情報処理学会論文誌 Vol.62 No.4 1142–1151 (Apr. 2021)

深層ニューラルネットワークの中間層出力を利用した半教師あり分布外検知

岡本 弘野[1,a)] 鈴木 雅大[1] 松尾 豊[1]

受付日 2020年7月28日, 採録日 2021年1月12日

概要：分布外検知はあるデータが入力されたときに，そのデータが特定の分布からのデータ（分布内データ）かそれ以外の分布からのデータ（分布外データ）かに分類するタスクである．分布外検知の問題設定は 2 種類あり，訓練データとして分布内データしか用いることができない教師なし分布外検知と，一部の分布外データを訓練データとして利用できる半教師あり分布外検知が存在する．近年提案された最も検知精度が高い半教師あり分布外検知の手法は，深層ニューラルネットワーク（DNN）を用いて分布内データのクラス分類を行い，分布外データを入力としたときには DNN の出力が一様分布になるように学習を行う．モデルの学習後，DNN の最終層の出力が一様分布に近いものを分布外データであるとして検出を行う．しかし，DNN の出力が一様分布に近いものになる分布内データが存在するため，この手法にはこのようなデータと分布外データの区別がつかなくなってしまう問題がある．筆者らはこの問題が DNN の最終層の出力だけを用いて分布外検知を行うことを困難にする点に着目する．この問題を解決するために，筆者らは DNN の複数の中間層の出力を特徴量として利用し，これらを同時に入力とするための DNN を新たに用意し，分布内外のデータを分類するように学習することを提案する．この提案は，分布外データは分布内データと異なりクラス分類のための特徴が抽出されないため，中間層での挙動が異なり分布内外のデータを分類するのに役立つという仮説に基づく．実験では，少量（16 枚）の訓練用分布外データの利用したとき，提案手法は先行研究と比較して AUROC で約 0.2 の改善がみられた．

キーワード：半教師あり分布外検知，深層ニューラルネットワーク

Semi-supervised Out-of-distribution Detection Using Output of Intermediate Layer in Deep Neural Networks

Hirono Okamoto[1,a)] Masahiro Suzuki[1] Yutaka Matsuo[1]

Received: July 28, 2020, Accepted: January 12, 2021

Abstract: In this paper, we study a method for semi-supervised out-of-distribution (OOD) detection. Recently, a semi-supervised OOD detection method with the highest detection accuracy has been proposed, which uses deep neural networks (DNNs) to classify the data, and then trains the DNNs so that the output of the DNNs is uniformly distributed when OOD data is input. After training the model, the output of the last layer of DNN is detected as OOD data if it is close to a uniform distribution. However, there are some in-distribution data that make the output of the DNN close to a uniform distribution, so this method has the problem of not being able to distinguish between such data and OOD data. We point out that it is not sufficient to use only the output of the last layer of DNN as a feature to perform OOD detection. To solve this problem, we propose to train a new DNN to classify in-distribution and OOD data by using the outputs of several intermediate layers of the DNN as the features. This proposal is based on the hypothesis that when OOD data is input, unlike in-distribution data, the features for classifying are not extracted by DNNs, and thus the behavior at the intermediate layer is different, which helps to classify in-distribution and OOD data. In experiment, the proposed method showed an improvement of about 0.2 in AUROC compared to previous studies when using a small amount of OOD (16) for training.

Keywords: semi-supervised out-of-distribution detection, deep neural network

© 2021 Information Processing Society of Japan

1. はじめに

機械学習モデルはテストデータの分布が訓練データの分布と異なるとき意図しない挙動をすることが知られている [1]．たとえば，英語（アルファベット）の手書き文字の認識を行うソフトは，入力としてある日本語（ひらがな）が与えられた場合，そのひらがなを無理やりいずれかのアルファベットとして分類してしまう．このように，機械学習の分類器の実応用を考えたとき，テスト時の入力は一般的にコントロールすることはできないため，訓練時に与えられたクラス以外のクラスのデータ（分布外データ）が入力となる場合がある．この問題に対する解決策として，分類する前に分布外データを検知する方法があり，これを分布外検知 [2], [3], [4], [5], [6] と呼ぶ．

分布外検知の問題設定には，訓練データとして分布内データしか用いることができない教師なし分布外検知 [2], [3], [4], [7] と，一部の分布外データを訓練データとして利用できる半教師あり分布外検知 [6], [8], [9], [10] の 2 種類が存在する．たとえば**図 1** のように，分布内データを猫の画像とすると，犬，馬，パンダなどの猫以外の画像はすべて分布外データということになる．機械学習モデルを学習するための訓練データは，教師なし分布外検知の場合猫の画像だけしか利用できないが，半教師あり分布外検知の場合，カラス，犬の画像などの分布外データも利用できる．ただし，テストデータの分布外データは訓練データで用いる分布外データと異なるクラスのものを用いることが多く，図 1 の例ではフクロウや馬の画像を用いる．実世界の問題を解くときに，訓練時に分布外データが前もって得られることは多いため，近年ではこれを利用する半教師あり分布外検知の手法が注目されている．筆者らは特に画像のような高次元データのための半教師あり分布外検知の研究を行う．

半教師あり分布外検知の単純なアプローチとして，分布内データと分布外データを異なるクラスと見なし，通常の識別モデルを利用した 2 クラス分類と同じように解くことが考えられる．しかしこのアプローチは，機械学習モデルの訓練時にある分布外データを用いたとしても，テスト時の分布外データがそれとは異なる種類の分布外データであった場合，これを検知できる保証はない．一方で，高次元データを入力とする場合は，訓練用分布外データを大量に用いるという条件では，このような手法でもある程度検知できると報告されている [10]．

近年では画像における教師なし分布外検知の最も成功したモデルとして，自己教師あり学習を使った深層学習の

[1] 東京大学大学院工学系研究科技術経営戦略学専攻
Department of Technology Management for Innovation, Graduate School of Engineering, The University of Tokyo, Bunkyo, Tokyo 113–8656, Japan
[a)] h-okamoto@weblab.t.u-tokyo.ac.jp

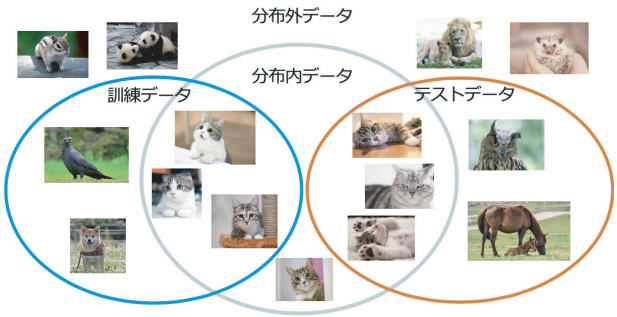

図 1 半教師あり分布外検知のときの各データの関係．上図では例として猫を分布内データとし，それ以外のデータを分布外データとしている．訓練データとして分布外データを利用することができる

Fig. 1 Relationship between each data distribution in semi-supervised out-of-distribution detection. In this figure, cats are defined as in-distribution data and others are defined as out-of-distribution data. The out-of-distribution data can also be used as training data.

