深い最近傍の異常検出

Liron Bergman [*1] **Niv Cohen** [*1] **Yedid Hoshen** [1]

Abstract

ニアレストネイバーは、異常検出のための成功した長年の手法である。最近、自己教師付きディープメソッド（例えばRotNet）によって大きな進歩が達成された。しかし、自己教師付き特徴量は、一般的にImagenetで事前に訓練された特徴量を下回る。本研究では、最近の進歩が、Imagenetで事前に訓練された特徴空間上で動作する最近傍手法を本当に上回ることができるかどうかを調査する。単純な最近傍ベースのアプローチは、画像分布に関する仮定が少ない一方で、精度、数ショット汎化、学習時間、ノイズ頑健性において、自己教師あり手法を上回ることが実験的に示されている。

1. Introduction

世界と相互作用するエージェントは常に連続的なデータの流れにさらされている。エージェントは特定のデータを異常、つまり特に興味深い、あるいは予期せぬものとして分類することで利益を得ることができる。このような識別は、必要な観測にリソースを割り当てるのに役立つ。このメカニズムは、人間がチャンスを発見したり、危険を警告するために使用される。人工知能による異常検知は、詐欺検知、サイバー侵入検知、重要な産業機器の予知保全など、多くの重要な応用がある。

機械学習では、異常検出のタスクは、データポイントを正常または異常としてラベル付けできる分類器を学習することから成る。教師あり分類では、異常なデータはノイズとみなされるのに対して、正常なデータに対してはうまく機能しようとする。異常検知手法の目標は、変動が激しく予測が困難な極端なケースを特別に検知することである。このため、異常検知のタスクは困難である（そしてしばしば仕様が不十分である）。

異常検知の3つの主な設定は、教師あり、半教師あり、教師なしである。

ヘブライ大学コンピューターサイエンス・工学部、イスラエル。宛先 Yedid Hoshen <yedid.hoshen@mail.huji.ac.il>.

教師あり設定では、正常データと異常データに対してラベル付けされた訓練例が存在する。従って、他の分類タスクと基本的な違いはない。この設定はまた、多くの異常検出タスクにとっては制約が多すぎる。例えば、新しい病気の出現のように、興味のある異常の多くは過去に見たことがないからである。より興味深い半教師あり設定では、全ての学習画像は正常であり、異常は含まれない。正常-異常分類器を学習するタスクは1クラス分類となる。最も困難な設定は教師なし設定で、正常データと異常データの両方からなるラベル付けされていない訓練セットが存在する。典型的な仮定は、異常データの割合が正常データよりも有意に小さいことである。本稿では、半教師あり設定と教師なし設定の両方を扱う。典型的な異常検知手法は、距離、分布、または分類に基づく。ディープニューラルネットワークの出現は、それぞれのカテゴリーに大きな改善をもたらした。この2年間で、ディープな分類に基づく手法は、他の全ての手法を大幅に凌駕した。これは主に、正常なデータに対してあるタスクを実行するように訓練された分類器は、未見の正常なデータに対してはこのタスクをうまく実行するが、異常なデータに対しては、異なるデータ分布に対する汎化がうまくいかないために失敗するという原理に依存している。

最近の論文で，Gu ら（2019）は，生データ上の K 最近傍（kNN）アプローチが，表データ上の最先端手法と競合することを実証した．驚くべきことに、kNNは現在のほとんどの画像異常検出の論文では使われていないし、比較もされていない。本論文では、生の画像データに対するkNNの性能は良くないが、強力な市販の汎用特徴抽出器と組み合わせると、最新技術を凌駕することを示す。具体的には、Imagenetで事前学習されたResNet特徴抽出器を用いて、全ての（訓練とテストの）画像を埋め込む。各テスト画像の埋め込みとトレーニングセットの間のK最近傍（KNN）距離を計算し、単純な閾値ベースの基準を使用して、データが異常かどうかを判断します。

我々は、一般的に使用されるデータセットと、Imagenetとは全く異なるデータセットの両方で、このベースラインを広範囲に評価した。我々は、このベースラインが既存の手法と比較して大きな利点があることを発見した。

Deep Nearest Neighbor Anomaly Detection

Liron Bergman [*1] **Niv Cohen** [*1] **Yedid Hoshen** [1]

Abstract

Nearest neighbors is a successful and long-standing technique for anomaly detection. Significant progress has been recently achieved by self-supervised deep methods (e.g. RotNet). Self-supervised features however typically underperform Imagenet pre-trained features. In this work, we investigate whether the recent progress can indeed outperform nearest-neighbor methods operating on an Imagenet pretrained feature space. The simple nearest-neighbor based-approach is experimentally shown to outperform self-supervised methods in: accuracy, few shot generalization, training time and noise robustness while making fewer assumptions on image distributions.

1. Introduction

Agents interacting with the world are constantly exposed to a continuous stream of data. Agents can benefit from classifying particular data as anomalous i.e. particularly interesting or unexpected. Such discrimination is helpful in allocating resources to the observations that require it. This mechanism is used by humans to discover opportunities or alert of dangers. Anomaly detection by artificial intelligence has many important applications such as fraud detection, cyber intrusion detection and predictive maintenance of critical industrial equipment.

In machine learning, the task of anomaly detection consists of learning a classifier that can label a data point as normal or anomalous. In supervised classification, methods attempt to perform well on normal data whereas anomalous data is considered noise. The goal of an anomaly detection methods is to specifically detect extreme cases, which are highly variable and hard to predict. This makes the task of anomaly detection challenging (and often poorly specified).

The three main settings for anomaly detection are: supervised, semi-supervised and unsupervised. In the *supervised* setting, labelled training examples exist for normal and anomalous data. It is therefore not fundamentally different from other classification tasks. This setting is also too restrictive for many anomaly detection tasks as many anomalies of interest have never been seen before e.g. the emergence of new diseases. In the more interesting *semi-supervised* setting, all training images are normal with no included anomalies. The task of learning a normal-anomaly classifier is now one-class classification. The most difficult setting is *unsupervised* where an unlabelled training set of both normal and anomalous data exists. The typical assumption is that the proportion of anomalous data is significantly smaller than normal data. In this paper, we deal both with the semi-supervised and the unsupervised settings. Anomaly detection methods are typically based on distance, distribution or classification. The emergence of deep neural networks has brought significant improvements to each category. In the last two years, deep classification-based methods have significantly outperformed all other methods, mainly relying on the principle that classifiers that were trained to perform a certain task on normal data will perform this task well on unseen normal data, but will fail on anomalous data, due to poor generalization on a different data distribution.

In a recent paper, Gu et al. (2019) demonstrated that a K nearest-neighbours (kNN) approach on the raw data is competitive with the state-of-the-art methods on tabular data. Surprisingly, kNN is not used or compared against in most current image anomaly detection papers. In this paper, we show that although kNN on raw image data does not perform well, it outperforms the state of the art when combined with a strong off-the-shelf generic feature extractor. Specifically, we embed every (train and test) image using an Imagenet-pretrained ResNet feature extractor. We compute the K nearest neighbor (KNN) distance between the embedding of each test image and the training set, and use a simple threshold-based criterion to determine if a datum is anomalous.

We evaluate this baseline extensively, both on commonly used datasets as well as datasets that are quite different from Imagenet. We find that it has significant advantages over existing methods: i) higher than state-of-the-art accuracy ii) extremely low sample complexity iii) it can utilize

[*]Equal contribution [1]School of Computer Science and Engineering, The Hebrew University of Jerusalem, Israel. Correspondence to: Yedid Hoshen <yedid.hoshen@mail.huji.ac.il>.

例えば、画像は回転不変であり、任意のサイズであることが可能である。

もう一つの一般的なアプローチは、1クラスSVM (Scholkopf et al., 2000)と関連するSVDD (Tax & Duin, 2004)である。SVDDは、少なくともある割合の正常データ点を含む最小体積球を当てはめるものと見なすことができる。この方法は特徴空間に対して非常に敏感であるため、効果的な特徴空間を学習するためにカーネル法を用いた。

異常検出にkNNを使用することは新しい手法ではないが、最近の画像異常検出の研究ではあまり使用されず、比較もされていない。我々の目的は、この単純だが非常に効果的で一般的な画像異常検知手法を認知させることである。そのシンプルさ、ロバスト性、低いサンプル複雑性、一般性から、全ての新しい研究はこのシンプルな手法と比較されるべきだと我々は考えている。

2. Previous Work

プリディープ学習法 異常検出のための2つの古典的なパラダイムは、再構成ベースと分布ベースである。再構成ベースの手法は、学習セットを使用して、正常データを効果的な方法で表現する基底関数のセットを学習する。テスト時には、学習した基底関数を用いて新しいサンプルの再構成を試みる。この方法では、正常なデータはうまく再構成されるが、異常なデータは再構成されないと仮定する。再構成コストを閾値化することで、サンプルが正常か異常かに分類される。さまざまな基底関数の選択には、他のサンプルの疎な組み合わせ（例：kNN）（Eskin et al.、2002）、主成分（Jolliffe、2011; Candès et al.、2011）、K-means（Hartigan & Wong、1979）などがある。再構成メトリックには、ユークリッド距離、L_1距離、あるいはSSIM（Wang et al.） 再構成に基づく手法の主な弱点は、i）識別基底関数の学習が困難である ii）効果的な類似性尺度を見つけることが自明でないこと、である。半教師付き分布ベースのアプローチは、正規データの確率密度関数（PDF）を学習しようとするものである。新しいサンプルが与えられると、その確率が評価され、確率がある閾値より低ければ異常と判定される。このような手法には次のようなものがある：ガウス混合（GMM）などのパラメトリック・モデル。非パラメトリック手法には、Kernel Density Estimation (Latecki et al., 2007)やkNN (Eskin et al., 2002)がある（我々は再構成ベースも考慮する）。

