Workshop Challenge Report

Participant Information

- Name(s):USTC-IAT-United
- Affiliation(s):University of Science and Technology of China
- Contact Information:yrq@mail.ustc.edu.cn

- Track II - VLM Anomaly

Abstract

In this report, we briefly introduce our solution for the second track of the VAND Challenge 2024. With the development of the visual language model VLM, the ability to detect anomalies may reach an exciting new level, such as discovering that it is not only necessary to detect logical anomalies that structure defects. For the VLM anomaly detection track, we refer to the template information of prompt learning based on AprilGan. When there are reference images, we use multiple repositories to store their features and compare them with the features of test images during the testing phase.

Introduction

- **Background:** Industrial anomaly detection datasets tend to favor local structural anomalies such as scratches, dents, or contamination. In particular, they lack abnormal forms that violate logical constraints. MVTec Loco, as a dataset exploring logical anomalies, evenly covers both logical and structural anomalies. Track 2 of the Vand2024 challenge is dedicated to creating models using small sample learning and VLM to find and locate structural and logical anomalies. This demonstrates that these models can handle structural defect detection and logical reasoning.
- **Challenge Description:** This track uses the MVTec LOCO AD dataset, and models can be pre-trained on any public dataset other than the MVTec LOCO dataset. For each scenario of few sample learning, k normal images are randomly sampled from the training set of the MVTec LOCO dataset, focusing on the model's ability to learn from small samples. The final evaluation metric will be the area under the F1-max curve AUFC.

Methodology

Model Design

- **Approach:** We use a binary zero-shot anomaly classification framework based on AprilGan to describe both normal and abnormal objects. Detection accuracy is further improved by utilizing a text prompt ensemble strategy, which includes both state and template levels.
- **Architecture:** The model is based on a binary zero-shot classification framework that requires descriptions of normal and abnormal objects. We further improve detection accuracy by using a text prompt integration strategy, including state and template levels. We also refer to the Domain-aware State Prompting proposed by the AnoVL model to decompose the prompt engineering for zero-sample anomaly localization into three parts: base prompts, contrastive-state prompts, and domain-aware prompts. We combine the template with the state of the object, and use the text features extracted by the text encoder as the final text features. Finally, we use the probability corresponding to the anomaly as the zero-shot anomaly score for the image. In the few-shot setting, the anomaly score for the image comes from two parts. The first part is guided by text prompts, which is the same as the zero-sample setting. The second part follows the traditional method used in many anomaly detection methods, which considers the maximum value in the anomaly graph. Adding these two parts together

gives us the final anomaly score. We use the CLIP model and ViT-L/14 as the backbone network for feature extraction, with an image resolution of 336. It has a total of 24 layers, which can be arbitrarily divided into four stages, each containing six layers. At the same time, four additional linear layers are required to map the image features of the four stages into a joint embedding space. Finally, we store the reference image features of these four stages separately.

