Transfer-Based Semantic Anomaly Detection

Lucas Deecke ¹  Lukas Ruff ²  Robert A. Vandermeulen ³  Hakan Bilen ¹

Abstract

Detecting semantic anomalies is challenging due to the countless ways in which they may appear in real-world data. While enhancing the robustness of networks may be sufficient for modeling simplistic anomalies, there is no good known way of preparing models for all potential and unseen anomalies that can potentially occur, such as the appearance of new object classes. In this paper, we show that a previously overlooked strategy for anomaly detection (AD) is to introduce an explicit inductive bias toward representations transferred over from some large and varied semantic task. We rigorously verify our hypothesis in controlled trials that utilize intervention, and show that it gives rise to surprisingly effective auxiliary objectives that outperform previous AD paradigms.

1. Introduction

The goal of anomaly detection (AD) is the identification of unusual samples within data (Edgeworth, 1887; Chandola et al., 2009; Pang et al., 2020a; Ruff et al., 2021). For data types that are semantically rich such as images, “unusualness” can be caused by a variety of high-level (or semantic) factors, for example the appearance of new objects classes, or unexpected shapes or poses. For these settings, there has been continued interest in developing new deep AD methods (Zhai et al., 2016; Schlegl et al., 2017; Sabokrou et al., 2018; Deecke et al., 2018; Ruff et al., 2018; Golan & El-Yaniv, 2018; Pidhorskyi et al., 2018; Hendrycks et al., 2019b;c; Goyal et al., 2020; Tack et al., 2020) that utilize end-to-end learning, a defining property amongst deep learning approaches (Krizhevsky et al., 2012; He et al., 2016; Goodfellow et al., 2016).

Because of the sheer number of factors that can potentially cause an anomaly, there is no feasible way of a priori describing or anticipating them. As a result, for deep AD there exists no established principal learning objective. Several auxiliary solutions have been proposed: one line of work utilizes self-supervision (Golan & El-Yaniv, 2018; Hendrycks et al., 2019c; Bergman & Hoshen, 2020; Tack et al., 2020; Sohn et al., 2021), for example learning representations from the task of predicting simple geometric transformations (rotations, translations, etc.) applied to non-anomalous examples. A second popular approach broadly resembles weak supervision, and uses large unstructured collections of data as auxiliary anomalies (Hendrycks et al., 2019b;c; Ruff et al., 2020a).

Considering the rather ad-hoc nature of many of these approaches, especially given the semantic richness present in natural images, one may wonder whether they learn particularly meaningful features from such auxiliary objectives. This is problematic since anomalies can manifest themselves in ways that require a good semantic understanding — for example when anomalies appear in crowded scenes (Mahadevan et al., 2010).

Here we propose a different perspective and hypothesize that, because there is simply no way of anticipating all potential semantic anomalies for unseen images in advance, the best bet is to follow a transfer-based approach that utilizes the semantically rich features obtained from some semantic task solved on a large, semantically varied dataset. We systematically evaluate different strategies to introduce such an inductive bias in AD, and identify simple strategies that yield surprisingly powerful AD methods.

Our work builds on the recently increased availability and utilization of networks pretrained on semantically rich tasks that incorporate different variations commonly seen in data (edges, color, semantic categories, etc.). While He et al. (2019) fundamentally questioned whether actual benefits are achieved from the use of pretrained models, Hendrycks et al. (2019a) painted them in a more positive light, showing they improve robustness and uncertainty calibration.

Rich semantic representations have been shown to boost the performance in many machine learning problems, including image classification (Donahue et al., 2014; Guo et al., 2019), object detection (Girshick et al., 2014; Girshick, 2015), the transfer between large numbers of tasks (Zamir et al., 2018), or from one domain to another (Rebuffi et al., 2017; 2018). In another example, a surge of papers has recently elevated
the role of pretrained models in natural language processing (Mikolov et al., 2018; Devlin et al., 2019; Howard & Ruder, 2018; Adhikari et al., 2019; Hendrycks et al., 2020).

Our central hypothesis is that transferring features from semantic tasks such as ILSVRC image classification (Deng et al., 2009) provides very powerful and generic representations for various AD problems, even when the pretraining task is only loosely related to the task of AD. In doing so, it is important to ensure that the change in representation is not excessive, as this risks catastrophic forgetting (Kirkpatrick et al., 2017). For AD in particular, it is crucial to preserve variations incorporated during pretraining that, even though they potentially don’t exist in the training data, can nonetheless be meaningful for inferring anomalous semantics at test time (Tax & M¨uller, 2003; Rippel et al., 2020). Opposed to mere feature extraction (Bergman et al., 2020), our experiments show that it is critical to let the network have some flexibility to learn new variations important for AD.

To the best of our knowledge, a rigorous analysis and evaluation of transfer-based approaches for AD is still lacking in the literature. Our experiments show that such strategies provide very powerful methods for AD that outperform previous approaches in the deep AD literature on a set of common benchmarks (Section 5). Moreover they are straightforward to train and deploy, and can be coupled with any modern network architecture.

Besides experiments on the predominant AD benchmarks, we propose the use of disentanglement datasets (Gondal et al., 2019) to evaluate the semantic detection performance of AD models. In doing so, we verify that our proposed method is able to robustly detect anomalies under interventions (e.g. a change of object color), showing it preserves meaningful semantic variations in its representations.

