で混在し、標本の所属ドメインを知ることができないデータである。

推論根拠に関する研究では、特徴量に着目して推論根拠を示し、深層学習モデルのエンジニアリング性を高めることに取り組んだ。推論時の特徴量に着目して深層学習モデルの推論根拠を解析する手法を構築し、クラウドソーシングの評価で妥当性を示すとともに、生成した推論根拠により、エンジニアリングのためのヒントが得られる可能性を考察した。本研究により、深層学習モデルのブラックボックス性が完全に解決するものではないが、深層学習モデルの中間特徴量を分析するアプローチが、深層学習モデルのエンジニアリング性を高めるために有効とわかった。

未知不均衡ドメイン学習に関する研究では、未知不均衡ドメインのデータに対し、多数派ドメイン性能を維持しつつ、少数派ドメイン性能を向上することに取り組んだ。未知不均衡ドメイン機械学習の問題設定を行い、セントロイド距離を用いた提案手法（機械学習の手法）の有効性を示すとともに、現実世界（母集団）から標本を抽出し学習データを作成する能動学習にも未知不均衡ドメインを拡張した。本研究により、未知不均衡ドメインに対して、適切なモデル学習と能動学習を組み合わせたアプローチが有効とわかった。

以上から、本研究は、深層学習を用いた画像認識における推論根拠と未知不均衡ドメイ

ン学習の有効なアプローチを提案しており、研究学術的にも社会実装上も意義がある。以下、順を追って各章を振り返る。

第 2 章「深層学習の推論根拠の解析」では、深層学習ネットワークの特徴量に着目し、構造的特徴分析、言語的特徴分析、整合性分析の 3 種類の分析法を構築した。構造的特徴分析として、特徴量 ID の形式で推論根拠を提示した。言語的特徴分析として、人間が特徴量 ID の意味を理解できる特徴量ラベルを付加した。整合性分析として、提案手法で生成した推論根拠の整合性をクラウドソーシングで評価した。その結果、生成した推論根拠は、70% 以上の人間から理解可能と同意が得られたので、提案手法は、人間への説明としてある程度利用可能と類推される。これを受け、推論根拠の個別結果を検討した結果、ラベル間違いパターン、クラス混在パターン、特徴混在パターン、状況不足パターン、特徴抽出不足パターンなど、深層学習モデルのエンジニアリングに有用な解析ができ、深層学習モデルの改善への示唆が得られた。

第 3 章「未知不均衡ドメイン機械学習」では、未知不均衡ドメイン機械学習の問題設定を定式化し、特徴量の空間における標本間距離を利用した学習手法を提案した。未知不均衡ドメイン機械学習の問題設定では、学習データに様々なドメインが含まれ、それらのドメインが未知かつ多数派と少数派が混在するという状況で、多数派ドメインの性能を維持しつつ少数派ドメインの性能を向上させることを目指す。提案手法では、特徴空間における各標本と全標本の重心の距離（セントロイド距離）を近づける center loss と、セントロイド距離の遠い標本を重点的に選択するミニバッチ抽出を組み合わせた。実験により、提案手法は、多数派ドメインのドメイン精度を損なうことなく、少数派ドメインで大幅なドメイン精度向上を達成することを示した。

第 4 章「未知不均衡ドメイン能動学習」では、能動学習の現実的な実験設定の整理と、未知不均衡ドメインの問題設定による最適な能動学習とモデル学習の手法の組み合わせを構築した。先行研究の中には、能動学習のソースとなる未アノテーションデータ（データプール）の多様性がないなどの現実と乖離した能動学習の実験設定があった。そこで本研究では、深層学習が実応用で使われる状況を想定し、現実的な実験設定を整理した。次に、未知不均衡ドメイン機械学習を能動学習に拡張し、様々なモデル学習の手法と能動学習の手法の組み合わせを評価した。実験により、第 3 章「未知不均衡ドメイン機械学習」で提案したモデル学習の手法と、softmax 確率を利用した単純な能動学習の手法の組み合わせが、未知不均衡ドメイン能動学習での各ドメイン性能を向上させることを示した。