モデルがある [11]．この手法は，分布外データは，分布内データと異なり，データの幾何変換を行ったときにそれがどのような変換か分からないというアイデアを用いている．具体的には，分布内データに回転のような変換を行い，この変換をあてるタスクを深層ニューラルネットワーク（DNN）を利用して解く．テスト時の DNN の出力値を指標とし，正しい変換をあてられなければ分布外データとして検知することができる．この手法を半教師あり分布外検知として拡張した GEOM+ [8] は，上記の幾何変換を利用した分布外検知において，訓練時に分布外データを入力としたときに，DNN の予測値が一様になるように学習する．モデルの学習後，DNN の出力が一様分布に近くなるような入力を分布外データと判定することで．GEOM の検知精度をさらに向上させた．しかしこの手法は，出力が一様分布に近いものになるような分布内データが存在するため，このようなデータと分布外データの区別がつかなくなってしまうという問題があり，文献 [5] でも同様のことが論じられている．

筆者らはこの問題が DNN の最終層の出力だけを用いて分布外検知を行うことを困難にする点に着目する．上記の幾何変換を利用した分布外検知モデル [11] の最終層の出力の情報は幾何変換されたか否かであり，分布外データを検知するための特徴量が必ずしも集約されているわけではない．一方，モデルの中間層では分布内データを分類するための特徴量を抽出していると考えられるが，分布外データを入力とした場合，見たことがないデータなので分布内データを分類するための特徴量を抽出することができず，分布内データとは異なる出力をすることが予測される．これが分布内データと分布外データの識別のための情報として役立つことが考えられる．

以上のことから，筆者らは DNN の複数の中間層の出力

© 2021 Information Processing Society of Japan

を分布外検知の特徴量として利用することを提案する．また筆者らはDNNのそれぞれの中間層の出力はベクトルの次元が異なるため，これらを同時に入力とするために，新たにDNNを用意する．このDNNは，分布内データの特徴量もしくは訓練用分布外データの特徴量を分類するように学習する．

実験では，分布外検知の手法の検証のためによく用いられるデータセットであるMNIST [12] とCIFAR10 [13] を利用する．それぞれのデータセットにおいて，ある特定のクラスを分布内データとしたときに，同データセットの残りのクラスを分布外データと定義して分布外検知を行う．また，訓練用分布外データセットとして，EMNIST [14]，CIFAR100 [13]，80 Million Tiny Images [15] を利用した．これらの数が数十枚程度であっても提案手法は，既存の半教師あり学習の手法が約千枚の分布外データを利用するときに匹敵する検知性能がでることを示す．また，同じ少量の枚数（16枚）を利用したとき，提案手法は先行研究と比較してAUROCで約0.2の改善を達成した．

本研究の主な貢献は以下のとおりである．

- DNNを用いた教師なし分布外検知モデルにおいて，最終層の出力だけを分布外検知の特徴量として利用することは不十分であると指摘し，複数の中間層の特徴量を入力とし，訓練用分布外データを用いて新たにDNNを学習させるという，半教師あり分布外検知の手法を提案した．
- 提案手法は16枚の分布外データを利用するだけで，先行研究の半教師あり分布外検知が約千枚の分布外データを利用するときに匹敵する検知性能を達成し，同量の訓練用分布外データを用いたときは先行研究に比べてAUROCで約0.2改善することを実験的に示した．

2. 関連研究

深層学習を利用した高次元データの分布外検知の手法は大きく，訓練用分布外データを用いない教師なし分布外検知の手法と，訓練用分布外データを用いる半教師あり分布外検知の手法の2種類に分類される．

2.1 教師なし分布外検知

まず最初に訓練用分布外データを用いない教師なし分布外検知の手法が存在する．その中でも，識別モデルベースの手法と生成モデルベースの手法が存在する．

2.1.1 識別モデルベースの手法

Baseline [2] は，DNNを用いて分布内データのクラス分類を行い，最終層の出力値であるソフトマックス出力の最大値を異常スコアとしている．これは，入力が分布外データの場合，その最大値は小さくなるという性質を利用して分布外検知を行っている．文献 [3] は上記手法に対し，ソフトマックスの出力値のキャリブレーションと入力に対する摂動を利用することで，分布外検知の性能を高めることに成功した．同様に，文献 [4], [5], [16] も上記手法の拡張といえる．これらの論文の手法はすべて，分布内データにおいて複数のクラスがあることを前提にしており，同時にクラスラベルを必要とする．

一方で，データセットのクラスラベルを必要としない分類器を利用した手法も存在する．たとえば，SVDD [17] は訓練データを，カーネル関数を利用して特徴空間の超球内に押し込み，テストデータがその超球内に写像されなければ分布外データであると判断する．しかし，SVDDは高次元データには対応できないため，Deep SVDD [7] はSVDDにおいて，カーネル関数ではなく深層モデルを利用することでこれを解決している．

また，補助タスクの分類を利用し，分布外検知を行う手法も提案されており，近年の教師なし分布外検知では最も有望な手法である．GEOM [11] では訓練データに対して回転，反転などの幾何変換を行い，その変換をあてるような補助タスクを解くようにモデルを学習させる．テスト時には分布外データはそのような幾何変換をあてることができないという性質を利用して分布外検知を行う．GOAD [18] は上記手法の学習方法は安定しないと指摘し，距離学習を使うことでこれを解決し，さらに検知精度を向上させた．

2.1.2 生成モデルベースの手法

次に，生成モデルを利用した分布外検知の手法がある．たとえば，autoencoder（AE）の利用により，再構成誤差を異常スコアとして利用する方法がある [19], [20]．これは，分布外データは訓練データに存在しないため，元のデータを再構成するのが難しいという考えに基づいたスコアであり，分布外データを入力とすると異常スコアは大きくなる．発展型として，DAGMM [21] は再構成誤差と潜在空間での負の対数尤度の両方合わせたものを異常スコアとしている．再構成ベースの手法は分布外検知だけなく，異常箇所を特定する異常部位検知にも有効である [22]．

