The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

類似度学習を用いた敵対的訓練による特徴表現の検閲

Similarity Learning Based Adversarial Training for Censoring Representatoins

岩澤 有祐 *1　　松尾 豊 *1
Yusuke Iwasawa　　Yutaka Matsuo

*1東京大学工学系研究科
The University of Tokyo

Deep neural networks (DNN) continuously demonstrates excellent performance in various application domains. However, how to control the representations is critical issues to use DNN in real-world scenarios. Notably, control the invariance of the representations is essential to incorporate social constraints, such as privacy-protection and fairness. This paper proposes a novel way to control the representations learned by DNN, called "similarity confusion training". Empirical validations on an task of learning anonymous representations from the data of wearables shows that the proposed method successfully remove unwanted information with less performance degradation compared to the existing methods.

1. はじめに

データから有用な表現を学習する深層ニューラルネットワーク（以降 DNN と表記する）は，画像認識や音声認識をはじめ多種多様なタスクで予測精度を向上させることが報告されており，実世界での利用に期待が集まっている．一方，実世界への適用を考えると，認識精度だけでなくさまざまな社会的な制約が問題になる場合がある．例えば社会システムの一部として機械学習を利用する場合は，人種や性別など，公平性に問題があるような情報を含む表現を判断に利用しないことが求められる [Louizos 16, Edwards 16]．あるいは，プライバシーの観点からはデータにあるユーザ個別の情報に過剰にアクセスしないことが望ましい場合もあるだろう [Edwards 16, Iwasawa 17]．しかし，深層学習は最適化の結果としてブラックボックスに表現を獲得するだけであり，獲得した表現がどのような情報を持つかを事前に予期することは容易ではない．このような背景から，DNN が学習した/する表現が特定の情報を利用していないかを検閲するような技術（英語では *censoring representations* と呼ばれる）に関する研究が進められている．

表現の検閲を実現する最も単純な方法の 1 つは，本来最適化したい目的関数（例えば，予測したい情報 y に関する交差エントロピー）に,「違反度合い」を定量化した指標を正則化項として陽に組み込む方法である．このクラスに属する手法の 1 つに，近年提案された敵対的特徴学習（*adversarial feature leaning; AFL*）がある（図 1-a）[Edwards 16, Iwasawa 17, Xie 17]．敵対的特徴学習では，現在の特徴表現 z が望ましくない因子 s についてどの程度の情報を持っているかを，識別器 D による z から s の分類しやすさとして定量化し，識別器 D が予想しにくくなる（敵対的な）方向に特徴表現 z を修正する（より厳密な定式化については．2.1 節で言及する）．敵対的特徴学習は，Maximum Mean Discrepancy といった統計的な尺度を利用した手法群 [Zemel 13, Louizos 16] と比較して高精度な検閲が報告されている．

本稿では，z と s の単一ペアに依存した従来の敵対的特徴学習の代わりに，2 つのサンプルに紐付いた特徴表現 z_1, z_2 が同じグループに属するかを判別する類似度識別器を用いた敵対的訓練の枠組み（類似度学習を用いた敵対的訓練）を提案する（図 1-b）．既存のクラス分類に基づく敵対的特徴学習（以降 AFL と表記する）の問題点として，s が複雑になると分類タスクが複雑になりかつ騙すのが簡単になるという問題がある（例えば，s が 100 クラスのユーザ ID であった場合，この識別器を騙すためには 99 クラスのどれかに予想させるようにすれば良い）一方で類似度学習を用いた敵対的訓練では，1 つのサンプル z から対応する s を判別する識別器を騙すように E を更新する代わりに，2 つのサンプル z_1 と z_2 が同じ s に属するかを判別する識別器 D_{sim} の情報を利用した学習をおこなう．類似度学習を用いた敵対的訓練では s がどのような形式であっても識別器のタスクは 2 クラス分類であり，結果として s が複雑な場合でも安定した学習が行えることが期待できる．