本論文の研究はあくまで基礎検討であり、今後は基礎検討の結果に基づいた応用研究が期待される。本論文では、ImageNet や MNIST などのトイデータを用いて研究を勧めてきた。推論根拠の解析や未知不均衡ドメイン学習・能動学習を、自動運転システムなどの

実データに適用するには、様々な技術課題がある。以下、4 つの考えられる技術課題（推論根拠解析の課題 3 件、未知不均衡ドメイン学習の課題 1 件）を述べる。

推論根拠解析に関して、第一に、スケーラビリティ向上を狙った特徴量アノテーション（人間による作業）の効率化が必要である。本研究で提案した推論根拠の解析方法は、人間による特徴量アノテーションが必要であり、深層学習の推論過程と人間の思考過程の乖離については、多数の作業者を使ってアンケートをとることで評価した。人間によるタスクを実行するコストは巨大であり、また、各モデルを学習するとパラメータはすべて変化するので、モデルが変わるごとにやり直しが発生する。そこで、特徴量ラベルや対応する視覚属性の（社内）共有などで、コストがかかる特徴量アノテーションの作業を効率化できる可能性がある。視覚属性と特徴量ラベルの大規模なデータセットや、特徴量の組み合わせと推論の論理的な対応関係を記述したデータセットを構築するなど、特徴量アノテーションと整合性評価の自動化は応用に向けた重要な課題である。例えば、computer vision のテストデータに含めるべき困難な認識条件をインターネット上で共有する CV-HAZOP[108] がある。同様に、特徴量ラベルや対応イメージの（社内）共有することで、最もコストがかかる特徴量アノテーションの作業を効率化することが考えられる。

推論根拠解析に関して、第二に、深層学習は画像分類だけでなく、様々なタスクで利用されている。今回は画像分類に特化したが、他のタスクへも拡張が必要である。今回の問題設定である分類問題では、一枚の画像に対して 1 つの推論結果があり、推論根拠特徴量 ID と推論結果の関係を結び付けられた。しかし、例えば検出問題（bounding box 回帰）では 4 つの推論結果（座標）がある。そのため、推論根拠特徴量 ID がどの値に影響しているかを解析する必要があり、また、複雑な解析結果を人間に理解できる形式で提示するインタフェースの検討など、技術的課題がある。

推論根拠解析に関して、第三に、解析結果に基づいたエンジニアリング方法の整理が必要である。今回の研究では、それぞれの推論根拠解析の結果を見て、テストデータのラベル更新や特定の学習データの増加などを個別検討した。本研究では、推論根拠解析は最終目的ではなく、解析結果を用いた深層学習のエンジニアリングで、改善を行うことを目指していた（図 5.1）。そこで、今後は、多数の事例をもとにエンジニアリングの活動を類型化し、推論根拠解析の結果に基づいて性能向上の手が打てるエンジニアリング事例集などを作る課題がある。実用化に向けて、推論根拠解析の結果のどこを見てどのようなアクションをとるべきか、多数の事例をもとに類型化し、推論根拠解析の結果が得られれば誰でも深層学習モデルの性能向上できるような、深層学習エンジニアリングの事例集やガイドラインなどを作ることが必要である。

