Student–Teacher Anomaly Detection with Discriminative Latent Embeddings

Anass Denguir

Outline

1. Problem statement
2. Proposed solution
3. Results
4. Conclusions

Outline

1. Problem statement
2. Proposed solution
3. Results
4. Conclusions

What is anomaly detection?

Anomaly detection is a computer vision problem that consists of:

- **segmenting** all the regions of an input image that present anomalies (i.e. defective regions)
- **scoring** each pixel of an input image to output an anomaly map

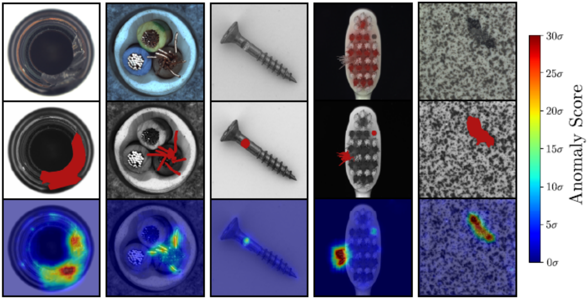

Figure: Anomaly map

Problem motivation

Problem

- State-of-the-art AD algorithms are not well fitted to treat high-resolution images
- Use of shallow machine learning algorithms

Solution

- Bergmann et al. propose a **Student-Teacher** learning framework that leverages the power of deep neural network
- In this approach, anomaly detection is presented as a regression problem where a set of Students networks are trained to mimic a Teacher network

Outline

1. Problem statement
2. Proposed solution
3. Results
4. Conclusions

Student-Teacher Anomaly Detection - Overview

The student-teacher framework is composed of two types of neural networks:

- The **Teacher**, which is trained on a large dataset of images. It will output a description vector for each pixel. These are used as a label for the Students networks.
- The **Students**, which are trained on an **anomaly-free** dataset to mimic the Teacher's output.

The idea behind this approach is that we expect the Students to make poor predictions on images presenting anomalies.

Student-Teacher Anomaly Detection - Illustration

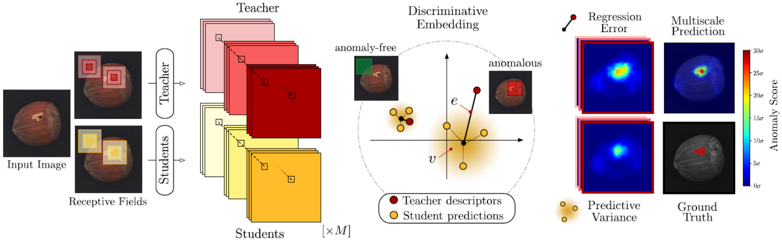

Figure: Student-Teacher Anomaly Detection

Training the Teacher (1/4) - Network Architecture

We train the Teacher network $\hat{T}(p)$ on patches **p** of **fixed size** and we apply a deterministic network transformation to infer $T(I)$ from $\hat{T}$

Layer	Output Size	Parameters	
		Kernel	Stride
Input	65×65×3		
Conv1	61×61×128	5×5	1
MaxPool	30×30×128	2×2	2
Conv2	26×26×128	5×5	1
MaxPool	13×13×128	2×2	2
Conv3	9×9×128	5×5	1
MaxPool	4×4×256	2×2	2
Conv4	1×1×256	4×4	1
Conv5	1×1×128	3×3	1
Decode	1×1×512	1×1	1

Figure: Network architecture of $\hat{T}$ with a receptive field $p = 65$

Training the Teacher (2/4) - Knowledge Distillation

- Let us consider a very deep pre-trained network **P** trained on image classification dataset
- The CNN architecture of *P* provides a very precise (deep) description of each pixel **BUT** at the expense of high time complexity
- Hence, we distill the knowledge of the powerful network *P* into $\hat{T}$ by training a decoded version of $\hat{T}$, i.e $D(\hat{T})$ against *P*:

$$L_k(\hat{T}) = ||D(\hat{T}(p)) - P(p)||^2 \tag{1}$$

where *D* is an extra fully connected layer that is added to match the output dimension of $\hat{T}$ (128) with *P* (512)

Training the Teacher (3/4) - Metric Learning

Alternatively, we can train $\hat{T}$ using self-supervised learning techniques such as **triplet-learning**.