本稿の主要な貢献は次の 2 つである．(1) 類似度学習を敵対的特徴学習に組み合わせた新たな枠組みを提案した．本手法は，類似度学習と敵対的特徴学習を橋渡しする初めての試みであり，類似度学習の発展的な手法を利用した敵対的特徴学習の高精度化に繋がるものである．(2) 提案した枠組みに基づく具体的な方法として，Similarity Confusion Training（SCL）と Learned Similarity Maximization（LSM）の 2 つを提案した．また，SCL および LSM が既存手法と比較してよい性能を達成できることを，ウェアラブルセンサを用いた行動認識におけるユーザ匿名化タスクにおいて確認した．

2. 関連研究

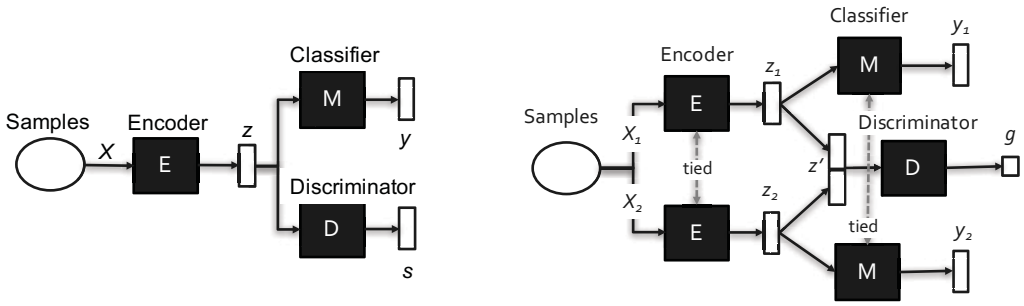
2.1 敵対的特徴学習

1 章で言及したように，敵対的特徴学習 *(adversarial feature leaning)* とは，DNN が学習する表現 z が本来予測したかった対象 y についての情報を含みながら別の因子 s に関する情報を持たないように明示的に制約を加える方法の 1 つである．以降，X を入力データ，E を X を z に変換する符号化器，M を z から y を予測する分類器とする．定式的には，AFL は次の最適化問題を解く．

$$\min_{E,M}\max_{D}\mathbb{E}[-\log q_M(y|z=E(X))+\lambda\log q_D(s|z=E(X))]. \tag{1}$$

ただし，q_M および q_D はそれぞれ M と D でパラメタ化された条件付き分布であり，λ は制約の強さを制御するハイパーパ

連絡先: 岩澤有祐，東京大学工学系研究科技術経営戦略学専攻，iwasawa@weblab.t.u-tokyo.ac.jp



(a) クラス識別器を用いた敵対的特徴学習.

(b) 類似度学習を用いた敵対的特徴学習.

図 1: クラス識別器を用いた敵対的特徴学習と（a，既存手法）と，類似度学習を用いた敵対的特徴学習（b，提案手法）の比較．(a) クラス識別器を用いた敵対的特徴学習では，不変性を確保したい因子 s を予測する識別器 D を騙すように符号化器 E を更新する．(b) 敵対的類似度学習では，2 つのサンプル X_1 と X_2 が同じ s に属するかに基づくグループラベル g を予測する類似度識別器 D を利用する．具体的な実装例については 3 章を参照されたい．

ラメタである．(1)E を固定した上での D の更新（式 1 の第 2 項の最大化），(2) D を固定した上で式 1 全体の最小化，の 2 つのステップを交互に繰り返す（あるいは，Gradient reversal layer[Ganin 16] を利用した場合は同時に学習する）ことにより AFL では z が持つラベル y に関する情報量の最大化と s に関する情報量の最小化を同時に行う．

2.2 類似度学習

類似度学習 (*Similarity learning*) とは，2 つのサンプルを入力にそれらのサンプルが同一のグループに属するかを判定する機械学習のタスクの 1 種である．類似度学習を定式化する方法には単純なクラス分類を用いるもの [Vo 16]，Saiamese 構造を用いるもの [Lin 15, Wang 15]，トリプレット損失を用いるもの [Wang 14, Parkhi 15]，より最近の方法では Angular 損失を利用したもの [Wang 17] などいくつかの種類がある．ここでは画像マッチングのタスクで良い精度が出ることが報告されており，本稿でも利用する Saiamese 構造とクラス分類を利用したハイブリットな方法の定式化 [Vo 16] について説明する．