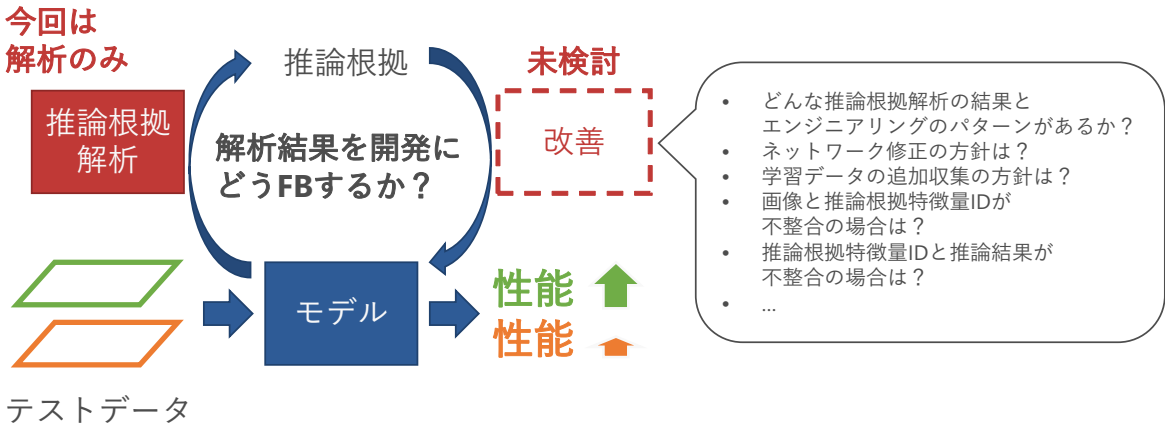



図 5.1: 推論根拠解析を用いた深層学習モデルの改善

最後に、未知不均衡ドメイン学習に関して、運転データなど現実的なデータへの応用が必要である。今回は未知不均衡ドメインのデータを模擬しやすい問題設定として、手書き文字認識を基礎検討した。手書き文字認識には、MNIST、EMNIST、USPS の 3 類似データセット存在するため、それらの混合率を変化させることで、未知不均衡ドメインデータを模擬することができた。次のステップとして、BDD100K [107]、ImageNet [23, 84, 80]、CIFAR10 [54] などの自然画像を用いて、未知不均衡ドメインを評価できるデータセットを構築することが必要である（図 5.2）。図 3.2 の通り、現実的にはデータセットに多数のドメインが入り込むことは避けられないが、現在の実験データセットは基本的には単独ドメインを想定している。例えば、ImageNet の構築 [23, 84, 80] は、Collecting Candidate Images（Flickr 画像、YouTube 動画、Google Image Search データベースなど複数の画像検索エンジンから、英語、中国語、スペイン語、オランダ語、イタリア語を用いて各クラスの候補画像を検索して取得）と Cleaning Candidate Images（Amazon Mechanical Turk を用いたクラウドソーシングでクラスの正しさを多数決投票）の二段階のステップから成る。なお、ImageNet クラスは WordNet [68, 67] に階層的に定義されている名詞概念に対応するが、詳細は省略する。検索エンジンや各国の文化によって、検索できる画像の傾向（ドメイン）が異なると想定すると、ImageNet のデータ収集過程において、データソースである検索エンジンや検索言語をドメインとみなし、ドメイン別データセットを構築することもできたと考えられる。しかし、そのようなドメイン情報は失われ、ImageNet は単独ドメインのデータセットとして整備されている。深層学習は、人間が設計できない特徴をデータドリブンで抽出するため、データ収集の段階では、人間にはどのようなデータが重要かは完全にはわからない。未知不均衡ドメインに対する汎化能力を計

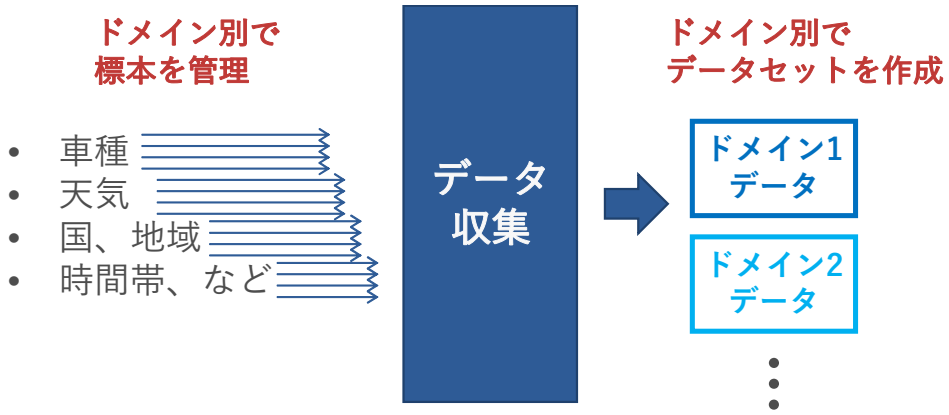



図 5.2: 未知不均衡ドメインデータの構築

測可能にし、深層学習の能力を引き出すためにも、データ収集において、設計者に今見えている問題だけではなく、ドメイン情報を残す仕組みが必要である。

# 参考文献


[1] *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015.

