様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ５ 月 ３１ 日現在

機関番号：１２６０１
研究種目：若手研究
研究期間：2018～2022
課題番号：１８Ｋ１８１０１

研究課題名（和文）敵対的訓練を用いた制御可能な表現学習に関する研究

研究課題名（英文）A Study on Controllable Representation Learning using Adversarial Training

研究代表者
岩澤　有祐（Iwasawa, Yusuke）

東京大学・大学院工学系研究科（工学部）・講師

研究者番号：７０８０８３３６

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：研究期間を通じて、下記のような技術的な成果を得た。(1) 既存手法であるAdversarial Feature Learningの不安定性について解析を行い、解決する方法を提案した（IJCAI2020などに採択）。(2) 新しい不変性基準の提案。ある予測したい因子については情報を既存しない範囲で不変性を最大化する十分不変性という基準および十分不変性を達成する手法を提案した（ECML2019などに採択）。(3) ユーザが消したい情報についての詳細を与えることなく、データからそのような情報を削除するグラフィカルモデルに基づく枠組みとその実現する手法を提案した（ECML2021などに採択）。

研究成果の学術的意義や社会的意義
大規模言語モデルの登場などにより深層学習の実世界での活用は本格化しているが、通常の学習アルゴリズムは内部にあるバイアスを増長してしまう可能性がある。また、意図しない状況で不安定な挙動をすることがある。本研究の目的は、深層NNの表現が特定の情報を持たないように制御する要素技術の開発である。本研究成果により、未知ユーザの行動を高精度に認識したり、深層NNの判断基準が特定の因子によらないことを保証（プライバシー保護、公平性配慮）できると考えられる。

研究成果の概要（英文）：Throughout the research period, I achieved the following technical accomplishments:
(1) Analyzed the instability of the existing method, Adversarial Feature Learning, and proposed a solution (accepted in IJCAI2020 and other conferences). (2) Proposed a new criterion for invariance, called Sufficient Invariance, which maximizes invariance with respect to a factor of interest in an informationally novel range, and suggested methods to achieve Sufficient Invariance (accepted in ECML2019 and other conferences). (3) Proposed a framework based on graphical models to remove information from data without providing detailed information about the specific aspects the user wants to eliminate, and presented the corresponding methodology (accepted in ECML2021 and other conferences).

研究分野：Deep Learning

キーワード：Deep Learning　Firaness　Privacy　Transfer　Adversarial Training

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属します。

様　式　C－１９、F－１９－１、Z－１９（共通）

１．研究開始当初の背景

(1) 近年，人工知能の分野で深層学習が目覚ましい発展を見せている．深層学習とはデータの背後に階層構造を仮定することで，有用な表現（言い換えればタスクに関係のない情報が捨象され圧縮された）を効率的に獲得する技術である．
(2) しかし，ニューラルネット（以降 NN と表記する）がどの情報を捨象するべきかを明示的に設計したり，あるいは事後に確認するのは容易ではない．すなわち NN は通常最適化を介してデータやタスクをよく表す表現をブラックボックスに獲得するだけであり，「身体的特徴によらずロバストに識別してほしい」「肌の色といった社会通念上問題になる属性を判断に使ってほしくない」といった非依存性を持つ保証はない

２．研究の目的

本研究の問いは「特定の情報に依存した表現を学習しないように制御できるか？」である．以降，特定の情報に依存しないことを指すより専門的な言葉として，「因子に対して不変」という言葉を利用する．本技術により，何に不変であるべきかというドメイン知識を取り入れたロバストな分類や公平性へ配慮した分類が行えるようになり，実世界データへの深層学習応用の進展が期待できる．

３．研究の方法

本研究では特に，画像生成の領域で近年注目を集めている敵対的訓練を応用した方法を模索する．敵対的訓練を利用した手法とは，NN の中間表現 R から因子 S を予測する分類器 D を構築し，この D が因子 S を予測できない方向に（=S に不変になる方向に）NN を更新する方法である（図 1-a 参照）．つまり，分類器がどの程度正確に S を予測できるかを不変性の指標に利用してその指標をもとに学習する．このとき，片方の NN は S に関する予測性を最大化しようとし，片方の NN は予測性を最小化しようとするという敵対的な関係を有している（より形式的には，L を予測性の損失としたときに min max L という形式を取る）ため，敵対的訓練と呼ばれる．

この敵対的学習の方法は，本質的に大きく (1) どのような基準を最適化するのか，(2) その基準をどのように最適化するのかの 2 つの観点で整理することができる．本研究プロジェクトでは，この両輪で研究を行った．

４．研究成果

研究期間を通じて、下記のような技術的な成果を得た。

■ 新基準１：十分不変性
(1)　ドメイン汎化性能を高めることを念頭に置いた不変性のより適切な基準についての提案。ある予測したい因子については情報を既存しない範囲で最大の不変性を達成する十分不変性という基準を提案し（下図）、また十分不変性を達成する手法を提案した。本成果は ECML2019 に採択された。

