Transformaly - Two (Feature Spaces) Are Better Than One

Matan Jacob Cohen
Blavatnik School of Computer Science
Tel-Aviv University
Tel-Aviv, Israel
matanyaakovc@mail.tau.ac.il

Shai Avidan
School of Electrical Engineering
Tel-Aviv University
Tel-Aviv, Israel
avidan@tauex.tau.ac.il

Abstract

Anomaly detection is a well-established research area that seeks to identify samples outside of a predetermined distribution. An anomaly detection pipeline is comprised of two main stages: (1) feature extraction and (2) normality score assignment. Recent papers used pre-trained networks for feature extraction achieving state-of-the-art results. However, the use of pre-trained networks does not fully-utilize the normal samples that are available at train time. This paper suggests taking advantage of this information by using teacher-student training. In our setting, a pre-trained teacher network is used to train a student network on the normal training samples. Since the student network is trained only on normal samples, it is expected to deviate from the teacher network in abnormal cases. This difference can serve as a complementary representation to the pre-trained feature vector. Our method - Transformaly - exploits a pre-trained Vision Transformer (ViT) to extract both feature vectors: the pre-trained (agnostic) features and the teacher-student (fine-tuned) features. We report state-of-the-art AUROC results in both the common unimodal setting, where one class is considered normal and the rest are considered abnormal, and the multimodal setting, where all classes but one are considered normal, and just one class is considered abnormal1.

1. Introduction

Anomaly detection is a long-standing field of research that has many applications in computer vision. The realm of anomaly detection is broad and involves different types of problems.

In the case of multi-class classification, the term "Anomaly Detection" is often used to describe out-of-distribution (OOD) detection or novelty detection; where

the task is to determine at inference time, if a test sample belongs to one of the classes the model was trained to classify or not. This task relies on what is known as the open-set assumption, where the class of a query sample can be outside the set of classes used in training.

Aside from OOD detection, one can consider two anomaly detection variants: (i) semantic anomaly detection, in which the normal and the abnormal samples differ in their semantic meaning; (ii) defect detection, in which the normal and the abnormal samples differ in their local appearance (i.e., defect), but are semantically identical. We consider the case of semantic anomaly detection.

At a high level, anomaly detection involves the combination of representation and modelling. Solutions to the problem can be divided into three categories: the unsupervised approach, the self-supervised approach and the pretrained-based approach.

Unsupervised methods use only normal data, without any form of labeling. These methods include methods
for reconstructing the normal data [8, 11, 21, 40, 42], estimating its density [12], or concentrating it into one manifold [30, 32, 35].

In the self-supervised approach, a model is trained on an auxiliary task. Hopefully, the model learns meaningful features that reflect the normal nature of the data. The construction of an auxiliary task that motivates the model to learn these relevant features is not trivial, and several suggestions have been made such as geometric transformations classification [13], rotation classification [19], puzzle-solving [31] and CutPaste [20].

Recently, significant progress has been made in the self-supervised domain, with the use of contrastive learning [14, 15, 36]. Studies have shown that contrastive learning can be useful for semantic anomaly detection and produce good results [33, 34].

The combination of feature extraction from a pre-trained model and simple scoring algorithm on top of it, is an effective approach for anomaly detection [2, 19, 41]. Bergman et al. [2] used a pre-trained ResNet model, and applied a kNN scoring method on the extracted features. That alone surpassed almost all unsupervised and self-supervised methods. Fine-tuning the model using either center loss or contrastive learning, leads to even better results [27, 28].

On the downside, pre-trained features are agnostic to the normal data that is available at the training stage. It is a loss of valuable information, and we propose to address it by using teacher-student training. Specifically, our work combines both pre-trained features and teacher-student training. In both cases, we use a pre-trained Vision Transformer (ViT) network as our backbone [10].

Teacher-student training was already used for pixel-precise anomaly segmentation in high resolution images [5]. Their representation learns low level statistics that are suitable for defect detection tasks. They report results that are considerably sub-par for the case of semantic anomaly detection.

We, on the other hand, focus on semantic anomaly detection, where semantic representation is crucial. In contrast to [5], we utilize not only the teacher-student discrepancy representation, but also the raw pre-trained embedding, achieving SOTA results in detecting semantic anomalies.