[2] Collision between a car operating with automated vehicle control systems and a tractor-semitrailer truck near williston, florida. https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR1702.pdf, 2016.

[3] Collision between vehicle controlled by developmental automated driving system and pedestrian tempe, arizona. https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR1903.pdf, 2018.

[4] NEF-104 PE21-020. https://static.nhtsa.gov/odi/inv/2021/INIM-PE21020-84913P.pdf, 2021.

[5] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.

[6] Javad Azimi, Alan Fern, Xiaoli Z. Fern, Glencora Borradaile, and Brent Heeringa. Batch active learning via coordinated matching. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 307–314, Madison, WI, USA, 2012. Omnipress.

[7] Mayank Bansal, Alex Krizhevsky, and Abhijit Ogale. Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. In *Proceedings of Robotics: Science and Systems*, FreiburgimBreisgau, Germany, June 2019.

[8] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[9] Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. In *Artificial Neural Networks and Machine Learning - ICANN 2016 - 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6-9, 2016, Proceedings, Part II*, pages 63–71, 2016.

[10] Christopher M Bishop. Pattern recognition. *Machine learning*, 128(9), 2006.

[11] Mariusz Bojarski, Anna Choromanska, Krzysztof Choromanski, Bernhard Firner, Larry D. Jackel, Urs Muller, and Karol Zieba. Visualbackprop: visualizing cnns for autonomous driving. *CoRR*, abs/1611.05418, 2016.

[12] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, and Karol Zieba. End to end learning for self-driving cars. *CoRR*, abs/1604.07316, 2016.

[13] George EP Box and George C Tiao. *Bayesian inference in statistical analysis*. John Wiley & Sons, 2011.

[14] Paula Branco, Luis Torgo, and Rita P Ribeiro. Rebagg: Resampled bagging for imbalanced regression. In *Second International Workshop on Learning with Imbalanced Domains: Theory and Applications*, pages 67–81. PMLR, 2018.

[15] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[16] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.

[17] John Bridle. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. *Advances in neural information processing systems*, 2, 1989.

[18] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[19] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2921–2926. IEEE, 2017.

[20] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145, 1996.

[21] Hoa Khanh Dam, Truyen Tran, and Aditya Ghose. Explainable software analytics. In *Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results, ICSE (NIER) 2018, Gothenburg, Sweden, May 27 - June 03, 2018*, pages 53–56, 2018.

[22] Josiah Davis, Jason Zhu, Jeremy Oldfather, Samual MacDonald, and Maciej Trzaskowski. Quantifying uncertainty in deep learning systems. https://d1.awsstatic.com/APG/quantifying-uncertainty-in-deep-learning-systems.pdf, 2020.

[23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[24] Weiguang Ding, Ruoyan Wang, Fei Mao, and Graham Taylor. Theano-based large-scale visual recognition with multiple gpus. *arXiv preprint arXiv:1412.2302*, 2014.

[25] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2650–2658, 2015.

[26] Victor Escorcia, Juan Carlos Niebles, and Bernard Ghanem. On the relationship between visual attributes and convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015* [1], pages 1256–1264.

[27] Alhussein Fawzi, Horst Samulowitz, Deepak Turaga, and Pascal Frossard. Adaptive data augmentation for image classification. In *2016 IEEE international conference on image processing (ICIP)*, pages 3688–3692. Ieee, 2016.

[28] Sylvia Frühwirth-Schnatter. Data augmentation and dynamic linear models. *Journal of time series analysis*, 15(2):183–202, 1994.

[29] Kunihiko Fukushima. Visual feature extraction by a multilayered network of

analog threshold elements. *IEEE Transactions on Systems Science and Cybernetics*, 5(4):322–333, 1969.

[30] Kunihiko Fukushima and Sei Miyake. Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern Recognition*, 15(6):455–469, 1982.

[31] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[32] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian active learning with image data. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1183–1192. PMLR, 06–11 Aug 2017.

[33] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[34] Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. Hybrid speech recognition with deep bidirectional LSTM. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, December 8-12, 2013*, pages 273–278, 2013.

[35] Geoffrey Grimmett and David Stirzaker. *Probability and random processes*. Oxford university press, 2020.

