

深層ニューラルネットワークによるクラスと幾何変換の 同時分類確率を利用した分布外検知

岡本 弘野^{1,a)} 鈴木 雅大¹ 松尾 豊¹

受付日 2020年10月28日, 採録日 2021年3月2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. 近年の研究により, 深層ニューラルネットワークを使った分類器は, 分布外データを入力としたとき, 分布内データを入力とする場合と比較して, (1) クラス分類の出力が一様分布に近づき, (2) 幾何変換を予測することが難しいという2つの仮説のもとで, 分布外検知が行えることが明らかになっている. しかし, (1) において, クラス分類するのが難しいデータセットを用いる場合, 分布内データにも分類器の出力が一様分布になるようなものが存在する. また, (2) において, near-distribution outliers と呼ばれる分布外データを使った場合, 幾何変換が予測できてしまう場合がある. このとき, (1) または (2) の仮説をもとにした先行研究は, 分布外データと分布内データの見分けがつかなくなってしまう問題が発生する. 筆者らは, それぞれの先行研究において分布外データの検知精度が低くなる条件が異なるので, 両者の手法を組み合わせることで, 両者の欠点を補えると仮定した. この仮定に基づき, 筆者らは (1) と (2) の仮説を同時に活用する指標を用いることを提案する. 具体的には, 筆者らは入力を与えられたときのクラスと幾何変換の同時確率を求め, これをもとにした新しいスコアを提案する. 実験では, 様々なネットワーク構造とデータセットを使って実験を行い, (1) と (2) のそれぞれの仮説に基づく先行研究よりも, 両方の仮説に基づく本手法のほうが, AUROC を指標として平均 6.7%, 分布外データの検知精度が高くなることを示す.

キーワード: 分布外検知, 深層ニューラルネットワーク, 自己教師あり学習, 多クラス分類

Out-of-distribution Detection Using Joint Probability between Class and Geometric Transformation

HIRONO OKAMOTO^{1,a)} MASAHIRO SUZUKI¹ YUTAKA MATSUO¹

Received: October 28, 2020, Accepted: March 2, 2021

Abstract: Out-of-distribution detection is a task to categorize the input data into data from a specific distribution (in-distribution data) or data from another distribution (out-of-distribution data). Recent studies have shown that a classifier using deep neural networks can detect out-of-distribution data under the following two hypotheses: (1) the output of the classifier approaches a uniform distribution, and (2) it is difficult to predict geometric transformations, when the inputs are out-of-distribution data. However, in (1), when using datasets that are difficult to classify, there are some in-distribution data that result in a uniform distribution of the classifier's output. In (2), the use of near-distribution outliers, which are out-of-distribution data, can lead to the correct prediction of geometric transformations. In this case, previous studies based on hypotheses (1) or (2) cause problems in distinguishing between out-of-distribution and in-distribution data. We hypothesized that the combination of the two methods could compensate for the low detection accuracy of out-of-distribution data in each previous study. Therefore, we propose to use a metric that utilizes both hypotheses (1) and (2) simultaneously. Specifically, we calculate the joint probability of class and geometric transformations given the input and propose a new anomaly score based on it. We conduct experiments using various network structures and datasets and show that the method based on both hypotheses is more accurate in detecting out-of-distribution data than the method based on either (1) or (2) by 6.7% on average.

Keywords: out-of-distribution detection, deep neural network, self-supervised learning, multi-class classification

1. はじめに

近年、人工知能を支える技術である深層学習は様々な分野において急速に発展している。特に画像認識分野では、深層学習のモデルの1つである ResNet [1] は、2015 年に ImageNet [2] という大規模データセットの 1,000 クラス分類を行うコンペティションである ILSVRC [3] において、人間を超える分類精度を達成した。ここで、上記のような深層学習のモデルに、テスト時に訓練データとは異なる種類のデータを与えた場合どうなるだろうか。人間の場合、見たことのない種類のデータに対して「知らない」と答えることができる。一方で、学習済みの画像分類モデルは、訓練データに存在しなかったクラスのデータを入力として与えると、「知らない」と答えることはできず、高い確信度で訓練データに存在するクラスに分類することが知られている [4]。この問題は実用上で様々な不都合をもたらす。たとえば自動運転において、道路標識の STOP サインが何らかの原因で読みづらい状態になっていたとする。このとき、自動運転車が「知らない」標識であると判断し、運転者に判断を任せるとはせず、勝手に進むという判断を行ってしまうと衝突事故を起こしてしまうだろう。この問題を解決する方法として、間違った分類をする前にあらかじめ訓練に用いられていないクラスのデータ、すなわち分布外データ (Out-of-distribution, OOD) を検知する方法があり、これを分布外検知 [5], [6], [7], [8], [9] 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 [5], [6], [7], [10] 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 [9], [11], [12], [13] の 2 種類が存在する。教師なし分布外検知の中には、テストデータの一部の分布外データの情報にアクセスできる問題設定のものがある。筆者らは、より現実的な問題設定である、訓練データおよびテストデータの分布外データにアクセスしない教師なし分布外検知の手法に焦点を当てる。

訓練データに複数クラスが存在する場合を考える。このときの分布外検知において、深層ニューラルネットワーク (DNN) を用いた多クラス分類器を利用し、その出力された表現をもとにしたスコアを用いる手法は有望であることが知られている [5], [6], [7]。特に、単純な手法として、Hendrycks ら (2017) の手法 (以降 Hendrycks らの手法と呼ぶ) [5] がある。この手法は 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分布内データであることを示すスコア

(正常スコア^{*1}) としている。このスコアは、入力が分布外データの場合、その出力が一様分布に近づくという仮説に基づいている。しかし、多クラス分類器のクラス分類精度が落ちるデータセットを用いた場合、明確に特定のクラスに属さないような、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。そのため、多クラス分類器ベース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。本研究でこの問題を実験的に明らかにする。