分布手法の主な弱点は、高次元データの密度推定が難しいことである。

ディープネットワークによる古典的手法の補強： ディープニューラルネットワークの成功は、ディープに学習された特徴を古典的な手法と組み合わせる研究を促した。PCA法はディープオートエンコーダーに拡張され（Yang et al., 2017）、再構成コストはディープ知覚損失に拡張された（Zhang et al.） GANは画像における再構成のための基底関数としても用いられた。分布モデルを改善する1つのアプローチ（Zong et al., 2018）は、まずデータを意味的な低次元空間に埋め込むことを学習し、次にGMMなどの標準的な手法を用いてその分布をモデル化することである。SVDDはRuffら(2018)によって、カーネル法の優れた代替法として深層特徴を学習するために拡張された。この方法は「モード崩壊」の問題に悩まされており、後続の研究の対象となっている。この論文で調査されたアプローチは、古典的なkNNが深く学習された特徴で拡張されているため、このカテゴリーに属すると見なすことができる。

自己教師付きディープメソッド： 深層表現を学習するために監視を使用する代わりに、自己教師ありの方法は、データを入手するのが無料であるか、少なくとも非常に安価である補助的なタスクを解決するためにニューラルネットワークを訓練する。自己教師ありの表現は、通常、Imagenetのような大規模な教師ありデータセットから学習されたものを下回ることに注意すべきである。高品質な画像特徴を学習するための補助的なタスクとしては、ビデオフレーム予測（Mathieu et al.、2016）、画像カラー化（Zhang et al.、2016；Larsson et al.、2016）、パズル解法（Noroozi & Favaro、2016）などがある。最近、Gidarisら(2018)は一連の画像処理変換（画像軸を中心に0、90、180、270度回転）を用い、真の画像の向きを予測した。彼らはこれを用いて高品質の画像特徴を学習した。Golan & El-Yaniv(2018)は、画像の異常を検出するために同様の画像処理タスク予測を用いた。この手法は、異常クラスからの画像検出において良好な性能を示した。この手法の性能はHendrycksら(2019)によって改善され、一方Bergman & Hoshen(2020)によってopenset分類と組み合わされ、表形式データに拡張された。本研究では、画像異常検出タスクにおいて、自己教師付き手法が、強力な汎用特徴抽出器を用いた、より単純なkNNベースの手法を下回ることを示す。

本稿のもう一つの貢献は、深層学習コミュニティからほとんど注目されていなかったタスクである、画像グループの異常検出へのkNNの新しい適応を提示することである。

我々は、画像異常検出のための単純なK最近傍（kNN）ベースの手法を研究する。この手法をDeep Nearest Neighbors (DN2)と呼ぶ。

very strong external feature extractors, at minimal cost iv) it makes few assumptions on the images e.g. images can be rotation invariant, and of arbitrary size v) it is robust to anomalies in the training set i.e. it can handle the unsupervised case (when coupled with our two-stage approach) vi) it is plug and play, does not have a training stage.

Another contribution of our paper is presenting a novel adaptation of kNN to image group anomaly detection, a task that received scant attention from the deep learning community.

Although using kNN for anomaly detection is not a new method, it is not often used or compared against by most recent image anomaly detection works. Our aim is to bring awareness to this simple but highly effective and general image anomaly detection method. We believe that every new work should compare to this simple method due to its simplicity, robustness, low sample complexity and generality.

2. Previous Work

Pre-deep learning methods: The two classical paradigms for anomaly detection are: reconstruction-based and distribution-based. Reconstruction-based methods use the training set to learn a set of basis functions, which represent the normal data in an effective way. At test time, they attempt to reconstruct a new sample using the learned basis functions. The method assumes that normal data will be reconstructed well, while anomalous data will not. By thresholding the reconstruction cost, the sample is classified as normal or anomalous. Choices of different basis functions include: sparse combinations of other samples (e.g. kNN) (Eskin et al., 2002), principal components (Jolliffe, 2011; Candès et al., 2011), K-means (Hartigan & Wong, 1979). Reconstruction metric include Euclidean, L_1 distance or perceptual losses such as SSIM (Wang et al., 2004). The main weaknesses of reconstruction-based methods are i) difficulty of learning discriminative basis functions ii) finding effective similarity measures is non-trivial. Semi-supervised distribution-based approaches, attempt to learn the probability density function (PDF) of the normal data. Given a new sample, its probability is evaluated and is designated as anomalous if the probability is lower than a certain threshold. Such methods include: parametric models e.g. mixture of Gaussians (GMM). Non-parametric methods include Kernel Density Estimation (Latecki et al., 2007) and kNN (Eskin et al., 2002) (which we also consider reconstruction-based) The main weakness of distributional methods is the difficulty of density estimation for high-dimensional data. Another popular approach is one-class SVM (Scholkopf et al., 2000) and related SVDD (Tax & Duin, 2004). SVDD can be seen as fitting the minimal volume sphere that includes at least a certain percentage of the normal data points. As this method is very sensitive to the feature space, kernel methods were used to learn an effective feature space.

Augmenting classical methods with deep networks: The success of deep neural networks has prompted research combining deep learned features to classical methods. PCA methods were extended to deep auto-encoders (Yang et al., 2017), while their reconstruction costs were extended to deep perceptual losses (Zhang et al., 2018). GANs were also used as a basis function for reconstruction in images. One approach (Zong et al., 2018) to improve distributional models is to first learn to embed data in a semantic, low dimensional space and then model its distribution using standard methods e.g. GMM. SVDD was extended by Ruff et al. (2018) to learn deep features as a superior alternative for kernel methods. This method suffers from a "mode collapse" issue, which has been the subject of followup work. The approach investigated in this paper can be seen as belonging to this category, as classical kNN is extended with deep learned features.

Self-supervised Deep Methods: Instead of using supervision for learning deep representations, self-supervised methods train neural networks to solve an auxiliary task for which obtaining data is free or at least very inexpensive. It should be noted that self-supervised representation typically underperform those learned from large supervised datasets such as Imagenet. Auxiliary tasks for learning high-quality image features include: video frame prediction (Mathieu et al., 2016), image colorization (Zhang et al., 2016; Larsson et al., 2016), and puzzle solving (Noroozi & Favaro, 2016). Recently, Gidaris et al. (2018) used a set of image processing transformations (rotation by $0, 90, 180, 270$ degrees around the image axis), and predicted the true image orientation. They used it to learn high-quality image features. Golan & El-Yaniv (2018), have used similar image-processing task prediction for detecting anomalies in images. This method has shown good performance on detecting images from anomalous classes. The performance of this method was improved by Hendrycks et al. (2019), while it was combined with openset classification and extended to tabular data by Bergman & Hoshen (2020). In this work, we show that self-supervised methods underperform simpler kNN-based methods that use strong generic feature extractors on image anomaly detection tasks.

3. Deep Nearest-Neighbors for Image Anomaly Detection

We investigate a simple K nearest-neighbors (kNN) based method for image anomaly detection. We denote this method, Deep Nearest-Neighbors (DN2).

3.1. 半教師付き異常検知

DN2 は入力画像の集合 X_{train} = x_1 , x_2 ..x_N を取ります。
半教師付き設定では、全ての入力画像が正常であると仮定します。DN2は、訓練済みの特徴抽出器Fを用いて、訓練セット全体から特徴を抽出する：

$$f_i = F(x_i) \tag{1}$$

本稿では、Imagenetデータセットで事前学習されたResNet特徴抽出器 を使用する。一見すると、この監視は強い要求であるように見えるかもしれないが、このような特徴抽出器は広く利用可能である。正常な画像や異常な画像がImagenetと特に密接に関連している必要はないことは、後で実験的に示す。

f_{train} = f_1 , f_2 ..f_N . 初期段階の後、埋め込みは保存され、学習セットの推論を償却することができる。

新しいサンプルyが異常かどうかを推論するために、我々はまずその特徴埋め込みを抽出する： f_y = F（y). f_y = F（y）# 次にそのkNN距離を計算し、異常スコアとして利用する：

$$d(y) = \frac{1}{k} \sum_{f \in N_k(f_y)} \|f - f_y\|^2 \tag{2}$$

N_k（f_y ）は、学習集合 F_{train} における f_y に最も近い k 個の埋め込みを表す。我々はユークリッド距離を使用することにした。これはディープネットワークによって抽出された特徴量に対してしばしば強力な結果をもたらすが、他の距離尺度も同様の方法で使用することができる。距離d(y)が閾値より大きいかどうかを検証することで、画像yが正常か異常かを判定する。

3.2. 教師なし異常検知

完全に教師なしな場合、全ての入力画像が正常であると仮定することはできなくなり、代わりに、ごく一部の入力画像のみが異常であると仮定する。このより困難な設定に対処するために（そして教師なし異常検出に関する先行研究にも沿う）、我々はまず入力画像に対してクリーニング段階を行うことを提案する。特徴抽出段階の後、各入力画像と残りの入力画像との間のkNN距離を計算する。異常画像は低密度の領域にあると仮定して、kNN距離が最大の画像の一部を除去する。この割合は、推定される異常な入力画像の割合よりも大きくなるように選ぶべきである。DN2が非常に少ない学習画像しか必要としないことは、後の実験で示される。従って、除去する画像の割合に非常に積極的になり、正常である可能性が最も高い画像だけを残すことができる（実際にはトレーニング画像の50%を除去する）。異常が疑われる入力画像を除去した後、画像は正常な画像の割合が非常に高いと仮定する。したがって、半教師ありの場合と全く同じように進めることができる。

3.3. グループ画像の異常検出

グループ異常検知は、入力サンプルが画像の集合からなる設定に取り組む。特定の組み合わせは重要だが、順番は重要ではない。集合内の各画像は個々には正常であるが、集合全体としては異常である可能性がある。例として、M個の画像からなる正常な集合を想定しよう。点（画像単位）の異常検出器を学習させた場合、点的に異常な画像を含む異常集合を検出することができる。しかし、あるクラスからの複数の画像を含み、別のクラスからの画像を含まない異常集合は、全ての画像が個々に正常であるため、正常集合として分類される。以前、画像におけるグループ異常検出に取り組むために、いくつかのディープオートエンコーダ手法が提案された（例えばDOroら（2019））。このような手法は、i）高いサンプル複雑性 ii）再構成メトリックに対する感度 iii）グループに対する感度の潜在的欠如、という複数の欠点に苦しんでいる。我々は効果的なkNNベースのアプローチを提案する。提案手法は、集合内の画像の全ての特徴に対して無秩序にプール（我々は平均化を選んだ）することで集合を埋め込む：