[36] Felix Grün, Christian Rupprecht, Nassir Navab, and Federico Tombari. A taxonomy and library for visualizing learned features in convolutional neural networks. *CoRR*, abs/1606.07757, 2016.

[37] David Gunning. Explainable artificial intelligence (xai). *https://www.darpa.mil/program/explainable-artificial-intelligence*, 2016.

[38] Yuhong Guo and Dale Schuurmans. Discriminative batch mode active learning. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

[39] Isabelle Guyon, Gavin C. Cawley, Gideon Dror, and Vincent Lemaire. Results of the active learning challenge. In Isabelle Guyon, Gavin Cawley, Gideon Dror, Vincent Lemaire, and Alexander Statnikov, editors, *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, volume 16

of *Proceedings of Machine Learning Research*, pages 19–45, Sardinia, Italy, 16 May 2011. PMLR.

[40] Elmar Haussmann, Michele Fenzi, Kashyap Chitta, Jan Ivanecky, Hanson Xu, Donna Roy, Akshita Mittel, Nicolas Koumchatzky, Clement Farabet, and Jose M. Alvarez. Scalable active learning for object detection. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1430–1435, 2020.

[41] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, pages 3–19, 2016.

[42] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International workshop on similarity-based pattern recognition*, pages 84–92. Springer, 2015.

[43] Steven CH Hoi, Rong Jin, Jianke Zhu, and Michael R Lyu. Batch mode active learning and its application to medical image classification. In *Proceedings of the 23rd international conference on Machine learning*, pages 417–424, 2006.

[44] Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

[45] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

[46] Dana Hull. The tesla advantage: 1.3 billion miles of data. *Bloomberg, December*, 20, 2016.

[47] Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554, 1994.

[48] Rami Ibrahim and M Omair Shafiq. Explainable convolutional neural networks: A taxonomy, review, and future directions. *ACM Computing Surveys*, 2022.

[49] Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002.

[50] Jack Kiefer, Jacob Wolfowitz, et al. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

[51] Yoon-Yeong Kim, Kyungwoo Song, JoonHo Jang, and Il-chul Moon. LADA: Look-ahead data acquisition via augmentation for deep active learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 22919–22930. Curran Associates, Inc., 2021.

[52] Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. BatchBALD: Efficient and diverse batch acquisition for deep bayesian active learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[53] Philip Koopman and Michael Wagner. Challenges in autonomous vehicle testing and validation. *SAE International Journal of Transportation Safety*, 4(2016-01-0128):15–24, 2016.

[54] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[55] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[56] Hiroshi Kuwajima and Fuyuki Ishikawa. Adapting square for quality assessment of artificial intelligence systems. In *2019 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*, pages 13–18. IEEE, 2019.

[57] Hiroshi Kuwajima, Masayuki Tanaka, and Masatoshi Okutomi. Improving transparency of deep neural inference process. *Progress in Artificial Intelligence*, 8:273–285, 2019.

[58] Hiroshi Kuwajima, Masayuki Tanaka, and Masatoshi Okutomi. Machine learning with blind imbalanced domains. *Electronic Imaging*, 34:1–6, 2022.

[59] Hiroshi Kuwajima, Masayuki Tanaka, and Masatoshi Okutomi. Evaluating active learning for blind imbalanced domains. *Electronic Imaging*, 35:1–8, 2023.

[60] Hiroshi Kuwajima, Hirotoshi Yasuoka, and Toshihiro Nakae. Engineering problems in machine learning systems. *Mach. Learn.*, 109(5):1103–1126, 2020.

[61] Yann LeCun. The mnist database of handwritten digits.

*http://yann.lecun.com/exdb/mnist/*, 1998.

[62] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[63] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[64] David J. C. MacKay. A practical bayesian framework for backpropagation networks. *Neural Comput.*, 4(3):448–472, may 1992.

[65] David J. C. MacKay. Probable networks and plausible predictions-a review of practical bayesian methods for supervised neural networks. *Network: computation in neural systems*, 6(3):469, 1995.

[66] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.

[67] George A Miller. *WordNet: An electronic lexical database*. MIT press, 1998.

[68] George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244, 1990.

[69] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *CoRR*, 2017.