もう 1 つの DNN の分類器を用いた分布外検知のための有望な手法として、自己教師あり学習を利用したモデル [11], [14], [15] が提案されている。自己教師あり学習とは、ラベルを利用せずに補助のタスクを教師あり学習のやり方で解くことで、下流タスクが解けるような良い表現を得るための学習方法である。これらの手法は、分布内データの幾何変換された画像を分類することで、分類器が幾何的な特徴を学習し、その特徴が学習したクラスに特有であるため、DNN が分布外データの幾何変換を行ったときにそれがどの幾何変換か分からないという仮説に基づいている。具体的な手法例として、GEOM [14] はまず、分布内データに回転のような幾何変換を行い、この変換をあてられるように学習する。次にテスト時の 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。文献 [11] によると、CIFAR100 が分布内データの時、CIFAR10 は near-distribution outliers と呼ばれており、自己教師あり学習を用いた手法は、near-distribution outliers を検知するのに役立つといわれている。一方で、筆者らは分布内データセットとして CIFAR100、分布外データセットとして CIFAR10 を用いたとき、ほとんどの分布外データの幾何変換である回転を予測できてしまい、分布外データを検知できなくなってしまうことを実験的に発見した。

上記の Hendrycks らの手法は、クラス分類精度が落ちるデータセットを用いた場合に検知精度が落ちる一方で、GEOM は幾何変換を予測できるような分布外データが存在し、このとき検知精度が落ちる。筆者らは上記 2 つの手法は検知精度が低くなる条件が異なるので、両者の正常スコアの性質を同時に反映したスコアを作成することで、両者の欠点を補い合えると仮定した。この仮定に基づき、筆者らは 1 つの画像が入力となったときに、クラスラベルの分類確率と幾何変換の分類確率を 1 つのモデルで計算し、それぞれの正常スコアを組み合わせることを提案する。具体的にはクラスラベルと幾何変換の同時確率を求め、それをもとにした新しい正常スコアを提案する。このスコアは先行研究である Hendrycks らの手法や GEOM と同様に、分

¹ 東京大学工学系研究科技術経営戦略学専攻松尾研究室
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

^{*1} ここでの正常スコアとは、分布内データを表す尺度であり、このスコアが大きいかほど分布内データであると判定されやすい。これ以降、本文中で同様の意味として用いる。

布内データのみで学習することができ、テスト時に分布外データを利用する必要もない。ここで、提案手法はクラスラベルと幾何変換の同時確率を求める必要があるため、分布内データセットは複数のクラスを持っているような画像データを対象としていることに注意されたい。

実験では、分布外検知の手法の検証のための標準的な複数のデータセットとネットワーク構造を用いる。提案手法は、仮説 (1) をもとにした Hendrycks らの手法と仮説 (2) をもとにした手法である GEOM 両方と比べて、分布外データの検知性能が向上することを確認し、それぞれの手法の問題が軽減されていることを確認する。具体的には、仮説 (1) をもとにした手法が苦手とする、クラス分類精度が落ちる分布内データを用いたときや、仮説 (2) をもとにした手法が苦手とする near-distribution outliers を用いたときにおいて、精度が大きく向上していることを確認する。さらに、クラス分類と幾何変換の2つの確率を直接出力し、それらをもとにしたスコアを単純に足し合わせるようなアンサンブル法を用いるよりも、同時確率を出力するモデルを利用して分布外検知の指標とするほうが検知精度が高くなることを示す。

2. 関連研究

分布外検知における多くの先行研究は<1>多クラス分類器を使う方法、<2>自己教師あり学習を用いる方法、<3>1クラス分類を用いる方法、<4>密度推定ベースの方法、<5>再構成誤差を用いる方法の5種類に分類できる。従来<3>、<4>、<5>が研究されていたが、2019年以降ではDNNにおいて<1>、<2>が主流になっており [16]、今回の研究はこれらの研究に焦点をあてている。<1>と<2>の方法の大まかな分類は表 1 にまとめた。

2.1 多クラス分類器を用いる方法

多クラス分類器を用いる分布外検知の手法は、分布内

表 1 <1>多クラス分類器を用いる手法と<2>自己教師あり学習を用いる手法の分類

Table 1 Classification of methods using multi-class classifiers <1> and using self-supervised learning <2>.

手法	訓練データに 分布外データを 使う必要がない	テストデータの 分布外データを 使う必要がない
<1>ODIN [6]	✓	✗
<1>Mahalanobis [7]	✓	✗
<1>Prior Networks [8]	✗	✓
<1>Hendrycks ら (2019) [9]	✗	✓
<2>GEOM+ [11]	✗	✓
<1>Hendrycks ら (2017) [5]	✓	✓
<2>GEOM [14]	✓	✓
<1, 2>Ours	✓	✓

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、クラスラベルを必要とする。Hendrycks らは、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正常スコアとしている。これは、入力分布外データの場合、その最大値は小さくなるという仮説のもと分布外検知を行っている。ODIN [6] 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。マハラノビス距離に基づく手法 (以降、Mahalanobis) [7] は中間層表現がガウス分布に従うことを仮定し、その分布から離れたサンプルを分布外データとする。これら2つの手法は [5] の性能を非常に高めた一方で、テストデータの分布外データを少量使う必要がある。

一方、テストデータの分布外データを使う必要はないが、訓練時にテストデータとは異なる種類の分布外データを必要とするモデルも存在する。たとえば、[8] はディリクレ分布のパラメータを DNN の出力として用いてデータの不確かさと分布の不確かさを分類する。このようにすることで、ソフトマックスの出力として一様分布を出力するような分布内データと分布外データを見分けることができる。文献 [9] は文献 [5] における学習に加え、訓練用分布外データが入力のときに、出力が一様分布になるように学習する。これらの訓練用分布外データを用いる問題点として、テストデータの分布外データの種類の異なるため、必ずしも汎化しないことがあげられる。実際に、少量の訓練用分布外データで学習した場合はその他の種類の分布外データに汎化しないという実験結果もある [13]。筆者らは、多クラス分類器を用いる方法の中でも訓練中に分布外データを用いず、テストデータの分布外データにもアクセスしない手法に焦点を当てる。