以降，X_1 および X_2 を訓練集合 $\mathcal{X}$ から得られたサンプルであるとする．この時，クラス分類に基づく類似度学習では，X_1 と X_2 が同じグループに属する（あるいは，近くであると判定して欲しい）場合を正例，そうでない場合には 0 を出力するように訓練する．定式的には，g を X_1 と X_2 が同じグループに属する場合には 1，そうでない場合には 0 を取るようなグループラベルであるとすると，

$$-\mathbb{E}[g \log q_{D_{sim}}(g|z') + (1-g)\log(1-q_{D_{sim}}(g|z'))] \quad (2)$$

を最小化するように特徴抽出器 E および類似度識別器 D_{sim} を学習する．ただし，z' は 2 つの特徴量 z_1 と z_2 を特徴方向に結合したものとする（すなわち，z' の次元は z_1 および z_2 の 2 倍になる）．

類似度学習は主に顔画像の同一性判定や検索に利用されてきたが，敵対的訓練の文脈で類似度学習を利用することを提案する本稿と類似した試みもいくつか存在する．Li らは，2 サンプルが同じ分布から得られたかを判定する 2 サンプル識別器を敵対的生成ネットワーク（GAN）の学習やドメイン適応に利用することを提案している [Li 17]．同様の試みとして，Dou は，類似度学習と関連性の強いメトリック学習を GAN の学習に利用することを提案している [Dou 17]．Motiian らは，類似度学習と同様のペアリング・グループ化技術を利用することにより，教師ありドメイン適応がより効率的に（サンプルが少ない場合でも）行えることを報告している [Motiian 17]．これらの研究は，クラス分類に基づく従来の敵対的訓練とは異なり複数のサンプルのペアを利用するという点で提案法と類似しているものの，1) いずれの研究も s がそもそも 2 値の問題設定であり，応用先が異なる，2)Motiian らの研究はサンプル数が少ない問題を補うためにペアリングを利用することに主眼が当たっており学習された類似度自体に基づく敵対的訓練を提案しているわけではないという点で本論文とは本質的に異なる．

3. 類似度学習に基づく敵対的訓練

本稿の提案は，z と s の単一ペアに依存した従来の敵対的特徴学習の代わりに，2 つのサンプルに紐付いた特徴表現 z_1, z_2 が同じグループに属するかを判別する類似度識別器に基づいた E の更新を行うことである．図 1 に，クラス分類に基づく敵対的訓練および類似度学習に基づく敵対的訓練を対比した図を示す．本章では提案した枠組みに基づいた 2 つの具体例として，Similarity Confusion Training と Learned Similarity Maximization を説明する．

3.1 Similarity Confusion Training

類似度学習を敵対的訓練に利用するもっとも単純な方法は，通常の敵対的特徴学習が識別器 D が判別不能な方向に E を更新するのと同様に，類似度識別器 D_{sim} が正しく類似度を判別不能な（すなわち，同じ s であるかどうかの 2 クラス分類が正しく行えない）方向に E を更新する方法である．具体的には，SCT では次の 2 つのステップを交互に行う．

ステップ 1: 類似度識別器の更新

類似度識別器 D_{sim} は通常の類似度学習と同様に，式 2 を最小化するように訓練する．ただし，通常の類似度学習と異なり符号化器 E が類似度をよく表すように更新したいわけではないので，E は更新したまま D_{sim} のみ更新する点に留意されたい．また，グループラベル g は，s が同じ場合は 1，異なる場合は 0 であるとする．

ステップ 2: 符号化器の更新

ステップ 1 で更新した類似度識別器を利用して，類似度識別器が正しく類似度を測定できない用になる方向に符号化器を更新する．定式的には，次の similarity confusion loss を最小

化するように E を更新する．

$$L_{sc} = -\mathbb{E}[(1-g)\log q_{D_{sim}}(g|z') + g\log(1-q_{D_{sim}}(g|z'))]. \tag{3}$$