We modify the standard teacher-student setting by using blocks instead of the entire network. Specifically, we use the Vision-Transformer architecture (ViT) and exploit the nature of its block structure. We construct a student block that corresponds to each block of the teacher backbone. At train time, each student block is trained to mimic its corresponding teacher block. The student blocks are trained independently, and are exposed only to normal samples. At test time, each sample is represented by a vector of the differences between the teachers' outputs and the students' outputs. We term this representation teacher-student discrepancy.

Figure 1 gives a high-level overview of our approach. Each sample is mapped to two different feature spaces: one created by a pre-trained ViT network (the agnostic features) and another created by the discrepancy between student and teacher blocks (fine-tuned features).

The likelihood of a sample is taken to be the product of its likelihood in both feature spaces. To model the likelihoods we experimented with several options that include kNN, a single Gaussian, and a Gaussian Mixture Model.

We evaluate our method on several data sets, and find that in most cases this combined representation outperforms existing state-of-the-art methods. The main contributions of this paper are:

- Transformaly - a first use of dual feature representation for anomaly detection: agnostic and fine-tuned.
- A novel use of teacher-student differences using Visual Transformers (ViT) for semantic anomaly detection.
- State of the art results on multiple datasets and multiple settings: Cutting the error by 40 – 65% on competitive benchmarks.
- We are the first to report a hard-to-detect failure in the common unimodal setting, which we called "Pre-training Confusion", justify our additional multimodel evaluation process.

2. Background

The term “Anomaly Detection” encapsulate several different tasks in the literature. It has been used to describe Out-of-Distribution Detection, Defect Detection, and Semantic Anomaly Detection (the topic of this work). We will briefly cover these tasks here.

Out Of Distribution Detection: Out-of-distribution (OOD) detection, also known as novelty detection, considers the case of multi-class classification. In this task, in addition to training an accurate model, we would also like to detect when it encounters a sample that does not belong to any of the known classes.

Several approaches took advantage of the power of supervised multi-class classification. For example, the model predictions of out-of-distribution samples have lower values than those of in-distribution samples, making anomaly detection possible [16]. Another notable approach has shown that the gap between temperature-scaled softmax scores of a sample and a perturbed version of it can be measured and used as a normality score [22]. It turns out that a larger gap is apparent in anomalous samples than in normal ones.

In addition, using a small dataset of possible anomalies can boost the detection performance [18]. More recent
Defect Detection: In defect detection the normal and the abnormal samples differ in local appearance, but are semantically identical. For example, defects in printed circuits, cables or medicinal pills.

Researchers proposed several approaches for defect detection [4, 5, 37]. Two recent papers use the Vision Transformer (ViT) architecture; Mishra et al. [24] suggest using encoder-decoder architecture in order to reconstruct the normal data. Pirany and Chai [26] proposed to train ViT using the auxiliary task of patch-inpainting. At inference time both methods use the discrepancy between the input image patches and the reconstructed image patches as an indication of possible defects.

Neither of these solutions includes a pre-training phase for the ViT. Both are best suited to detecting local defects reflected in patches, not semantic anomalies.

Bergmann et al. [5] suggested a teacher-student architecture for defect detection. The teacher model is based on a ResNet model pre-trained on a large dataset of patches from natural images. Then, an ensemble of student networks is trained on anomaly-free training data using regression loss with the teacher's ultimate outputs.

This patch-based method is suited for defect detection, where the anomaly is in appearance and not semantic. Furthermore, this method uses only the last layer outputs for the teacher-student mechanism, and does not utilize the semantic embedding of the pre-trained network. Transformaly, on the other hand, takes advantage of different blocks in the model, by training the student blocks using intermediate outputs. Additionally, the semantically pre-trained embedding is used along with fine-tuned embedding to obtain SOTA results.

Semantic Anomaly Detection: We can classify semantic anomaly detection solutions into three categories: Unsupervised Learning, self-supervised, and pretrained-based approaches.

Unsupervised anomaly detection solutions fall into three approaches: reconstruction-based, density estimation, and one-class classifiers.