一方，generative adversarial networks（GAN）[23] ベースの手法も生成モデルベースの手法の1つであり，文献 [24], [25] は分布外検知および異常部位検知が可能である．しかし，推論時間の長さや訓練の不安定さというGANならではの欠点をかかえており，近年ではあまり発展していない．

また，深層学習を用いた生成モデルによって直接推定された対数尤度（確率密度）を異常スコアとして分布外検知を行う方法がある．たとえば，pixelCNN [26]，glow [27] を使ったモデルは訓練データの対数尤度をモデル化し，対数尤度を最大化することによって学習を行っている．また，VAEは対数尤度を直接求めることはできないものの，対数尤度の下限を得ることができる．しかし，これらの深層生成モデルを利用した方法は，分布内データと大きく異なるような分布外データ（たとえばデータセットが異なる場合

© 2021 Information Processing Society of Japan

など）に対しては正しく密度推定できないことも報告されており，分布外検知に利用することは難しい [28].

2.2 半教師あり分布外検知

次に，訓練用分布外データを用いた半教師あり分布外検知の手法について述べる．半教師あり分布外検知は教師なし分布外検知の手法をベースに発展させたものが多い．たとえば，Deep SAD [9] は Deep SVDD [7] を半教師あり分布外検知に拡張している．この手法は見たことのない分布外データを検知するために，訓練用分布外データが入力のときにエントロピーが増大するような正則化を加えて学習する．Baseline+ [6] は Baseline [2] の拡張であり，分布外データを入力としたときにモデルの出力が一様になるように学習を行う．テスト時には Baseline [2] と同様に，最終層の出力値であるソフトマックス出力の最大値を異常スコアとしている．また，GEOM+ [8] は GEOM [11] において，Baseline+ [6] と同様に，分布外データを入力としたときに，モデルの出力が一様になるように学習する．提案手法は GEOM+と異なり，もとの GEOM のネットワークに関して学習を行わず，最終層の出力のみを分布外検知の特徴量としない．代わりに，GEOM の複数の中間層の出力を分布外検知の特徴量とし，これを新たな DNN の入力とする．この DNN は訓練用分布外データと分布内データを用いて 2 クラス分類を行うことで学習する．その他 BCE [10] は，教師なし分布外検知の手法を発展させたものではなく，単純に訓練用分布外データと分布内データを 2 クラス分類を行い，テスト時に分布外データを検知するという手法であるが，最も精度が高い手法の 1 つであり比較手法として用いる．

3. 既存手法と提案手法

この章ではまず，幾何変換を利用した分布外検知の手法である GEOM [11] を説明したあと，この手法を半教師あり分布外検知の手法として拡張した提案手法について説明する．

3.1 GEOM

$x \in \mathcal{X}$ を入力，$\mathcal{T} = \{T_1, \cdots, T_K\}$ を幾何変換の集合とする．ここで，$1 \leq y \leq K, T_y : \mathcal{X} \to \mathcal{X}$ である．$y \in \mathcal{Y} = \{1, \cdots, K\}$ はどの幾何変換を利用するかを示すラベルである．分類器は深層ニューラルネットワーク（DNN）である f_ϕ を用意する．f_ϕ は x を入力とし，K 次元のベクトルを出力する．K 次元のベクトルはロジットと呼ぶことにし，これにソフトマックス関数を通すことで確率 $\hat{y}$ を出力する．GEOM のロス関数は式 (1) のようになり，このロス関数を最小化するように分類器 f_ϕ の学習を行う．

$$\mathbb{E}_{p(x,y)}[\mathcal{L}_{CE}(\mathcal{S}(f_\phi(T_y(x))), y)]. \tag{1}$$

ここで，$\mathcal{L}_{CE}$ はクロスエントロピーロス関数であり，$\mathcal{S}$ はソフトマックス関数とする．また，$p(x, y) = p(x)p(y)$ であり，$p(x)$ はデータの経験分布，$p(y)$ は一様なカテゴリカル分布とする．テスト時は正常度を測るスコアとして，$\sum_{j=1}^{K}[\mathcal{S}(f_\phi(T_j(x)))]_j$ を用いる．このスコアは，幾何変換をどれだけあてられたかを意味している．

3.2 提案手法

GEOM は分布外データを入力としたときに幾何変換をあてることができないというアイデアをもとにした手法である．しかし，実際には幾何変換に対して不変な画像はたとえ分布内データだとしてもあてることができないため，このような分布内データは，分布外データと区別がつかない問題がある．筆者らはこの問題が DNN の最終層の出力だけを用いて分布外検知を行うことを困難にする点に着目する．この問題を解決するために，筆者らは DNN の複数の中間層の出力を特徴量として利用し，これらをまとめて入力とするために，**図 2** のように，新たな DNN を用いることを提案する．この提案は，分布外データは分布内データと異なりクラス分類のための特徴抽出が行われないため，浅い層での挙動が異なり，これらの挙動の違いが分布内外の識別に役立つという仮説に基づく．

まず 3.1 節で説明した方法で，あらかじめ分類器 f_ϕ を学習しておく．ここで，$l \in \{1, \cdots, L\}$ を DNN の層のブロックのインデックスとする．層のブロックとは複数の DNN の層を意味している．図 2 の Conv Block は畳み込み層が複数個，FC Block は全結合層が複数個，Conv Layer は畳み込み層が 1 つ含まれていることを意味する．このとき，$f_\phi = f_{\phi_L} \circ f_{\phi_{L-1}} \circ \cdots \circ f_{\phi_1}$ のように表記できる．また，中間層の出力は $z_l = f_{\phi_l}(z_{l-1})$ とかくことができる．ここで，$z_0 = x$ とし，z_L はソフトマックス関数を通す前のロジットとする．また，幾何変換を行った画像の潜在表現を z_l^y と表記することにする．y は 3.1 節と同様にどの幾何変換を利用するかを示すラベルであり，$z_0^y = T_y(x)$ である．これらの中間層の出力を結合させるために，新たに DNN である g_θ を用意する．ここで提案した g_θ のネットワーク構造は任意であるが，抽出した形の異なる中間層の出力を結合させるため，いくつかの層が必要となる．g_θ は複数の f_ϕ の中間層の出力である z_l（$l = 1, \cdots, L$）を入力とする．z_l（$l = 1, \cdots, L$）はそれぞれ各次元の数が異なるため，図 2 のように，g_θ の各層に入れる．g_θ の出力は 0 から 1 の間のスカラー値 $\hat{d}_y$ である．また，分布内データであるとき 0，分布外データであるとき 1 を返すようなラベルを d とする．このとき，提案手法のロス関数は式 (2) のようになり，このロス関数を最小化するように g_θ を学習する．g_θ の学習時には f_ϕ のパラメタは固定し，学習を行わない．

