

自然言語での異常解釈：LLM を用いた AI 説明モデルの提案

山科 勇輔¹ 須賀 圭一¹ 白井 祐典¹ 市川 佳彦¹

¹ 株式会社 Insight Edge

{yusuke.yamashina, keichi.suga, yusuke.shirai, yoshihiko.ichikawa}@insightedge.jp

概要

本研究では、画像異常検モデルと大規模言語モデル (LLM) と組み合わせることで、より解釈しやすくする新しいアプローチを提案する。異常検知は、製造業や医療分野で不可欠であり、迅速かつ正確な判断が求められるが、その出力が抽象的で理解づらいことがある。そこで画像処理アルゴリズムによって検知された異常を、LLM を用いて自然言語で説明する言語駆動型説明可能 AI (Language-Driven Explainable AI) を提案し、異常検知結果の信頼性と透明性の向上を図る。また、その出力が実業務に有用であるかも検証する。

1 はじめに

近年、深層学習技術の進展に伴い、画像認識における異常検知の精度は飛躍的に向上している。特に製造業や医療分野において、製品の欠陥検出や病変の早期発見など、様々な応用が期待されている [1, 2]。しかし、異常は一般的に稀な事象であるため、正常データに比べて異常データの収集に課題がある。また、異常のパターンは多様で、明確に定義することが難しく、各異常はそれぞれ異なる特徴を持つため、一貫して検出するためのモデル設計も困難なものとなっている。さらに、ラベリングが複雑で誤りが発生しやすく、アノテーションコストが高いことも課題となっている。このため、異常検知の精度向上に向けた多方面での研究が引き続き行われている [3]。また、既存の深層学習モデルはブラックボックスであることが多く、異常と判断した根拠を人間が理解することが困難なケースもあった。このため、AI の判断に対する信頼性の担保や、さらなる性能向上のために、AI の意思決定プロセスを説明可能なものにする Explainable AI (XAI、説明可能な AI) 技術 [4, 5, 6] が注目されているが、XAI の手法で生成される説明がデータ分析に長けた専門家でない

と解釈が難しい場合や、説明が多すぎて重要な情報が埋もれてしまったり、逆に説明が限定的で詳細に欠けることがあり、適切な情報量のバランスを取ることが難しいといった課題がある。

一方で、近年の LLM の発展は XAI の可能性も大幅にひろがっている。例えば、画像キャプション生成タスク [7] のように、画像に対する異常検知結果を LLM の入力として説明文を生成することで、画像内の異常の理由を人間が理解しやすく自然な言葉で説明することが可能となる。さらには、異常検知結果が誤っていた場合にもその間違いを正すことで異常検知精度の向上や、異常の種類や発生箇所から発生要因をフィードバックすることで異常検知後の作業の優先順位なども指示することが期待できる。

本研究では、画像に対する異常検知結果を LLM へのインプットとすることで説明性を付与する言語駆動型説明可能 AI (Language-Driven XAI) を提案する。各種画像に対する異常検知結果 (2 値分類、クラス分類、セグメンテーション、物体検知) に対してキャプション生成を行い、正確性や詳細度、異常のバリエーションに対するキャプションの柔軟性などの評価を行う。

2 関連研究

2.1 画像に対する異常検知

異常検知は、通常データセット内で稀なパターンや事象を識別する技術である。製品の欠陥や医療画像における病変の早期発見などが含ま、多種多様な分野での重要な課題として認識されており、統計モデルや特徴量エンジニアリングに基づき、特定の指標やルールに従って異常を判断するものが提案されている [1]。近年では深層学習を用いた手法が、従来の手法が抱える課題を克服するためのアプローチとして注目されている。特に、Convolutional Neural Network (CNN) を利用した手法は、画像データから自動的に特徴を抽出し、2 値分類やクラス分類で高い異常検出精度を達成することが実証されてい

る。また、ImageNet などの大規模な画像データセットで事前学習されたモデルを使用して、特徴を抽出した後、複数の層の特徴を結合してピクセルレベルで異常検知を行う PaDiM[8] や PatchCore[9] なども提案されている。他にも YOLO[10] などの物体検知を用いた手法も提案されている。

2.2 説明可能な AI

XAI は、AI モデルの決定理由を人間が理解できる形で説明することを目指す技術である。例えば、医療診断システムでは、XAI が診断の根拠を明示することで、医師がその判断を信用することを助ける。しかし、生成された説明が専門外の人には難解であったり、説明が冗長で重要な情報が埋もれてしまうといった課題も存在する。特に画像処理においては、モデルがどの特徴（例えば、ピクセル）に注目しているかを視覚的に示す手法が重要であり、代表的な手法として、特徴マップを用いる Grad-CAM[11] や LPR[12] が提案されている。また、LIME[4] は、画像の特定の部分やピクセルを抜き出して入力し、それに基づいた予測結果とのペアを作成し、そこからより単純なモデルに学習させることで説明力の高い結果を出力できるようにすることで、重要な特徴量を明確に取得することを可能にしている。