1. グループg内の全画像から特徴抽出 f_g^i = F（x ig ）# 4.1.
2. グループ全体にわたる特徴の無秩序なプール：
   $f_g = \frac{\sum_i f_g^i}{number\ of\ images}$

上述のグループ特徴を抽出した後、DN2を用いて異常の検出に進む。

4. Experiments

このセクションでは、上述の単純なkNNアプローチが、最先端の性能よりも優れた性能を達成することを示す、広範な実験を紹介する。この結論は、タスクやデータセットを超えて一般化される。我々はこの手法をノイズに対してより頑健に拡張し、教師なし設定に適用できるようにする。さらに、この方法をグループ異常検出に有効なように拡張する。

4.1. 単峰性異常検出

異常検知手法を評価するための最も一般的な設定は単峰性である。この設定では，1つのクラスを正常とし，他のクラスを異常とすることで，分類データセットが適応される．正常な訓練セットは手法の訓練に使用され、全てのテストデータは手法の推論性能を評価するために使用される。先行研究と同様に、ROC曲線下面積（ROCAUC）を報告する。

3.1. Semi-supervised Anomaly Detection

DN2 takes a set of input images $X_{train} = x_1, x_2..x_N$. In the semi-supervised setting we assume that all input images are normal. DN2 uses a pre-trained feature extractor F to extract features from the entire training set:

$$f_i = F(x_i) \tag{1}$$

In this paper, we use a ResNet feature extractor that was pretrained on the Imagenet dataset. At first sight it might appear that this supervision is a strong requirement, however such feature extractors are widely available. We will later show experimentally that the normal or anomalous images do not need to be particularly closely related to Imagenet.

The training set is now summarized as a set of embeddings $F_{train} = f_1, f_2..f_N$. After the initial stage, the embeddings can be stored, amortizing the inference of the training set.

To infer if a new sample y is anomalous, we first extract its feature embedding: $f_y = F(y)$. We then compute its kNN distance and use it as the anomaly score:

$$d(y) = \frac{1}{k} \sum_{f \in N_k(f_y)} \|f - f_y\|^2 \tag{2}$$

$N_k(f_y)$ denotes the k nearest embeddings to f_y in the training set F_{train}. We elected to use the euclidean distance, which often achieves strong results on features extracted by deep networks, but other distance measures can be used in a similar way. By verifying if the distance $d(y)$ is larger than a threshold, we determine if an image y is normal or anomalous.

3.2. Unsupervised Anomaly Detection

In the fully-unsupervised case, we can no longer assume that all input images are normal, instead, we assume that only a small proportion of input images are anomalous. To deal with this more difficult setting (and inline with previous works on unsupervised anomaly detection), we propose to first conduct a cleaning stage on the input images. After the feature extraction stage, we compute the kNN distance between each input image and the rest of the input images. Assuming that anomalous images lie in low density regions, we remove a fraction of the images with the largest kNN distances. This fraction should be chosen such that it is larger than the estimated proportion of anomalous input images. It will be later shown in our experiments that DN2 requires very few training images. We can therefore be very aggressive in the percentage of removed image, and keep only the images most likely to be normal (in practice we remove 50% of training images). After removal of the suspected anomalous input images, the images are now assumed to have a very high-proportion of normal images. We can therefore proceed exactly as in the semi-supervised case.

3.3. Group Image Anomaly Detection

Group anomaly detection tackles the setting where the input sample consists of a set of images. The particular combination is important, but not the order. It is possible that each image in the set will individually be normal but the set as a whole will be anomalous. As an example, let us assume normal sets consisting of M images, a randomly sampled image from each class. If we trained a point (per-image) anomaly detector, it will be able to detect anomalous sets containing pointwise anomalous images e.g. images taken from classes not seen in training. An anomalous set containing multiple images from one seen class, and no images from another will however be classified as normal as all images are individually normal. Previously, several deep autoencoder methods were proposed (e.g. DOro et al. (2019)) to tackle group anomaly detection in images. Such methods suffer from multiple drawbacks: i) high sample complexity ii) sensitivity to reconstruction metric iii) potential lack of sensitivity to the groups. We propose an effective kNN based approach. The proposed method embeds the set by orderless-pooling (we chose averaging) over all the features of the images in the set:

1. Feature extraction from all images in the group g, $f_g^i = F(x_g^i)$
2. Orderless pooling of features across the group: $f_g = \frac{\sum_i f_g^i}{number\ of\ images}$

Having extracted the group feature described above we proceed to detect anomalies using DN2.

4. Experiments

In this section, we present extensive experiments showing that the simple kNN approach described above achieves better than state-of-the-art performance. The conclusions generalize across tasks and datasets. We extend this method to be more robust to noise, making it applicable to the unsupervised setting. We further extend this method to be effective for group anomaly detection.

4.1. Unimodal Anomaly Detection

The most common setting for evaluating anomaly detection methods is unimodal. In this setting, a classification dataset is adapted by designating one class as normal, while the other classes as anomalies. The normal training set is used to train the method, all the test data are used to evaluate the inference performance of the method. In line with previous works, we report the ROC area under the curve (ROCAUC).

表1. Cifar10 における異常検知精度（ROCAUC %）

	OC-SVM	Deep SVDD	GEOM	GOAD	MHRot	DN2
0	70.6	61.7 ± 1.3	74.7 ± 0.4	77.2 ± 0.6	77.5	**93.9**
1	51.3	65.9 ± 0.7	95.7 ± 0.0	96.7 ± 0.2	96.9	**97.7**
2	69.1	50.8 ± 0.3	78.1 ± 0.4	83.3 ± 1.4	**87.3**	85.5
3	52.4	59.1 ± 0.4	72.4 ± 0.5	77.7 ± 0.7	80.9	**85.5**
4	77.3	60.9 ± 0.3	87.8 ± 0.2	87.8 ± 0.7	92.7	**93.6**
5	51.2	65.7 ± 0.8	87.8 ± 0.1	87.8 ± 0.6	90.2	**91.3**
6	74.1	67.7 ± 0.8	83.4 ± 0.5	90.0 ± 0.6	90.9	**94.3**
7	52.6	67.3 ± 0.3	95.5 ± 0.1	96.1 ± 0.3	**96.5**	93.6
8	70.9	75.9 ± 0.4	93.3 ± 0.0	93.8 ± 0.9	95.2	**95.1**
9	50.6	73.1 ± 0.4	91.3 ± 0.1	92.0 ± 0.6	93.3	**95.3**
Avg	62.0	64.8	86.0	88.2	90.1	**92.5**

表2. ファッションMNISTとCIFAR10における異常検出精度（ROCAUC %）

	OC-SVM	GEOM	GOAD	DN2
FashionMNIST	92.8	93.5	94.1	**94.4**
CIFAR100	62.6	78.7	-	**89.3**

我々は最先端の手法、OCSVMと深層特徴学習を組み合わせたdeep-SVDD（Ruff et al. Geometric（Golan & El-Yaniv, 2018)、GOAD（Bergman & Hoshen, 2020)、Multi-Head RotNet（MHRot）(Hendrycks et al., 2019)。後者の3つはすべてRotNetのバリエーションを使用している。

DN2を除くすべての手法については、利用可能であれば原著論文の結果を報告した。Geometric（Golan & El-Yaniv, 2018)とmulti-head RotNet（MHRot）(Hendrycks et al., 2019)の場合、著者から報告されていないデータセットについては、低解像度の実験ではGeometricのコードリリースを実行し、高解像度の実験ではMHRotを実行した（低解像度の実験ではコードがリリースされていないため）。

Cifar10: これはユニモーダル異常検出を評価するための最も一般的なデータセットである。CIFAR10は10のオブジェクトクラスからなる32×32のカラー画像を含む。各クラスには5000枚のトレーニング画像と1000枚のテスト画像がある。結果を表1に示す。DN2の性能は、与えられた訓練セットとテストセットに対して決定論的である（実行間の変動はない）ことに注意。OC-SVMとDeep-SVDDの性能が最も低いことがわかる。これは、Deep-SVDDによって学習された特徴量だけでなく、生のピクセルも、正規分布の中心への距離が成功するのに十分な識別力がないためである。幾何学的アプローチとそれ以降のアプローチGOADとMHRotは、かなり良いパフォーマンスを示すが、90%のROCAUCを超えない。DN2は他のすべての手法より有意に優れている。

本論文では、データセットとシミュレートされた異常値（DN2を含む全ての手法で性能を向上させる）の間の微調整を行わない場合の性能を評価することにする。外れ値露出は、そのような微調整のための1つの手法である。単体ではトップクラスの性能は得られないが、MHRotと組み合わせることで改善し、CIFAR10において平均95.8%のROCAUCを達成した。この手法や他のアンサンブル手法もDN2の性能を向上させることができるが、本稿の範囲外である。

ファッションMNIST： クラスあたり6000枚のトレーニング画像とクラスあたり1000枚のテスト画像からなるファッションMNISTデータセットでGeometric、GOAD、DN2を評価する。DN2とOCSVM、Deep SVDD、Geometric、GOADの比較を示す。特徴量を抽出したImagenetとは視覚的にかなり異なるデータであるにも関わらず、DN2が他の全ての手法を凌駕していることがわかる。

CIFAR100: Geometric、GOAD、DN2 を CIFAR100 データセットで評価する。CIFAR100には、それぞれ500枚の訓練画像からなる100の細粒度クラス、またはそれぞれ2500枚の訓練画像からなる20の粗粒度クラスがある。先行論文に従い、我々は粗視化クラスを用いる。プロトコルはCIFAR10と同じである。DN2とOCSVM、Deep SVDD、Geometric、GOADの比較を行う。結果はCIFAR10で得られたものとほぼ同じです。

MHRotとの比較：

DN2とMHRot（Hendrycks et al. この実験は、RotNetベースの手法が低解像度や回転に対する画像の不変性によって制限されないデータセットにおいて、DN2の一般性をさらに証明するものである。

最初の20カテゴリー（20未満の場合は全カテゴリー）それぞれについて、アルファベット順にROCAUCスコアを計算する。標準的な訓練とテストの分割が使用される。からのすべてのテスト画像。