多クラス分類器を用いる分布外検知の手法に共通して、クラス分類精度が落ちるデータセットを用いた場合、明確に特定のクラスに属さないような、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。このとき、多クラス分類器ベースの分布外検知の手法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り、この問題は先行研究によって確かめられていないため、筆者らはこの問題を実験で明らかにする。

2.2 自己教師あり学習を用いる方法

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特に有望な手法である。GEOM では訓練データに対して回転、並行移動などの幾何変換を行い、その変換を当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仮説のもとで分布外検知を行う。GOAD [15] 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。GEOM+ [11] は GEOM において、Hendrycks ら (2019) の手法 [9] 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。

自己教師あり学習を用いた分布外検知は検知するのが難しい near-distribution outliers に役立つといわれている [11]。しかし、筆者らは CIFAR100 を分布内データとし、near-distribution outliers のデータセットである CIFAR10 を分布外データとして実験を行ったとき、ほとんどの分布外データの幾何変換（回転）を予測できてしまい、分布外データと分布内データを見分けることができなくなることを実験によって発見した。

2.3 1 クラス分類を用いる方法

1 クラス分類の手法は、分布内データと分布外データを直接分類できるようなモデルを構築する方法である。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 [10] 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。Deep SAD [12] は訓練用分布外データを用いることで文献 [10] の手法の精度をさらに向上させた。

2.4 密度推定ベースの方法

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）を正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 [18]、glow [19] 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、対数尤度の下界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 [20]。

2.5 再構成ベースの方法

autoencoder の利用により、再構成誤差を分布外データであることを示すスコア（異常スコア）として利用する方法がある [21], [22]。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すると異常スコアは大きくなる。発展型として、DAGMM [23] 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

外検知だけでなく、異常箇所を特定する異常部位検知にも有効である [24]。

3. 既存手法

この章では、多クラス分類器を利用した Hendrycks らの手法と幾何変換を利用した分布外検知の手法である GEOM を説明する。これらの手法は実験で比較手法としても用いる。次の章で、これらの手法をベースにした提案手法について説明する。

3.1 Baseline

$x \in \mathcal{X}$ を入力、 $c \in \{1, \dots, C\}$ をクラスラベルとする。DNN によるクラス分類器は $f_\phi(x)$ と表記し、 ϕ は DNN のパラメータとする。 $f_\phi(x)$ は x を入力とし、 C 次元のベクトルを出力する。 C 次元のベクトルは出力層のソフトマックス関数によって正規化されており、それぞれの要素は 0 以上で和が 1 となる。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 $f_\phi(x)$ の学習を行う。

$$\mathbb{E}_{p(x,c)}[\mathcal{L}(f_\phi(x), c)], \quad (1)$$

ここで、 \mathcal{L} はクロスエントロピーロス関数であり、これは通常の識別モデルの学習と同じである。また、 $p(x, c)$ はデータの経験分布とする。テスト時の正常スコアとして、式 (2) の最大ソフトマックス確率（maximum softmax entropy, MSP）を用いる。

$$S_{msp}(x) = \max_c f_\phi(x). \quad (2)$$

この手法は、多クラス分類器を分布外検知のために再学習する必要がなく、分布外データを訓練時およびテスト時に利用する必要がない。

3.2 GEOM

$\mathcal{T} = \{T_1, \dots, T_K\}$ を幾何変換の集合とする。ここで、 $1 \leq y \leq K, T_y: \mathcal{X} \rightarrow \mathcal{X}$ である。 $y \in \mathcal{Y} = \{1, \dots, K\}$ はどの幾何変換を利用するかを示すラベルである。幾何変換の分類器は $g_\phi(x)$ と表記する。 $g_\phi(x)$ は x を入力とし、 K 次元のベクトルを出力する。 K 次元のベクトルは出力層のソフトマックス関数によって正規化されており、それぞれの要素は 0 以上で和が 1 となる。GEOM のロス関数は式 (3) のようになり、このロス関数を最小化するように分類器 $g_\phi(x)$ の学習を行う。

$$\mathbb{E}_{p(x,y)}[\mathcal{L}(g_\phi(T_y(x)), y)]. \quad (3)$$

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。テスト時の正常スコアとして、以下の式を用いる。

$$S_{geom}(x) = \frac{1}{K} \sum_{y=1}^K [g_\phi(T_y(x))]_y, \quad (4)$$

ここで, $[\cdot]_y$ はベクトルの y 番目の成分を抽出する操作とする. このスコアは, 幾何変換をどれだけ当てられたかを意味している.

4. 提案手法

Hendrycks らは 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正常スコアとしている. このスコアは, 入力分布外データの場合, その最大値は小さくなるという仮説に基づいている. しかし, 多クラス分類器のクラス分類精度が落ちるデータセットを用いた場合, 明確に特定のクラスに属さないような, 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. そのため, この仮説をもとにした手法は上記の分布内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り, この問題は先行研究によって明らかにされていないため, 本研究で実験的に明らかにする.

GEOM は, DNN が分布外データの幾何変換を行ったときに, それがどの幾何変換か分からないはずという仮説に基づいている. 筆者らは分布内データセットとして CIFAR100, near-distribution outliers と呼ばれる分布外データセットとして CIFAR10 を用いたとき, ほとんどの分布外データの幾何変換である回転を予測できてしまい, 分布外データを検知できなくなってしまうことを実験的に発見した. このように, モデルが分布外データの幾何変換を予測できてしまう場合には分布外データを検知することができない問題がある.