In addition to verifying the suitability of our method on semantic benchmarks (Section 5.2), models trained on semantic tasks have been shown to learn elements required for non-semantic decisions in early parts of the network (Zeiler & Fergus, 2014). Indeed we find that our methods are suitable for tasks considered non-semantic (Ahmed & Courville, 2020), such as the popular CIFAR-10 one-versus-rest benchmark (Section 5.3).

2. Related Work

Anomaly detection has a long history (Edgeworth, 1887) and has been extensively studied in the machine learning literature, e.g. through hidden Markov models for detecting network attacks (Ourshton et al., 2003), active learning of anomalies (Pelleg & Moore, 2005), or dynamic Bayesian networks for traffic incident detection (Singliar & Hauskrecht, 2006). An overview over traditional AD methods can be found in Chandola et al. (2009) and Emmott et al. (2013).

Previous deep AD methods utilized autoencoders (Zhou & Paffenroth, 2017; Zong et al., 2018), hybrid methods (Erfani et al., 2016), one-class classification (Ruff et al., 2018; Sabokrou et al., 2018; Ghafouri & Leckie, 2020; Goyal et al., 2020), or GANs (Goodfellow et al., 2014; Schlegl et al., 2017; Akcay et al., 2018; Deecke et al., 2018; Perera et al., 2019; Ngo et al., 2019; Berg et al., 2019). Another line of work explores detecting anomalous videos (Sultani et al., 2018; Ionescu et al., 2019; Ngo et al., 2019; Pang et al., 2020b).

A recent focus has been on developing auxiliary tasks for AD, often following the paradigm of self-supervision, for example predicting geometric transformations of normal data (Golan & El-Yaniv, 2018; Hendrycks et al., 2019c). Different from this, Hendrycks et al. (2019b) propose carrying out AD in a weakly supervised manner in what they call outlier exposure (OE), where one utilizes large unstructured sets of data as auxiliary outliers to improve detection performance. While our approaches also leverage large corpora, they establish inductive biases as a separate crucial element for semantic AD.

Several recent publications have investigated unsupervised mechanisms to learn disentangled representations (Kulka- rni et al., 2015; Higgins et al., 2017; Bouchacourt et al., 2018; Burgess et al., 2018; Chen et al., 2018; Kim & Mnih, 2018; Kumar et al., 2018; Locatello et al., 2019; 2020). We propose using image datasets developed for disentangle-ment (Gondal et al., 2019) for gaining better insights into AD methods, letting us show that an inductive bias seems necessary to detect anomalies on a semantic level.

3. Motivation

To motivate our approach we investigate the semantic viability of features learned under different auxiliary AD objectives. We employ a strategy inspired by linear probing (Zhang et al., 2017) in a two-stage setup very commonly used in AD applications (Erfani et al., 2016; Sohn et al., 2021): after a feature extraction phase $f$ over some data $x \sim p(x)$, a subsequent one-class model $g$ is learned to encapsulate the normal class w.r.t. the push-forward $f_g(p(x))$.

For the data we make use of a standard AD benchmark (Ruff et al., 2018): single classes from CIFAR-10 (e.g. dogs) constitute the normal class, and the one-class model $g$ is learned over all embedded training examples of this class. At test time, we measure whether the two-stage model can successfully identify the appearance of the remaining object classes (cats, deers, etc.) as anomalous. The metric reported in Table 1 results from repeating this procedure for all ten classes, and recording the area under the ROC curve (AUC) relative to that of a random baseline (AUC of 0.5).

The initial extraction phase occurs at one of three layers...
Table 1. Percent improvement in AUC relative to a random baseline on CIFAR-10 AD for one-class SVMs on top of features extracted from LeNet layers (conv1–3) trained through different learning paradigms (transfer-based etc.). Standard deviations in parentheses were computed over five random seeds.

<table>
<thead>
<tr>
<th>Layer</th>
<th>(i) Self-sup.</th>
<th>(ii) Weakly-sup.</th>
<th>(iii) Transfer-b.</th>
</tr>
</thead>
<tbody>
<tr>
<td>f_{conv,1}</td>
<td>1.44 (0.19)</td>
<td>1.58 (0.20)</td>
<td>2.02 (0.21)</td>
</tr>
<tr>
<td>f_{conv,2}</td>
<td>4.60 (0.17)</td>
<td>3.83 (0.16)</td>
<td>5.48 (0.19)</td>
</tr>
<tr>
<td>f_{conv,3}</td>
<td>4.63 (0.15)</td>
<td>5.12 (0.12)</td>
<td>6.72 (0.15)</td>
</tr>
</tbody>
</table>

In AD, a (semantic) understanding of normal examples is extracted from a set $S_n = \{x_i\}_{i=1}^{n}$ assumed to have been sampled i.i.d. from the normal data distribution $P^+$ over some space $\mathcal{X}$. The goal is to learn a one-class model $f_\theta : \mathcal{X} \rightarrow [0, 1]$ with parameters $\theta \in \Theta$ that decides whether a previously unseen $x \in \mathcal{X}$ is normal (s.t. $f_\theta(x) \approx 0$) or anomalous ($f_\theta(x) \approx 1$).