Table 1. Anomaly Detection Accuracy on Cifar10 (ROCAUC %)

	OC-SVM	Deep SVDD	GEOM	GOAD	MHRot	DN2
0	70.6	61.7 ± 1.3	74.7 ± 0.4	77.2 ± 0.6	77.5	**93.9**
1	51.3	65.9 ± 0.7	95.7 ± 0.0	96.7 ± 0.2	96.9	**97.7**
2	69.1	50.8 ± 0.3	78.1 ± 0.4	83.3 ± 1.4	**87.3**	85.5
3	52.4	59.1 ± 0.4	72.4 ± 0.5	77.7 ± 0.7	80.9	**85.5**
4	77.3	60.9 ± 0.3	87.8 ± 0.2	87.8 ± 0.7	92.7	**93.6**
5	51.2	65.7 ± 0.8	87.8 ± 0.1	87.8 ± 0.6	90.2	**91.3**
6	74.1	67.7 ± 0.8	83.4 ± 0.5	90.0 ± 0.6	90.9	**94.3**
7	52.6	67.3 ± 0.3	95.5 ± 0.1	96.1 ± 0.3	**96.5**	93.6
8	70.9	75.9 ± 0.4	93.3 ± 0.0	93.8 ± 0.9	95.2	**95.1**
9	50.6	73.1 ± 0.4	91.3 ± 0.1	92.0 ± 0.6	93.3	**95.3**
Avg	62.0	64.8	86.0	88.2	90.1	**92.5**

Table 2. Anomaly Detection Accuracy on Fashion MNIST and CIFAR10 (ROCAUC %)

	OC-SVM	GEOM	GOAD	DN2
FashionMNIST	92.8	93.5	94.1	**94.4**
CIFAR100	62.6	78.7	-	**89.3**

We conduct experiments against state-of-the-art methods, deep-SVDD (Ruff et al., 2018) which combines OCSVM with deep feature learning. Geometric (Golan & El-Yaniv, 2018), GOAD (Bergman & Hoshen, 2020), Multi-Head RotNet (MHRot) (Hendrycks et al., 2019). The latter three all use variations of RotNet.

For all methods except DN2, we reported the results from the original papers if available. In the case of Geometric (Golan & El-Yaniv, 2018) and the multi-head RotNet (MHRot) (Hendrycks et al., 2019), for datasets that were not reported by the authors, we run the Geometric code-release for low-resolution experiments, and MHRot for high-resolution experiments (as no code was released for the low-resolution experiments).

Cifar10: This is the most common dataset for evaluating unimodal anomaly detection. CIFAR10 contains 32×32 color images from 10 object classes. Each class has 5000 training images and 1000 test images. The results are presented in Tab. 1, note that the performance of DN2 is deterministic for a given train and test set (no variation between runs). We can observe that OC-SVM and Deep-SVDD are the weakest performers. This is because both the raw pixels as well as features learned by Deep-SVDD are not discriminative enough for the distance to the center of the normal distribution to be successful. Geometric and later approaches GOAD and MHRot perform fairly well but do not exceed 90% ROCAUC. DN2 significantly outperforms all other methods.

In this paper, we choose to evaluate the performance of without finetuning between the dataset and simulated anomalies (which improves performance on all methods including DN2). Outlier Exposure is one technique for such finetuning. Although it does not achieve the top performance by itself, it reported improvements when combined with MHRot to achieve an average ROCAUC of 95.8% on CIFAR10. This and other ensembling methods can also improve the performance of DN2 but are out-of-scope of this paper.

Fashion MNIST: We evaluate Geometric, GOAD and DN2 on the Fashion MNIST dataset consisting of 6000 training images per class and a test set of 1000 images per class. We present a comparison of DN2 vs. OCSVM, Deep SVDD, Geometric and GOAD. We can see that DN2 outperforms all other methods, despite the data being visually quite different from Imagenet from which the features were extracted.

CIFAR100: We evaluate Geometric, GOAD and DN2 on the CIFAR100 dataset. CIFAR100 has 100 fine-grained classes with 500 train images each or 20 coarse-grained classes with 2500 train images each. Following previous papers, we use the coarse-grained version. The protocol is the same as CIFAR10. We present a comparison of DN2 vs. OCSVM, Deep SVDD, Geometric and GOAD. The results are inline with those obtained for CIFAR10.

Comparisons against MHRot:

We present a further comparison between DN2 and MHRot (Hendrycks et al., 2019) on several commonly-used datasets. The experiments give further evidence for the generality of DN2, in datasets where RotNet-based methods are not restricted by low-resolution, or by image invariance to rotations.

We compute the ROCAUC score on each of the first 20 categories (all categories if there are less than 20), by alphabetical order, designated as normal for training. The standard train and test splits are used. All test images from

表3. 花、鳥、猫対犬におけるMHRotとDN2の比較（平均クラスROC AUC %）

Dataset	MHRot	DN2
Oxford Flowers	65.9	**93.9**
UCSD Birds 200	64.4	**95.2**
CatsVsDogs	88.5	**97.5**

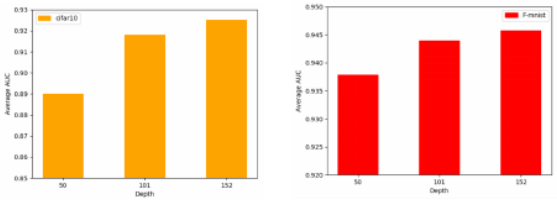


図1. ネットワークの深さ（ResNetの層数）は、Cifar10とFashionMNISTの結果を向上させた。

すべてのクラスが推論に使用され、適切なクラスが正常とされ、残りはすべて異常とされる。表示を簡潔にするため、テストされたクラスの平均ROCAUCスコアが報告されている。

102 Category Flowers (Nilsback & Zisserman, 2008)： このデータセットは102カテゴリの花からなり、それぞれ10枚の学習画像から構成される。テストセットは1クラスあたり30枚から200枚以上の画像で構成される。

Caltech-UCSD Birds 200 (Wah et al., 2011)： このデータセットは、鳥類の200のカテゴリーからなる。クラスは通常55から60の画像を含み、trainとtestで均等に分割される。

CatsVsDogs (Elson et al., 2007)： このデータセットは、犬と猫の2つのカテゴリから構成され、それぞれ10,000枚のトレーニング画像がある。テストセットは各クラス2,500枚の画像から構成される。各画像には、様々なシーンで様々な角度から撮影された犬か猫が含まれている。データはASIRRAデータセットから抽出され、各クラスを最初の10,000画像をトレーニング、最後の2,500画像をテストとして分割した。

結果を表3に示す。3. DN2は全てのデータセットにおいてMHRotを大きく上回る。

ネットワークの深さの効果：

Imagenetのような大規模データセットで訓練された深いネットワークは、浅いネットワークよりも汎化する特徴を学習する。異なる深さのネットワークからの特徴を使用した場合のDN2の性能を調査した。具体的には、50層、101層、152層のResNetの平均ROCAUCを図1にプロットする。DN2はすべてのネットワークでうまく機能するが、ネットワークの深さが深くなるほど性能は向上する。

近傍探索数の効果

The only free parameter in DN2 is the number of neigh-

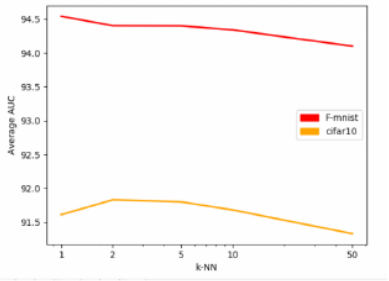
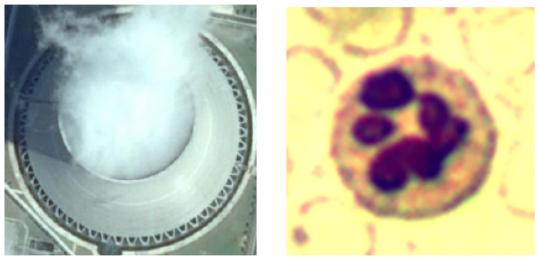


図2. 最適なKの数は約2である。



左）DIORデータセットからの煙突画像 （右）WBCデータセットからの画像。

kNNで使用されるボル 異なる最近傍数に対するCIFAR10とFashionMNISTの平均ROCAUCの比較を図2に示す。差は特に大きくないが、通常2近傍が最良である。

データ不変性の効果

幾何学的変換の予測に依存する手法、例えば（Golan & El-Yaniv, 2018; Hendrycks et al., 2019; Bergman & Hoshen, 2020）は、画像が（回転予測のための）所定の方向と（並進予測のための）中心を持つという強力なデータ事前分布を用いる。この仮定は実際の画像ではしばしば誤りである。この仮定を満たさない2つの興味深いケースは、空撮画像と顕微鏡画像で、これらは好ましい方向を持たないため、回転予測は効果がない。

DIOR (Li et al., 2020)： 航空画像データセット。画像は登録されているが、好ましい方向はない。このデータセットは、120 × 120以上の解像度を持つ50枚以上の画像を持つ19のオブジェクト・カテゴリから構成される（1クラスあたりの画像数の中央値は578枚）。バウンディング

表4. DIORとWBCの異常検出精度（ROCAUC）

Dataset	MHRot	DN2
DIOR	83.2	**92.2**
WBC	60.5	**82.9**

Table 3. MHRot vs. DN2 on Flowers, Birds, CatsVsDogs (Average Class ROCAUC %)

Dataset	MHRot	DN2
Oxford Flowers	65.9	**93.9**
UCSD Birds 200	64.4	**95.2**
CatsVsDogs	88.5	**97.5**

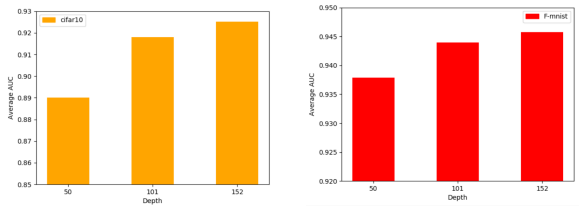


Figure 1. Network depth (number of ResNet layers) improves both Cifar10 and FashionMNIST results.

all classes are used for inference, with the appropriate class designated normal and all the rest as anomalies. For brevity of presentation, the average ROCAUC score of the tested classes is reported.