Reconstruction-based methods attempt to capture the main characteristics of a normal training set by measuring reconstruction success. By assuming that only normal data will reconstruct well at test time, these methods attempt to detect anomalies. Previous papers have suggested $k$ nearest neighbors ($k$-NN) [11], autoencoder [40] and GANs [8, 21, 42] in order to reconstruct and classify the samples.

Density-based methods estimate the density of the normal data. These methods predict the samples’ likelihood as their normality scores. Previous papers suggested parametric density estimation, such as mixture of Gaussians (GMM) [12], and nonparametric, such as $k$-NN [25].

One-class classification methods map normal data to a manifold, leaving the abnormal samples outside. A few modifications have been made to SVM to adjust it for this purpose, with training only on one class [30, 32, 35].

Self-supervised methods use auxiliary tasks in order to learn relevant features of the normal data. These methods train a neural network to solve an unrelated task, using just the normal training data. At inference time, the model’s auxiliary task performance on the test set is considered as its normality score.

A number of papers have proposed applying transformations to the normal data and predicting which transformations have been applied. Predicting predefined geometric transformations [13], rotation [19], puzzle-solving [31] and CutPaste [20] are a few of the auxiliary tasks that have been suggested. In addition, (random) general transformations can be applied not only to images but also to tabular data, enabling anomalies to be detected in this domain as well [3].

Recent papers demonstrate the effectiveness of contrastive learning as a self-supervised method for learning visual representations, achieving SOTA results [6, 14, 15]. A contrastive learning approach, such as SimCLR, produces uniformly distributed outputs, making anomalies difficult to spot [6, 33]. Using contrasting shifted instances, along with rotation or transformation prediction, managed to surpass this challenge and achieved high performance [34]. Additionally, training a feature extractor using shifted contrastive learning and applying one class classification or kernel density estimation has been shown to be effective for anomaly detection [33].

Pretrained-based methods use backbones that are trained on large datasets, such as ImageNet, to extract features [9]. While in the past obtaining pre-trained models was a limitation; these days they are readily available and commonly used across many domains. These pre-trained models produce separable semantic embeddings and, as a result, enable the detection of anomalies by using simple scoring methods such as $k$-NN or Gaussian Mixture Model [2, 41].

Surprisingly, the embeddings produced by these algorithms lead to good results also on datasets that are drastically different from the pretraining one. A follow-up paper improves the pre-training model detection performance by fine-tuning it using center loss [27]. Recent publication has suggested to fine-tuned the pre-trained network using an additional dataset as outlier exposure, to further boost the results [7].
3. Method

We exploit the power of pre-trained ViT by constructing two features spaces for each sample; the pre-trained and the fine-tuned. A sample $x$ is embedded into a pre-trained feature vector $z_p$ and a fine-tuned feature vector $z_f$.

**Pre-trained Features** The pre-trained vector $z_p$ is obtained by passing the input $x$ through a pre-trained ViT network. This produces an embedding that is agnostic to the actual normal and abnormal data at hand.

We set $z_p = \text{ViT}(x)$ and fit a Gaussian to it. At inference time, each sample is scored according to its log probability as induced by the fitted model. Normal samples are assumed to have higher probability than abnormal samples.

**Fine-tuned Features** The fine-tune feature embedding is inspired by the knowledge distillation domain [39]. It is calculated as the difference between the output of a teacher and student ViT blocks. Specifically, we train the student block only on normal data, such that it produces an output that is similar to the teacher output. The output of the student block is expected to be quite different when the data is abnormal.

We followed this process for $m$ different teacher-student blocks. We train each student block $h'_j$ independently to mimic $h_j$ using an $MSE$ loss:

$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^{n} |h_j(x_i) - h'_j(x_i)|^2_2$$

At inference time, sample $x$ is represented with:

$$z_f = [z_f^{(0)}, \ldots, z_f^{(m-1)}]$$

where $m$ is the number of blocks in the ViT network, and $z_f^{(j)} = ||h_j(x_i) - h'_j(x_i)||^2_2$ is the difference between the $j$-th teacher block $h_j$ and the $j$-th student block $h'_j$. We typically use $m = 10$ to model the last 10 blocks in the ViT network. We empirically observed that using the first two blocks does not improve the model’s performance. The first two layers may have learned low-level features that appear in both normal and abnormal samples. As such, these features are useless for detecting semantic anomalies.

