**Neural Transformation Learning for Deep Anomaly Detection Beyond Images**

In self-supervised learning, data transformations are crucial. Images are frequently translated into many viewpoints, and neural networks trained on tasks requiring multiple views offer relevant feature representations for subsequent tasks, such as anomaly detection. The notion of data augmentation is the foundation of many recent advancements in anomaly detection. Predefined transformations like rotations, reflections, and cropping are employed in the self-supervised setting, particularly for image data, to produce different views of the data. However, it is frequently unclear which transformations to employ for anomaly detection beyond image data, such as time series data and tabular data. This study explores self-supervised anomaly detection for non-image data types. Here a neural transformation learning for anomaly detection is proposed. A simple end-to-end process for anomaly detection called NeuTraLAD uses learnable transformations is developed. A single objective function is designed for jointly learning useful data transformations and anomaly thresholding, which eliminates the need for manually designing data transformations to build auxiliary prediction tasks that may be used for anomaly detection.

A fixed set of learnable transformations and an encoder model are the only two main components of NeuTraLAD. A noise-free, deterministic contrastive loss (DCL) that is intended to learn faithful transformations is used to jointly train both elements. Other contrastive losses used in representation learning and image anomaly detection use negative samples from a noise distribution. However, DCL differs in this aspect. The suggested technique provides a non-stochastic objective that doesn't require any further regularization or adversarial training and can be utilized straight as the anomaly score. DCL is tuned during training to determine the encoder's parameters and transformations. The DCL is also utilized to classify each sample as an inlier or an anomaly during testing. The assumption is that the transformations are learnable, that is, that any parameterized function with parameters available for gradient-based optimization can represent them. Each transformed sample is encouraged by the DCL to resemble its original sample while also being encouraged to differ from other transformed versions of the same sample. The numerator in the loss function brings each converted sample's embedding relatively near to the original sample's embedding. This facilitates the preservation of essential semantic information by the transformations. All the embeddings of the altered samples are pushed apart by the denominator in the loss function, which promotes varied transformation. The proposed methodology has the advantage over existing approaches in that training loss is also the anomaly score. The score for training examples (inliers) is minimized by minimizing the DCL. The probability that a sample has an abnormality increases with the anomaly score. The model consists of three residual blocks of 1d convolutional layers with instance normalization layers and ReLU activations, with one additional 1d convolutional layer on top of each stack of three residual layers. The masks are given a sigmoid activation for the multiplicative parameterization. The instance normalization layers' learnable affine parameters are frozen, and all bias terms are fixed to zero.

The suggested method for self-supervised anomaly identification includes learnable transformations. The important component is a novel training objective built on a deterministic contrastive loss that promotes the learnt changes to generate a variety of perspectives that are distinct from one another while still sharing semantic information with the original sample. Several empirical studies reveal that NeuTraL AD outperforms state-of-the-art methods for learning transformations and identifying outliers on several data types. The main contribution of this work is that DCL can be used as a loss function and as a score. The score can be directly evaluated for new data points without the use of negative samples because it is deterministic. Negative samples are created deterministically from sample space rather than being randomly selected from a noise distribution.

**Investigating Why Contrastive Learning Benefits Robustness against Label Noise**

Deep neural networks have achieved significant success in a variety of fields, including vision and natural language processing, thanks to large datasets. However, the effectiveness of the training labels is crucial to this achievement. With the expansion of datasets, manual labeling of data becomes impractical, and the widespread use of web crawling, crowdsourcing, and automated data labeling techniques leads to a proliferation of noisy labels in sizable real-world datasets. Over-parameterized networks that have been trained using first order gradient techniques can fit any labeling of the training data, including random labels. So noisy labels significantly decrease deep models' ability to generalize. The primary strategies employed in traditional work on robust learning from noisy labels are estimating the noise transition matrix, developing robust loss functions, correcting noisy labels, employing explicit regularization techniques, and choosing or reweighting training samples. However, these methods become incredibly ineffective as the noise level rises. There is very little theoretical knowledge about how contrastive learning might increase the robustness of deep networks against noisy labeling. In this work, the limitations of the previous studies by characterizing the beneficial properties of representations obtained by contrastive learning theoretically for enhancing robustness against noisy labels are addressed.

By maximizing agreement between variously augmented views of the same example and decreasing agreement between variously augmented views of various examples, self-supervised contrastive learning learns representations of various data points. A contrastive loss in the latent space is used to achieve this. Certain aspects of the augmentation graph are assumed in order to understand the robustness offered by contrastive learning, and the low-rank structure of the resulting representation matrix is analyzed. The assumptions that formalize the following two properties on the data augmentation are sub-augmented class's examples are like one another, and another sub-augmented class's examples differ from those of other sub-classes, respectively. The lower singular values of the Jacobian are effectively reduced, and the overfitting of the noisy labels is slowed down by pre-training the network with contrastive learning. Existing robust training techniques use the initial level of robustness provided by contrastive learning to produce greater performance under extremely high noise levels. Intuitively, the aforementioned three properties of the representation matrix have an impact on the downstream training in the following ways: (1) the size of the largest singular values determines the rate of evolution and the degree to which the model can fit the training data; (2) the alignment between prominent singular vectors and clean labels shows whether the model evolves in the right direction; and (3) the size of the smaller singular values determines how well the model fits the training data. The components of error are bias and variation. The bias encapsulates the discrepancy between the model's average prediction and the labels assigned to the real world data. It depends on both the size of the salient singular values and how closely the related singular vectors match the labels from the ground truth. By matching the first K singular vectors with the ground truth labels, contrastive learning minimizes the bias and creates a minor second term in the bias. The amount of the non-prominent singular values controls the variance, which measures the sensitivity to label noise. Even in the presence of significant noise, contrastive learning can keep the linear model from memorizing any incorrect labels. The suggested method also demonstrates that when the sub-class structure is more compact, that is, smaller, when the noise is more symmetric, or when the sub-classes are more balanced, the model can withstand more noise.

The proposed method has shown that contrastive learning generates a representation matrix with a prominent singular value corresponding to each sub-class in the data, and a significant alignment between the prominent singular vectors and the ground-truth labels. The analysis of the scenario where a linear model is trained on the acquired representations with labels modified with Gaussian noise or randomly switched to other classes is done. The suggested approach demonstrates that the model can rarely memorize the incorrect labels and that noise has little impact on learning the clean labels. Additionally, it is demonstrated that deep networks that have been trained with contrastive learning beforehand and fine-tuned on noisy labels might initially perform better before overfitting the noise. Finally, it is shown that robust approaches can further benefit from the initial robustness offered by contrastive learning to obtain cutting-edge performance under extremely high noise levels. Even after several training rounds, such approaches prevent the low-rank Jacobian matrix from overfitting the noise.