2.3 キャプション生成

画像キャプション生成は、画像のテキスト説明やキャプションを自動的に生成するタスクであり、近年では LLM を活用した芸術作品に対する説明の生成に関する研究もなされている [13]。

3 問題設定

図 1 に提案アプローチの概要図を示す。異常検知モデルとしては、例えば、多クラス分類やセグメンテーション、物体検知モデルなどを想定している。

本研究では画像に対する異常検知結果に対して LLM で説明性を付与した際のキャプションの正確性、個数や位置、大きさ、形状などの詳細度、異常のバリエーションに対するキャプションの柔軟性、を評価する。キャプションの生成結果は定量的に評価する手法もあるが、ここでは、設定した 5 段階の評価スケールを用いて人間による定性的な評価とした。「正確性」と「詳細度」は一枚一枚採点し、その平均点を算出する。「柔軟性」は異常のバリエーションに対する評価となるので各手法での全結果を

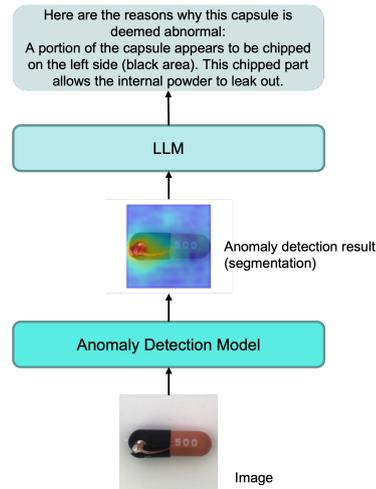


図 1 提案アプローチ手法：Language-Driven XAI

表 1 異常検知モデル

2 値分類	CNN
多クラス分類	CNN
セグメンテーション	PatchCore
物体検知	Grounding DINO [14]

鑑みて採点した。

4 実験設定

データセットとしては、産業用異常検出のベンチマークとしてよくも用いられている MVTec AD[5] で評価する。MVTec AD は、合計 5354 枚の画像を含む 15 のクラスで構成されており、各クラスは特定の製品に対してさまざまな欠陥タイプを持つ異常データとなっている。本研究では、テストデータである 1725 枚を対象とした。

異常検知モデルとしては、正常か異常かの 2 値分類、欠けや印刷ミスといった複数の異常の種類を分類する多クラス分類、画像内の異常箇所をピクセルレベルで評価しヒートマップで表すセグメンテーション、画像内の異常箇所をバウンディングボックス (bbox) で示す物体検知、の 4 種類を適用し、これらの手法によるキャプション生成の結果を比較評価する。本研究で用いた各種法のモデルを表 2 に示す。なお、各パラメータに関しては、誤検知や検知漏れの結果に対しても評価するため、ある程度ランダムとした。また、モデル毎に出力される結果も異なるのでモデル毎に指示文を調整している。表 2 に指示文の例を示す。また、LLM には gpt-4o (OpenAI) を適用した。

表2 各モデルへの指示文例（クラス：hazelnut の場合）

2値分類	これは hazelnut の画像で、異常と判断されたものです。この画像がなぜ異常と判断されたか説明してください。
多クラス分類	これは hazelnut の画像で、crack と判断されたものです。この画像がなぜ crack と判断されたか説明してください。
セグメンテーション	これは hazelnut の画像に異常度を表すヒートマップを重ね合わせた画像です。ヒートマップの色が濃い（赤い）ほど異常度が高いことを示します。この画像がなぜ異常と判断されたか説明してください
物体検知	この画像は hazelnut の画像です。画像内の枠内の箇所は異常と判断された箇所です。この枠内がなぜ異常と判断されたか説明してください。



クラス	hazelnut
異常パターン	crack
2値分類結果	異常

この画像のヘーゼルナッツは、以下の理由で異常と判断できます：
亀裂と割れ目：ヘーゼルナッツの殻に明らかな亀裂と割れ目が見られます。通常、ヘーゼルナッツの殻は滑らかであり、亀裂や割れ目はナッツの品質や保存状態が良くないことを示唆します。
以上の点から、このヘーゼルナッツは異常とされています。

図2 二値分類によるキャプション生成結果例



クラス	capsel
異常パターン	squeeze
多クラス分類結果	誤分類 (crack)