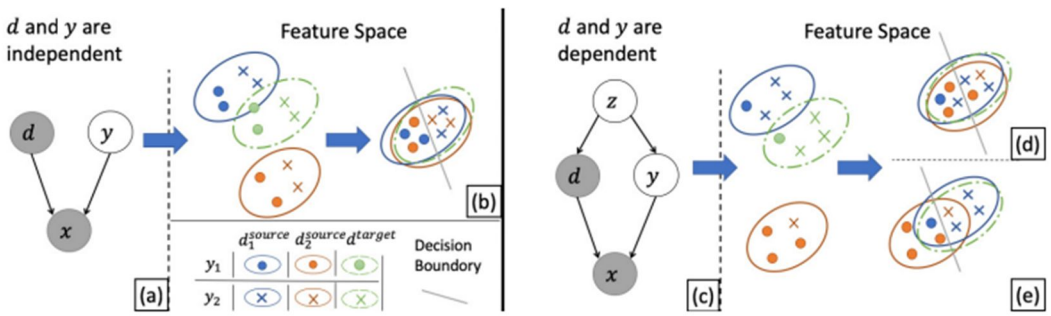

Figure 1 十分不変性基準の概要.

■　新基準２：教師なしデータによる代理不変性基準

（２）既存の不変表現学習は「どの情報を表現から消すか」を明示する必要があり、消したい情報についての教師データが必要である。本研究では、ユーザが消したい情報についての詳細を与えることなく、データからそのような情報を削除するグラフィカルモデルに基づく枠組みとその実現する手法を提案した（下図）。本成果は ECML2021 に採択された。

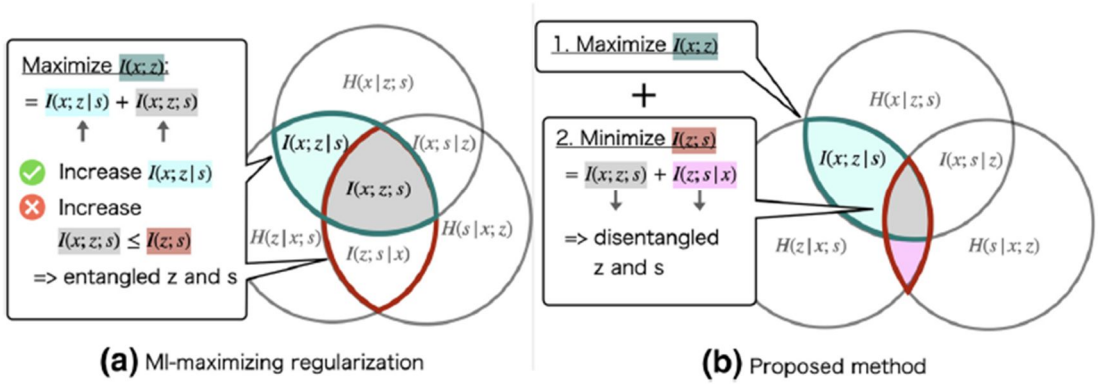

Figure 2 情報理論に基づいた教師なしでの表現の分離方法の概要.

■　安定な学習手法の提案

（３）既存手法である Adversarial Feature Learning (AFL)は有望なアプローチであるものの、実際的な挙動は不安定であり（下図）、利用者の細かいチューニングなしには表現の制御は困難である。AFL の不安定性について解析を行い、解決する方法を提案した。本成果は IJCAI2020 に採択された。

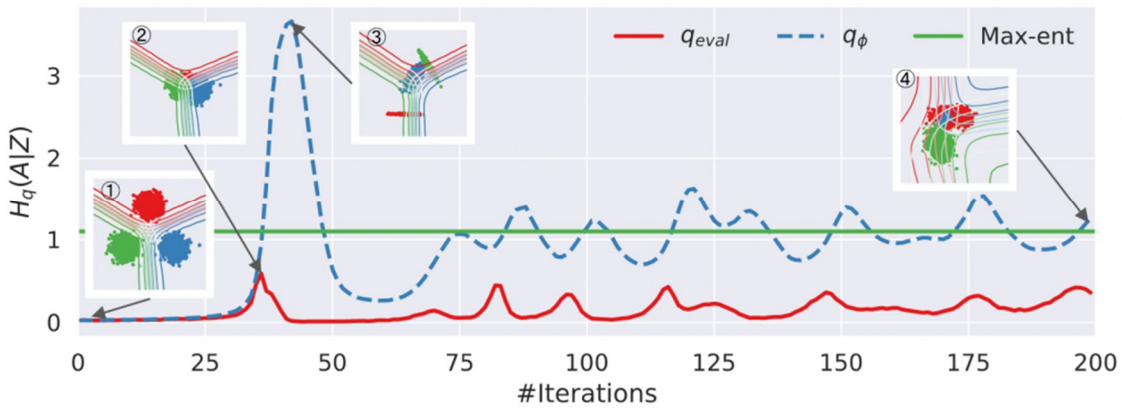

Figure 3　AFL の不安定性の可視化.