筆者らは上記 2 つの手法は検知精度が低くなる条件が異なるので, 両者の正常スコアの性質を同時に反映したスコアを作ることで, 両者の欠点を補い合えると仮定した. この仮定に基づき, 筆者らはある画像が入力のとき, クラスラベルの分類確率と幾何変換の分類確率を 1 つのモデルで計算し, それぞれの正常スコアを組み合わせることを提案する. 具体的にはクラスラベルと幾何変換の同時確率を求め, それをもとにした新しい正常スコアを提案する. ここで, 提案手法はクラスラベルと幾何変換の同時確率を求めなければならないため, 分布内データセットは複数のクラスを持っているような画像データを対象としていることに注意されたい.

クラスと幾何変換の同時確率を求める最も単純な手法として, 先行研究のクラス分類と幾何変換分類を同時に解く分類器 [25] を利用する. この分類器は構造と学習方法はこれまでのモデルと同じであるが, ラベルはクラス c とどの幾何変換を利用するかを示すラベル y の結合ラベル $(c, y) \in \{1, \dots, C\} \times \{1, \dots, K\}$ を教師とするところのみ異なる. 実装上は, $C \times K$ 個の分類問題を考えるだけで良い. ここで, 2次元の結合ラベル (c, y) を 1次元のラベル z にまとめ $z \in \{1, \dots, C \times K\}$ とし, DNN の多クラス

分類器として $h_\phi(x)$ を用意する. $h_\phi(x)$ は x を入力とし, $C \times K$ 次元のベクトルを出力する. このベクトルは出力層のソフトマックス関数によって正規化されており, それぞれの要素は 0 以上で和が 1 となる. このとき, 提案手法のロス関数は式 (5) のようになる.

$$\mathbb{E}_{x,c,y \sim p(x,c,y)} [\mathcal{L}(h_\phi(T_y(x)), z)], \quad (5)$$

ここで, $z = c + C \times y$ である. また, $p(x, c, y) = p(x, c)p(y)$ であり, $p(x, c)$ は経験分布, $p(y)$ は一様なカテゴリカル分布である.

テスト時には, $h_\phi(x)$ の出力であるベクトルの形を変え, 1 つの x に対してベクトルではなく, 同時確率である行列を返すようにする. ここで, 行列を返す関数を $p_{YC}(x)$ と表記することにする. このとき, テスト時の正常スコアとして, 以下を提案する.

$$S_{ours}(x) = \frac{1}{K} \sum_{y=1}^K \max_c [p_{YC}(T_y(x))]_{y,\cdot}, \quad (6)$$

ここで, $[\cdot]_{y,\cdot}$ は y 番目の成分すべてを返す操作であり, ここでは C 次元のベクトルを返している. 式 (6) のスコアは, 式 (2) と式 (4) を組み合わせたものになっている. 提案したスコアは, 分布外データが (1) クラス分類器の出力が一様分布に近づき, (2) 幾何変換を予測することが難しいという 2 つの仮説が同時に反映されるように作られている. このスコアはベースとなった Hendrycks らの手法や GEOM と同様に, 分布内データのみで学習することができ, テスト時に分布外データを利用する必要もない. また, 片方のどちらかのスコアのスケールに依存しないため, 単純に 2 つのモデルでクラス分類と幾何変換の分類の確率を導出し, そのスコアの和を取る場合と比較して, 片方のスコアだけを着目してしまう可能性が低い. 片方のスコアだけを着目しないようにスケールの調整をすることは, テスト時の分布外データがどのようなものが与えられるか分からないので, 基本的には困難である.

提案手法のアルゴリズム (Algorithm 1) を以下に説明する. まず, 訓練データ x , 訓練データのクラスラベル c , 画像の回転角 y をランダムにサンプリングする. 次に, 結合ラベル z を y, c をもとに作り, 結合分類器 $h_\phi(x)$ のターゲットとする. $h_\phi(x)$ はクロスエントロピー損失関数を用いて学習する. テスト時は予測された結合ラベルのクラス方向の最大値を求め, 回転角方向に平均した値を最終的なスコアとする.

5. 実験

DNN を用いた分類器は, 分布外データを入力としたとき, (1) クラス分類の出力が一様分布に近づき, (2) 幾何変換を予測することが難しいという 2 つの仮説のもとで, 分布外検知が行えることが知られている. 実験では上記の仮

Algorithm 1 algorithm for proposed method**Require:** joint classifier $h_\phi(x)$, training set, test set.Initialize parameters of the classifiers: ϕ ▷ training joint classifier $h_\phi(x)$ **repeat** $x, c \in$ training set $y \sim \text{Uni}(1, K)$ $z = y \times C + c$ $\phi \leftarrow \text{Update}(\mathcal{L}(h_\phi(T_y(x)), z))$ **until** convergence of parameters ϕ

▷ estimating test score

for $x \in$ test set **do** $S_{ours}(x) = \frac{1}{K} \sum_{y=1}^K \max_c [p_{YC}(T_y(x))]_{y.}$ **end for**

説に基づく手法に問題があることを確かめる。まず (1) の仮説に基づく手法において、クラス分類精度が低くなるときに分布外データの検知精度が低くなるという問題があり、このことを実験で明らかにする。次に (2) の仮説に基づく手法において、分布外データが near-distribution outliers のとき分布外データの検知精度が下がる問題があることを確かめる。これらの問題が発生する理由はそれぞれの手法において異なる。そのため、(1) と (2) を同時に考慮した提案手法によって、上記の問題が軽減されるということを示す。最後に、提案手法で用いる同時確率の有効性を確認するための実験を行う。

5.1 実験設定

5.1.1 データセット

筆者らは以下の分布外検知のための標準的なデータセットを利用した。SVHN [26] と CIFAR10・CIFAR100 [27] は分布内データまたは分布外データとして利用した。上記のデータセットを分布内データとしたときに、Tiny ImageNet (TIM) [2], LSUN [28], iSUN [29], Gaussian noise と Uniform noise は分布外データとして利用した。これらのデータセットの分布内外の設定は [6], [7], [30] の設定にあわせた。