この画像のカプセルは、「crack」と判断されたことですが、画像を評価した結果、以下の理由で異常が認定されたと考えられます。
1. 変形: カプセルの右側（オレンジ色の部分）が明らかに変形しています。通常のカプセル形状が滑らかな円筒形であるべきところ、この部分には凹みが見られます。
2. 表面の傷や裂け目: カプセルのオレンジ色の部分に微細な傷が見えることから、これが「crack（ひび）」と誤判定された可能性があります。しかし、実際にはひび割れではなく表面の傷や変形が原因と考えられます。したがって、この画像から「crack」と判断されたのは、ひび割れではなく変形と表面の傷が要因である可能性が高いです。この変形はカプセル製品としての品質に大きな影響を及ぼし、製造工程中または輸送・保管中に何らかの物理的な圧力が加わったと推察されます。
したがって、このカプセルは「crack」とは誤認されているように見えるものの、「変形」という異常が確認できるため、依然として異常品であると判断できます。

図3 多クラス分類によるキャプション生成結果例

5 実験結果

二値分類 図2に入力画像と検知結果、生成したキャプションの例を示す。亀裂と割れ目から異常であることを正確に説明できており、また、保存状態にも言及することができている。しかし位置や傷の数などについての説明はなく詳細度はやや低い。

多クラス分類 図3に入力画像と多クラス分類結果、生成したキャプションの例を示す。異常の種類としては squeeze が正しいが crack と分類してしまったケースだが、右側に凹みがあることを説明しており、誤判定である可能性があることも説明できている。右側やオレンジ色の部分といった位置の情報も示されており詳細度が高く、ユーザーの理解が容易であり、検知精度向上へのフィードバックにも活用が期待できる結果となった。

セグメンテーション 図4に異常検知モデルへの入力画像と検知結果を示す。赤枠で示す通り、左側に縦のほつれがあるが、セグメンテーションではほつれ付近の異常度が高くなっているものの、左上と左下が特に異常度が高い結果となっている。LLMへの入力画像と生成したキャプションの例を図5に示す。ヒートマップで基の図が見にくくなっているためか、ほつれだけでなく、色ムラや擦り傷の可能性についても言及できており、正確性の高い結果と言える。また、異常度が高い箇所以外にも言及できている詳細度は高い。

物体検知 図6に物体検知結果と生成したキャプションの例を示す。異常箇所2箇所のうち、1箇所を検知漏れしているが、キャプションではもう1箇所についても異常であることが言及されている。検知できている異常箇所に対しては、テキストチャ、色、形の違いなど、様々な要因から異常の理由を詳述しており、検知漏れの箇所に関しても同様の要因から異常であることを言及できている。キャプションは異常の理由を多角的に説明できており、正確性と詳細度が高い結果となっている。

表3 比較検証の採点結果

	正確性	詳細度	柔軟性
2値分類	4.17	4.00	5.00
多クラス分類	4.83	4.92	5.00
セグメンテーション	3.75	4.17	3.00
物体検知	4.17	4.75	4.00

他のクラスなど全テストデータに対して検証を行った評価結果を表3に示す。二値分類や多クラス

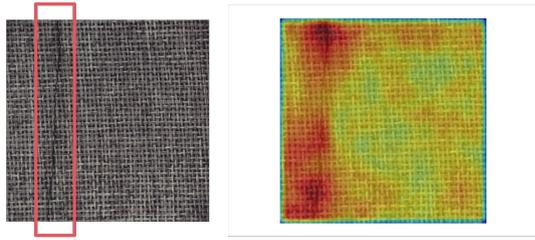
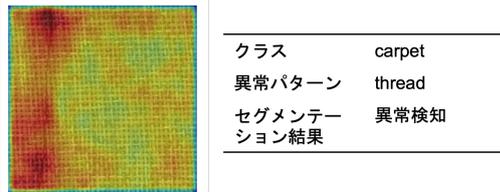


図4 セグメンテーションによる異常検知結果例
(クラス: carpet)



このcarpetの画像に対して異常度を示すヒートマップが重ねられています。左上および左下の異常点（赤く表示されている部分）は、次のような理由で異常と判定された可能性があります：

1. 色ムラや染み：赤く表示されている部分に色のムラや染みがある可能性があります。カーペットの表面に色の不均一性があると、それが正常な状態と異なるため異常として検出されます。
2. 構造の乱れ：カーペットの織り方やパターンに乱れが生じている場合も異常と判定されます。画像上の赤い部分には、繊維の密度が異常に高い、または低い場所があるかもしれません。
3. 物理的なダメージ：カーペットに物理的なダメージ（例えば、引っ掛け傷や擦り切れなど）がある場合、その部分が正常な状態とは異なるテクスチャや色合いを示すため、異常として検出されることがあります。

したがって、このカーペット画像が異常と判定される主な原因は、左上と左下の部分に集中する色ムラや構造の乱れ、または物理的なダメージであると推定されます。他の部分については、特に赤く表示されていることで明らかな異常が示されている部分以外には、大きな異常は見受けられません。