**Final Scoring method** We fit two Gaussians to both the pre-trained embedding $z_p$ and the fine-tuned embedding $z_f$.

$$Pr(z_p) = \mathcal{N}(z_p; \mu_p, \Sigma_p)$$

$$Pr(z_f) = \mathcal{N}(z_f; \mu_f, \Sigma_f)$$

where $\mu_p, \Sigma_p$ are the mean and covariance of the pre-trained embeddings, and $\mu_f, \Sigma_f$ are the mean and covariance of the fine-tuned embeddings. The final score of sample $x$ is simply the product of the two (or the sum of their
(5)

Despite the fact that the likelihoods of the two Gaussians are not independent, we chose their product as our normality scoring method. This method and scoring procedure is used in all the experiments in the next section, unless otherwise stated.

4. Experiments

We describe implementation details, the benchmark settings and datasets that we used, and present the results of the experiments in the following sub-sections.

4.1. Implementation Details

We use a PyTorch implementation of ViT, trained on ImageNet-21k and fine-tuned on ImageNet-1k. ViT has 12 heads, 16 × 16 input patch size, dropout rate of 0.1 and its penultimate layer outputs 768-dimensional vectors, which form the pre-trained features of our method. All the input images are normalized according to the pre-training phase of the ViT. Unless otherwise specified, the fine-tune features are 10D vectors that are taken to be the result of applying teacher-student training to the last ten blocks of ViT. In each feature space we model the data with a single Gaussian using its mean and full covariance. Since pre-trained features live in a 768D space, we first whiten and reduce the dimensionality of these features by keeping the number of components that explain 90% of the data variance (Typically, this results in vector a of 300 dimensions).

4.2. Datasets

Transormaly is evaluated on commonly used datasets: Cifar10, Cifar100, Fashion MNIST, and Cats vs Dogs. We evaluated Transformaly’s robustness using additional datasets: aerial images (Dior), blood cell images (Blood Cells), X-ray images of Covid19 patients (Covid19), natural scenes images (View Recognition), weather image (Weather Recognition) and images of plain and cracked concrete (Concrete Crack Classification). We show a representative image from each dataset in Figure 3. As can be seen, the datasets are quite diverse. Please refer to the supplementary material for more details on the datasets used.

4.3. Benchmark Settings

We examine Transformaly in the unimodal and multimodal settings. In the unimodal case, one semantic class is randomly chosen as normal while the rest of the classes are treated as abnormal. During training, the model is only exposed to the normal class’ samples. At inference time, all samples of the test set are evaluated, while samples that are not from the normal class are considered anomalous. In the multimodal case the roles are reversed, and all classes but one are considered normal.

We evaluate both settings because the unimodal setting, while common in the literature, does not adequately reflect all real-life scenarios; where normal data might contain multiple semantic classes. Moreover, unimodal setting leads to a peculiar evaluation process, where we have much more anomalies than normal samples (nine and nineteen times more for Cifar10 and Cifar100, respectively). This is not in line with the common anomaly detection use-case where the normal samples are the majority of the data that the model encounters and the anomalies are the rare events.

It should be noted that this problem was recently discussed in the context of OOD detection by Courville and Ahmed who criticized the standard benchmark in OOD detection. We adopt their proposed multimodal paradigm for anomaly detection as complementary evaluation process to the common unimodal setting.

4.4. Results

Table 1 shows results of the common unimodal setting, in which one class is considered normal while all other classes are abnormal. A threshold-free area under the receiver operating (AUROC) characteristic curve is used to evaluate the models. We report the performance of our method and compare it to unsupervised, self-supervised and pre-trained based methods. Each of the scores presented in following tables is the average of the AUROC scores across all classes in each dataset. One can observe that our method outperforms all other methods on all datasets, except for Fashion MNIST dataset.

We further compared our work against some of the leading methods on additional datasets and report results in Table 2. Our method outperforms other methods on most datasets, often by a large margin (over 18% and 12% on
Table 1. **AUROC scores of the unimodal setting:** We compare our method (rightmost column) against the alternatives. We outperform all methods on all datasets, except for FMNIST (All values, except for our method, are taken from [27], except the value 94.21 of MSAD [28] on FMNIST that was computed by us using the official code released by the authors). ♯ taken from [27], ♮ taken from [28].