5.1.2 比較手法

著者らは比較手法として、多クラス分類器を用いる Hendrycks らの手法と、自己教師あり学習を用いた手法として GEOM を用いる。また、提案手法で用いる同時確率の有効性を示すため、周辺確率を出力するモデルとも比較する。このモデルは、クラス分類と幾何変換の分類を同時に学習し、クラス分類確率と幾何変換分類確率を同時に出力する。同時確率を出力するわけではないので、出力サイズはクラス数 × 幾何変換数ではなく、クラス数 + 幾何変換数である。

5.1.3 訓練方法の詳細

提案手法と比較手法のネットワーク構造は WideResNet (WRN-28-10) [31] と DenseNet (Dense-BC-100-12) [32] を利用した。それぞれのモデルはモメンタム項 0.9 の確率的

勾配法を使って訓練を行った。学習率の初期値は 0.1 とし、epoch 数が総 epoch 数の 50% を超えたときに 0.01、75% を超えたときは 0.001 とした。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。具体的には、[5] のモデル、周辺確率を出力するモデル、同時確率を出力するモデルの訓練では、SVHN を利用したとき 60 epoch、CIFAR10 と CIFAR100 を利用したとき 400 epoch とした。GEOM のモデルの訓練では SVHN のとき 100 epoch、CIFAR10・CIFAR100 を用いたときは 200 epoch とした。各訓練時は、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い、クロッピングやフリッピングといった標準的なオーグメンテーションを行った。

5.1.4 モデルの評価

評価指標について、筆者らは分布外データの検知精度と分布内データを分類精度をテストするための標準的な評価指標を利用した。具体的には、Area Under the Receiver Operating Characteristics (AUROC)、分布内データの正解率 (ACC) を用いた。この評価指標は [6], [7] でも用いられている。分布外検知の検証の際は、テスト用の分布内データと分布外データを利用し、それぞれの正常スコアを求めた。提案手法と比較手法の正常スコアとして、3 章で説明したものをそれぞれ用いた。比較手法のクラス分類の検証の際は、それぞれの学習したモデルで、テスト用の分布内データを利用し、予測の正解率を計算した。提案手法のクラス分類の検証の際は、文献 [25] と同じように、幾何変換とクラス分類の同時確率を計算し、幾何変換を与えたときのクラス分類の確率をアンサンブルして予測した。

5.2 先行研究における問題の確認

この節では、2 つの先行研究 [5], [11] における問題を本実験によって明らかにする。まず、多クラス分類器を用いる Hendrycks らの手法において、クラス分類精度が低くなるときに分布外データの検知精度が低くなるということを確認する。結果は図 1 にまとめた。縦軸は Baseline による分布外検知の検知精度 (AUROC)、横軸は分布外データセットを意味する。上図と下図の結果はネットワーク構造として、それぞれ WRN-28-10, Dense-BC-100-12 を用いている。分布内データセットのクラス分類の正解率は表中の凡例に記載した。

図の結果より、クラス分類の精度が低くなるようなデータセット、特に CIFAR100 において分布外データの検知精度が低下することが分かる。分布内データが入力だとしても、予測が間違ってしまう場合出力が一様分布に近づき、最大ソフトマックス確率が低くなってしまうのが理由であると考えられる。

次に自己教師あり学習を用いることで分布外検知を行う手法である GEOM において、分布外データが near-distribution outliers のとき分布外データの検知精度が下が

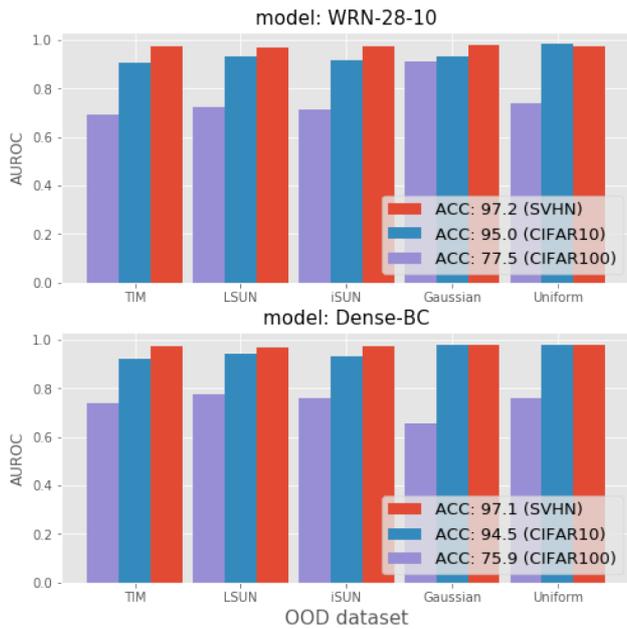


図 1 分布内データセットの正解率と Hendrycks らの手法による検知精度の関係

Fig. 1 Relationship between the classification accuracy using in-distribution dataset and the detection accuracy using the method of Hendrycks et al.

表 2 幾何変換予測の正解率と GEOM の AUROC の関係. near-distribution outliers は太字にした. また, 最も AUROC が低いものを赤字にした

Table 2 Relationship between the classification accuracy of geometric transformation predictions and the AUROC of GEOM. Near-distribution outliers are in bold, and the lowest AUROC scores are in red.

In-dist (ACC)	OOD	ACC	AUROC
CIFAR100 (84.1) (WRN-28-10)	SVHN	17.6	92.7
	CIFAR10	83.6	49.5
	TIM	43.9	82.7
	LSUN	52.7	77.8
	iSUN	50.7	78.8
CIFAR100 (82.3) (Dense-BC)	SVHN	22.4	90.9
	CIFAR10	83.9	47.7
	TIM	37.2	84.7
	LSUN	42.4	81.9
	iSUN	43.1	81.7

ることを確かめる.