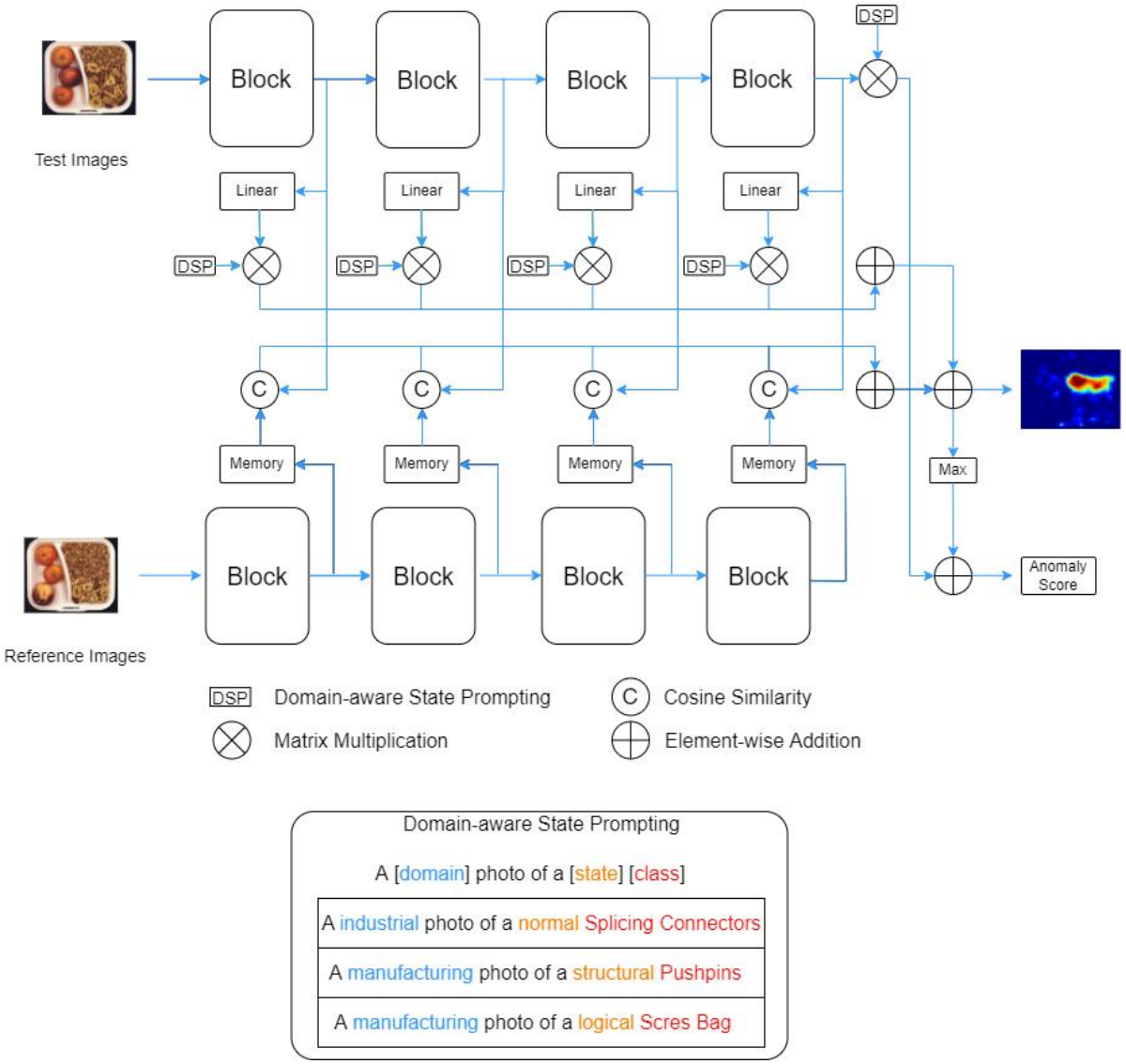

Figure 1. **Overall diagram of our solution**

- **Training:** We use an image resolution of 518 × 518 and combine it with a proposed data augmentation method. Specifically, we stitch four images of the same category from the MVTec AD dataset together with a probability of 20% to create a new synthetic image. We use an Adam optimizer with a fixed learning rate of 1e-3. In order to enable the model to recognize normal and abnormal objects while preventing overfitting, we train for 3 epochs on a single GPU with a batch size of 16.

Dataset & Evaluation

- **Dataset Utilization:** We used the MVTec training set for pre-training and conducted experiments on the standard MVTec Loco dataset. In the setting of a small number of samples, we did not perform any additional fine-tuning on the linear layer
- **Evaluation Criteria:** The evaluation metric for each k-shot setting in the MVTec LOCO subset will be the F1-max score for the anomaly classification task, and the arithmetic mean of the average metrics will be the evaluation metric for the k-normal-shot setting in each category. The final evaluation metric will be the area under the F1-max curve AUFC

Results

- **Performance Metrics:** We achieved the following performance on our local original dataset

objects	auroc_px	f1_px	ap_px	aupro	auroc_sp	f1_sp	ap_sp
breakfast_box	76.8	28.6	24.1	60	76.5	80.3	86.2
juice_bottle	82.8	32	25.9	64.6	74.3	84.7	87.6
pushpins	49.1	3.6	2	44.6	61.3	72.3	69.6
screw_bag	64.7	12.5	8.3	56.4	62.1	78.4	77
splicing_connectors	67.1	20.5	15.3	64.3	58.3	76.8	67.8
mean	68.1	19.4	15.1	58	66.5	78.5	77.6

Figure 2. **Performance on MVTec Loco**

Discussion

- **Challenges & Solutions:** Firstly, the format issue has become a major problem, as the local training framework only retains the weight parameters of the linear layer, and requires retraining and format matching for the submission format. Secondly, due to the inclusion of logically abnormal data in the MVtec loco dataset, matching the prompts is also a key optimization issue.
- **Model Robustness & Adaptability:** We try to optimize the classification of logical anomalies by adjusting the prompt words of contrastive-state, such as adding contrastive states such as misplacement, loss, redundancy, and geometric constraints.

- **Future Work:** Although there have been many discussions on industrial logic anomaly detection in the unsupervised field, the optimization of logic anomaly detection for VLM is still a relatively unexplored topic. In addition to exploring the hint engineering, the adjustment and optimization of the VLM model structure is still a direction that we need to explore in depth in the future.

Conclusion

In Track 2, we conducted a few-shot anomaly industrial detection exploration on the MVtec loco with reference to the Aprilgan model and anovl's domain-aware state prompts. The anomaly logic detection based on VLM is still a topic that needs to be explored. We hope that we can make progress in this field through this competition.

References

Deng, Hanqiu, Zhaoxiang Zhang, Jinan Bao and Xingyu Li. “Bootstrap Fine-Grained Vision-Language Alignment for Unified Zero-Shot Anomaly Localization.” (2023).

Chen, Xuhai, Yue Han and Jiangning Zhang. “A Zero-/Few-Shot Anomaly Classification and Segmentation Method for CVPR 2023 VAND Workshop Challenge Tracks 1&2: 1st Place on Zero-shot AD and 4th Place on Few-shot AD.” ArXiv abs/2305.17382 (2023): n. pag.