<table>
<thead>
<tr>
<th>Dataset</th>
<th>OC-SVM</th>
<th>DeepSVDD</th>
<th>MHRot</th>
<th>CSI</th>
<th>DN2</th>
<th>PANDA</th>
<th>MSAD</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIFAR10</td>
<td>64.7</td>
<td>64.8</td>
<td>90.1</td>
<td>94.3</td>
<td>92.5</td>
<td>96.2</td>
<td>97.2</td>
<td><strong>98.31</strong></td>
</tr>
<tr>
<td>CIFAR100</td>
<td>62.6</td>
<td>67.0</td>
<td>80.1</td>
<td>89.6</td>
<td>94.1</td>
<td>94.1</td>
<td>96.4</td>
<td><strong>97.34</strong></td>
</tr>
<tr>
<td>FMNIST</td>
<td>92.8</td>
<td>84.8</td>
<td>93.2</td>
<td>-</td>
<td>94.5</td>
<td><strong>95.6</strong></td>
<td>94.21</td>
<td>94.43</td>
</tr>
<tr>
<td>CatsVsDogs</td>
<td>51.7</td>
<td>50.5</td>
<td>86.0</td>
<td>86.3</td>
<td>96.0</td>
<td>97.3</td>
<td>99.3</td>
<td><strong>99.52</strong></td>
</tr>
<tr>
<td>DIOR</td>
<td>70.7*</td>
<td>70.0*</td>
<td>73.3*</td>
<td>78.5*</td>
<td>92.2</td>
<td>94.3</td>
<td>97.2</td>
<td><strong>98.08</strong></td>
</tr>
</tbody>
</table>

Table 2. **AUROC scores of the unimodal setting on various datasets:** We compare our method against some of the alternatives. We outperform other methods on most datasets, sometimes by a large margin (over 18% and 12% on "Blood Cells" and "Weather Recognition" respectively), while we underperform only slightly on "Concrete Crack Classification" and "Covid19", where we come in second.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>DeepSVDD</th>
<th>DN2</th>
<th>PANDA</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blood Cells</td>
<td>52.28</td>
<td>54.91</td>
<td>56.25</td>
<td><strong>74.85</strong></td>
</tr>
<tr>
<td>Covid19</td>
<td>97.32</td>
<td>97.63</td>
<td><strong>99.33</strong></td>
<td>98.87</td>
</tr>
<tr>
<td>Weather</td>
<td>73.55</td>
<td>80.05</td>
<td>81.53</td>
<td><strong>94.32</strong></td>
</tr>
<tr>
<td>View</td>
<td>60.31</td>
<td>90.86</td>
<td>93.63</td>
<td><strong>95.80</strong></td>
</tr>
<tr>
<td>Concrete</td>
<td>92.27</td>
<td>99.81</td>
<td><strong>99.93</strong></td>
<td>99.77</td>
</tr>
</tbody>
</table>

"Blood Cells" and "Weather Recognition", respectively. Our method under-performs only slightly on "Concrete Crack Classification" and "Covid19", in which it comes in second. The different characteristics of these datasets, which belong to very different domains, demonstrate the robustness and flexibility of our method.

We further analyze the contribution of pre-trained and fine-tuned features to the final outcome. Results are reported in Table 3. We also report in Table 3 the performance of the algorithm using different numbers of teacher-student blocks. In all cases, the data in each feature space is modeled with a single Gaussian.

As shown in the first four columns, each feature space independently yields good results (left column for pre-trained features, middle three columns for various number of teacher-student blocks used to produce the fine-tuned features). Combined (the rightmost three columns) we report the best results on most datasets. Using the pre-trained features combined with only the last ViT block for the fine-tuned features yields SOTA results in most cases, suggesting a more compact version of our method. We focus on the pre-trained features combined with 10 blocks teacher-student fine-tuned features, as it achieved the best results on most datasets.

We next considered several modeling functions for the pre-trained features, including kNN, a single Gaussian, and a Gaussian Mixture Model (GMM). Results, for the pre-trained features only, are found in Table 4. The main observation we draw from this table is that no particular modeling function is consistently better than others. Therefore, we prefer the use of a single Gaussian, which requires less memory to store and is faster to compute.