© 2021 Information Processing Society of Japan

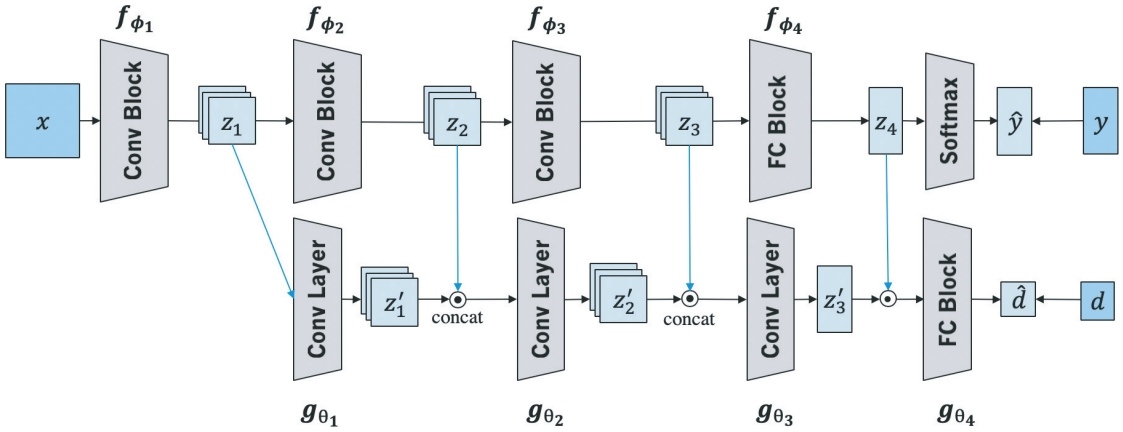

図 2 教師なし分布外検知モデル f の中間層を利用する分布外データ分類器 g の概念図．x は画像，y は x のクラスラベル，d は分布内データか分布外データかを示すラベルを表す．ϕ と θ はそれぞれ DNNf，g のパラメタである

Fig. 2 Conceptual diagram of the out-of-distribution data classifier g that uses the intermediate layers of the unsupervised out-of-distribution detection model f. x is the image, y is the class label of x, and d is the label indicating whether the data is in-distribution or not. ϕ and θ are parameters of f and g, respectively.

Algorithm 1 algorithm for proposed method

Require: base classifier f_ϕ, out-of-distribution classifier g_θ, training set, test set.
Initialize parameters of the classifiers: ϕ, θ
▷ training base classifier f_ϕ like GEOM
repeat
  $x \in$ training set
  $y \sim Uni(1, K)$
  $\phi \leftarrow$ Update $(\mathcal{L}_{CE}(\mathcal{S}(f_\phi(T_y(x))), y))$
until convergence of parameters ϕ
▷ training out-of-distribution classifier g_θ
fix parameter ϕ
repeat
  $x \in$ training set
  $y \sim Uni(1, K)$
  $z_0^y = T_y(x)$
  $z_l^y = f_{\phi_l}(z_{l-1}^y) \quad l = 1, \cdots, L$
  $\theta \leftarrow$ Update $(L_{BCE}(g_\theta(z_1^y, \cdots, z_L^y), d))$
until convergence of parameters θ
▷ estimating test score
for $x \in$ test set **do**
  for $y \in \{1, \cdots, K\}$ **do**
    $z_0^y = T_y(x)$
    $z_l^y = f_{\phi_l}(z_{l-1}^y) \quad l = 1, \cdots, L$
    $\hat{d}_y = g_\theta(z_1^y, \cdots, z_L^y)$
  end for
  estimate score with $\sum_{y=1}^{K} \hat{d}_y$
end for

$$
\begin{aligned}
&z_0^y = T_y(x), \\
&z_l^y = f_{\phi_l}(z_{l-1}^y) \quad l = 1, \cdots, L, \\
&\mathbb{E}_{p(x,y)}[\mathcal{L}_{BCE}(g_\theta(z_1^y, \cdots, z_L^y), d)]. \qquad (2)
\end{aligned}
$$

ここで，$\mathcal{L}_{BCE}$ はバイナリークロスエントロピーである．テスト時に正常度を測るスコアとして，g_θ の出力の y における総和 $\sum_{y=1}^{K} \hat{d}_y$ を利用する．最終的に提案手法のアルゴリズムは Algorithm 1 のようになる．

4. 実験

この章では分布外データを検出するための特徴量として，教師なし分布外検知の手法の DNN の最終層の出力だけを用いることは不十分であり，DNN の複数の中間層の出力を用いることで精度が大きく上昇することを示す．また，訓練用分布外データの数や種類を変化させ，既存の半教師あり分布外検知の手法との精度比較を行い，提案手法の訓練用分布外データに対する頑健性を確かめる．

4.1 実験設定

データセットとして，MNIST [12] と CIFAR10 [13] を利用した．MNIST は 10 クラスの手書きの数字画像データ，CIFAR10 は 10 クラスの乗り物や動物の画像データである．各データセットは訓練画像が 50,000 枚，テスト用画像が 10,000 枚存在する．データセットのある特定のクラスのデータを分布内データとし，それ以外のクラスのデータを分布外データと定義した．訓練時には CIFAR10 を利用するときのみデータオーグメンテーションを行った．具体的には，画像をクロップしもとの大きさに戻す操作，色相・彩度・明度を動かす操作，グレースケールに変える操作，ガウスぼかしを行う操作をそれぞれ確率的に行った．さらに，データセットの平均が 0，分散が 1 になるように正規化を行った．訓練用分布外データセットとして，MNIST データセットを用いるときは EMNIST-Letters [14]，CIFAR10 データセットを用いるときは 80 Million Tiny Images（80MTI）[15] を用いた．ただし，4.5 節のみ，CIFAR10 の訓練用分布外

© 2021 Information Processing Society of Japan

表 1 分布外検知分類器 g の構造．Conv2d はコンボリューション，BN はバッチノーマライゼーション，Linear は線形層を意味する．z_1 を入力とし，中間層で z_2，z_3，z_4 を結合する

Table 1 Structure of the out-of-distribution detection classifier g, where Conv2d means convolution, BN means batch normalization, and Linear means linear layer. z_1 is used as input, and z_2, z_3, z_4 are combined in the intermediate layer.