102 *Category Flowers (Nilsback & Zisserman, 2008):* This dataset consists of 102 categories of flowers, consisting of 10 training images each. The test set consists of 30 to over 200 images per-class.

Caltech-UCSD Birds 200 *(Wah et al., 2011):* This dataset consists of 200 categories of bird species. Classes typically contain between 55 to 60 images split evenly between train and test.

CatsVsDogs (Elson et al., 2007): This dataset consists of 2 categories; dogs and cats with $10,000$ training images each. The test set consist of $2,500$ images for each class. Each image contains either a dog or a cat in various scenes and taken from different angles. The data was extracted from the ASIRRA dataset, we split each class to the first $10,000$ images as train and the last $2,500$ as test.

The results are shown in Tab. 3. DN2 significantly outperforms MHRot on all datasets.

Effect of network depth:

Deeper networks trained on large datasets such as Imagenet learn features that generalize better than shallow network. We investigated the performance of DN2 when using features from networks of different depths. Specifically, we plot the average ROCAUC for ResNet with 50, 101, 152 layers in Fig. 1. DN2 works well with all networks but performance is improved with greater network depth.

Effect of the number of neighbors:

The only free parameter in DN2 is the number of neighbors used in kNN. We present in Fig. 2, a comparison of average CIFAR10 and FashionMNIST ROCAUC for different numbers of nearest neighbors. The differences are not particularly large, but 2 neighbors are usually best.

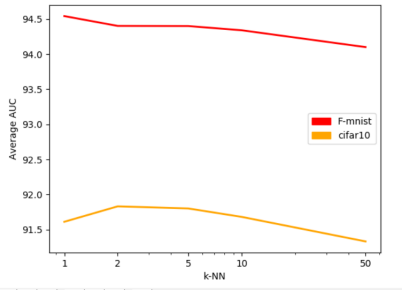


Figure 2. Number of neighbors vs ROCAUC, the optimal number of K is around 2.

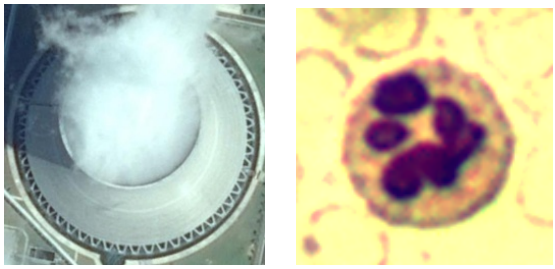


Figure 3. (left) A chimney image from the DIOR dataset (right) An image from the WBC Dataset.

Effect of data invariance:

Methods that rely on predicting geometric transformations e.g. (Golan & El-Yaniv, 2018; Hendrycks et al., 2019; Bergman & Hoshen, 2020), use a strong data prior that images have a predetermined orientation (for rotation prediction) and centering (for translation prediction). This assumption is often false for real images. Two interesting cases not satisfying this assumption, are aerial and microscope images, as they do not have a preferred orientation, making rotation prediction ineffective.

DIOR (Li et al., 2020): An aerial image dataset. The images are registered but do not have a preferred orientation. The dataset consists of 19 object categories that have more than 50 images with resolution above 120×120 (the median number of images per-class is 578). We use the bounding

Table 4. Anomaly Detection Accuracy on DIOR and WBC (ROCAUC %)

Dataset	MHRot	DN2
DIOR	83.2	**92.2**
WBC	60.5	**82.9**

表5. マルチモーダル正規画像分布における異常検出精度（ROCAUC %） # 表5.

Dataset	Geometric	DN2
CIFAR10	61.7	**71.7**
CIFAR100	57.3	**71.0**

各オブジェクトのバウンディングボックスは各軸に少なくとも120ピクセルある。それを256×256ピクセルにリサイズする。以前のデータセットと同じプロトコルに従う。画像は高解像度であるため、自己教師付きベースラインとしてHendrycks（Hendrycks et al. 結果をTab. 4. DN2がMHRotを大幅に上回っていることがわかる。これは、特徴抽出器の性能が一般的に強いことと、RotNetタイプの手法で強く使われる回転事前分布がないことの両方によるものである。画像はセンタリングされていることに注意してください。

WBC（Zheng et al., 2018）： 困難な実世界データでの性能をさらに調査するために、我々は、白血球の異なるカテゴリの高解像度顕微鏡画像からなるWBC画像データセットで実験を行った。このデータには好ましい方向がない。さらにこのデータセットは非常に小さく、1クラスあたり数十枚の画像しかない。我々は、中国のJiangxi Telecom Science Corporationから入手したデータセット1を使用し、それぞれ20枚以上の画像を含む4つの異なるクラスに分割する。各クラスの最初の80%の画像を学習セットに、最後の20%をテストセットに設定する。結果を表4に示す。4. 予想通り、DN2はMHRotを大差で上回り、実データへの適用性の高さを示している。

4.2. マルチモーダル異常検知

実際には正規分布には複数のクラスが含まれるため，単峰性の異常検出は現実的でないと主張されている（Ahmed & Courville（2019）など）。我々は、両方の設定が実際に発生すると考えているが、異常とみなされる1つのクラス（例えば、「Cat」を除くすべてのCIFAR10クラスが正常である）を除いて、すべてのクラスが正常であると指定されるシナリオの結果も提示する。我々は正常クラスを構成する異なるクラスのクラスラベルを提供せず、むしろそれらを1つのマルチモーダルなクラスとみなしていることに注意してください。これは、多くの異なるラベルのないタイプのデータからなる複雑な正常クラスを持つ現実的なケースをシミュレートしていると考えている。

CIFAR10とCIFAR100において、DN2とGeometricを比較した。全てのクラスにおける平均ROCAUCを表5に示す。5. DN2はGeometricよりも大幅に高い性能を達成しています。

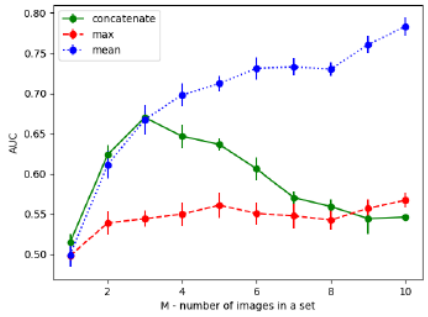


図4. 図4.グループ毎の画像数と検出ROCAUCの比較。平均プーリングを用いたグループ異常検出は、3枚以上の画像を持つグループに対して、単純な特徴連結よりも優れている。

これはGeometricが異常データに対してネットワークを汎化させないことを要求しているためであると考えられる。しかし、学習データが十分に変化すると、ネットワークは未見のクラスでも汎化できるようになり、この手法が有効でなくなる。これは特にCIFAR100で顕著である。

4.3. 小さな訓練データセットからの汎化

正常データセットでの学習を利用しないDN2の利点の1つは、非常に小さなデータセットからの汎化能力である。これは自己教師あり学習ベースの手法では不可能であり、通常のテスト画像に汎化するのに十分な一般的特徴を学習しない。CIFAR10におけるDN2とGeometricの比較を図5に示す。学習画像数と平均ROCAUCをプロットした。DN2は10枚の画像からでも非常に正確に異常を検出できるのに対し、Geometricは学習画像の枚数が減るにつれて急速に悪化することがわかる。同様のプロットをFashionMNISTについても図5に示す。Geometricは画像数が少ないと数値的な問題に悩まされるため示していない。DN2はまたもや非常に少ない画像から強力な性能を達成した。

4.4. 教師なし異常検知

学習セットが純粋な正常画像から構成されるのではなく、ラベル付けされていない正常画像と異常画像が混在する設定もある。その代わりに、異常画像は正常画像のごく一部であると仮定する。学習セット中の異常画像の割合に対するDN2の性能を図5に示す。訓練セットの不純物の割合が存在するにつれて、性能はやや低下する。性能を向上させるために、我々はクリーニングステージを提案し、トレーニングセット内の最も離れたkN Nを持つトレーニングセット画像の50%を除去する。その後、通常通りDN2を実行する。性能は図5にも示されている。我々のクリーニング手順により、DN2が大幅に改善されたことがわかる。

Table 5. Anomaly Detection Accuracy on Multimodal Normal Image Distributions (ROCAUC %)

Dataset	Geometric	DN2
CIFAR10	61.7	**71.7**
CIFAR100	57.3	**71.0**

boxes provided with the data, and take each object with a bounding box of at least 120 pixels in each axis. We resize it to 256×256 pixels. We follow the same protocol as in the earlier datasets. As the images are of high-resolution, we use the public code release of Hendrycks (Hendrycks et al., 2018) as a self-supervised baseline. The results are summarized in Tab. 4. We can see that DN2 significantly outperforms MHRot. This is due both to the generally stronger performance of the feature extractor as well as the lack of rotational prior that is strongly used by RotNet-type methods. Note that the images are centered, a prior used by the MHRot translation heads.

WBC (Zheng et al., 2018): To further investigate the performance on difficult real world data, we performed an experiment on the WBC Image Dataset, which consists of high-resolution microscope images of different categories of white blood cells. The data do not have a preferred orientation. Additionally the dataset is very small, only a few tens of images per-class. We use Dataset 1 that was obtained from Jiangxi Telecom Science Corporation, China, and split it to the 4 different classes that contain more than 20 images each. We set the first 80% images in each class to the train set, and the last 20% to the test set. The results are presented in Tab. 4. As expected, DN2 outperforms MHRot by a significant margin showing its greater applicability to real world data.

4.2. Multimodal Anomaly Detection

It has been argued (e.g. Ahmed & Courville (2019)) that unimodal anomaly detection is less realistic as in practice, normal distributions contain multiple classes. While we believe that both settings occur in practice, we also present results on the scenario where all classes are designated as normal apart from a single class that is taken as anomalous (e.g. all CIFAR10 classes are normal apart from "Cat"). Note that we do not provide the class labels of the different classes that compose the normal class, rather we consider them to be a single multimodal class. We believe this simulates the realistic case of having a complex normal class consisting of many different unlabelled types of data.