A single Gaussian is used in order to model teacher-student fine-tuned features as well, based on empirical distribution of those features (see supplemental for details).

**Multimodal Setting** We further tested our algorithm in the multimodal setting, where one of the classes is considered abnormal while all other classes are considered normal. That is, all samples of the normal classes are used as single multimodal class, without using their original labels.

We report the AUROC results in Table 5. As can be seen, the proposed method achieved SOTA results on cifar10 (AUROC score of 90.24) and Cifar100 (AUROC score of 83.05), outperforming alternative methods by approximately 5% and 12%, respectively.

The performance of our method degrades when using the grayscale Fashion MNIST dataset. We suspect that this might be due to the fact that the grayscale dataset is not aligned with the pretraining phase of ViT, which used color images.

We observe a sharp drop in the performance of our method on the DIOR dataset when switching from the unimodal to the multimodal setting. We thoroughly discuss the details of this drop in sub-section 4.5.

Interestingly, in the multimodal setting, the performance of the algorithm does not change much as we try different modeling functions, see Table 6. We observe that using kNN (with different values of $k$), as well as a Gaussian Mixture Model (GMM) with varying number of Gaussians gives similar results.

**Whitening:** Finally, in the last experiment we test the robustness of our algorithm to the dimensionality reduction parameter. Since we use a Gaussian with full covariance to model the pre-trained features, we reduce the dimensionality...
Table 3. **AUROC scores of pre-trained and fine-tuned features in the unimodal setting:** We compare the performance of pre-trained (leftmost column), fine-tuned features with different number of teacher-student blocks (middle 3 columns), and the combined effect of pre-trained and various fine-tuned features (rightmost three columns). In all cases, the data in each feature space is modeled with a single Gaussian. Fine-tuned features (generated by Teacher-Student discrepancy) provide, by themselves, satisfactory results. Fine-tuned features using only to the last ViT block (“Fine-Tuned, Last 1 block”) represent a lightweight variant of our method that gives good results. On almost all datasets, fine-tuned features boost the performance of pre-trained features, sometimes by a large margin (i.e., weather recognition).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Pre-trained</th>
<th>Fine-tuned</th>
<th>Full Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Last 1 block</td>
<td>Last 3 blocks</td>
<td>Last 10 blocks</td>
</tr>
<tr>
<td>CIFAR10</td>
<td>97.81</td>
<td>95.02</td>
<td>97.18</td>
</tr>
<tr>
<td>CIFAR100</td>
<td>96.21</td>
<td>90.71</td>
<td>94.79</td>
</tr>
<tr>
<td>FMNIST</td>
<td>93.94</td>
<td>88.41</td>
<td>92.39</td>
</tr>
<tr>
<td>CatsVsDogs</td>
<td>99.60</td>
<td>98.30</td>
<td>97.45</td>
</tr>
<tr>
<td>DIOR</td>
<td>93.97</td>
<td>94.72</td>
<td>95.26</td>
</tr>
<tr>
<td>Blood Cells</td>
<td>72.19</td>
<td>74.80</td>
<td>75.43</td>
</tr>
<tr>
<td>Covid19</td>
<td>97.06</td>
<td>92.67</td>
<td>96.24</td>
</tr>
<tr>
<td>Weather</td>
<td>81.06</td>
<td>80.62</td>
<td>80.32</td>
</tr>
<tr>
<td>View</td>
<td>95.48</td>
<td>94.40</td>
<td>94.41</td>
</tr>
<tr>
<td>Concrete</td>
<td>99.72</td>
<td>99.47</td>
<td>99.13</td>
</tr>
</tbody>
</table>

