マスク誘導型 Visual Encoder による異常検知の言語説明と可視化

ER22075 森田悠斗　　指導教授：藤吉　弘亘

1. はじめに

工業製品の品質保証において，異常検知は重要工程である．異常検知は正常・異常の分類だけでなく異常領域の特定や判断根拠の可視化など，説明性も重要である．Anomaly-OneVision (Anomaly-OV) [1] は大規模言語モデル (LLM) を用いることで，異常内容を自然言語で説明することを可能とした．しかし，Anomaly-OV は異常内容の説明が異常領域と一致しないことがある．本研究では，異常領域を強調する学習を Visual Encoder に導入することで，異常内容の言語説明と異常領域の一致性を向上させる．

2. Anomaly-OV

Anomaly-OV は，入力画像から異常内容を自然言語で説明するマルチモーダルモデルである．本モデルは，Look-Twice Feature Matching (LTFM) と Visual Token Selector (VTS) を中核として構成される．LTFM は元画像と 4 分割して得られる局所画像の特徴を二段階で照合し，異常度スコアを算出する．VTS はそのスコアに基づき異常領域を強調した重み付き視覚特徴を生成し，LLM へ入力する．本モデルは，未知データの場合に異常を正確に捉えられず，説明文が異常内容から逸れる場合がある．

3. 提案手法

本研究では，教師マスクにより異常領域への注目を強化した Anomaly-OV に基づく異常説明と，その判断根拠の可視化手法を提案する．

3.1. 教師マスクによる Visual Encoder の拡張

Anomaly-OV の説明能力を特定の領域に適応させるため，欠陥箇所を明示したマスク画像を用いて追加学習を行う．提案手法のアーキテクチャを図 1 に示す．Visual Encoder には，元画像 I_0 およびそれを 4 分割した局所画像 I_n を入力する．Visual Encoder の出力である視覚特徴 v_j^0 を LTFM に入力し，異常度スコア m_j を出力する．そして，異常度スコア m_j と教師マスク y_j との誤差を最小化するよう LTFM を学習する．教師マスクは異常を 1，正常を 0 とする 2 値ラベルである．誤差関数には，Binary Cross Entropy (BCE) Loss を用いる．VTS では，異常度スコア m_j と視覚特徴 v_j^0 の要素積をクエリ，視覚特徴 v_j^0 をキー，バリューとして Q-Former で処理する．これを元画像 I_0 および局所画像 I_n に対して行う．異常領域を強調した特徴 v_j^s を画像ごとに LLM に入力し，異常説明文を生成する．

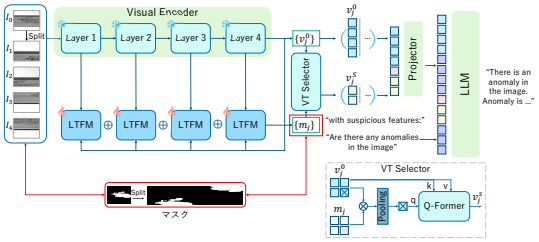


図 1: 提案手法のアーキテクチャ

3.2. 言語説明に対する可視化

異常説明文の生成時にモデルがどの画像および領域を参照しているか可視化する．各画像から得られる異常強調特徴の単語生成への寄与を比較するため，LLM の Attention weight を抽出し，注目画像を特定する．そして，異常度スコアに基づく異常強調特徴を生成する VTS の Q-Former における Cross-Attention から得られる注目度をパッチ単位で可視化することで，画像内の注目領域を特定する．

4. 評価実験

提案手法の有効性を示すために，異常ラベルの分類精度と説明性能の 2 つの観点で比較を行う．

4.1. 実験条件

評価には 10 種類の異常を含む電線データセットを用い，学習データ 8976 枚，評価データ 563 枚を使用する．学習条件は，学習率 $1e^{-4}$，バッチサイズ 16，10 エポックとする．また，異常説明生成に用いる LLM は Anomaly-Instruct125k でファインチューニングされたものを使用する．異常分類性能の評価では，従来のマルチモーダル手法である LLaVA，Anomaly-OV と提案手法を比較対象とする．評価指標として，生成された説明文と各異常カテゴリの正解文との一致度を用いる．生成文が正解文と 80%以上一致した場合を正解とみなし，正解率 (Accuracy) を算出した．異常説明性能の評価では，Anomaly-OV と提案手法を比較し，生成された異常説明文が異常特徴を適切に表現できているか定性的に評価する．

4.2. 異常分類結果

異常分類精度の比較結果を表 1 に示す．提案手法は LLaVA より 11.4 ポイント，Anomaly-OV より 0.5 ポイント高い精度を示した．従来手法では正常領域に注目することがあり，異常の識別が不十分となる場合があったが，提案手法は異常領域を強調することで誤分類が低減したと考えられる．

表 1: 異常分類精度の比較

	LLaVA	Anomaly-OV	提案手法
Accuracy	81.5	92.4	**92.9**

4.3. 異常説明性能の定性的評価

提案手法による説明文の質の向上について，異物_繊維の事例を用いて比較する．異物_繊維の出力結果を図 2 に示す．Anomaly-OV は，異物繊維の特徴を十分に反映できていない説明文である．一方，提案手法は,「波状」など繊維特有の異常形状を表す表現が生成され，異常特徴に基づく説明文が得られた．

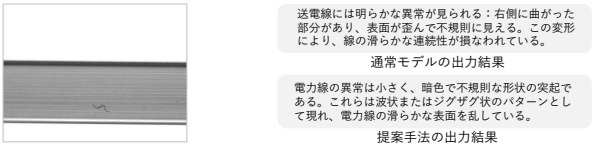
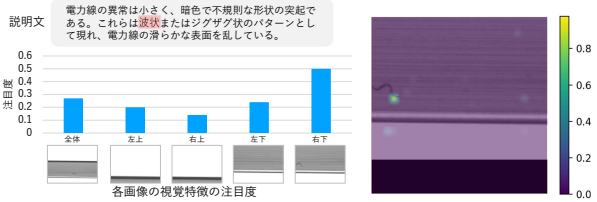


図 2: 異物_繊維の出力結果

4.4. 特定単語の注目領域の可視化

異常説明文内の「波状」という単語に着目し，単語生成時にモデルが参照する画像および領域を可視化した．注目画像および注目領域の可視化結果を図 3 に示す．図 3(a) より，欠陥箇所を含む右下の画像に注目度が集中しており，異常説明時に異常領域を多く含む画像を重視していることが分かる．図 3(b) の注目領域の可視化結果から，異常領域を含むパッチに注目が集中していることが分かる．これらの結果から，異常位置を明示的に学習させることで，モデルは異常領域に基づく単語生成を行うことを確認した．



(a) 注目画像の特定　　(b) 注目領域の可視化

図 3: 注目画像および注目領域の可視化結果

5. おわりに

本研究では，Anomaly-OV を電線データに適用し，異常領域の強調と説明文生成の改善を行った．異常度スコアに対してマスク画像を用いた追加学習を実施し，モデルは異常領域をより正確に捉え，説明文も適切な異常表現へと改善された．今後は，異常領域の推定精度向上や，単語生成時の注目領域をより詳細に可視化する手法の検討を進める．

参考文献

[1] Y. Xu, *et al.*, “Anomaly-OV: Towards Zero-Shot Anomaly Detection and Reasoning with Multimodal Large Language Models”, In CVPR, 2025