We compared DN2 against Geometric on CIFAR10 and CIFAR100 on this setting. We provide the average ROCAUC across all the classes in Tab. 5. DN2 achieves significantly stronger performance than Geometric. We believe this is occurs as Geometric requires the network not to generalize on the anomalous data. However, once the training data is sufficiently varied the network can generalize even on unseen classes, making the method less effective. This is particularly evident on CIFAR100.

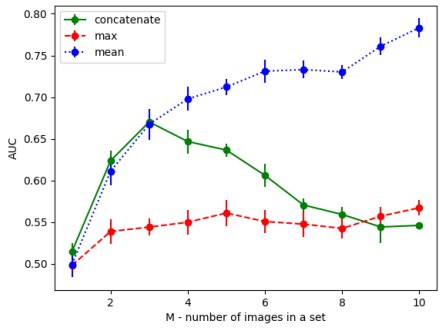


Figure 4. Number of images per group vs. detection ROCAUC. Group anomaly detection with mean pooling is better than simple feature concatenation for groups with more than 3 images.

4.3. Generalization from Small Training Datasets

One of the advantage of DN2, which does not utilize learning on the normal dataset is its ability to generalize from very small datasets. This is not possible with self-supervised learning-based methods, which do not learn general enough features to generalize to normal test images. A comparison between DN2 and Geometric on CIFAR10 is presented in Fig. 5. We plotted the number of training images vs. average ROCAUC. We can see that DN2 can detect anomalies very accurately even from 10 images, while Geometric deteriorates quickly with decreasing number of training images. We also present a similar plot for FashionMNIST in Fig. 5. Geometric is not shown as it suffered from numerical issues for small numbers of images. DN2 again achieved strong performance from very few images.

4.4. Unsupervised Anomaly Detection

There are settings where the training set does not consist of purely normal images, but rather a mixture of unlabelled normal and anomalous images. Instead we assume that anomalous images are only a small fraction of the number of the normal images. The performance of DN2 as function of the percentage of anomalies in the training set is presented in Fig. 5. The performance is somewhat degraded as the percentage of training set impurities exist. To improve the performance, we proposed a cleaning stage, which removes 50% of the training set images that have the most distant kNN inside the training set. We then run DN2 as usual. The performance is also presented in Fig. 5. Our cleaning procedure is clearly shown to significantly improve

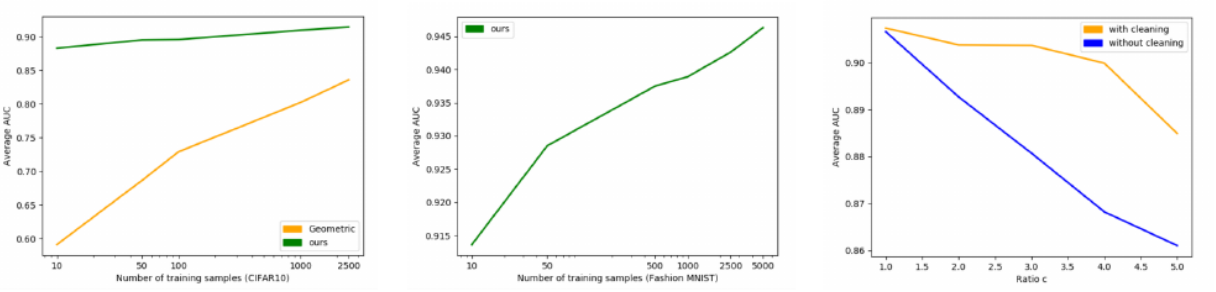


図5. 学習画像数とROCAUCの比較（左）CIFAR10 DN2は10枚の画像からでも高い精度を達成しているが、Geometricは致命的に悪化している。（中央）FashionMNISTも同様にDN2により高い性能を達成。（右）CIFAR10における不純物比率とROCAUCの比較。トレーニングセットのクリーニングにより、性能が大幅に向上。

不純物の割合による性能劣化

4.5. グループ異常検知

既存のベースラインと比較するために、まずDOro et al. (2019)のタスクで我々の手法をテストした。データは同じ桁のMNIST画像10～50枚を含む正常セットと、異なる桁の画像10～50枚を含む異常セットからなる。各画像セットの画像毎のResNet特徴量の共分散行列のトレース対角線を計算するだけで、前論文の0.81に対して0.92のROCAUCを達成した（学習セットを全く使用せず）。

順不同の画像集合におけるグループ異常検出の困難なタスクとして、CIFAR10のM個のクラス（具体的にはID 0..M − 1のクラス）のそれぞれからちょうど1つの画像からなる集合を正常クラスとし、各異常集合は同じクラスの中からランダムに選択されたM個の画像から構成される（クラスによっては1つ以上の画像を持つものもあれば、ゼロのものもある）。単純なベースラインとして、セット内の個々の画像の連結特徴量に対するDN2を用いた異常検出の平均ROCAUC（図4.2）を報告する。予想されるように、このベースラインは、クラス順序のすべての可能な順列の十分な例を持っているMの小さな値ではうまく機能するが、Mが大きくなるにつれて（M > 3）、順列の数が指数関数的に成長するため、その性能は低下する。学習用に1000の画像セットを用いて、この方法を順序なしの最大プール特徴量と平均プール特徴量の最近傍と比較すると、Mが大きい値では平均プール特徴量がベースラインを大きく上回ることがわかる。訓練セットのすべての可能な順序でデータセットを補強することによって、連結された特徴のパフォーマンスを向上させることができますが、Mの数が自明でない場合は指数関数的に成長するため、効果的なアプローチではありません。

4.6. Implementation

DN2のすべてのインスタンスでは、まず入力画像を256 × 256にリサイズし、サイズ224 × 224の中央のクロップを取る。

表6.

C=1	C=3	C=5	C=10	kNN
91.94	92.00	91.87	91.64	92.52

Imagenetで事前に訓練されたResNet（特に指定がない限り101層）を使って、グローバルプーリング層の直後の特徴を抽出する。この特徴は画像埋め込みである。

5. Analysis

このセクションでは、kNNと他の分類手法との比較、および事前学習されたネットワークによって抽出された特徴と自己教師あり手法によって学習された特徴との比較によって、DN2の分析を行う。

kNN vs. 1クラス分類

我々の実験では、kNNは異常検知タスクで非常に強力な性能を達成することがわかった。この強力な性能の理由をより良く理解するために、試行錯誤してみよう。図6に、CIFAR10のテストセットの特徴量のt-SNEプロットを示す。正常クラスは黄色で表示され、異常データは青色で表示されています。事前学習された特徴量は、同じクラスの画像をかなりコンパクトな領域に埋め込んでいることがわかります。したがって、正常なテスト画像の周辺では、異常なテスト画像の周辺よりも正常な学習画像の密度がはるかに高くなることが予想される。これがkNN法の成功の原因である。

kNNは訓練データサンプル数に対して線形な複雑性を持つ。One-Class SVMやSVDDのような手法は、単一の超球を学習し、超球の中心までの距離を異常の尺度として使用しようとする。この場合、推論の実行時間はkNNの場合のように線形ではなく、訓練セットのサイズに対して一定である。欠点は典型的な性能の低さである。推論時間を短縮するもう一つの一般的な方法（Fukunaga & Narendra, 1975）。

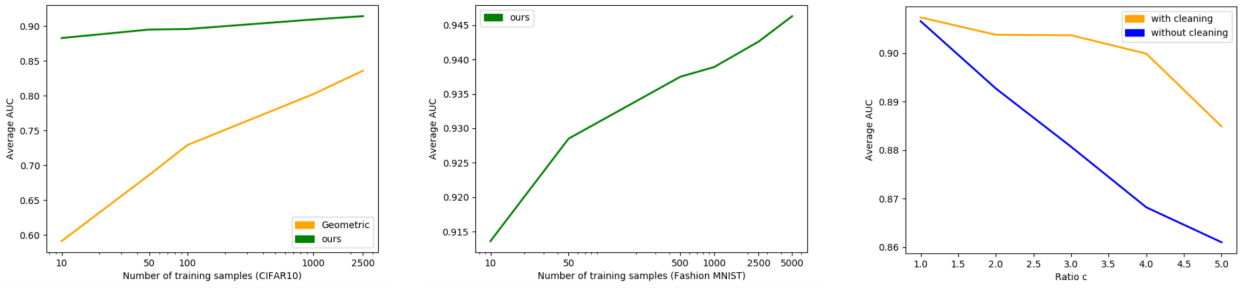


Figure 5. Number of training images vs. ROCAUC (left) CIFAR10 - Strong perfromance is achieved by DN2 even from 10 images, whereas Geometric deteriorates critically. (center) FashionMNIST - similarly strong performance by DN2. (right) Impurity ratio vs ROCAUC on CIFAR10. The training set cleaning procedure, significantly improves performance.

the performance degradation as percentage of impurities.

4.5. Group Anomaly Detection

To compare to existing baselines, we first tested our method on the task in DOro et al. (2019). The data consists of normal sets containing $10-50$ MNIST images of the same digit, and anomalous sets containing $10-50$ images of different digits. By simply computing the trace-diagonal of the covariance matrix of the per-image ResNet features in each set of images, we achieved 0.92 ROCAUC vs. 0.81 in the previous paper (without using the training set at all).

As a harder task for group anomaly detection in unordered image sets, we designate the normal class as sets consisting of exactly one image from each of the M CIFAR10 classes (specifically the classes with ID $0..M-1$) while each anomalous set consisted of M images selected randomly among the same classes (some classes had more than one image and some had zero). As a simple baseline, we report the average ROCAUC (Fig, 4.2) for anomaly detection using DN2 on the concatenated features of each individual image in the set. As expected, this baseline works well for small values of M where we have enough examples of all possible permutations of the class ordering, but as M grows larger ($M > 3$), its performance decreases, as the number permutations grows exponentially. We compare this method, with 1000 image sets for training, to nearest neighbours of the orderless max-pooled and average-pooled features, and see that mean-pooling significantly outperforms the baseline for large values of M. While we may improve the performance of the concatenated features by augmenting the dataset with all possible orderings of the training sets, it is will grow exponentially for a non-trivial number of M making it an ineffective approach.