Let us randomly crop a patch p from a training image I and compute the triplet (p, p^+, p^-), where:

- p^+ is obtained after a small translation of p and AWGN
- p^- is a random crop chosen from a different image

The idea is to minimize the following loss function:

$$L_m(\hat{T}) = \max(0, \delta^+ - \delta^- + \delta) \tag{2}$$

$$\delta^+ = ||\hat{T}(p) - \hat{T}(p^+)||^2$$
$$\delta^- = \min(||\hat{T}(p) - \hat{T}(p^-)||^2, ||\hat{T}(p^+) - \hat{T}(p^-)||^2)$$

Training the Teacher (4/4)- Descriptor Compactness

In addition, we want to minimize the correlation between the descriptors $\hat{T}(p)$ of each patch p within a mini-batch.

The goal is to obtain the most compacted feature vector to describe a patch p. To do that, we minimize the correlation $c_i j$ between the feature descriptor of two different patches $\hat{T}(p_i)$ and $\hat{T}(p_j)$

$$L_c(\hat{T}) = \sum_{i \neq j} c_{ij} \tag{3}$$

Training the Teacher - Summary

We train the teacher using the 3 loss functions we defined so far:

$$L(\hat{T}) = \lambda_k L_k(\hat{T}) + \lambda_m L_m(\hat{T}) + \lambda_c L_c(\hat{T}) \tag{4}$$

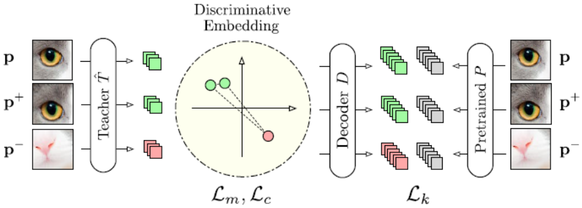

Figure: Teacher training procedure

Training the Students

- Let us train a set of M Student networks $\{S_1, \dots, S_M\}$ on an **anomaly-free** dataset D
- The Students S_i have the same network architecture as the Teacher T and they are **randomly** initialized
- The goal of the Students is to predict the descriptor vector of the teacher $y^T_{(r,c)}$ for each pixel (r, c) of the image I
- Hence, the Student minimize the following loss function:

$$L(S_i) = \frac{1}{wh} \sum_{(r,c)} ||\mu^{S_i}_{(r,c)} - (y^T_{(r,c)} - \mu) diag(\sigma)^{-1}||^2 \qquad (5)$$

$\mu^{S_i}_{(r,c)}$ being the prediction made by S_i for pixel (r, c)

Anomaly Scoring Function (1/2)

For each pixel (r, c), a scoring function is computed to tell how likely this pixel lies in an anomalous region. This scoring function can be broken into two parts:

- the error of the Students prediction mean $\mu_{(r,c)}$ w.r.t Teacher:

$$e_{(r,c)} = ||\mu_{(r,c)} - (y^{T}_{(r,c)} - \mu) diag(\sigma)^{-1}||^2 \tag{6}$$

- the variance of the Students predictions:

$$v_{(r,c)} = \frac{1}{M} \sum_{i=1}^{M} ||\mu^{S_i}_{(r,c)}||^2 - ||\mu_{(r,c)}||^2 \tag{7}$$

Anomaly Scoring Function (2/2)

We combine the two scores $e_{(r,c)}$ and $v_{(r,c)}$ by normalizing them:

$$score_{(r,c)} = \frac{e_{(r,c)} - e_{\mu}}{e_{\sigma}} + \frac{v_{(r,c)} - v_{\mu}}{v_{\sigma}} \quad (8)$$

where the subscripts μ and σ denotes the mean and std over a validation set of anomaly-free images