Layers	In	Out	Ksize	Stride	Pad
Conv2d	64	64	4	1	2
BN	-	-	-	-	-
ReLU	-	-	-	-	-
MaxPool	-	-	-	2	-
concat(z_2)	64	64+128	-	-	-
Conv2d	192	64	4	1	2
BN	-	-	-	-	-
ReLU	-	-	-	-	-
MaxPool	-	-	-	2	-
concat(z_3)	64	64+256	-	-	-
Conv2d	320	64	4	1	2
BN	-	-	-	-	-
ReLU	-	-	-	-	-
MaxPool	-	-	-	2	-
reshape	64	1,024	-	-	-
concat(z_4)	1,024	1,024+4	-	-	-
Linear	1,028	1,024	-	-	-
BN	-	-	-	-	-
ReLU	-	-	-	-	-
Linear	1,024	1	-	-	-
Sigmoid	-	-	-	-	-

データセットとして CIFAR100 を用いている．80 Million Tiny Images は約 8 千万の画像データであり，CIFAR10 と CIFAR100 を含むため，これら 2 つを除外したものを利用する．EMNIST-Letters は各クラス 4,800 枚，26 クラスのアルファベットの画像データである．

比較手法として，教師なし分布外検知の手法の中で最も検知性能が高い GEOM [11] を利用した．さらに，半教師あり分布外検知の手法としては，最も検知性能が高い BCE [10] と GEOM+ [8] の 2 つの手法と比較することにする．BCE は訓練用分布内データと分布外データを単純に 2 クラス分類することで，テスト用分布外データを検知する手法である．GEOM+は GEOM の学習の際に訓練用分布外データも用いて，ソフトマックスの出力が一様になるように学習する手法である．

GEOM および GEOM+のネットワーク構造はドロップ割合 0.3 の 16-4WideResNet [29] を利用した．提案手法である中間層の特徴量を出力するネットワーク f_ϕ も上記と同じものを利用した．そのうえで，上記の中間層の特徴量を結合し，訓練用分布外データを利用して学習するネットワーク（g_θ）として，**表 1** のような構造のネットワークを利用した．BCE においても提案手法と同じ構造のネットワークを利用した．提案手法において，中間層の特徴量は 16-4WideResNet [29] における 3 つの残差ブロックの出力と平均プーリング後の出力であり，それぞれを z_1，z_2，z_3，z_4 と表記する．表 1 のように 4 層のネットワーク構造にした理由は，これらの中間層の出力（z_1，z_2，z_3，z_4）を結合させるためである．どのモデルにおいても，最適化は Adam [30] を利用した．学習率の初期値は 0.001 とし，epoch 数が総 epoch 数の 80%を超えたときに 0.0001 に変更した．総 epoch 数は各実験において訓練誤差が十分に収束する数に設定した．

検証の際は，テスト用の分布内データと分布外データを利用し，それぞれの正常度スコアを求めた．各手法の正常スコアとして，GEOM と GEOM+は $\sum_{j=1}^{K}[\mathcal{S}(f_\phi(T_j(x)))]_j$，BCE は $\sum_{j=1}^{K} g_\theta(T_j(x))$，提案手法は $\sum_{j=1}^{K} g_\theta(z_1^j, z_2^j, z_3^j, z_4^j)$ を用いた．評価はしきい値に依存しない測り方であり，分布外検知に一般的に使われている Area Under the Receiver Operating Characteristics（AUROC）を利用した．

4.2 DNN の各層の出力を特徴量としたときの精度比較

まず提案手法において，最終層の出力ではなく，中間層の出力を利用することによって実際に検知精度あがるのかを確認する．MNIST の訓練用分布外データセットとして EMNIST，CIFAR10 の訓練用分布外データセットとして 80MTI を利用した．まず，3.1 節の GEOM のネットワーク f_ϕ の中間層を用いたのが提案手法であるため，ベースラインとして GEOM と比較する．このときのスコアは 3.1 節と同様に $\sum_{j=1}^{K}[\mathcal{S}(f_\phi(T_j(x)))]_j$ である．次に，最終層の出力結果である $\mathcal{S}(f_\phi(T_j(x)))$（$=\hat{y}_j$）を特徴量として入力とし，中間層の出力を利用せずに DNN(g_θ) を訓練用分布外データを使って学習する手法を OURS(softmax) とする．このときのスコアは $\sum_{j=1}^{K} g_\theta(\hat{y}_j)$ である．さらに，それぞれの中間層の出力を特徴量として分布外検知する手法を OURS(z_4)，OURS(z_3)，OURS(z_2)，OURS(z_1) とする．このときのスコアはそれぞれ，$\sum_{j=1}^{K} g_\theta(z_4^j)$，$\sum_{j=1}^{K} g_\theta(z_3^j)$，$\sum_{j=1}^{K} g_\theta(z_2^j)$，$\sum_{j=1}^{K} g_\theta(z_1^j)$ である．最後に提案手法である，$f_\phi(T_j(x))$ の各中間層の出力すべてを利用したものである OURS(z_1, z_2, z_3, z_4) との比較を行う．

結果は**表 2** のようになった．AUROC はデータセットのある 1 つのクラスのデータを分布内データとし，それ以外のクラスのデータを分布外データとしたときの分離度である．ここでは，データセットのすべてのクラス（10 クラス）においてそれぞれ AUROC を求めその平均を表示した．両データセットにおいて最終層の出力のみを特徴量として利用した OURS(softmax) はもとの教師なし分布外検知である GEOM の結果とほぼ変わらないことが分かる．一方で，それぞれの中間層の特徴量 z_4，z_3，z_2，z_1 を利用した OURS(z_4)，OURS(z_3)，OURS(z_2)，OURS(z_1) は，GEOM と同等かそれ以上の性能がでていることが分か

© 2021 Information Processing Society of Japan

表 2 特徴量を変えたときの提案手法の AUROC の比較
Table 2 Comparison of AUROC of the proposed method for different feature values.