従来の AFL と同様に，ラベル予測に関する損失の最小化を同時に行うとすると，全体の最適化問題は．

$$\min_{E,M} L_y + \lambda L_{sc}, \tag{4}$$

となる．ただし，L_y はラベル予測に関する交差エントロピーであり，形式的には

$$L_y = -\mathbb{E}[\log q_M(y_1|z_1) + \log q_M(y_2|z_2)],$$

となる．ただし，y_i はサンプル X_i に対応したクラスラベルである．

3.2 Learned Similarity Maximisation

類似度学習を敵対的特徴学習に利用するもうひとつの方法として，学習された類似度関数に基づき，異なる s に対応したサンプルのペアの類似度が高くなる方向に符号化器 E を更新する方法が考えられる．定式的には，式 3.1 の代わりに次の式を最小化するように E を更新する．

$$L_{sim} = -[\mathbb{E}_{\mathcal{G}_{np}}[\log q_{D_{sim}}(g|z')] + \mathbb{E}_{\mathcal{G}_{nn}}[\log(1-q_{D_{sim}}(g|z'))].] \tag{5}$$

ただし，$\mathcal{G}_{np}$ は s が異なるが y が同じサンプルのペアの集合，$\mathcal{G}_{nn}$ は s も y も異なるサンプルのペアの集合を表す．すなわち，第 1 項目は s が異なるが y が同じペアについては近づくようにすることを，第 2 項目は s も y も異なるサンプルのペアについては離れたままになるように符号化器を更新する．ここで，第 2 項目は y が異なるサンプル同士を近づけてしまうとラベル予測精度を毀損する可能性があるために加えている制約である．

SCT と LSM はどちらも類似度識別器を利用した学習方法であるものの，(1) SCT は $\mathcal{G}_{pp}$ のサンプルを異なる s のペアであると誤分類する方向に E を更新する，(2) LSM は $\mathcal{G}_{nn}$ の類似度が近くならないように E の更新に制約をつけるという点で異なっている．4 章では，実験的に SCT と LSM の性能を比較することで，これらの差分の影響について検討する．

4. 実験・考察

4.1 実験設定

検証タスク： [Iwasawa 17] で利用されていたウェアラブルセンサデータにおけるユーザ匿名化タスクのうち 1 つ（USC-HAD）を利用した．ユーザ匿名化タスクとは，複数のセンサデータ系列 X から，対応する時刻における行動 y を予測でき，かつユーザ ID を予測出来ないような特徴量 z を獲得することを目標とするタスクである．データセットとしては，USC-HAD データセットを利用した．USC-HAD には，14 名分の計測データ（3 軸加速度＋ジャイロセンサの合計 6 次元）に 12 クラスの行動ラベルが付与されている．分類問題を解くにあたり，一般に利用されるスライディングウィンドウ方式（スライド幅 30，重なり率 0.5）により学習サンプルを生成し，結果として 14 ユーザ合計で 172,169 のサンプルを得た．この内，ユーザ ID が 1 から 10 のユーザのうち 90%を訓練データとし，残りの 10%および残りのユーザのデータをテスト用のデータとして利用することで，既知ユーザに対する予測精度と未知ユーザに対する予測精度の双方を計測した．

ネットワーク構造： 符号化器 E の構造および分類器 M の構造としては，既存研究に基づき，E: 3 層の畳み込み層（畳み込み＋最大プーリング）+1 層の全結合層，M: ロジスティック回帰，とした．活性化関数は出力層にソフトマックス関数を利用したほかは ReLU を利用し，各畳み込み層および全結合層のあとにはドロップアウトを行った．また，敵対的ドメイン分類器 D_{sim} としては，800-200 の中間層を持つ全結合ニューラルネットを利用した．

最適化： ラベル予測精度を最大化するように E を事前学習した後に，敵対的訓練を利用して表現を検閲する後処理を行った．事前学習の段階では Adam（学習率 0.0001）を利用した．後処理の段階では，RMSprop（学習率 0.000025）を用いて最適化した．バッチサイズは 100 とし，200,000 サンプルの学習ごとにステップ 1 と 2 を交互に 30 回繰り返し学習した．