The way $S_n$ is used to learn $f_\theta$ defines how different approaches in the AD literature can be categorized, e.g. in an unsupervised way (Ruff et al., 2018), or through self-supervision (Golan & El-Yaniv, 2018) (c.f. Section 2). The concept of outlier exposure (OE) (Hendrycks et al., 2019b) utilizes a large number of unlabelled images from some unstructured corpus of data $Q_m$ (where commonly $m \gg n$), for example 80 Million Tiny Images (Torralba et al., 2008), on which models are trained to identify whether samples belong to the corpus or the normal data $P^+$. Importantly, this is a form of weak supervision via existing resources (Zhou, 2018), and not equivalent to supervised classification: images from the auxiliary corpus are not necessarily true anomalies (and may even contain samples from $P^+$). For an $N$-sized batch of samples, the associated learning objective can be formulated as

$$\arg\min_\theta \left\{ L[f_\theta] = L_{S_n}[f_\theta] + L_{Q_m}[f_\theta] = \frac{1}{N} \left[ \sum_{x \in S_n} \log f_\theta(x) + \sum_{x \in Q_m} \log(1 - f_\theta(x)) \right] \right\}. \quad (1)$$

Recent state-of-the-art AD methods have proposed to modify this objective by using radial functions Ruff et al. (2020a), which is in line with the so-called concentration assumption common in AD (Schölkopf & Smola, 2002; Steinwart et al., 2005). We include such radial functions in our ablation (Table 4), however empirically observed that – when paired with an explicit inductive bias – standard classifiers typically performed better.

4.2. Transfer-Based AD

Following work that investigated the prospects of large pretrained networks (Zamir et al., 2018; Adhikari et al., 2019; Hendrycks et al., 2020), a recent study proposed carrying out AD through a nearest neighbor search on top of features extracted from a large pretrained residual network (Bergman et al., 2020). However as can be seen from the experimental results in Table 5, simply transferring over fixed representations to an unrelated task seems subpar for semantic AD. Next, we outline how parameters obtained from some pretraining task can be more effectively transferred to AD.
Pretraining itself tends to follow a simple standard protocol: a model’s parameters are randomly initialized with some distribution, for example Xavier initialization (Glorot & Bengio, 2010). Optimization of a suitable transfer task $T$ (e.g. ImageNet object classification) yields a set of general-purpose parameters $\theta_0 \in \Theta$. The so-obtained model $f_{\theta_0}$ is then ready to be transferred to some downstream task $B$.

The traditional methodology for leveraging pretrained models is to continue to optimize the model parameters (or a subset thereof) on $B$. One crucial limitation of this learning protocol is that when learning on $B$ isn’t carried out carefully through the introduction of some explicit inductive bias (Li et al., 2018), this risks catastrophic forgetting of information previously extracted from $T$. To alleviate this issue, a common approach is to use regularization, e.g. as in continual learning (Kirkpatrick et al., 2017; Lopez-Paz & Ranzato, 2017). In the following two sections, we introduce two new AD-specific learning methods that prevent catastrophic forgetting.

4.3. Anomaly Detection with an Inductive Bias

Due to the complex ways in which semantic anomalies may manifest themselves in images, we hypothesize that our best bet for robust semantic AD is to introduce a semantic inductive bias into models.

To achieve this, we augment the learning criterion introduced in eq. (1) with an additional regularizer $\Omega$: $\Theta \rightarrow \mathbb{R}_+$ that constrains models, resulting in the following objective:

$$\arg \min_{\theta \in \Theta} \{ \mathcal{L}_{S_n} [f_{\theta}] + \mathcal{L}_{Q_m} [f_{\theta}] + \Omega(\theta) \}. \quad (2)$$

As our ablations (Table 4) show, an inductive bias such as $L_2$ regularization toward initial pretrained parameters $\theta_0 \in \Theta$ is crucial for robust semantic AD performance. Motivated by this finding, in Anomaly Detection with an Inductive Bias (ADIB) we set $\Omega(\theta) = \alpha \| \theta - \theta_0 \|^2$ scaled by $\alpha \in \mathbb{R}_+$.

We find that ADIB outperforms previous state-of-the-art AD methods on semantic anomaly benchmarks. For the CIFAR-10 semantic AD benchmark, for example, it raises the state of the art to 74.6 versus 41.6 mean AP reported previously by Ahmed & Courville (2020). Moreover, ADIB sets a new state of the art on the widespread one-versus-rest AD benchmark, raising the bar from 96.1 (Ruff et al., 2020a) to 99.1 mean AUC.

4.4. Anomaly Detection with Residual Adaptation

Regularization can also be formulated to bolster parameter efficiency. For this, we propose constraining the underlying generating function of residual networks (He et al., 2016) $\Phi(x) = x + f(x)$ to allow at most a linear change from the pretrained mapping $\Phi_0$ with $f_0$ in every layer, whereby $\Phi(x) - \Phi_0(x) = Vx$. This is then rearranged to:

$$\Phi(x) = x + f_0(x) + Vx,$$

where $V$ linearly corrects from adjacent layers (and can be implemented via 1x1 convolutions), and $f_0$ is the residual $3x3$ convolution obtained from some transfer task $T$. Assuming that pretrained models will have obtained strong general-purpose representations that should require only minimal changes to adapt to new tasks, only the $V$ are learned, while $f_0$ is left unchanged. Similar strategies have been used in multi-task (Rebuffi et al., 2017; 2018; Deecke et al., 2020) and NLP (Stickland & Murray, 2019) settings to restrict the number of learnable parameters there.