Method	MNIST	CIFAR10	mean
GEOM [11]	92.6	89.0	90.8
OURS(softmax)	92.1	87.3	89.7
OURS(z_4)	94.5	88.7	91.6
OURS(z_3)	**98.5**	94.1	96.3
OURS(z_2)	97.8	95.6	96.7
OURS(z_1)	96.4	**95.7**	96.1
OURS(z_1, z_2, z_3, z_4)	98.0	95.6	**96.8**

る．よって，教師なし分布外検知モデルの分類器の最終層の出力のみを特徴量として利用することは不十分であり，中間層の出力を利用すべきであるという仮説が実験的に示された．

ここで，1 部の層だけを使ったほうが性能が良くなる場合がある．具体的には，MNIST においては OURS(z_3)，CIFAR10 においては OURS(z_1) を利用したほうが提案手法である OURS(z_1, z_2, z_3, z_4) よりも精度が高いことが分かる．しかし，それら中間層の出力をすべて使った提案手法 OURS(z_1, z_2, z_3, z_4) は両データセットの平均の AUROC において最も精度が高い．また，どの層を使えば精度が良くなるかはテストデータで実験してみないと分からない．そのため，筆者らはすべての中間層の出力を利用するものを提案手法としている．

4.3 先行研究との比較

ここでは，提案手法とこれまでに提案された分布外検知手法の中でも検知精度が最も高い手法と比較する．MNIST の訓練用分布外データセットとして EMNIST，CIFAR10 の訓練用分布外データセットとして 80MTI を利用した．先行研究の文献 [6], [10] を参考に，十分に精度がでる枚数として，80MTI のすべてのデータのうち 65,536 枚をランダムに抽出したものを利用した．

結果は**表 3** のようになった．GEOM は教師なし分布外検知の手法ではあるが，提案手法と GEOM+は GEOM をベースのモデルとして，半教師ありの手法として拡張しているため，参考のために結果を載せている．提案手法は他の手法と比べて，MNIST データセットにおいては最も検知精度が高く，CIFAR10 データセットにおいては BCE に匹敵する精度であった．

教師なし分布外検知のモデルである GEOM は幾何変換が当てられないものを分布外検知と見なすため，回転したときに似た画像となる MNIST の 0 と 8 のクラスのとき，他のクラスと比較して精度が低いことが分かる．たとえば GEOM は**図 3** のように，回転不変な画像は回転を予測することができず，分布外データとして見なしてしまう．

GEOM+は MNIST のクラスが 8 のデータのとき，GEOM と比較してむしろ精度を落としている．GEOM+は分布外データが入力のとき出力が一様分布に近づくように学習する手法であり，分布内データである 8 も上記で述べた性質上出力が一様分布に近くなってしまう．結果的に分布外データとの区別がつかなくなってしまい，検知精度を落としてしまったと推測される．

BCE に関しては，MNIST において他の手法に比べて最も精度が低い一方で，CIFAR10 において最も精度が高い結果となった．その理由として次のことが考えられる．BCE は比較のため提案手法と同じ構造の DNN(g_θ) を用いているが，提案手法と異なり特徴量として GEOM の中間層の出力ではなく，入力画像のみを用いている．そのため，BCE は教師なし分布外検知の手法の拡張ではなく，BCE の精度は訓練用分布外データに何を用いるかということに大きく依存する．訓練用分布外データである EMNIST の

表 3 分布外検知の手法の比較
Table 3 Comparison of methods for out-of-distribution detection.

Dataset	Indist	GEOM	GEOM+	BCE	ours
MNIST	0	81.5	81.8	98.7	99.2
	1	92.4	91.7	96.5	99.8
	2	90.3	95.7	81.3	94.1
	3	99.7	99.3	87.5	99.3
	4	98.0	98.8	79.9	99.3
	5	91.4	94.6	89.5	98.1
	6	99.8	99.4	92.3	99.2
	7	93.4	96.4	86.1	96.6
	8	82.0	79.6	89.2	97.6
	9	98.0	96.1	90.1	96.5
	mean	92.6	93.3	89.1	**98.0**
CIFAR10	plane	79.9	68.0	96.3	95.4
	car	96.5	96.6	99.0	98.9
	bird	86.1	87.9	91.9	93.2
	cat	79.3	81.3	89.2	87.6
	deer	89.3	89.5	96.0	96.4
	dog	89.5	88.1	93.8	94.2
	frog	87.7	88.8	97.5	97.6
	horse	96.0	96.6	97.1	97.3
	ship	93.9	94.4	98.3	97.9
	track	93.1	94.5	97.7	97.3
	mean	89.0	88.6	**95.7**	95.6

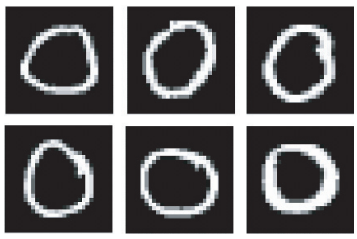

図 3 回転に不変なクラス 0 の画像
Fig. 3 Class 0 image which is invariant to rotation.

© 2021 Information Processing Society of Japan

種類は 26 クラスしか存在しないため，他の分布外データである MNIST に汎化しなかった一方で，80MTI は様々な種類のデータが含まれるため，他の分布外データである CIFAR10 にも汎化し，検知精度が高かったと考えられる．

提案手法は GEOM の中間層の出力といった，入力画像よりも抽象度が高く他の分布外データに汎化するような特徴量を利用したため，両方のデータセットにおいて，どのクラスが分布内データであっても高い検知精度を獲得できたと考えられる．

筆者らは GEOM と提案手法を用いて，回転したとしても同じような画像になるデータ（図 3）のスコア（s）を求めた．それぞれの手法において，分布外データだと判定する閾値（τ）として，テストデータの分布内データと分布外データが最も分離される値を用いた．ここで，$s(x) < \tau$ のとき，入力データ x を分布外データとして判定することにする．結果として，GEOM は図 3 のデータをすべて分布外データとして判定する一方で，提案手法はすべて正しく分布内データとして判定しており，提案手法の検知精度向上に寄与しているといえる．

4.4 少数の訓練分布外データを利用したときの検知精度比較

4.3 節では十分に訓練用分布外データが利用できる状況を想定している．しかし，実際には分布外データが少量しか使えない状況は十分にあり，大量の分布外データにアクセスできるとは限らない．そのため，利用できる分布外データが少量であっても提案手法が分布外データを検知できるかどうかを検証する．