図5 セグメンテーションによるキャプション生成結果例

分類では異常画像をそのまま入力としているためか、より詳細な説明をしやすい、また、誤検知や誤分類の際に訂正するようなキャプションを生成できるケースが多く見られた。異常の種類や発生要因についても言及できており、全体的に高い評価結果となった。セグメンテーションでは、正しく検知できている場合は異常箇所の位置や個数なども詳述できているケースが多く得られた。しかし、ヒートマップを重ねているため異常の種類について詳細にキャプションすることが難しく、また、ヒートマップがあることによって検知漏れしている場合に検知漏れ箇所を言及できているケースも少なく、全体的にやや劣る結果であった。物体検知の場合、異常箇所の位置や個数に関して言及するケースが多く、検知漏れしている場合でも異常箇所を指摘できたりと、正確性、詳細度、柔軟性とも高いキャプションが生成されやすい結果であった。しかし、bboxによっては異常箇所に被ってしまい、異常の種類を間違えてしまったり、検知漏れしている場合に、bbox



クラス	leather
異常パターン	glue
物体検知結果	検知漏れ

この画像は革 (leather) の画像で、赤い枠内に「glue」と判断された箇所があります。

枠内が異常と判断された理由について説明します。革の表面は通常、均一で特定の質感があり、シワや独特のパターンが見られます。枠内部分には他の部分と明らかに異なる特徴があります。具体的には、以下の点が指摘できます：

1. テクスチャの違い：枠内部分は周囲の革のテクスチャと比べて滑らかで異なっています。革本来のシワやパターンが見られなくなっています。
 2. 色の違い：この部分の色は周囲の色と若干異なるように見え、反射の仕方も異なります。
 3. 形の違い：この部分は他の部分と比べて隆起しており、異物が付着しているように見えます。
- これらの理由から、機械学習モデルがこの部分を「glue」、つまり接着剤のような異物として検出したのは妥当だと言えます。

また、枠の箇所以外に異常があるかどうかについても検討します。画像全体を見たところ、下の中央近くの一つの異常な箇所が見受けられます。これも他の部分と比べて色が異なり、滑らかな隆起が確認できます。この部分も異常、具体的には別の異物（おそらく同様に接着剤）として検出されるべきです。

従って、枠内の部分が異常と判断されたのは正当であり、さらに枠の外にも異常と判断されるべき箇所があります。

図6 物体検知によるキャプション生成結果例

が無いのにあるような説明をしてしまったりと、プロンプトの改善などが必要なケースもいくつか確認できた。

しかしながら、いずれの異常検知手法においても、正しく検知できている場合には正確性、詳細度が高いキャプションが生成できており、誤検知や検知漏れしている場合にも訂正したり、保存や輸送状態に起因するなどといった異常の発生した要因についても言及したりと、詳細かつユーザビリティが高いキャプションが生成でき、本アプローチの有効性を確認できた。

6 まとめ

本研究では、画像に対する異常検知結果に説明性を付与する Language-Driven XAI を提案し、産業用異常検出のベンチマークとしてよくも用いられている MVTecAD を対象に評価を行った。異常についての説明だけでなく、発生要因や誤検知時に訂正できることも確認でき、よりユーザビリティの高い検知システムに寄与できることを確認した。今後は、検知とキャプションを組み合わせた新しい評価手法の開発や、実データを対象とした検証による作業工程へのフィードバックの可否、や複数の異常検知結果を入力とすることでより安定性の高いキャプションの生成可否について検討を行っていく予定。

参考文献

- [1] Arindam Banerjee Varun Chandola and Vipin Kumar. Anomaly detection : A survey. **ACM Computing Surveys**, Vol. 41, No. 3, pp. 1–8, 2009.
- [2] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In **Information Processing in Medical Imaging**, 2017.
- [3] Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. In **International Conference on Learning Representations**, 2020.
- [4] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “why should i trust you”’: Explaining the predictions of any classifier. In **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, 2016.
- [5] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In **Advances in Neural Information Processing Systems 30**, 2017.
- [6] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In **International Conference on Learning Representations**, 2018.
- [7] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hua, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In **European Conference on Computer Vision**, 2020.
- [8] Thomas Defard, Aleksandr Setkov, Angélique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In **International Conference on Pattern Recognition**, 2021.
- [9] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In **Computer Vision and Pattern Recognition Conference**, 2022.
- [10] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In **Computer Vision and Pattern Recognition Conference**, 2016.
- [11] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In **International Conference on Computer Vision**, 2017.
- [12] Alexander Binder, Grégoire Montavon, Sebastian Bach, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. In **Computer Vision and Pattern Recognition Conference**, 2017.
- [13] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In **arXiv preprint arXiv:2301.12597**, 2023.
- [14] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Liand Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In **European Conference on Computer Vision**, 2024.