4.6. Implementation

In all instances of DN2, we first resize the input image to 256×256, we take the center crop of size 224×224, and using an Imagenet pre-trained ResNet (101 layers unless otherwise specified) extract the features just after the global pooling layer. This feature is the image embedding.

Table 6. Accuracy on CIFAR10 using K-means approximations and full kNN (ROCAUC %)

C=1	C=3	C=5	C=10	kNN
91.94	92.00	91.87	91.64	92.52

5. Analysis

In this section, we perform an analysis of DN2, both by comparing kNN to other classification methods, as well as comparing the features extracted by the pretrained networks vs. features learned by self-supervised methods.

5.1. kNN vs. one-class classification

In our experiments, we found that kNN achieved very strong performance for anomaly detection tasks. Let us try to gain a better understanding of the reasons for the strong performance. In Fig. 6 we can observe t-SNE plots of the test set features of CIFAR10. The normal class is colored in yellow while the anomlous data is marked in blue. It is clear that the pre-trained features embed images from the same class into a fairly compact region. We therefore expect the density of normal training images to be much higher around normal test images than around anomalous test images. This is responsible for the success of kNN methods.

kNN has linear complexity in the number of training data samples. Methods such as One-Class SVM or SVDD attempt to learn a single hypersphere, and use the distance to the center of the hypersphere as a measure of anomaly. In this case the inference runtime is constant in the size of the training set, rather than linear as in the kNN case. The drawback is the typical lower performance. Another popular way (Fukunaga & Narendra, 1975) of decreasing the inference

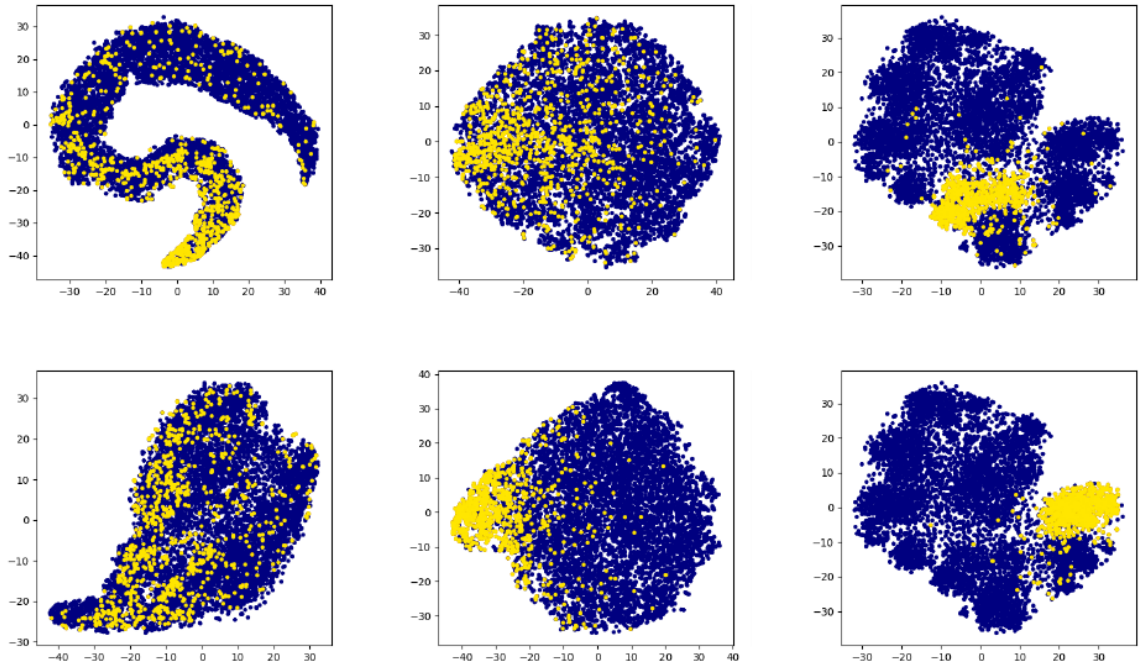


図6. CIFAR10でSVDD（左）、Geometric（中央）、Imagenet事前学習（右）により学習された特徴量のt-SNEプロット（通常クラスはAirplane（上）、Automobile（下））。Imagenetで事前学習された特徴量は、正常クラス（黄色）と異常クラス（青色）を明確に分離していることがわかります。Geometricは、Airplaneについては不十分な特徴を学習し、Automobileについては妥当な特徴を学習する。Deep-SVDDはきれいな分離を可能にする特徴を学習しない。

K}^{N}の比率で推論が高速化される。これは$\frac{N}{K}$の比率で推論を高速化する。従って、学習特徴量をK個のクラスタにクラスタリングし、元の特徴量ではなくクラスタに対してkNNを実行することで、DN2を高速化することを提案する。表6にDN2の性能比較を示す。表6は、DN2とそのK-means近似の性能を、異なる平均数（2つの最近傍への距離の和を用いる）で比較したものである。わずかな精度の低下で、検索速度が大幅に低下することがわかる。

5.2. 事前学習された特徴量と自己教師付き特徴量の比較

事前訓練された特徴抽出器によるパフォーマンスの向上を理解するために、Deep-SVDD、Geometric、DN2（Imagenet上で事前訓練されたResnet50）によって抽出された正常なテスト特徴と異常なテスト特徴のt-SNEプロットを示します。上のプロットは、中程度の検出精度を達成した正常クラスのプロットであり、下のプロットは、高い検出精度を達成した正常クラスのプロットである。Deep-SVDDの正常クラスが異常クラスの中に散在していることがすぐに観察でき、その性能の低さを説明している。Geometricでは、正常クラスの特徴はもう少し局在していますが、正常領域の密度はまだ中程度にしか集中していません。

我々は、Geometricのかなり優れた性能は、それが実行する大規模なアンサンブル（72のオーグメントの組み合わせ）によって達成されると考えている。Imagenetが事前に学習した特徴量は、非常に強い局所性を保持していることがわかる。これはDN2の強力なパフォーマンスを説明する。

6. Discussion

異常検出のための一般的なパラダイム： 最近の論文（Golan & El-Yaniv(2018)など）は、自己監視のパラダイムを提唱している。本論文の結果は、別のパラダイムに強い証拠を与える：i）曖昧に関連するデータセット上で利用可能な全ての監視を使用して一般的な特徴を学習する ii）学習された特徴は、標準的な異常検出手法（例えばkNNやk-means）を使用できるように十分に一般的であることが期待される。事前学習されたパラダイムは、自己教師ありの手法よりもはるかに高速に導入でき、その他にも多くの利点がある。Imagenetとの類似性が全くない画像データに対しては、事前に訓練された特徴量を使用することはあまり効果的でない可能性がある。しかし、我々の実験では、Imagenetで事前学習された特徴量は、顕微鏡画像だけでなく、航空写真にも有効であることがわかった。

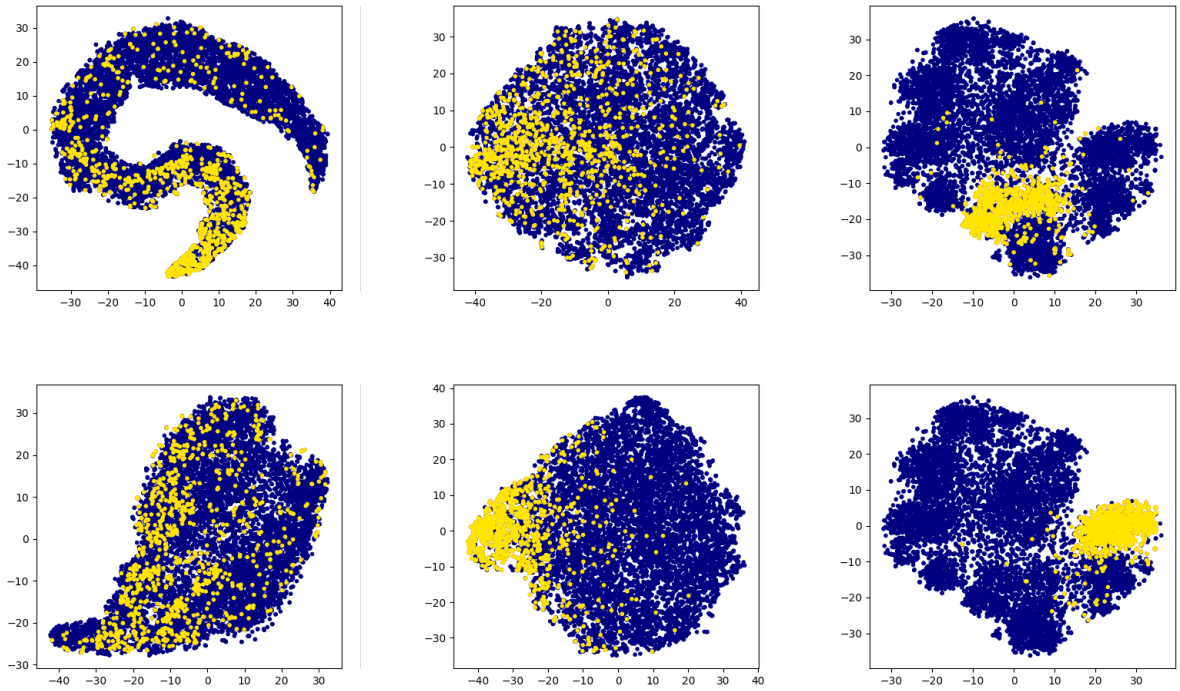


Figure 6. t-SNE plots of the features learned by SVDD (left), Geometric (center) and Imagenet pre-trained (right) on CIFAR10, where the normal class is Airplane (top), Automobile (bottom). We can see that Imagenet-pretrained features clearly separate the normal class (yellow) and anomalies (blue). Geometric learns poor features of Airplane and reasonable features on Automobile. Deep-SVDD does not learn features that allow clean separation.

time is using K-means clustering of the training features. This speeds up inference by a ratio of $\frac{N}{K}$. We therefore suggest speeding up DN2 by clustering the training features into K clusters and the performing kNN on the clusters rather than the original features. Tab. 6 presents a comparison of performance of DN2 and its K-means approximations with different numbers of means (we use the sum of the distances to the 2 nearest neighbors). We can see that for a small loss in accuracy, the retrieval speed can be reduced significantly.