結果は表 2 にまとめた. 1 列目には分布内データセット, 分布内データセットの幾何変換分類の正解率, ネットワーク構造を表記した. 2 列目には分布外データの種類を表記した. 3 列目には分布外データの幾何変換分類の正解率 (ACC) を表記した. 4 列目には評価指標としてそれぞれ AUROC を利用したときの結果を表記した.

GEOM は near-distribution outliers のとき, 具体的には CIFAR100 が分布内データで CIFAR10 が分布外データのとき, 分布外データの検知精度 (AUROC) を大きく落

としていることが分かる. 文献 [11] によれば自己教師あり学習を用いた分布外検知の手法は検知するのが難しい near-distribution outliers に役立つといわれているため, これは先行研究とは逆の結果となった.

この結果の原因を以下に述べる. 文献 [11] での実験では, CIFAR10 のある特定のクラスを分布内データとし, その他のクラスを分布外データとしていた. ある特定のクラスの分布内データの幾何変換を学習したとき, 分布外データの幾何変換を予測することはできなかったため検知精度が高いものとなった. いい換えると, 1 つのクラスの分布内データのみを訓練データとしているため, その他のクラスの幾何変換が予測できるように汎化しなかったから分布外データを検知できていたと考えられる. 一方で, 今回の実験では CIFAR100 のすべてのクラスを分布内データとして訓練した. このとき, モデルは様々なクラスのデータの回転を学習しており, 似たようなクラスのデータである CIFAR10 のデータが予測できるように汎化してしまった. そのため, 自己教師あり学習を用いた分布外検知手法であるにもかかわらず, near-distribution outliers を検知できなかったと考えられる.

5.3 先行研究との比較

この実験では, 仮説 (1) と仮説 (2) のどちらかに基づいた手法を用いるよりも, これらを両方に基づいた指標を使ったほうが検知精度が高いことを示す. 結果は表 3 にまとめた. 1 列目には分布内データ, ネットワーク構造, 分布内データの正解率を表記した. 2 列目には分布外データの種類を表記した. 3, 4, 5 列目には評価指標としてそれぞれ AUROC を利用したときの結果を表記した. 3, 4, 5 列目の 2 段目は比較手法と提案手法を記した. Hendrycks らの手法は式 (2) のスコア, GEOM は式 (4) のスコア, Ours は同時確率のモデルを利用したときの式 (6) のスコアである.

結果よりほとんどの場合においてベースとなった分布外検知の手法よりも提案手法の検知精度は向上した. 特に提案手法は GEOM と比べて, GEOM が苦手とする near-distribution outliers の検知精度が高いことが分かる. 具体的には, CIFAR10 を分布内データ, CIFAR100 を分布外データとしたときの検知精度, および CIFAR100 を分布内データ, CIFAR10 を分布外データとしたときの検知精度が向上していることが分かる. また, 提案手法は Hendrycks らの手法と比べて, Hendrycks らの手法が苦手とする CIFAR100 のようなクラス分類精度が下がる難しいデータセットが分布内データのときに検知精度が高いことが分かる. これは, 提案手法は Hendrycks らの手法と GEOM の分布外データを見分ける基準を持つ手法であるため, 片方の基準で分布外データか分布内データを見分けるのに曖昧な入力があったとしても, もう片方の基準で明確にわけることができれば, 分布外データを検知することができるか

表 3 異なるテストデータとモデルを用いたときの分布外検知の結果. 各実験で最も良い結果は太字にした

Table 3 Results of out-of-distribution detection when using different test data and models. The best results for each experiment are in bold.

In-dist (model)	OOD	AUROC		
		GEOM	Hendrycks ら	Ours
SVHN (WRN-28-10)	CIFAR10	96.6	93.9	98.1
	CIFAR100	95.8	93.8	97.7
	TIM	96.7	95.6	98.4
	LSUN	95.9	94.6	98.0
	iSUN	96.7	95.4	98.3
	Gaussian	97.1	97.0	99.0
	Uniform	97.1	96.7	98.9
CIFAR10 (WRN-28-10)	SVHN	97.9	94.5	98.5
	CIFAR100	76.2	87.2	92.0
	TIM	90.7	83.5	96.5
	LSUN	85.5	87.9	96.0
	iSUN	86.6	85.9	96.1
	Gaussian	97.2	93.9	98.3
	Uniform	97.2	94.8	98.9
CIFAR100 (WRN-28-10)	SVHN	92.7	79.9	95.4
	CIFAR10	49.5	79.7	74.9
	TIM	82.7	67.4	91.4
	LSUN	77.8	66.0	90.4
	iSUN	78.8	65.8	89.9
	Gaussian	90.8	62.0	93.3
	Uniform	90.6	46.0	93.7
SVHN (Dense-BC)	CIFAR10	96.3	97.7	98.1
	CIFAR100	95.6	97.3	97.7
	TIM	96.4	97.8	98.4
	LSUN	94.3	97.5	97.7
	iSUN	95.6	97.9	98.3
	Gaussian	97.3	97.8	98.7
	Uniform	97.3	97.8	98.5
CIFAR10 (Dense-BC)	SVHN	97.7	92.3	98.2
	CIFAR100	79.4	88.1	90.9
	TIM	92.2	92.7	96.2
	LSUN	88.3	94.2	94.8
	iSUN	88.8	93.8	95.0
	Gaussian	97.2	99.2	98.1
	Uniform	97.2	96.7	98.0
CIFAR100 (Dense-BC)	SVHN	90.9	74.2	95.1
	CIFAR10	47.7	75.5	68.9
	TIM	84.7	68.3	87.9
	LSUN	81.9	72.5	87.2
	iSUN	81.7	70.2	86.0
	Gaussian	90.6	68.8	91.8
	Uniform	90.7	72.2	91.4

らだと考えられる.

