Brittle Features May Help Anomaly Detection

Kimberly T. Mai, Toby Davies, Lewis D. Griffin

University College London

Outcomes

- For good anomaly detection performance, representations are more important than the method.
- In instances where the representation is suitable for anomaly detection, knowledge distillation outperforms anomaly detectors that use the representation directly.
- Good representations require both separation of anomalies and normal data and the use of brittle features in the representation space.

Motivation

One-class anomaly detection uses only normal data at training time to build a representation.

- The novelty of a test datum is evaluated by comparing its features to the exemplar representation.
- Without some guidance about the nature of anomalies, it is difficult to learn a representation that distinguishes anomalies from normal data.
- Prior work indicates transfer learning is beneficial for anomaly detection as it provides guidance about anomalous characteristics, but it is unclear which representations are better candidates for transfer learning.

Figure 1: Examples of normal (left) and anomalous (right) X-ray parcel images. One-class anomaly detection is useful for detecting suspect items in X-ray images due to the lack of anomalous training examples. Previous work transferred features from ImageNet for anomaly detection, but it is unclear whether this is the best representation due to the domain shift from natural images.

Method

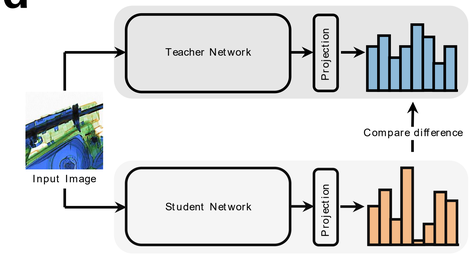


Figure 2: Schematic of the training architecture.

- Configure the anomaly detector as a knowledge distillation task and use mean squared error as the anomaly score.
- Train a student to match the internal representations of a frozen teacher network which is pre-trained on an auxiliary task.
- Representations from the student and teacher should agree for normal images and differ for anomalous images.

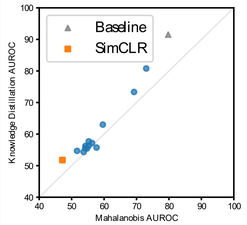
Evaluating Representations

- Knowledge distillation performance is compared to Mahalanobis scoring which uses the same auxiliary representations.
- The brittleness of auxiliary representations is compared using average L2 gradient norms of the normal training data with respect to the student network:

$$\frac{\mathbb{E} \parallel \partial_{x_{train}} L \parallel_2}{\mathrm{tr}(\Sigma_{train})}$$

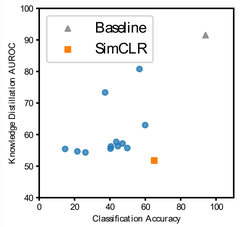
- L2 gradient norms are divided by the trace of the covariance matrix to account for the spread of different representations.

Results



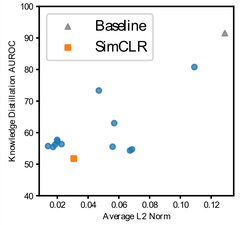
Anomaly detection performance is more dependent on the underlying representation than the anomaly detector.

- There is a clear correlation between knowledge distillation performance and Mahalanobis performance across all representations.



Separability between anomalies and normal data is not the sole factor for good anomaly detection.

- Correlation between classification accuracy and knowledge distillation is weaker.
- Although the SimCLR representation results in good classification performance, it is a poor representation for anomaly detection.



Better representations for anomaly detection are also correlated with larger average gradient norms.

- Corresponding to its knowledge distillation performance, SimCLR has a lower average L2 norm compared to better-performing representations.
- This suggests brittle features may be useful for indicating anomalies.

Figures 3-5: Scatter plots depicting the performance of different auxiliary representations on CIFAR-10, averaged over all configurations.

We also outperformed previous anomaly detection performance results on the X-ray anomaly detection dataset, increasing performance from 92.8% to 96.4% AUROC.

Contact: kimberly.mai@ucl.ac.uk

Women in Computer Vision (WiCV) @ CVPR 2021