Table 4. **Modeling functions in the unimodal setting:** We report AUROC results of our algorithm using different modelling functions on the pre-trained features only. In particular, we try k-NN with different values of k, the number of nearest neighbors. We also test Gaussian Mixture Model (GMM) with varying number of Gaussians. It can be observed that no modelling function is consistently better than the others. This leads us to use a single Gaussian because if offers an attractive trade-off between high accuracy, low memory footprint and fast computation time.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>k-NN</th>
<th>k-NN</th>
<th>GMM</th>
<th>GMM</th>
<th>GMM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>k = 2</td>
<td>k = 5</td>
<td>n = 1</td>
<td>n = 5</td>
<td>n = 20</td>
</tr>
<tr>
<td>CIFAR10</td>
<td><strong>97.81</strong></td>
<td><strong>97.84</strong></td>
<td>97.79</td>
<td>95.98</td>
<td></td>
</tr>
<tr>
<td>CIFAR100</td>
<td><strong>96.41</strong></td>
<td><strong>96.40</strong></td>
<td>96.25</td>
<td>95.18</td>
<td>91.00</td>
</tr>
<tr>
<td>FMNIST</td>
<td><strong>94.19</strong></td>
<td><strong>94.09</strong></td>
<td>93.94</td>
<td>93.69</td>
<td>93.04</td>
</tr>
<tr>
<td>CatsVsDogs</td>
<td><strong>99.59</strong></td>
<td><strong>99.63</strong></td>
<td>99.60</td>
<td><strong>99.63</strong></td>
<td>98.97</td>
</tr>
<tr>
<td>DIOR</td>
<td><strong>91.74</strong></td>
<td><strong>92.52</strong></td>
<td><strong>93.97</strong></td>
<td>91.27</td>
<td>88.78</td>
</tr>
</tbody>
</table>

Table 5. **AUROC score of the multimodal setting:** We compare our method (rightmost column) against the alternative. ♮ taken from [2], ♯ taken from [28].

<table>
<thead>
<tr>
<th>Dataset</th>
<th>DeepSVDD</th>
<th>DN2</th>
<th>PANDA</th>
<th>MSAD</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIFAR10</td>
<td>50.67</td>
<td>71.7g</td>
<td>78.5g</td>
<td>85.3g</td>
<td><strong>90.38</strong></td>
</tr>
<tr>
<td>CIFAR100</td>
<td>50.75</td>
<td>71.0g</td>
<td>62.47</td>
<td>67.67</td>
<td><strong>79.80</strong></td>
</tr>
<tr>
<td>FMNIST</td>
<td>70.85</td>
<td><strong>77.64</strong></td>
<td><strong>79.45</strong></td>
<td>72.26</td>
<td>72.53</td>
</tr>
<tr>
<td>DIOR</td>
<td>56.71</td>
<td>81.10</td>
<td>86.92</td>
<td><strong>90.11</strong></td>
<td>66.71</td>
</tr>
</tbody>
</table>

Table 6. **Modeling functions in the multimodal setting:** We report AUROC results of our algorithm using different modelling functions on the pre-trained features. In particular, we try k-NN with different values of k, the number of nearest neighbors. We also test Gaussian Mixture Model (GMM) with varying number of Gaussians. It can be observed that no modelling function is consistently better than the others. This leads us to use a single Gaussian because if offers an attractive trade-off between high accuracy, low memory footprint and fast computation time.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>k-NN</th>
<th>k-NN</th>
<th>GMM</th>
<th>GMM</th>
<th>GMM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>k = 2</td>
<td>k = 5</td>
<td>n = 1</td>
<td>n = 5</td>
<td>n = 20</td>
</tr>
<tr>
<td>CIFAR10</td>
<td><strong>88.76</strong></td>
<td><strong>89.16</strong></td>
<td>90.23</td>
<td><strong>90.81</strong></td>
<td>90.39</td>
</tr>
<tr>
<td>CIFAR100</td>
<td><strong>82.20</strong></td>
<td><strong>82.68</strong></td>
<td>78.76</td>
<td>79.42</td>
<td>77.66</td>
</tr>
<tr>
<td>FMNIST</td>
<td><strong>75.59</strong></td>
<td><strong>74.99</strong></td>
<td>72.29</td>
<td>75.43</td>
<td><strong>78.00</strong></td>
</tr>
<tr>
<td>DIOR</td>
<td><strong>66.66</strong></td>
<td>66.08</td>
<td>65.72</td>
<td>69.75</td>
<td><strong>69.82</strong></td>
</tr>
</tbody>
</table>