ベースライン：1) CNN：敵対的訓練を用いずに，単純にラベル予測精度を最大化するように訓練したもの．2) AFL-MLP：クラス分類を用いた敵対的特徴学習で，D を中間層が 800 の全結合層の多層パーセプトロンとしたもの．3) AFL-DNN：クラス分類を用いた敵対的特徴学習で，D を 800-200 の中間層を持つ全結合ニューラルネットとしたもの．4) SCT：式 4 に基づきに訓練したもの．

評価方法： 既存研究にならい，ラベル予測精度とユーザ予測精度のバランスを評価に利用した．ラベル予測精度としては，既知ユーザと未知ユーザそれぞれのデータに関する F 値のマクロ平均（それぞれ F_y^v，F_y^t と表記）を利用した．ユーザ予測精度としては，800-200 の中間層を持つ全結合ニューラルネットによる正解率（Acc_u）を利用した．

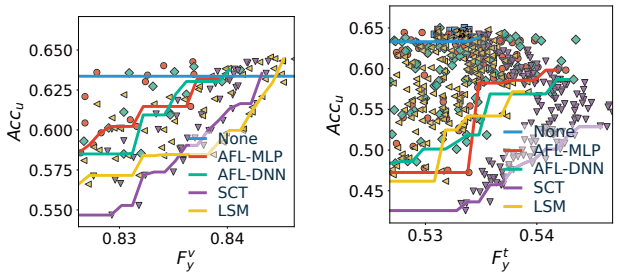
4.2 実験結果

図 2 に，各手法のラベル予測精度とユーザ予測精度のトレードオフに関する比較結果を示す．図中 X 軸はラベル予測精度（左側が F_y^v，右側が F_y^t）を，Y 軸がユーザ予測精度を表している．各点は λ=0.1, 0.2, 0.5, 1.0, 1.5, 2.0, 2.5 で訓練した場合の，各繰り返し終了時の性能を示しており，各線は，ラベル予測精度の棄損を一定以下に限定した場合に各手法が達成した最も低いユーザ予測精度を表しており，右下側を通るほどよい性能（ラベル予測精度を毀損することなく匿名性を高めることができている）ことを示す．それぞれの手法は，異なる色およびにマーカーで表されている．また，図 3 には提案手法とベースライン手法の学習曲線を比較した結果に示す．

これらの結果から，SCT および LSM が AFL と比較して同等以上の性能を達成していることがわかる．特に SCT は，ラベル予測に関するどちらの指標を利用した場合でもその他の手法と比較して良い性能を達成していた．また，F_y^t の精度については，SCT はほとんどその性能を棄損することなく，むしろ良い性能となる傾向が見られた．これは，SCT により得られた特徴が未知ユーザに対して良く汎化する方向に学習が進んでいたことを示している．LSM は F_y^t を指標として用いる場合には AFL と同程度の評価指標であったものの，F_y^v をラベル予測指標として用いる場合には AFL より優れた精度（SCT と同程度）を達成していた，

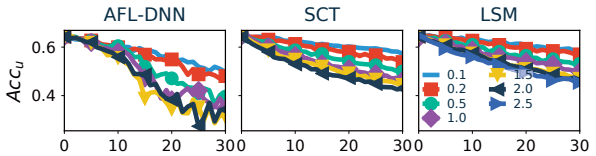
5. 考察・おわりに

本稿では，類似度学習を敵対的特徴学習に組み合わせた新たな枠組みを提案した．また，提案した枠組みに基づいた 2 つの具体例として Similarity Confusion Training（SCT）と Learned Similarity Maximization（LSM）を提案し，ウェアラブルセンサを用いた行動認識におけるユーザ匿名化タスクで

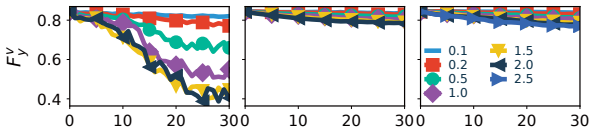


(a) F_y^v に関するトレードオフ　(b) F_y^t に関するトレードオフ

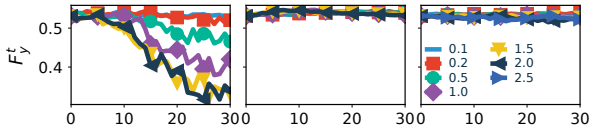
図 2: 提案手法とベースライン手法間でのラベル予測精度とユーザ予測精度のトレードオフに関する比較.