We apply this strategy in Anomaly Detection with Residual Adaptation (ADRA), and our experiments in Section 5 demonstrate that its performance is often comparable to that of regularizing all parameters via $\Omega(\theta)$. At the same time, as only $V$ gets learned at each layer, ADRA is highly efficient. Such savings are crucial for applications in which multiple normal datasets exist but memory footprints are restrictive (e.g. federated learning scenarios (Yang et al., 2019; Bhagoji et al., 2019)).

5. Experiments

For evaluation, we first propose a novel experiment that is based on disentanglement datasets that have been introduced recently (Section 5.1). We then evaluate ADIB and ADRA on two benchmark settings: semantic AD (Section 5.2), and the widely adopted one-versus-rest AD (Section 5.3). All experiments have been implemented with PyTorch (Paszke et al., 2019).

5.1. Examining Models through Interventions

As previous authors have emphasized, curating datasets with semantic anomalies is challenging (Ahmed & Courville, 2020). We here propose to achieve this via datasets originally developed for disentangled representations (Kulkarni et al., 2015; Higgins et al., 2017; Bouchacourt et al., 2018; Burgess et al., 2018; Chen et al., 2018; Kim & Mnih, 2018; Kumar et al., 2018; Locatello et al., 2019; 2020) that contain underlying ground-truth factors of images, in particular high-resolution, realistic datasets such as the recently released MPI3D (Gondal et al., 2019). In contrast to previous evaluations for semantic AD, for example those that modify CIFAR-10 to such a task (Ahmed & Courville, 2020), interventions on ground-truth factors allow for principled measurements of semantic capabilities of a given model, as for example the color of an object can be changed in a systematic fashion.

\footnote{Code available at https://github.com/VICO-UoE/TransferAD.}
Transfer-Based Semantic Anomaly Detection

MPI3D contains joint pairs of latent ground-truth factors $z$ (color, shape, angle, etc.), and corresponding images $x_z$ of a robot arm mounted with an object. The original dataset comes in three styles (photo-realistic, simple, or detailed animation); because the models we evaluate use rich deep architectures, we skip evaluation on simple and animated images (which are useful for simpler models) and focus on the photo-realistic images here.

All models use the same number of parameters, and differ only in which AD loss is optimized — DSVDD (Ruff et al., 2018) uses eq. (1) without any weak supervision (no $L_{Q_m}[\theta]$ term). SAD (Ruff et al., 2020a;b) differs from DSVDD only in that it uses OE. Our models combine both OE and an inductive bias, see eq. (2). To ensure fair comparison, we use the exact same ResNet26 for all methods, and initialize all of them in exactly the same way, i.e. with the same pretrained weights. Note however our proposed ADRA has less modeling power than DSVDD and SAD, due to having fewer learnable parameters.

For semantic AD experiments on MPI3D, we propose fixing a red cone as the normal object (chosen arbitrarily), and train models on all available views. Anomalies are obtained by interventions on three underlying factors: (i) changing color to blue, (ii) transforming shape to cube, and (iii) increasing size. Two additional degrees of freedom exist in the dataset: background color and camera height. Interventions on these have an outsized impact on images however, and do not provide any real challenge to a residual network (or any other modern vision architecture, for that matter), which is why we do not consider them here.

For weak supervision through OE we use all remaining images that do not belong to neither the normal nor the anomaly class. For example, white, green, brown, and olive all appear in the corpus $Q_m$.

**Optimization** The underlying model for DSVDD, SAD, ADRA, ADIB is the exact same ResNet26, optimized via stochastic gradient descent (momentum parameter of 0.9, weight decay of $10^{-4}$) for a total of 100 epochs, with learning rate reductions by 1/10 after 60 and 80 epochs. The batch size is fixed to 128, and we only use standard augmentations. For all models, we initialize parameters via $\theta_0$ obtained from pretraining on ImageNet and then train them further on the downstream AD task.

In ADRA only linear corrections $V$ are learned, while $\theta_0$ is fixed. For an explicit inductive bias in ADIB, we scale the regularization term $\Omega(\theta)$ with $\alpha = 10^{-2}$, as recommended by Li et al. (2018). Results are averaged over 5 seeds.

**Results** AUCs for different interventions are displayed in Fig. 1. Detecting even the most simple semantic anomaly, such as a change in object color $z_{\text{color}} = \text{red} \rightarrow \text{blue}$ is impossible when learning without any weak supervision, as is the case for DSVDD (11.7 AUC).

Our proposed intervention protocol confirms that it is beneficial to introduce a concept of differentness via OE. In other words, exposing models to the concept of red being normal, while also showing it examples of other colors (brown, green, etc.) prepares the model for potential anomalous shifts — although SAD has never seen a blue example, OE enables it to identify it as “not red”, and hence an anomaly.