[70] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017.

[71] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *CoRR*, abs/1706.07979, 2017.

[72] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5715–5725, 2017.

[73] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted

boltzmann machines. In *Icml*, 2010.

[74] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[75] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Attentive explanations: Justifying decisions and pointing to the evidence. *CoRR*, abs/1612.04757, 2016.

[76] Amin Parvaneh, Ehsan Abbasnejad, Damien Teney, Gholamreza Reza Haffari, Anton van den Hengel, and Javen Qinfeng Shi. Active learning by feature mixing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12237–12246, 2022.

[77] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[78] Luca Pizzuto, Christopher Thomas, Arthur Wang, and Ting Wu. How china will help fuel the revolution in autonomous vehicles. *McKinsey & Company, January*, 2019.

[79] Ce Qi and Fei Su. Contrastive-center loss for deep neural networks. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 2851–2855. IEEE, 2017.

[80] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019.

[81] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[82] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[83] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[84] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean

Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[85] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3723–3732, 2018.

[86] Ahmad El Sallab, Mohammed Abdou, Etienne Perot, and Senthil Kumar Yogamani. Deep reinforcement learning framework for autonomous driving. *electronic imaging*, 2017(19):70–76, 2017.

[87] Burr Settles. Active learning literature survey. *CS Technical Reports*, 2009.

[88] Burr Settles. Active learning. *su lectures on artificial intelligence and machine learning*, 6(1):1–114, 2012.

[89] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[90] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *CoRR*, abs/1605.01713, 2016.

[91] Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, Pascal Fua, and Francesc Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 118–126, 2015.

[92] Frederik J Simons, FA Dahlen, and Mark A Wieczorek. Spatiospectral concentration on a sphere. *SIAM review*, 48(3):504–536, 2006.

[93] Mahito Sugiyama and Karsten Borgwardt. Rapid distance-based outlier detection via sampling. *Advances in Neural Information Processing Systems*, 26:467–475, 2013.

[94] Ryo Takahashi, Takashi Matsubara, and Kuniaki Uehara. Data augmentation using random image cropping and patching for deep cnns. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(9):2917–2931, 2019.

[95] Yves Tillé. *Sampling algorithms*. Springer, 2006.

[96] Simon Tong. *ACTIVE LEARNING: THEORY AND APPLICATIONS*. PhD thesis, STANFORD UNIVERSITY, 2001.

[97] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001.

[98] Kazusa Uchida, Masayuki Tanaka, and Masatoshi Okutomi. Coupled convolution layer for convolutional neural network. *Neural Networks*, 105:197–205, 2018.

[99] David A Van Dyk and Xiao-Li Meng. The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50, 2001.

[100] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015* [1], pages 3156–3164.

[101] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016.

[102] Sebastien C Wong, Adam Gatt, Victor Stamatescu, and Mark D McDonnell. Understanding data augmentation for classification: when to warp? In *2016 international conference on digital image computing: techniques and applications (DICTA)*, pages 1–6. IEEE, 2016.

[103] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2840–2848, 2017.

[104] Binhui Xie, Longhui Yuan, Shuang Li, Chi Harold Liu, and Xinjing Cheng. Towards fewer annotations: Active learning via region impurity and prediction uncertainty for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8068–8078, 2022.

[105] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 2048–2057, 2015.

[106] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Re-

cent trends in deep learning based natural language processing. *CoRR*, abs/1708.02709, 2017.

[107] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2(5):6, 2018.

[108] Oliver Zendel, Markus Murschitz, Martin Humenberger, and Wolfgang Herzner. Cv-hazop: Introducing test data validation for computer vision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2066–2074, 2015.

[109] Peilin Zhao and Tong Zhang. Stochastic optimization with importance sampling for regularized loss minimization. In *international conference on machine learning*, pages 1–9. PMLR, 2015.

[110] Zhong-Qiu Zhao, Peng Zheng, Shou-Tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE Transactions on Neural Networks*, 30(11):3212–3232, 2019.

[111] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020.

[112] 桑島 洋, 平田 雄一, and 中江 俊博. 自動車業界における機械学習システムの品質確保の事例. **システム／制御／情報**, 66(5):187–194, 2022.