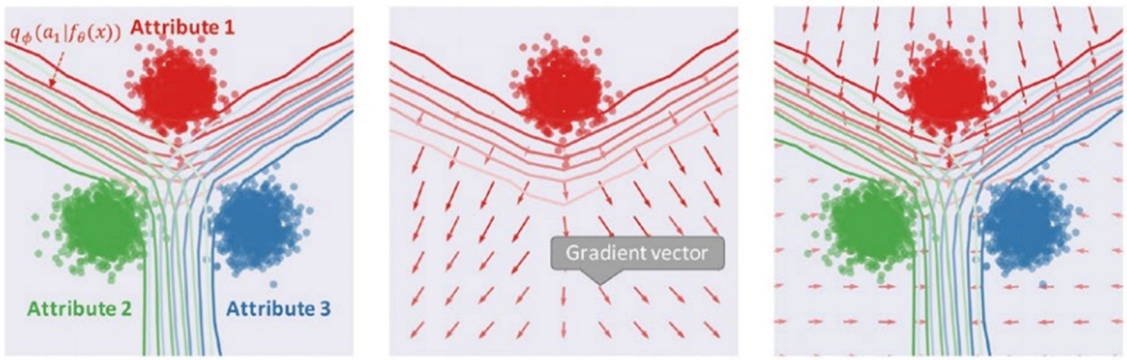

Figure 4 AFL の不安定性の原因とその解決方法.

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Akuzawa Kei、Iwasawa Yusuke、Matsuo Yutaka	４．巻 110
２．論文標題 Information-theoretic regularization for learning global features by sequential VAE	５．発行年 2021年
３．雑誌名 Machine Learning (Springer US)	６．最初と最後の頁 2239～2266
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s10994-021-06032-4	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計12件（うち招待講演　1件／うち国際学会　6件）

１．発表者名 Yusuke Iwasawa, Yutaka Matsuo
２．発表標題 Test-time classifier adjustment module for model-agnostic domain generalization
３．学会等名 Advances in Neural Information Processing Systems（国際学会）
４．発表年 2021年

１．発表者名 Yusuke Iwasawa, Kei Akuzawa, Yutaka Matsuo
２．発表標題 Stabilizing Adversarial Invariance Induction from Divergence Minimization Perspective
３．学会等名 IJCAI（国際学会）
４．発表年 2021年

１．発表者名 岩澤有祐
２．発表標題 制御可能な表現学習
３．学会等名 第5回 統計・機械学習若手シンポジウム（招待講演）
４．発表年 2020年

１．発表者名
阿久澤 圭，岩澤 有祐，松尾 豊

２．発表標題
分類性能による制約を考慮した敵対的不変表現学習によるドメイン汎化

３．学会等名
人工知能学会全国大会

４．発表年
2019年

１．発表者名
岩澤 有祐，阿久澤 圭，松尾 豊

２．発表標題
ペアワイズニューラルネット距離による不変表現学習

３．学会等名
人工知能学会全国大会

４．発表年
2019年

１．発表者名
阿久澤圭，岩澤有祐，松尾豊，

２．発表標題
大域的な潜在変数を持つ系列変分自己符号化器による状態遷移モデルのメタ学習

３．学会等名
情報論的学習理論ワークショップ

４．発表年
2019年

１．発表者名
Yusuke Iwasawa, Kei Akuzawa, Yutaka Matsuo

２．発表標題
Stabilizing Adversarial Invariance Induction from Divergence Minimization Perspective

３．学会等名
International Joint Conference of Artificial Intelligence (IJCAI)（国際学会）

４．発表年
2020年

１．発表者名
Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo

２．発表標題
Adversarial Invariant Feature Learning with Accuracy Constraint for Domain Generalization

３．学会等名
ECMLPKDD（国際学会）

４．発表年
2019年

１．発表者名
岩澤 有祐, 阿久澤 圭, 松尾 豊

２．発表標題
ペアワイズニューラルネット距離による不変表現学習

３．学会等名
人工知能学会全国大会

４．発表年
2019年

１．発表者名
阿久澤 圭, 岩澤 有祐, 松尾 豊

２．発表標題
分類性能による制約を考慮した敵対的不変表現学習によるドメイン汎化

３．学会等名
人工知能学会全国大会

４．発表年
2019年

１．発表者名
Yusuke Iwasawa, Kei Akuzawa, Yutaka Matsuo

２．発表標題
Invariant Feature Learning by Attribute Perception Matching

３．学会等名
International Conference of Learning Representations (Workshop)（国際学会）

４．発表年
2019年

１．発表者名 Kei Akuzawa, Yusuke Iwasawa, Yutaka Matsuo
２．発表標題 Adversarial Feature Learning under Accuracy Constraint for Domain Generalization
３．学会等名 International Conference of Learning Representations (Workshop)（国際学会）
４．発表年 2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関