一方で, クラスの分類精度が落ちるデータセット (CIFAR100) を使い, かつ, 分布外データが near-distribution outliers である CIFAR10 のときに提案手法の分布外デー

タの検知精度が落ちることが分かる. これはそれぞれの手法が分布外データを検知するのが苦手とする場合の組み合わせであるからだと考えられる.

5.4 単純なアンサンブル法との比較

この節では, クラス分類と幾何変換の2つの確率を直接出力するモデルを利用し, 単純にスコアを足し合わせるようなアンサンブル法 (以降これを Ensemble と表記する) よりも提案手法のほうが検知精度が高くなることを示すための実験を行う. 提案手法のスコアの計算方法として, Ensemble と条件を揃えるために, 同時分布を直接出力するモデルを利用して確率の周辺化を行い, スコアを足し合わせる (以降これを Ours2 と表記する).

まず, Ours2 と Ensemble の正常スコアは以下の式 (7) を利用した. 以下のスコアを利用するのは, 式 (6) は同時確率を出力するモデルのみが使えるスコアであり, 公平な比較のためである.

$$S_{Ours2}(x) = \frac{1}{K} \sum_{y=1}^K p_y(T_y(x)) + \frac{1}{K} \sum_{y'=1}^K \max_c p_c(T_{y'}(x)), \quad (7)$$

ここで, 第1項は GEOM, 第2項は Hendrycks らの手法に基づくスコアである. ただし, Ours2 を利用するときは,

$$p_y(x) = \sum_{c=1}^C [p_{YC}(x)]_{y,c}, p_c(x) = \sum_{y=1}^K [p_{YC}(x)]_y$$

とする. ここで, $[\cdot]_{y,c}$ は y, c 成分を返す操作であり, ここではスカラーを返している. また, $[\cdot]_y$ は y 番目の成分すべてを返す操作であり, ここでは C 次元のベクトルを返している. 上のスコアは, Ours2 を用いるとき, 同時確率をもとに幾何変換方向およびクラス方向に周辺化を行い, それぞれの項に対して式 (2) と式 (4) と同じ操作を行い, それぞれの項を足したスコアとなっている. 一方で Ensemble を用いるときは, クラス分類確率と幾何変換の分類確率がそれぞれ $p_c(x), p_y(x)$ となるため, 直接式 (7) を用いる.

結果は表 4 にまとめた. 1 列目には分布内データ, ネットワーク構造, 分布内データの正解率を表記した. 2 列目には分布外データの種類を表記した. 3, 4 列目には評価指標としてそれぞれ AUROC を利用したときの結果を表記した.

まずクラス分類に関して, Ensemble のモデルでクラス分類するよりも, Ours2 のように同時確率を学習してクラス分類を行う方が正解率が高い結果となった. このように同時確率を用いた方が正解率が高くなることは文献 [25] ですでに実験的に示されており, 文献 [25] と矛盾しない結果となった.

分布外検知の性能に関しては, Ours2 が Ensemble と比べて, それぞれのネットワーク構造, 分布外データセットに

表 4 異なるテストデータとモデルを用いたときの分布外検知の結果. 各実験で最も良い結果は太字にした

Table 4 Results of out-of-distribution detection when using different test data and models. The best results for each experiment are in bold.

In-dist (model) ACC (Ensemble/Ours2)	OOD	AUROC	
		Ensemble	Ours2
SVHN (WRN-28-10) 94.9 / 97.9	CIFAR10	98.3	98.4
	CIFAR100	97.8	97.9
	TIM	98.7	98.8
	LSUN	98.1	98.3
	iSUN	98.5	98.6
	Gaussian	99.6	99.7
	Uniform	99.4	99.6
CIFAR10 (WRN-28-10) 96.7 / 96.8	SVHN	98.6	99.2
	CIFAR100	91.1	92.1
	TIM	96.2	96.9
	LSUN	95.6	96.5
	iSUN	95.6	96.6
	Gaussian	98.7	97.7
	Uniform	98.4	100.0
CIFAR100 (WRN-28-10) 80.9 / 84.0	SVHN	93.0	95.2
	CIFAR10	72.3	75.1
	TIM	90.4	92.2
	LSUN	87.8	90.9
	iSUN	87.5	90.4
	Gaussian	91.2	93.8
	Uniform	89.6	97.4
SVHN (Dense-BC) 93.9 / 97.8	CIFAR10	98.3	98.3
	CIFAR100	97.7	97.8
	TIM	98.8	98.8
	LSUN	98.1	98.0
	iSUN	98.7	98.6
	Gaussian	99.5	99.3
	Uniform	99.5	99.1
CIFAR10 (Dense-BC) 94.9 / 95.8	SVHN	98.5	98.5
	CIFAR100	89.2	91.1
	TIM	96.1	96.8
	LSUN	94.1	95.3
	iSUN	94.3	95.5
	Gaussian	97.1	97.4
	Uniform	97.1	97.3
CIFAR100 (Dense-BC) 74.8 / 79.1	SVHN	91.0	94.3
	CIFAR10	65.8	69.2
	TIM	85.3	88.2
	LSUN	82.3	87.5
	iSUN	82.7	86.3
	Gaussian	86.3	91.0
	Uniform	85.5	88.4

において、ほとんどの場合で検知精度が上回る結果であることが分かる。この原因として、Ensemble のモデルは Ours2 のモデルよりも分布内データの分類精度が低くなることが考えられる。分布内データの精度が低くなると、図 1 から

分かるように、式 (2) を利用した分布外検知の精度は悪くなる傾向がある。Ours2 と Ensemble は式 (2) と式 (4) を組み合わせる手法であるため、結果的に分布外データの検知精度も低くなってしまったと考えられる。