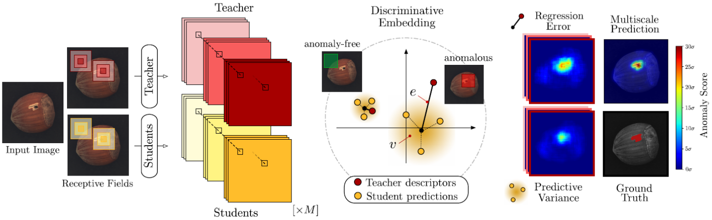

Figure: Student-Teacher Anomaly Detection

Multi-scale resolution

- In practice, anomaly detection performance depends on the size of the receptive field p
- If there is a small anomalous region within a too big receptive field, the description vector might be seen as an anomaly-free region
- the solution is to perform inference at different scales

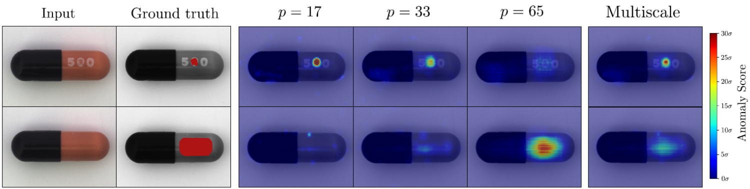

Figure: Multi-scale resolution

Outline

1. Problem statement
2. Proposed solution
3. Results
4. Conclusions

Evaluation metric

- We evaluate the algorithm with the **MVTec** AD Dataset. It contains images with their corresponding ground-truth anomalous regions segmented
- The chosen evaluation metric is the **PRO** (per-region-overlap):
    1. Each pixel is classified into **Anomalous** or **Not anomalous** by comparing its score with a threshold t
    2. The relative overlap between detected anomalous regions and the ground thruth is evaluated
    3. We redo (1) and (2) for lower values of the threshold t until the false-positive rate reaches 30%
    4. We evaluate the area under the PRO curve (normalized to 1)

Results on MVTec

	Category	Ours $p = 65$	1-NN	OC-SVM	K-Means	ℓ_2-AE	VAE	SSIM-AE	AnoGAN	CNN-Feature Dictionary
Textures	Carpet	**0.695**	0.512	0.355	0.253	0.456	0.501	0.647	0.204	0.469
	Grid	0.819	0.228	0.125	0.107	0.582	0.224	**0.849**	0.226	0.183
	Leather	**0.819**	0.446	0.306	0.308	**0.819**	0.635	0.561	0.378	0.641
	Tile	**0.912**	0.822	0.722	0.779	0.897	0.870	0.175	0.177	0.797
	Wood	0.725	0.502	0.336	0.411	**0.727**	0.628	0.605	0.386	0.621
Objects	Bottle	**0.918**	0.898	0.850	0.495	0.910	0.897	0.834	0.620	0.742
	Cable	**0.865**	0.806	0.431	0.513	0.825	0.654	0.478	0.383	0.558
	Capsule	**0.916**	0.631	0.554	0.387	0.862	0.526	0.860	0.306	0.306
	Hazelnut	**0.937**	0.861	0.616	0.698	0.917	0.878	0.916	0.698	0.844
	Metal nut	**0.895**	0.705	0.319	0.351	0.830	0.576	0.603	0.320	0.358
	Pill	**0.935**	0.725	0.544	0.514	0.893	0.769	0.830	0.776	0.460
	Screw	**0.928**	0.604	0.644	0.550	0.754	0.559	0.887	0.466	0.277
	Toothbrush	**0.863**	0.675	0.538	0.337	0.822	0.693	0.784	0.749	0.151
	Transistor	0.701	0.680	0.496	0.399	**0.728**	0.626	0.725	0.549	0.628
	Zipper	**0.933**	0.512	0.355	0.253	0.839	0.549	0.665	0.467	0.703
	Mean	**0.857**	0.640	0.479	0.423	0.790	0.639	0.694	0.443	0.515

Figure: Area under the PRO curve with a FPR limited to 30%

Outline

1. Problem statement
2. Proposed solution
3. Results
4. Conclusions

Conclusions

- Bergmann et al. proposed a Student-Teacher framework for the problem of anomaly segmentation
- Students network are trained to mimic a descriptive Teacher network, that serves as surrogate labels
- Anomaly scores are computed based on the regression error and the variance of the Students network
- The proposed algorithm can be extended to detect anomalies of different scales
- The proposed algorithm outperforms the state-of-the-art on MVTec AD dataset