To obtain more robust models that can pick up on less obvious interventions such as changing the shape $z_{\text{shape}} = \text{cone} \rightarrow \text{cube}$ or $z_{\text{size}} = \text{small} \rightarrow \text{large}$, adequate forms of regularization appear to be critical. While it has fewer learnable parameters, ADRA improves performance over SAD under all interventions. Some performance gap remains, however, which is likely a consequence of the parameter-efficiency of ADRA, letting it rely more on the weights of the base network which potentially aren’t particularly well suited for the task.

ADIB has a higher degree of flexibility, thus allowing for sample-efficient utilization of those features which are useful from the pretrained network. While ADIB might be a simple strategy for the transfer of rich semantic features to AD, the performance under all three interventions shows that it can robustly detect semantic anomalies.

We finally note that weak supervision through OE consistently increased disentanglement in the learned representations. DCI disentanglement (Eastwood & Williams, 2018) almost doubles from 0.068 for DSVDD to 0.103 for SAD, their only distinction being the absence and presence of weak supervision via $Q_m$, respectively. Locatello et al. (2020) made a similar observation in the context of unsupervised learning, finding that some weak supervision is required for disentanglement.

**Non-Semantic Shift** Recent work examined model robustness toward non-semantic shift, such as the appearance of color not contained in the training data, which can confuse models from their primary objective of detecting se-
mantic categories (Ahmed et al., 2021). In order to examine this setting on MPI3D we (w.l.o.g.) set \textit{cube} = anomalous and \textit{cone} = normal and consider color a non-semantic factor.

The experiment consists of a controlled sequence of trials: a single color (red) is included in the normal data at first, and the detection performance of models for \textit{cone} vs. \textit{cube} (of any color) is evaluated. Then another color is picked and added to the normal data (which now contains red & blue), and models are evaluated again. Repeating this for green, white, etc. yields \( P_{\text{red}} \), \ldots, \( P_{\text{all}} \) and gives precise control over the degree in which semantic context may be established.

Fig. 2 shows the extent to which our transfer-based approach improves robustness to non-semantic shifts and underlines the importance of preventing drift from the transfer task. SAD makes use of OE (which in this experiment includes shapes other than \textit{cone} and \textit{cube}, but never additional colors) and enhances performance relative to DSVDD (which does not use OE). A gap remains however when context is established only through OE (SAD vs. ours). Especially for few colors in \( P^+ \) transfer-based AD appears very useful to manifesting the right semantic context.

5.2. Semantic AD

In this section, we evaluate recently proposed benchmarks for semantic AD (Ahmed & Courville, 2020). This setup is equivalent to that presented in our motivation (Section 3), but here we evaluate a broad range of recent state-of-the-art AD models.

In the CIFAR-10 and STL-10 semantic AD benchmarks 9 out of 10 object classes form the normal data \( S_n \) (e.g. all classes except dogs), so that images from multiple classes form a multi-modal normal distribution \( P^+ \). The single class that is left out (i.e. dogs) is declared anomalous and never seen during training. At test time, the AD model has to identify the held-out class, i.e. we measure whether \( f_\theta(x) \approx 1 \) when \( x \) contains a dog. This requires that the AD model has a good semantic understanding of the objects in the normal distribution \( P^+ \), and this benchmark has been shown to be more difficult than the popular one-versus-rest AD benchmarks (Ahmed & Courville, 2020; Bergman et al., 2020) (which we also evaluate in Section 5.3).

Ahmed & Courville (2020) determine semantic anomalies via MSP (Hendrycks & Gimpel, 2017) and ODIN (Liang et al., 2018) using an auxiliary self-supervised criterion akin to RotNet (Gidaris et al., 2018), while Bergman et al. (2020) use a nearest neighbor search over fixed pretrained features. We include all existing results in Table 2.

**Optimization** As before, for ADIB we set \( \alpha = 10^{-2} \) following the suggestion of Li et al. (2018); in elastic weight consolidation (EWC) we set the Fisher multiplier to 400, as recommended by Kirkpatrick et al. (2017). For experiments on CIFAR-10 (Krizhevsky & Hinton, 2009), we introduce an inductive bias by regularizing the network weights towards those of ResNet26 trained on ImageNet at 32x32 resolution. We use the same architecture for STL-10 (Coates et al., 2011), but since images have a higher resolution the initial model weights were obtained by training on ImageNet at a resolution of 96x96.

For eqs. (1) and (2) we contrast \( S_n \) against images from an unstructured corpus \( Q_m \). Following previous work that uses OE (Hendrycks et al., 2019b; Ruff et al., 2020a), for CIFAR-10 we fix this to contain all samples from the CIFAR-100 training split. As already emphasized, \( Q_m \) equals weak supervision: CIFAR-100 gives a viable surrogate learning signal, however does not contain examples of the anomalous CIFAR-10 categories. STL-10 contains a large unlabeled split, which we use for OE.

**Results** There are discrepancies in how performance is reported in the semantic AD literature: some authors recommend average precision (AP) (Ahmed & Courville, 2020), while others report AUC (Bergman & Hoshen, 2020). We include both metrics in Table 2, and report AP for STL-10 (Table 3) as this benchmark was so far only evaluated by Ahmed & Courville (2020) who report AP as AUC is overly optimistic for the STL-10 semantic AD benchmark (Davis & Goadrich, 2006).