4.5. Limitations

Evaluating Transformaly in both the unimodal and multimodal settings reveals the strengths and limitations of our method and the pre-training approach. We outperform almost all methods in the unimodal case, and achieve SOTA results on Cifar10 and Cifar100, in the multimodal case. However, in the multimodal case we do fail on the DIOR dataset. This occurs because of "pre-training confusion", where the pre-trained model maps two semantically different classes to the same region in feature space.
Figure 4. *"Pre-training Confusion" in synthetic Setting, DIOR ViT embeddings and CIFAR10 ResNet embeddings:* (a) Synthetic Setting - each shape represents a different class in the dataset. In a unimodal setting only squares are considered normal. In a multimodal setting all classes except squares are considered normal. The situation is not symmetric and affects AUROC scores. See text for details. (b) tSNE of pre-trained ViT penultimate layer outputs of DIOR. As can be seen, although the sample embeddings are semantically separated for some classes, sample embeddings of classes 13 and 17 are mixed as well as sample embeddings of class 4 and class 8. (c) tSNE of pre-trained ResNet penultimate layer outputs of Cifar10. As can be seen, although the sample embeddings are semantically separated for some classes, sample embeddings of classes 3 and 5 are mixed. Best viewed in color. Zoom in for details.

<table>
<thead>
<tr>
<th>Explained Variance</th>
<th>CIFAR10</th>
<th>Weather Recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre-Trained</td>
<td>Full</td>
</tr>
<tr>
<td>85%</td>
<td>96.85</td>
<td>98.11</td>
</tr>
<tr>
<td>90%</td>
<td>97.81</td>
<td>98.31</td>
</tr>
<tr>
<td>95%</td>
<td>98.11</td>
<td>98.33</td>
</tr>
</tbody>
</table>

Table 7. **Sensitivity of Whitening hyperparameter:** We whiten and reduce the dimensionality of the pre-trained features by keeping 90% of the energy. In this experiment, we show the result of using either 85% or 95% of the energy on two different datasets. The left column, in each table, shows the performance of only the pre-trained features. The right column shows the performance of the full algorithm (pre-trained + fine-tuned features). As can be seen, the algorithm is not sensitive to this hyperparameter.

Figure 4a shows a toy example of this effect in the case of four semantically different classes (triangles, diamonds, circles, and squares). The triangles and diamonds are nicely separated, while the squares and circles are confused.

Consider the unimodal case, where only the blue squares are available as normal samples during training. In this case, at test time only the red circles will be confused as normal instead of abnormal. The red triangles and diamonds will be correctly classified as abnormal. The algorithm misses some of the abnormalities.

The situation is reversed in the multimodal case. Assume now that all red samples (triangles, diamonds, and circles) are normal. At test time, all the abnormal blue squares will be classified as normal. The algorithm misses all the abnormalities.

We suspect that the presented pre-training confusion happens in the DIOR case. To validate this, we plot a tSNE embedding of the features of DIOR in Figure 4b. One can observe that our pre-trained model confuses between class 13 and class 17 and between class 4 and class 8 (highlighted). That is, the model produces embeddings that are similar for both classes.

A similar phenomenon occurs with a ResNet architecture as well. This might explain the failure of recently suggested ResNet-based methods on Cifar10 and Cifar100 in the multimodal setting (such as DN2 [2] and PANDA [27]). A tSNE embedding of the pre-trained ResNet features of Cifar10 is plotted in Figure 4c. One can observe that pre-trained ResNet model confuses between class 3 and class 5 (highlighted).

The stress testing of anomaly detection algorithms in the multimodal settings helps to reveal their limitations. We believe that further analyzing anomaly detection in the multimodal setting is an important topic for future research.

**5. Conclusions**

Transformaly is an anomaly detection algorithm that is based on the Visual Transformer (ViT) architecture. The data is mapped to a pre-trained feature space and a fine-tuned feature space. The normality score of a query point is based on the product of its likelihood in both spaces.

Pre-trained features are obtained by running the samples through a pre-trained ViT. Fine-tuned features are obtained by training a student-network, on normal data only. The discrepancy between student and teacher networks forms the fine-tuned features.

We conduct extensive experiments on multiple datasets and obtain consistently good results, often surpassing the current state of the art.
References


