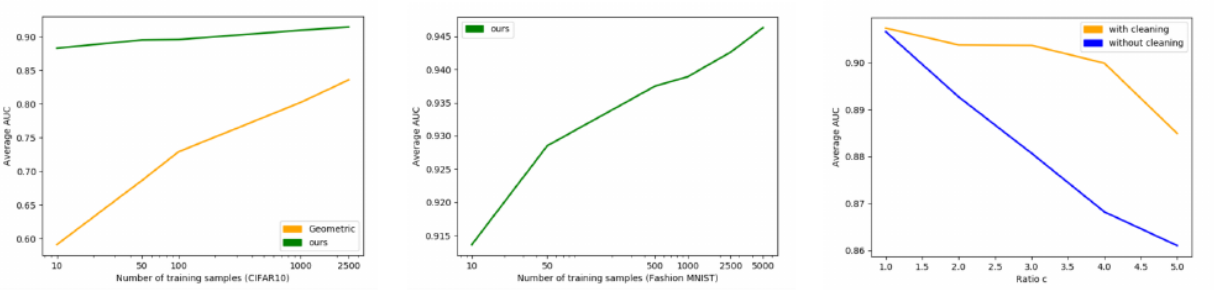

図5. 学習画像数とROCAUCの比較（左）CIFAR10 DN2は10枚の画像からでも高い精度を達成しているが、Geometricは致命的に悪化している。（中央）FashionMNISTも同様にDN2により高い性能を達成。（右）CIFAR10における不純物比率とROCAUCの比較。トレーニングセットのクリーニングにより、性能が大幅に向上。

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）。

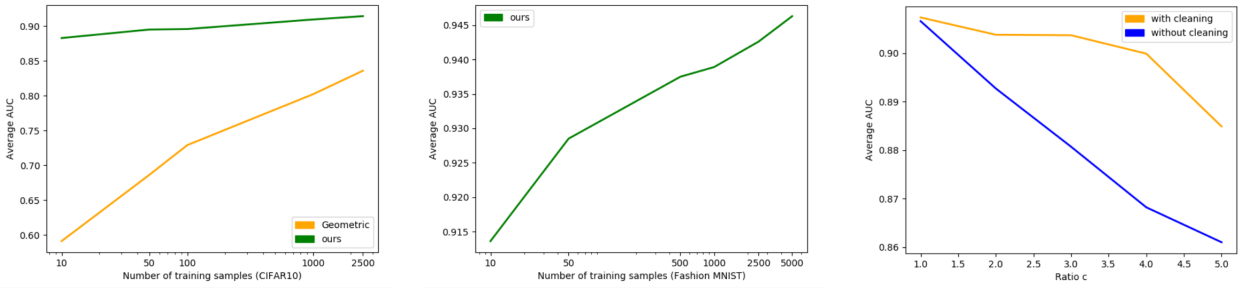

Figure 5. Number of training images vs. ROCAUC (left) CIFAR10 - Strong perfromance is achieved by DN2 even from 10 images, whereas Geometric deteriorates critically. (center) FashionMNIST - similarly strong performance by DN2. (right) Impurity ratio vs ROCAUC on CIFAR10. The training set cleaning procedure,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