この実験では，CIFAR10 の訓練用分布外データセットとして 80MTI を利用し，80MTI からランダムに少量のデータを抽出し，その数を 16，64，256，1,024 と変えたときの検知精度で評価した．モデルの精度は分布外データの選び方に依存するため，半教師あり分布外検知における各実験はそれぞれ 3 回行い，その平均と標準偏差を求めた．

実験結果は**図 4** のようになった．縦軸は AUROC であり，横軸は訓練用分布外データの数を表示している．提案手法は少量の訓練用分布外データを用いることでベースモデルである教師なしモデルの GEOM よりも高い検知精度を獲得していることが分かる．また，同じデータ数であればどの半教師あり分布外検知の手法に対してもより高い検知精度を獲得している．特に，提案手法は 16 枚の分布外データを使うだけで，1,024 枚の分布外データを利用した他の各手法の精度以上の検知精度を獲得していることが分かる．

GEOM+は提案手法と同じようにベースモデルとして教師なしモデルの GEOM を使っているが，検知精度がベースモデルよりも低くなってしまった．この原因として分布外データを入力としたときに出力を一様分布に近づけるという訓練を行うことで，分布内データの出力も一様分布に近づいてしまい，逆にこの 2 つを見分けられなくなってしまったからだと考えられる．BCE は教師なしモデルをベースにしておらず，BCE の精度は訓練用分布外データに大きく依存するため，訓練用分布外データの数が少ない場合はテスト時の他の分布外データに汎化しなかったと考えられる．提案手法は BCE で用いたような入力画像ではなく，GEOM の中間層の出力といった，より抽象度の高く他の分布外データに汎化するような特徴量を利用したため，少ない枚数でも高い検知精度を獲得できたと考えられる．

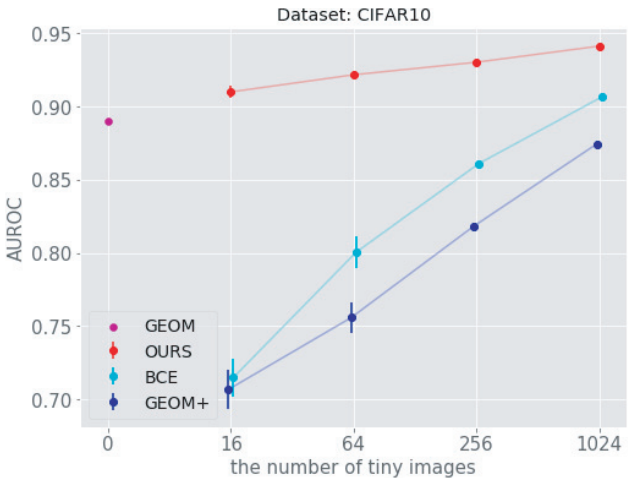

図 4 訓練用分布外データである 80MTI の数を変化させたときのそれぞれの手法における検知精度の比較

Fig. 4 Comparison of detection accuracy for each method when varing the number of out-of-distribution data for training, 80MTI.

4.5 訓練分布外データの種類における頑健性の検証

これまでの実験では CIFAR10 の訓練用分布外データとして 80MTI しか用いていない．80MTI は様々な種類の画像を含んでいるが，実際には分布外データとして十分な種類のデータが得られない場合がある．訓練用分布外データの種類が少ない場合，半教師あり分布外検知の精度は低下することが報告されている [6]．よって，この実験では提案手法において訓練用分布外データの種類を制限したときに分布外データを検知できるかを調べる．

この実験では CIFAR10 の訓練用分布外データセットとして CIFAR100 を利用した．訓練用分布外データの数をデータセットの 1 クラス分の数に固定して，クラスの数を変化させ検知精度を検証する．具体的には，CIFAR100 のデータ数を 500 に固定し，クラス数を 1，2，5，10，100 のように変えた．それぞれクラスとデータの選び方はランダムとし，実験は 3 回行い，検知精度の平均と標準偏差を求めた．

実験結果は**図 5** のようになった．縦軸は AUROC であり，横軸は CIFAR100 のクラスの数を示す．まず各手法

© 2021 Information Processing Society of Japan

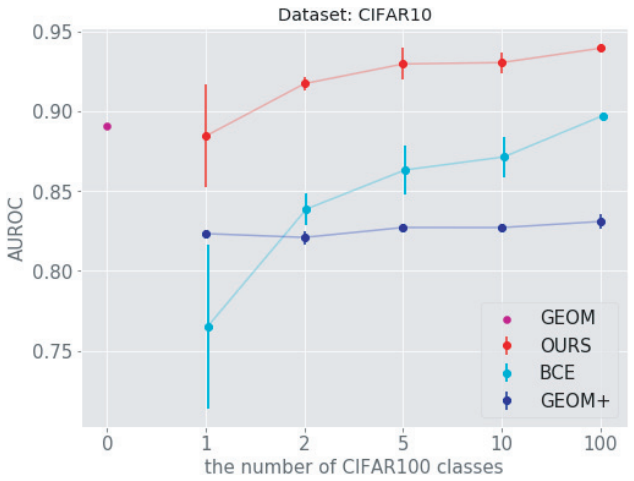

図 5 訓練用分布外データである CIFAR100 のクラス数を変化させたときの検知精度の比較

Fig. 5 Comparison of detection accuracy for each method when varying the number of classes of the out-of-distribution data for training, CIFAR100.

において，分布外データの種類が多いほど，検知精度は高く，分散は小さい傾向にあることが分かる．すなわち分布外データとして，その数よりもその種類数が重要であるということが示唆されており，文献 [6] の主張とも合致する．また，クラス数がどの場合でも提案手法は他の半教師あり分布外検知の手法と比べて，高い分布外検知精度を達成した．

5. 結論

本稿では既存の教師なし分布外検知を拡張し，検知精度向上のために訓練時に分布外データを利用することができる手法を提案した．提案手法は DNN の中間層の出力を分布外検知の特徴量として利用することで，他の半教師あり分布外検知の手法と比べてよりよい精度を達成することを示した．特に，訓練外分布外データの枚数が小さいときに他の分布外検知の手法と比べて大きく差をつけて精度が優れることが分かった．このことは，少量の分布外データだけが手に入るような実世界に即した問題設定において優れた検知精度が達成できることが期待できる．具体的な例として，アルファベットの手書き文字の認識を行うソフトにおいて，学習時に少量の分布外データとして数字を用いることで，テスト時に誤字やひらがなを分布外データとして検知することができると考えられる．今後は本手法を用いることで，GEOM 以外の教師なし分布外検知の手法に関しても検知精度が向上するのかを確かめる予定である．また，MNIST や CIFAR10 のベンチマークデータセットだけでなく，実世界のデータにおいても分布外データを検知できるのかを確かめる予定である．

参考文献

[1] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J. and Mané, D.: Concrete problems in AI safety, arXiv preprint arXiv:1606.06565 (2016).