(a) F_y^v



(b) F_y^v



(c) F_y^t

図 3: 提案手法とベースライン手法の学習曲線に関する比較.

その有効性を確認した．特に，類似度識別器が正しく類似度を識別できなくなる方向に符号化器を更新する SCT は，匿名化と性能維持の良いトレードオフを実現するし，また未知のユーザについての汎化性能も改善することを確認した．今回の検証はユーザ匿名化にとどまるものの，提案した枠組み自体は敵対的訓練を利用する多くのアプリケーションに応用可能であり，多様なアプリケーションでの検証を進めていきたい．

参考文献

[Dou 17] Dou, Z.: Metric Learning-based Generative Adversarial Network, *CoRR*, Vol. abs/1711.02792, (2017)

[Edwards 16] Edwards, H. and Storkey, A.: Censoring Representations with an Adversary, in *Proceedings of the International Conference on Learning Representations (ICLR)* (2016)

[Ganin 16] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V.: Domain-adversarial training of neural networks, *The Journal of Machine Learning Research (JMLR)*, Vol. 17, No. 1, pp. 2096–2030 (2016)

[Iwasawa 17] Iwasawa, Y., Nakayama, K., Yairi, I. E., and Matsuo, Y.: Privacy Issues Regarding the Application of DNNs to Activity-Recognition using Wearables and Its Countermeasures by Use of Adversarial Training, in *Proceedings of the IJCAI*, pp. 1930–1936 (2017)

[Li 17] Li, C., Alvarez-Melis, D., Xu, K., Jegelka, S., and Sra, S.: Distributional Adversarial Networks, *CoRR*, Vol. abs/1706.09549, (2017)

[Lin 15] Lin, T.-Y., Cui, Y., Belongie, S., and Hays, J.: Learning deep representations for ground-to-aerial geolocalization, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5007–5015 (2015)

[Louizos 16] Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R.: The variational fair auto encoder, in *Proceedings of the International Conference on Learning Representations (ICLR)* (2016)

[Motiian 17] Motiian, S., Jones, Q., Iranmanesh, S., and Doretto, G.: Few-Shot Adversarial Domain Adaptation, in *Advances in Neural Information Processing Systems(NIPS)*, pp. 6673–6683, Curran Associates, Inc. (2017)

[Parkhi 15] Parkhi, O. M., Vedaldi, A., Zisserman, A., et al.: Deep Face Recognition., in *BMVC*, Vol. 1, p. 6 (2015)

[Vo 16] Vo, N. N. and Hays, J.: Localizing and orienting street views using overhead imagery, in *European Conference on Computer Vision (ECCV)*, pp. 494–509Springer (2016)

[Wang 14] Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., and Wu, Y.: Learning Fine-Grained Image Similarity with Deep Ranking, in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1386–1393, Washington, DC, USA (2014), IEEE Computer Society

[Wang 15] Wang, F., Kang, L., and Li, Y.: Sketch-based 3d shape retrieval using convolutional neural networks, in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pp. 1875–1883IEEE (2015)

[Wang 17] Wang, J., Zhou, F., Wen, S., Liu, X., and Lin, Y.: Deep Metric Learning with Angular Loss, in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017)

[Xie 17] Xie, Q., Dai, Z., Du, Y., Hovy, E., and Neubig, G.: Controllable Invariance through Adversarial Feature Learning, in *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)* (2017)

[Zemel 13] Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C.: Learning fair representations, in *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pp. 325–333 (2013)