GENERATIVE AI AND LLMS PROJECT PROPOSAL

Anomaly Detection using Multimodal Models for Image Data

Student Name: Muhammad Bilal
Student ID: MSDS23007

Student Name: Ahmad Asif
Student ID: MSDS23025

Student Name: Usama Imdad
Student ID: MSDS22036

Student Name: Sarmad Bin Tahir
Student ID: MSCS21020

Supervisor Name: Dr. Mohsen Ali

Department of Computer Science

Information Technology University (ITU)

Contents

1 Title, Field of Research and Summary

Title: Anomaly Detection using Multimodal Models for Image Data
Research Area: Generative AI
Anomaly detection in image data plays a important role in multiple domains, such as industrial, healthcare and surveillance. Mostly traditional methods are used for visual anomaly but they struggle with accurately identifying anomalies, especially in diverse datasets. The implementation and combination of different multimodal models is proposed in this project, it has capabilities to maps the complex distributions of data, easily identifying and localizing visual anomalies with high accuracy results. Extensive experiments will be conducted on multiple datasets to assess the performance of proposed solution, with the goal of achieving state-of-the-art results in both anomaly detection and localization.

2 Introduction

Visual Anomaly Detection for image data plays an important role in our daily life, it is a technique used to identify any kind of abnormal behaviors and patterns. These anomalies can cause errors and frauds in multiple do mains such as: industrial, healthcare, surveillance and many more. There early identification can prevent significant issues. If it left undetected, it could leads to unfavorable effects such as financial loss, system failures, or security breaches. Timely detection and removal of anomalies are important for maintaining our security, integrity and reliability.
With the advancement, particularly few-shot learning and models like CLIP[1], there has been significant progress in detecting anomalies with minimal training data. Few-shot anomaly detection allows models to generalize from only a few examples, making it a more efficient and scalable solution. However, there is still room for improvement in the alignment between textual prompts and visual representations, which is crucial for enhancing model accuracy in such tasks. This work focuses on addressing these challenges and proposing a more effective approach for anomaly detection.

3 Problem Statement and Related Work

Problem Statement: To accurately identify visual anomalies in image data by applying and combining different vision language model architectures.
To detect anomaly, multiple approaches are used that includes generalist anomaly detection [2] using a few-shot approach. The authors utilize CLIP as the base model and test it on seven different datasets. Despite working in a few-shot setting, they retain the original text prompts of the model without any modification. Their method effectively identifies anomalies in images while using CLIP's pre-trained text prompts as-is, showcasing the model's generalization capabilities without altering the textual component. Another approach named PromptAD [3] also employs CLIP as the base model. However, the authors focus primarily on prompt learning rather than altering the image processing part of the model. They work on transforming a normal prompt into an anomaly-specific prompt using a learning-based approach. This method, which is based on the CoOp model [4] results in significantly improved performance by focusing on learning optimized prompts specifically tailored for detecting anomalies. A paper is AnomalyCLIP [5] introduces a slightly different methodology. While CLIP remains the base model, this paper not only applies prompt learning for text prompts but also fine-tunes the text encoder. The visual processing component of CLIP remains unchanged, but the authors argue that fine-tuning the text encoder can lead to better alignment between textual and visual representations, which helps in more accurately identifying anomalies. This approach suggests that both prompt learning and some degree of modification to the text encoder are necessary to enhance the model's anomaly detection performance.

4 Proposed Solution

We aim to enhance the approach presented in the generalist anomaly detection paper [2] which relies on few-shot learning and utilizes CLIP as the base model without modifying the text prompts. While their approach demonstrates effective anomaly detection, we believe that further improvements can be made by focusing on the text prompt component. Specifically, we propose to integrate PromptAD [3], a technique designed to optimize and adapt text prompts for anomaly detection tasks. By replacing the static, pre-trained text prompts with PromptAD, our goal is to achieve more accurate and tai-

lored anomaly detection results. This modification allows us to better align the textual cues with the anomalies present in the images, ultimately leading to superior performance compared to the original model's fixed prompts.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[2] J. Zhu and G. Pang, "Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 826–17 836.

[3] X. Li, Z. Zhang, X. Tan, C. Chen, Y. Qu, Y. Xie, and L. Ma, "Promptad: Learning prompts with only normal samples for few-shot anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 838–16 848.

[4] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[5] Q. Zhou, G. Pang, Y. Tian, S. He, and J. Chen, "Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection," *arXiv preprint arXiv:2310.18961*, 2023.