On the CIFAR-10 semantic AD benchmark, ADIB outperforms the previously reported methods by a substantial margin: 74.6 \text{ vs.} 41.6 \text{ mAP}, and 95.1 \text{ vs.} 71.7 \text{ mAUC}. Even though it requires a smaller number of learnable parameters ADRA comes very close: 95.0 mAUC, and 72.9 mAP. As our results confirm, inferring anomalies on STL-10 is significantly harder. In particular, even when using a state-of-the-art HSC classifier (Ruff et al., 2020a) initialized with pretrained \( \theta_0 \) but without regularization \( \Omega(\theta) \), this does not successfully address the semantic AD task (mAP of 27.7, etc.)
Table 2. Results on the CIFAR-10 semantic AD benchmark; GT reported in Bergman & Hoshen (2020).

<table>
<thead>
<tr>
<th>Class</th>
<th>ODIN (Ahmed &amp; Courville, 2020)</th>
<th>HSC</th>
<th>ADRA (ours)</th>
<th>ADIB (ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Airplane</td>
<td>23.4</td>
<td>23.1</td>
<td>49.3</td>
<td>41.4</td>
</tr>
<tr>
<td>Bird</td>
<td>40.1</td>
<td>13.8</td>
<td>18.9</td>
<td>44.0</td>
</tr>
<tr>
<td>Car</td>
<td>16.9</td>
<td>39.9</td>
<td>74.6</td>
<td>72.2</td>
</tr>
<tr>
<td>Cat</td>
<td>31.4</td>
<td>18.9</td>
<td>29.6</td>
<td>51.0</td>
</tr>
<tr>
<td>Deer</td>
<td>29.7</td>
<td>25.3</td>
<td>20.7</td>
<td>43.0</td>
</tr>
<tr>
<td>Dog</td>
<td>26.1</td>
<td>17.3</td>
<td>26.6</td>
<td>32.2</td>
</tr>
<tr>
<td>Horse</td>
<td>23.6</td>
<td>30.1</td>
<td>52.5</td>
<td>53.7</td>
</tr>
<tr>
<td>Monkey</td>
<td>28.3</td>
<td>18.4</td>
<td>23.0</td>
<td>46.6</td>
</tr>
<tr>
<td>Ship</td>
<td>15.4</td>
<td>49.2</td>
<td>69.2</td>
<td>51.7</td>
</tr>
<tr>
<td>Truck</td>
<td>16.6</td>
<td>40.7</td>
<td>46.3</td>
<td>58.7</td>
</tr>
<tr>
<td>mAP</td>
<td>25.1</td>
<td>27.7</td>
<td>42.9</td>
<td>49.5</td>
</tr>
</tbody>
</table>

Table 3. APs on the STL-10 semantic AD benchmark.

Ablation The transfer of features from rich semantic tasks to AD has to be carried out carefully. We examine this in an ablation shown in Table 4, for which we use the exact same model in each experiment \(a_1\)–\(a_7\), and only switch on and off individual components: starting from random \(a_1\) or pretrained models without regularization \(a_2\) is not sufficient, as also highlighted in our intervention experiments in Section 5.1. Using an HSC loss (\(\mathcal{L}\)) with the exact same explicit inductive bias through \(\Omega(\theta)\) that we use in ADIB reduces performance \(a_3\) vs. \(a_7\). DOC (Perera & Patel, 2019) is conceptually very similar to HSC, combining a radial compactness loss with a descriptiveness loss that requires ImageNet data. Our result \(a_3\) vs. \(a_4\) confirms they also behave very similarly performance-wise.

Table 4. Ablations in terms of mAP. Tr. indicates absence or presence of transfer learning; Reg. that of regularization. Included are comparisons against hyperspherical classifiers (Ruff et al., 2020a) in \(a_3\), and EWC (Kirkpatrick et al., 2017) in \(a_5\).

<table>
<thead>
<tr>
<th>(\mathcal{L})</th>
<th>Tr.</th>
<th>Reg.</th>
<th>CIFAR-10</th>
<th>STL-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a_1) eq. (1)</td>
<td>(\times)</td>
<td>(\times)</td>
<td>60.3</td>
<td>32.9</td>
</tr>
<tr>
<td>(a_2) eq. (1)</td>
<td>(\checkmark)</td>
<td>(\times)</td>
<td>64.9</td>
<td>38.6</td>
</tr>
<tr>
<td>(a_3) HSC</td>
<td>(\checkmark)</td>
<td>(L_2)</td>
<td>68.5</td>
<td>35.0</td>
</tr>
<tr>
<td>(a_4) DOC</td>
<td>(\checkmark)</td>
<td>(\times)</td>
<td>65.8</td>
<td>35.2</td>
</tr>
<tr>
<td>(a_5) eq. (2)</td>
<td>(\checkmark)</td>
<td>(\text{EWC})</td>
<td>66.4</td>
<td>39.7</td>
</tr>
</tbody>
</table>

In \(a_5\) we find that EWC, a popular strategy for continual learning that regularizes weights via the Fisher information (Kirkpatrick et al., 2017), performs poorly compared to ADIB. This makes perfect sense, as EWC was designed to slow down learning on model weights relevant for the pretraining task: when pretraining on a demanding task like ImageNet, this can restrain capacity. Crucially for our scenario we need model capacity to free up and focus on semantic AD instead. In other words, as we never return the model to ImageNet classification, there simply is no good reason why we would want to preserve it. This ablation shows that, while they may be simple, our proposed strategies are surprisingly effective AD strategies.