また、式 (7) による結果は式 (6) による結果に比べて、あまり差がでない、もしくは若干精度が高い結果となった。この理由として、式 (6) はスケールに依存しない提案手法である一方で、今回の対象としたデータのクラス分類と幾何変換の分類確率は、ほぼ同じスケールになったことが考えられる。

6. 本研究の限界と実応用例

この章では本研究の限界と実応用例について述べる。提案手法の特性上、データセットに対する条件がいくつか存在する。まず、分布内データのクラス分類を行う手法であるため、複数の種類のクラスのデータとそのラベルが必要である。次に、画像の回転の分類を利用している手法でもあるため、データのドメインは画像に限られる。また、回転対称性のある画像の場合、回転をあてることができないため、分布内外のデータを見分けるのが難しくなる可能性がある。さらに、実験結果より、分布内データのクラス数が多いと、似たようなクラスの分布外データを検知することが難しくなる。まとめると、高い検知精度が保証されるのは、数種類のクラスラベルがあり、回転対称性がないような画像データセットのときであると考えられる。具体的な実世界の適用例として、ある特定のグループの顔認証システムにおいて、グループ外の人の顔や、まったく顔とは関係のないものを高い精度で検知できると考えられる。

7. 結論

本稿では、DNN を使った分類器は分布外データを入力としたとき、(1) クラス分類の出力が一様分布に近づき、(2) 幾何変換を予測することが難しいという 2 つの仮説を元にした分布外検知の手法を提案した。具体的には、入力を与えられたときのクラスと幾何変換の同時確率を求め、これをもとにした新しい正常スコアを提案した。実験では、様々なネットワーク構造とデータセットを使って実験を行い、(1) と (2) のどちらかの仮説をもとに分布外検知した手法よりも、提案手法のほうが分布外データの検知精度が高くなることを示した。また、単純なアンサンブル法よりも提案手法のほうが検知精度が高くなることも示した。本研究では、自己教師あり学習の 1 つである画像の回転をあてる手法を用いているため、対象とするデータは画像ドメインに限る。一方で、自己教師あり学習の 1 つである対照推定によって、潜在空間上での表現を学習し、この空間のもとで分布内外をわかるような方法を用いれば、画像データ以外のドメインでも有効になる可能性があり、今後の調査課題である。

参考文献

- [1] He, K., Zhang, X., Ren, S. and Sun, J.: Deep residual learning for image recognition, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.770–778 (2016).
- [2] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L.: Imagenet: A large-scale hierarchical image database, *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp.248–255, IEEE (2009).
- [3] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge, *International Journal of Computer Vision*, Vol.115, No.3, pp.211–252 (2015).
- [4] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J. and Mané, D.: Concrete problems in AI safety, arXiv preprint arXiv:1606.06565 (2016).
- [5] 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).
- [6] 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).
- [7] 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).
- [8] Malinin, A. and Gales, M.: Predictive uncertainty estimation via prior networks, *Advances in Neural Information Processing Systems*, pp.7047–7058 (2018).
- [9] Hendrycks, D., Mazeika, M. and Dietterich, T.: Deep Anomaly Detection with Outlier Exposure, *Proc. International Conference on Learning Representations* (2019).
- [10] 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).
- [11] 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).
- [12] Ruff, L., Vandermeulen, R.A., Goernitz, N., Binder, A., Müller, E., Müller, K.-R. and Kloft, M.: Deep Semi-Supervised Anomaly Detection, *International Conference on Learning Representations* (2020).
- [13] 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).
- [14] Golan, I. and El-Yaniv, R.: Deep anomaly detection using geometric transformations, *Advances in Neural Information Processing Systems*, pp.9758–9769 (2018).
- [15] Bergman, L. and Hoshen, Y.: Classification-Based Anomaly Detection for General Data, *International Conference on Learning Representations* (2020).
- [16] Bulusu, S., Kailkhura, B., Li, B., Varshney, P.K. and Song, D.: Anomalous example detection in deep learning: A survey, *IEEE Access*, Vol.8, pp.132330–132347 (2020).
- [17] Tax, D.M. and Duijn, R.P.: Support vector data description, *Machine learning*, Vol.54, No.1, pp.45–66 (2004).
- [18] 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).
- [19] Kingma, D.P. and Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolutions, *Advances in Neural Information Processing Systems*, pp.10215–10224 (2018).
- [20] 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).
- [21] Sakurada, M. and Yairi, T.: Anomaly detection using autoencoders with nonlinear dimensionality reduction, *Proc. MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*, pp.4–11 (2014).
- [22] 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).
- [23] 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).
- [24] Dehaene, D., Frigo, O., Combexelle, S. and Eline, P.: Iterative energy-based projection on a normal data manifold for anomaly localization, *International Conference on Learning Representations* (2020).
- [25] Lee, H., Hwang, S.J. and Shin, J.: Self-supervised Label Augmentation via Input Transformations, *Proc. Machine Learning and Systems 2020*, pp.3537–3547 (2020).
- [26] Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B. and Ng, A.Y.: Reading digits in natural images with unsupervised feature learning (2011).
- [27] Krizhevsky, A. et al.: Learning multiple layers of features from tiny images, Technical Report, Citeseer (2009).
- [28] Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T. and Xiao, J.: Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop, arXiv preprint arXiv:1506.03365 (2015).
- [29] Xu, P., Ehinger, K.A., Zhang, Y., Finkelstein, A., Kulkarni, S.R. and Xiao, J.: Turkergaze: Crowdsourcing saliency with webcam based eye tracking, arXiv preprint arXiv:1504.06755 (2015).
- [30] DeVries, T. and Taylor, G.W.: Learning confidence for out-of-distribution detection in neural networks, arXiv preprint arXiv:1802.04865 (2018).
- [31] Zagoruyko, S. and Komodakis, N.: Wide residual networks, arXiv preprint arXiv:1605.07146 (2016).
- [32] Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q.: Densely connected convolutional networks, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.4700–4708 (2017).



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。2014~2018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。