Paper Review: Adapting Visual-Language Models for Generalizable Anomaly Detection in Medical Images

By: Lam Nguyen

1. SUMMARY

Advances in Visual Lange Models for zero-shot detection, classification and segmentation do not really transfer well to working with medical images. Much of the work in natural image domains focus on segmentation. However, medical images require detecting anomalies in images.

In order to adapt discoveries and improvements from natural image domains to medical image anomaly detection, a lightweight multi-level adaptation and framework was created to adapt these Visual Language Models from working mainly with natural images to working with medical images.

This adaption involves using multi-level pixel-wise visual-language feature alignment and on average improved anomaly detection by 6.24% and 7.33% using AUC metrics.

This framework which will be called a Multi-level Visual Feature Adapter (MVFA) addresses the challenge of overfitting due to a high number of parameters with a low number of training examples. The method appends a small set of learnable bottleneck linear layers to the visual branches of CLIP while keeping the original backbone unchanged.

To be able to effectively discern both global anomalies for classification and local anomalies for segmentation, Language Feature Formatting is used. This involves a two-tiered approach involving a state level and then a template level.

The state level involves using straightforward text descriptions. The template level involved a thorough examination of 35 templates

2. STRENGTHS

- This technique is able to preserve the weights of the original model, preserving all the enhancements and training from working with natural images. This very much reduces the need for expensive and specialized medical images and the accompanying experts needed to annotate this type of data.
- This technique can be used on consumer GPUs, or a single RTX 3090 GPU
- Technique is able to be generalized to other forms of medical imaging data that it wasn't trained on.

3. WEAKNESSES

- This method requires extra training and data to work. But not a lot of data is provided. Even with the mitigation methods, overfitting can still occur.
- Can add complexity and runtime to the tech stack. For highest performance, it might be best to train and fine tune fully on the medical image dataset.
- Not tested with other VLM architectures

4. TECHNICAL EXTENSIONS

To expand upon the findings in this paper, test on different datasets that were not tested in the experiment. Other ways to expand are to test on other VLMs besides CLIP. Possibly try Stable Diffusion.

5. OVERALL REVIEW

The technique presented in this paper allows a separate layer on top of a VLM that was trained on natural images to be usable with medical images. This means that any sort of discoveries and optimizations found on training with the more widely available Natural Images can be applied to Medical Images which are harder to obtain and to even annotate. The strength of this technique is that it can be trained and used on single RTX 3090 and any new discoveries in VLMs can be transferred over. The weakness is that more test need to be done on other VLMs and other datasets to confirm that this method can be generalized to more datasets and to more Visual Language Models.