Qualitative Analysis In Figure 3 we visualize the high semantic association of the representation learned by ADIB, and compute feature embeddings for samples from CIFAR-10, mapped to two dimensions using t-SNE (van der Maaten & Hinton, 2008). The representation was learned on the CIFAR-10 semantic AD setup, i.e. trained on a multiclass \(\mathcal{P}\) that contains 9 out of 10 classes •cat, •dog, etc.). At test time the singular anomalous category •bird gets revealed to the model.

While ADIB has no access to semantic categories or labels, it organizes the feature space in a highly semantic manner: •deer and •horses are semantically similar and cluster together, so do •cats and •dogs, while •frogs are separated from the other animals. Moreover, man-made objects such as •cars, •trucks, etc. are clearly separated from animal categories. This matches human intuition.

Even though the anomalous •bird category is never seen during training, it is located near other animals. We display one bird that has a similar feature representation to man-made objects: for this image, it is indeed difficult to identify it as a bird, and it should be no surprise that it is located far away from the •-cluster in feature space.

In Figure 4 we display examples from STL-10 that have been assigned a high anomaly score by ADIB. The anomaly...
Figure 3. ADIB learns a feature space that is semantically meaningful. It separates objects that occur naturally (cats, dogs, etc.) from man-made ones (airplanes, motorcycles). It even locates examples from the unseen bird category nearby other animals. The arrow highlights a bird that gets mapped close to man-made objects, and identifying it as one indeed requires a fair bit of imagination.

Figure 4. Anomalous examples from STL-10.

Table 5. AUCs for different methods on the CIFAR-10 one-versus-rest AD benchmark. Included are geometric transformations (GT) (Golan & El-Yaniv, 2018), kNN-AD (Bergman et al., 2020), self-supervised transformations (GT+) (Hendrycks et al., 2019c), and hyperspherical classifiers (HSC) (Ruff et al., 2020a).

<table>
<thead>
<tr>
<th>Class</th>
<th>GT</th>
<th>kNN</th>
<th>GT+</th>
<th>HSC</th>
<th>ADRA</th>
<th>ADIB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Airplane</td>
<td>74.7</td>
<td>93.9</td>
<td>90.4</td>
<td>96.7</td>
<td>99.0</td>
<td>99.2</td>
</tr>
<tr>
<td>Automobile</td>
<td>95.7</td>
<td>97.7</td>
<td>99.3</td>
<td>98.9</td>
<td>99.7</td>
<td>99.8</td>
</tr>
<tr>
<td>Bird</td>
<td>78.1</td>
<td>85.5</td>
<td>93.7</td>
<td>93.2</td>
<td>97.5</td>
<td>98.6</td>
</tr>
<tr>
<td>Cat</td>
<td>72.4</td>
<td>85.5</td>
<td>88.1</td>
<td>90.6</td>
<td>96.3</td>
<td>97.0</td>
</tr>
<tr>
<td>Deer</td>
<td>87.8</td>
<td>93.6</td>
<td>97.4</td>
<td>97.1</td>
<td>98.9</td>
<td>99.3</td>
</tr>
<tr>
<td>Dog</td>
<td>87.8</td>
<td>91.3</td>
<td>94.3</td>
<td>94.7</td>
<td>97.7</td>
<td>98.2</td>
</tr>
<tr>
<td>Frog</td>
<td>83.4</td>
<td>94.3</td>
<td>97.1</td>
<td>98.0</td>
<td>99.6</td>
<td>99.6</td>
</tr>
<tr>
<td>Horse</td>
<td>95.5</td>
<td>93.6</td>
<td>98.8</td>
<td>97.9</td>
<td>99.6</td>
<td>99.8</td>
</tr>
<tr>
<td>Ship</td>
<td>93.3</td>
<td>95.1</td>
<td>98.7</td>
<td>98.2</td>
<td>99.5</td>
<td>99.6</td>
</tr>
<tr>
<td>Truck</td>
<td>91.3</td>
<td>95.3</td>
<td>98.5</td>
<td>97.7</td>
<td>99.4</td>
<td>99.5</td>
</tr>
</tbody>
</table>

mAUC   | 86.0| 92.5| 95.6| 96.3| 98.7 | 99.1 |

Table 6. Ablations on the CIFAR-10 one-versus-rest AD benchmark for different choices of OE. Relative gain (vs. HSC) displayed in parentheses.