5.2. Pretrained vs. self-supervised features

To understand the improvement in performance by pretrained feature extractors, we provide t-SNE plots of normal and anomalous test features extracted by Deep-SVDD, Geometric and DN2 (Resnet50 pretrained on Imagenet). The top plots are of a normal class that achieves moderate detection accuracy, while the bottom plots are of a normal class that achieves high accuracy. We can immediately observe that the normal class in Deep-SVDD is scattered among the anomalous classes, explaining its lower performance. In Geometric the features of the normal class are a little more localized, however the density of the normal region is still only moderately concentrated. We believe that the fairly good performance of Geometric is achieved by the massive ensembling that it performs (combination of 72 augmentations). We can see that Imagenet pretrained features preserve very strong locality. This explains the strong performance of DN2.

6. Discussion

A general paradigm for anomaly detection: Recent papers (e.g. Golan & El-Yaniv (2018)) advocated the paradigm of self-supervision, possibly with augmentation by an external dataset e.g. outlier exposure. The results in this paper, give strong evidence to an alternative paradigm: i) learn general features using all the available supervision on vaguely related datasets ii) the learned features are expected to be general enough to be able to use standard anomaly detection methods (e.g. kNN, k-means). The pretrained paradigm is much faster to deploy than self-supervised methods and has many other advantages investigated extensively in Sec. 4. We expect that for image data that has no similarity whatsoever to Imagenet, using pre-trained features may be less effective. That withstanding, in our experiments, we found that Imagenet-pretrained features were effective on aerial images as well as microscope images, while both settings

Imagenetとは大きく異なる。そのため、DN2ライクな手法が非常に幅広く適用できることを期待している。

外部監視 DN2の成功の鍵となるのは、高品質の外部特徴抽出器を利用できることである。我々が使用したResNet抽出器は、以前Imagenetでトレーニングされたものです。一般的に、監視を使用することは、自己監視の方法よりも高価で手間がかかると考えられています。しかしこの場合、私たちはそれが不利だとは全く考えていません。私たちは、すでに訓練され、無料のオープンソースソフトウェアライブラリのように汎用化されているネットワークを使用した。それらは完全に無料で入手可能であり、新しいデータセットに対してそのようなネットワークを使用するために新たな監視は全く必要なく、トレーニングにかかる時間やストレージのコストも最小限である。全てのプロセスはPyTorchの1行で構成されているため、この場合、これらの手法が教師ありかどうかの議論は純粋に哲学的なものであると我々は考えている。

非常に大きなデータセットへのスケールアップ： ニアレストネイバーは大規模なデータセットに対して遅いことで有名である。この複雑さはニューラルネットワークのようなパラメトリック分類器ではそれほど深刻ではありません。これはニアレストネイバー分類でよく知られた問題であるため、これを回避するために多くの研究が行われた。1つの解決策は、kd-treeなどによる高速なkNN検索である。第5節で使用されるもう1つの解決策は、k-meansを計算し、その上でkNNを計算することによって訓練集合を削減することで、kNNを高速化することを提案した。これは、再帰的K-meansアルゴリズムによってNNを近似する確立された手法によってさらに一般化される（Fukunaga & Narendra, 1975）。実際には、実行時間のほとんどは、最近傍検索ではなく、テスト画像に対するニューラルネットワーク推論の結果になると予想されます。

非画像データ： 非画像データ：我々の調査により、画像異常検知のための非常に強力なベースラインが確立された。しかし、この結果は必ずしも全ての異常検知タスクがこの方法で実行できることを意味しない。一般的な特徴抽出器は画像上で非常に成功しており、他のタスク例えば自然言語処理（BERT（Devlin et al.） しかし、表データや時系列データなど、異常検出にとって最も重要な分野ではそうではない。これらの場合、一般的な特徴抽出器は存在せず、データセット間の分散が非常に大きいため、そのような特徴抽出器を作成するための明白な道筋はない。しかし、表データでは一般的にディープメソッドはあまり成功しないため、生データにおけるkNNのベースラインは非常に強力なものであることに留意されたい。それはともかく、我々はこれらのデータモダリティが、自己教師付き異常検知にとって最も有望な分野であると信じている。Bergman & Hoshen (2020)はこれに沿った方法を提案した。

7. Conclusion

我々は、単純な方法である深層画像特徴上のkNNを、半教師付き及び教師なし異常検知のための現在のアプローチと比較した。その単純さにもかかわらず、この単純な手法は、精度、学習時間、入力不純物に対する頑健性、データセットのタイプに対する頑健性、サンプルの複雑さの点で、最先端の手法を凌駕することが示された。我々は、より複雑なアプローチが最終的にこの単純なアプローチを凌駕すると信じているが、DN2は、異常検出の実践者にとって優れた出発点であると同時に、将来の研究のための重要なベースラインであると考える。

References

アーメッド、F.とクールヴィル、A.意味的異常の検出.arXiv preprint arXiv:1908.04388, 2019.

Bergman, l. and Hoshen, Y. Classification-based anomaly detection for general data. In *ICLR*, 2020.

Candès, E. J., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *JACM*, 2011.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

DOro, P., Nasca, E., Masci, J., and Matteucci, M. Group anomaly detection via graph autoencoders. 2019.

Elson, J., Douceur, J. R., Howell, J., and Saul, J. Asirra: a captcha that exploits interest-aligned manual image categorization. In *ACM Conference on Computer and Communications Security*, volume 7, pp. 366–374, 2007.

Eskin, E., Arnold, A., Prerau, M., Portnoy, L., and Stolfo, S. A geometric framework for unsupervised anomaly detection. In *Applications of data mining in computer security*, pp. 77–101. Springer, 2002.

Fukunaga, K. and Narendra, P. M. A branch and bound algorithm for computing k-nearest neighbors. *IEEE transactions on computers*, 100(7):750–753, 1975.

Gidaris, S., Singh, P., and Komodakis, N. Unsupervised representation learning by predicting image rotations. *ICLR*, 2018.

Golan, I. and El-Yaniv, R. Deep anomaly detection using geometric transformations. In NeurIPS, 2018.

Gu, X., Akoglu, L., and Rinaldo, A. Statistical analysis of nearest neighbor methods for anomaly detection. In *NeurIPS*, 2019.

are very different from Imagenet. We therefore expect DN2-like methods to be very broadly applicable.

External supervision: The key enabler for DN2's success is the availability of a high quality external feature extractor. The ResNet extractor that we used was previously trained on Imagenet. Using supervision is typically seen as being more expensive and laborious than self-supervised methods. In this case however, we do not see it as a disadvantage at all. We used networks that have already been trained and are as commoditized as free open-source software libraries. They are available completely free, no new supervision at all is required for using such networks for any new dataset, as well as minimal time or storage costs for training. The whole process consists of merely a single PyTorch line, we therefore believe that in this case, the discussion of whether these methods can be considered supervised is purely philosophical.

Scaling up to very large datasets: Nearest neighbors are famously slow for large datasets, as the runtime increases linearly with the amount of training data. The complexity is less severe for parametric classifiers such as neural networks. As this is a well known issue with nearest neighbors classification, much work was performed at circumventing it. One solution is fast kNN retrieval e.g. by kd-trees. Another solution used in Sec. 5, proposed to speed up kNN by reducing the training set through computing its k-means and computing kNN on them. This is generalized further by an established technique that approximates NN by a recursive K-means algorithm (Fukunaga & Narendra, 1975). We expect that in practice, most of the runtime will be a result of the neural network inference on the test image, rather than on nearest neighbor retrieval.

Non-image data: Our investigation established a very strong baseline for image anomaly detection. This result, however, does not necessarily mean that all anomaly detection tasks can be performed this way. Generic feature extractors are very successful on images, and are emerging in other tasks e.g. natural language processing (BERT (Devlin et al., 2018)). This is however not the case in some of the most important areas for anomaly detection i.e. tabular data and time series. In those cases, general feature extractors do not exist, and due to the very high variance between datasets, there is no obvious path towards creating such feature extractors. Note however that as deep methods are generally less successful on tabular data, the baseline of kNN on raw data is a very strong one. That withstanding, we believe that these data modalities present the most promising area for self-supervised anomaly detection. Bergman & Hoshen (2020) proposed a method along these lines.

7. Conclusion

We compare a simple method, kNN on deep image features, to current approaches for semi-supervised and unsupervised anomaly detection. Despite its simplicity, the simple method was shown to outperform the state-of-the-art methods in terms of accuracy, training time, robustness to input impurities, robustness to dataset type and sample complexity. Although, we believe that more complex approaches will eventually outperform this simple approach, we think that DN2 is an excellent starting point for practitioners of anomaly detection as well as an important baseline for future research.

References

Ahmed, F. and Courville, A. Detecting semantic anomalies. *arXiv preprint arXiv:1908.04388*, 2019.

Bergman, l. and Hoshen, Y. Classification-based anomaly detection for general data. In *ICLR*, 2020.

Candès, E. J., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *JACM*, 2011.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

DOro, P., Nasca, E., Masci, J., and Matteucci, M. Group anomaly detection via graph autoencoders. 2019.

Elson, J., Douceur, J. R., Howell, J., and Saul, J. Asirra: a captcha that exploits interest-aligned manual image categorization. In *ACM Conference on Computer and Communications Security*, volume 7, pp. 366–374, 2007.

Eskin, E., Arnold, A., Prerau, M., Portnoy, L., and Stolfo, S. A geometric framework for unsupervised anomaly detection. In *Applications of data mining in computer security*, pp. 77–101. Springer, 2002.

Fukunaga, K. and Narendra, P. M. A branch and bound algorithm for computing k-nearest neighbors. *IEEE transactions on computers*, 100(7):750–753, 1975.

Gidaris, S., Singh, P., and Komodakis, N. Unsupervised representation learning by predicting image rotations. *ICLR*, 2018.

Golan, I. and El-Yaniv, R. Deep anomaly detection using geometric transformations. In *NeurIPS*, 2018.

Gu, X., Akoglu, L., and Rinaldo, A. Statistical analysis of nearest neighbor methods for anomaly detection. In *NeurIPS*, 2019.