[2] Hendrycks, D. and Gimpel, K.: A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks, *Proc. International Conference on Learning Representations* (2017).

[3] Liang, S., Li, Y. and Srikant, R.: Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks, *International Conference on Learning Representations* (2018).

[4] Lee, K., Lee, K., Lee, H. and Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks, *Advances in Neural Information Processing Systems*, pp.7167–7177 (2018).

[5] Malinin, A. and Gales, M.: Predictive uncertainty estimation via prior networks, *Advances in Neural Information Processing Systems*, pp.7047–7058 (2018).

[6] Hendrycks, D., Mazeika, M. and Dietterich, T.: Deep Anomaly Detection with Outlier Exposure, *Proc. International Conference on Learning Representations* (2019).

[7] Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S.A., Binder, A., Müller, E. and Kloft, M.: Deep one-class classification, *International Conference on Machine Learning*, pp.4393–4402 (2018).

[8] Hendrycks, D., Mazeika, M., Kadavath, S. and Song, D.: Using self-supervised learning can improve model robustness and uncertainty, *Advances in Neural Information Processing Systems*, pp.15663–15674 (2019).

[9] Ruff, L., Vandermeulen, R.A., Görnitz, N., Binder, A., Müller, E., Müller, K.-R. and Kloft, M.: Deep Semi-Supervised Anomaly Detection, *International Conference on Learning Representations* (2020).

[10] Ruff, L., Vandermeulen, R.A., Franks, B.J., Müller, K.-R. and Kloft, M.: Rethinking Assumptions in Deep Anomaly Detection, arXiv preprint arXiv:2006.00339 (2020).

[11] Golan, I. and El-Yaniv, R.: Deep anomaly detection using geometric transformations, *Advances in Neural Information Processing Systems*, pp.9758–9769 (2018).

[12] LeCun, Y.: The MNIST database of handwritten digits, available from ⟨http://yann.lecun.com/exdb/mnist/⟩.

[13] Krizhevsky, A. et al.: Learning multiple layers of features from tiny images, Technical report, Citeseer (2009).

[14] Cohen, G., Afshar, S., Tapson, J. and Van Schaik, A.: EMNIST: Extending MNIST to handwritten letters, *2017 International Joint Conference on Neural Networks* (*IJCNN*), pp.2921–2926, IEEE (2017).

[15] Torralba, A., Fergus, R. and Freeman, W.T.: 80 million tiny images: A large data set for nonparametric object and scene recognition, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.30, No.11, pp.1958–1970 (2008).

[16] DeVries, T. and Taylor, G.W.: Learning confidence for out-of-distribution detection in neural networks, arXiv preprint arXiv:1802.04865 (2018).

[17] Tax, D.M. and Duin, R.P.: Support vector data description, *Machine Learning*, Vol.54, No.1, pp.45–66 (2004).

[18] Bergman, L. and Hoshen, Y.: Classification-Based Anomaly Detection for General Data, *International Conference on Learning Representations* (2020).

[19] Sakurada, M. and Yairi, T.: Anomaly detection using autoencoders with nonlinear dimensionality reduction,

© 2021 Information Processing Society of Japan

Proc. MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis, pp.4–11 (2014).
[20] Zhou, C. and Paffenroth, R.C.: Anomaly detection with robust deep autoencoders, *Proc. 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.665–674 (2017).
[21] Zong, B., Song, Q., Min, M.R., Cheng, W., Lumezanu, C., Cho, D. and Chen, H.: Deep Autoencoding Gaussian Mixture Model for Unsupervised Anomaly Detection, *International Conference on Learning Representations* (2018).
[22] Dehaene, D., Frigo, O., Combrexelle, S. and Eline, P.: Iterative energy-based projection on a normal data manifold for anomaly localization, *International Conference on Learning Representations* (2020).
[23] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y.: Generative adversarial nets, *Advances in Neural Information Processing Systems*, pp.2672–2680 (2014).
[24] Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U. and Langs, G.: Unsupervised anomaly detection with generative adversarial networks to guide marker discovery, *International Conference on Information Processing in Medical Imaging*, pp.146–157, Springer (2017).
[25] Zenati, H., Foo, C.S., Lecouat, B., Manek, G. and Chandrasekhar, V.R.: Efficient gan-based anomaly detection, arXiv preprint arXiv:1802.06222 (2018).
[26] Van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., et al.: Conditional image generation with pixelcnn decoders, *Advances in Neural Information Processing Systems*, pp.4790–4798 (2016).
[27] Kingma, D.P. and Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolutions, *Advances in Neural Information Processing Systems*, pp.10215–10224 (2018).
[28] Nalisnick, E., Matsukawa, A., Teh, Y.W., Gorur, D. and Lakshminarayanan, B.: Do Deep Generative Models Know What They Don't Know?, *International Conference on Learning Representations* (2019).
[29] Zagoruyko, S. and Komodakis, N.: Wide residual networks, arXiv preprint arXiv:1605.07146 (2016).
[30] Kingma, D.P. and Ba, J.: Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

岡本 弘野 （学生会員）

2015 年東京大学理学部物理学科卒業．2018 年同大学工学系研究科修士課程修了．同年同大学工学系研究科博士課程入学．機械学習の研究に従事．

鈴木 雅大 （正会員）

2013 年北海道大学工学部卒業．2015 年同大学大学院修士課程修了．2018 年東京大学工学系研究科博士課程修了．博士（工学）．2018 年より東京大学大学院工学系研究科技術経営戦略学専攻特任研究員．人工知能，深層学習の研究に従事．

松尾 豊 （正会員）

1997 年東京大学工学部卒業．2002 年同大学院博士課程修了．博士（工学）．産業技術総合研究所，スタンフォード大学を経て，2007 年より，東京大学大学院工学系研究科技術経営戦略学専攻准教授．2019 年より同大学院人工物工学研究センター/技術経営戦略学専攻教授．2014 年より 2018 年まで人工知能学会倫理委員長．2017 年より日本ディープラーニング協会理事長．人工知能学会論文賞，情報処理学会長尾真記念特別賞，ドコモモバイルサイエンス賞等受賞．専門は，人工知能，深層学習，Web 工学．

© 2021 Information Processing Society of Japan