<table>
<thead>
<tr>
<th>OE Dataset</th>
<th>HSC</th>
<th>ADRA</th>
<th>ADIB</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVHN</td>
<td>70.2</td>
<td>75.3</td>
<td>(+5.1)</td>
</tr>
<tr>
<td>CIFAR-100</td>
<td>96.3</td>
<td>98.7</td>
<td>(+2.4)</td>
</tr>
</tbody>
</table>

5.3. Non-Semantic AD

We evaluate the performance of ADIB and ADRA on the standard CIFAR-10 one-versus-rest AD benchmark, which has recently been deemed a non-semantic problem by Ahmed & Courville (2020). While this is a less complex benchmark that can be solved using shallower feature representations, it is reported across large parts of the AD literature (Ruff et al., 2018; Deecke et al., 2018; Golan & El-Yaniv, 2018; Hendrycks et al., 2019b; Abati et al., 2019; Hendrycks et al., 2019c; Perera et al., 2019; Bergman & Hoshen, 2020; Ruff et al., 2020b:a) and therefore is still meaningful for comparison of our proposed methods to previous AD models.

In some sense, this benchmark can be viewed as opposite of semantic AD: only a single object class is fixed as the normal class — say, dogs. All dogs in the CIFAR-10 training split are collected into $S_n$ (so 5 000 out of 50 000 total samples), from which models are trained. Models are then evaluated against the entire CIFAR-10 test split, and performance is measured by checking whether anomaly scores assigned to dogs are lower than scores assigned to all nine remaining non-dog classes.

For this benchmark previous works almost exclusively report AUC, and we follow this custom here. Note that the optimization settings remain unchanged from those in Section 5.2.

Results As shown in Table 5, ADIB raises the current state of the art to 99.1 mAUC, a marked gap to the previous best method with 96.1 mAUC. As demonstrated by the performance of kNN-AD (Bergman et al., 2020), simply using features from a large pretrained network is inferior when looking to detecting anomalies.

These results suggest that favorable inductive biases are critical for utilizing AD models to their full potential. ADRA once again comes very close in terms of performance, while requiring a much smaller number of learnable parameters.

Ablation Recent work examined the hierarchical relationship between distributions for out-of-distribution detection (Schirrmeister et al., 2020). We take inspiration from their work and critically examine the role of CIFAR-100 as OE
in an ablation that compares it to the use of SVHN as OE.\footnote{While some previous work such as Hendrycks et al. (2019b) or Hendrycks et al. (2019c) used 80 Million Tiny Images instead of CIFAR-100 for OE, the dataset was withdrawn and further use has been discouraged by the authors.}

Results in Table 6 make it evident that SVHN is less well suited for CIFAR-10, as performance drops for all methods. For HSC (the current state-of-the-art AD method using OE) we obtain 70.2 mAUC here. A sizeable drop, but still improving from 64.8 mAUC for DSVDD (the mathematical equivalent to using no OE).

Our method obtains 79.8 mAUC when coupled with SVHN, a gain of +9.6 over HSC. This is a considerably larger difference than the gain for CIFAR-10 coupled with CIFAR-100 of +2.8 reported in Table 5 (ours: 99.1 vs. HSC: 96.3 mAUC), indicating that transfer-based AD benefits performance more when not using CIFAR-100 as OE (albeit it is better suited overall). In other words, the importance of the transfer task increases as the suitability of OE decreases.

5.4. Robustness to Small Modes

An ideal AD model has the ability to incorporate information from normal examples even if they form only a minor mode of $P^+$, in the sense that only few samples from this class are contained in $S_n$—for example a rare dog breed. Since AD is concerned with low-probability events, the ability to robustly incorporate such small modes from few examples is of special importance.

To measure AD robustness, we let the normal class be constituted by samples associated with two classes $(y_a, y_b)$, such that $S_n \sim \frac{1}{r+1}P_{y_a} + \frac{r}{r+1}P_{y_b}$, where the minor mode amplitude $r \in [0, 1]$ controls the number of examples from $y_b$ in the normal data. For a robust AD model, even as $S_n$ is relaxed to contain only examples from $y_a$, its ability to identify the smaller category $y_b$ as non-anomalous would remain intact.

We use CIFAR-10 here, and report primary and secondary AUCs as a function of $r$ for $y_a =$ “ship” and $y_b =$ “truck” in Figure 5. We compare our methods to SAD (Ruff et al., 2020b) with pretrained weights, which corresponds to ADIB with $\alpha = 0$, i.e. without a regularization term $\Omega(\theta)$.

For SAD performance for the secondary class decreases much faster than for our methods. This trend is consistent across class pairings (more pairs are displayed in the supplements), and indicates that adequate transfer-based regularization as in ADRA and ADIB is crucial to robustly incorporating small modes of data in AD.

6. Conclusion

Detecting semantic anomalies is a difficult task, due to the infinite and complex ways these can manifest themselves. We have proposed two methods to account for such complexities: ADIB sets a new state of the art in semantic AD tasks and ADRA provides a highly efficient, yet surprisingly effective learning protocol.

We used interventions to show that our methods can detect subtle semantic anomalies, and verified that ADIB and ADRA offer high AD robustness. An interesting question for future research is whether detecting anomalies requires disentanglement, and if it can benefit from the ongoing development of disentangled representations.

7. Acknowledgments

We thank the anonymous reviewers for their help in improving this paper. LR acknowledges support by the German Federal Ministry of Education and Research (BMBF) in the project ALICE III (01IS18049B). RV acknowledges support by the Berlin Institute for the Foundations of Learning and Data (BIFOLD) sponsored by the German Federal Ministry of Education and Research (BMBF). HB is funded by the EPSRC programme grant Visual AI EP/T028572/1.

References


