LLM を用いた胸部 X 線画像および所見データに対するラベル付与時のノイズ軽減手法

有馬伊織　【 知能情報学研究室 】

A Noise Reduction Method for Label Annotation in Chest X-ray Images and Clinical Reports Using Large Language Models

ARIMA, Iori　【 Intelligent Informatics Lab. 】

1　はじめに

胸部 X 線画像診断は，呼吸器や循環器疾患のスクリーニングにおいて重要な役割を果たしている．近年, 深層学習を用いた自動診断モデルの研究が盛んであるが, 大規模な画像データセットに対し，専門医が全ての症例に正確な教師ラベルを付与することはコストの観点から困難である．そのため, 放射線科医が作成した読影レポート（報告文）から，自動テキストマイニングツールを用いて付与された MeSH ラベルを教師データとして利用することが標準となっている [1]．しかし，既存の自動抽出ツールには，診断モデルの精度を阻害するラベルノイズ問題がある．従来の単語ベースの抽出アルゴリズムは，文章の医学的な文脈を解釈する能力が不十分であり，例えば「心拡大は認められない」といった否定表現や,「胸水の疑い」といった不確定な推測表現を，単語の出現のみを根拠に一律で「陽性」と判定する傾向がある. このようなラベルノイズは，画像モデルに対して「疾患が存在しない領域を疾患部位として学習させる」という誤った信号を送り，モデルの識別精度と説明性を低下させる要因となる. 本研究では，高度な自然言語理解能力を有する大規模言語モデル（LLM）である GPT-4o mini を活用し，読影レポートの文脈を考慮した高精度な再ラベリング手法を提案する. 特に，LLM が出力する判定確率に対して確信度 0.9 以上とすることで，曖昧な判定を徹底的に排除した高品質な教師データセットを構築する．さらに，ラベルの置換だけではなく，ラベルノイズが画像分類モデルの学習過程における疾患特徴の抽出能力に与える影響を，Grad-CAM[2] によって検証を行う.

2　提案手法

本研究の提案手法は，大規模言語モデル（LLM）による放射線報告文の解釈を通じたラベルノイズ除去と，その高品質ラベルを用いた画像分類モデルの構築からなる.

2.1　LLM による再ラベリングとコンテキスト理解

既存の MeSH ラベルが抱える「否定表現の誤認」や「文脈無視」の問題を解決するため，推論能力を持つ GPT-4o mini を再ラベリングエンジンとして採用した. 具体的には，放射線報告文を入力とし，対象となる 20 の疾患ラベルそれぞれについて, 医学的文脈に基づいた「陽性・陰性」の判定を指示するプロンプトを設計した. 図 1 は再ラベリングに使用したプロンプトである.

2.2　確信度スコアに基づくラベルフィルタリング

LLM の判定結果の信頼性をさらに担保するため，各判定に対する確信度を用いたフィルタリングを導入した. LLM に対し, 各疾患の存在確率を 0.0 から 1.0 の範囲で算出させ，本研究では確信度 0.9 以上の判定のみを最終的な教師ラベルとして採用した.

2.3　ラベル浄化データの学習への適用

浄化された高品質なラベルを用い，画像分類モデルとして Vision Transformer (ViT) を学習させた. 学習プ

```
あなたは胸部 X 線の放射線診断 AI です。
入力されたレポートを読み、以下の 20 項目について
「存在確率（0.0〜1.0）」を推定してください。

【非常に重要】
- 出力は JSON のみ。
- 下記の 20 ラベル以外を出力してはいけません。
- ラベル名は 1 文字たりとも変更してはいけません。
【出力すべき 20 ラベル】
Cardiomegaly, Scoliosis, Bone Fractures,
Pleural Effusion, Pleural Thickening,
Pneumothorax, Hernia, Calcinosis, Emphysema,
Pneumonia,Edema, Atelectasis, Cicatrix,
Opacity, Lesion, Airspace Disease,Hypoinflation,
Medical Device, Other Finding, Normal

【確率のルール】
- 明確に否定されている場合（no...） → 0.01〜0.05
- 明確に存在する → 0.8〜1.0
- 曖昧（possible, may represent） → 0.3〜0.6
- 記載なし → 0.0
```

図 1: 再ラベリングに使用した指示プロンプト

ロセスでは, 既存の MeSH ラベルで学習したモデルと, 本提案手法で浄化したラベルで学習したモデルを同一のアーキテクチャ・ハイパーパラメータで構築した。これにより，モデルの性能差や識別根拠の変化が，純粋に「ラベルの質」に起因するものであることを比較検証できる設計とした.

2.4　Grad-CAM による識別根拠の定性評価

ラベルノイズの軽減がモデルの挙動に与える影響を視覚化するため，Grad-CAM を用いた解析を行い，モデルが画像上のどの領域（解剖学的特徴）を根拠に予測を行ったかを検査した.

3　実験設定

3.1　データセット

本研究では，インディアナ大学が公開しているオープンデータセット（IU-dataset）を使用した. 全読影レポート約 3,851 件の中から，正面画像とレポートが対になっている症例を厳選して抽出した. 分析対象とする疾患ラベルは，Cardiomegaly（心拡大），Pleural Effusion（胸水），および Normal（異常なし）を含む計 20 項目とした. また，比較検証のためのベースライン（Baseline）として，データセットに付随する MeSH（Medical Subject Headings）タームから自動抽出された既存ラベルを採用した.

3.2　評価指標と可視化手法

既存の MeSH ラベルと LLM によって浄化されたラベルの間で，疾患ごとの陽性判定件数の推移を算出し，ラベル分布の適正化を検証した. また，未知のテストデータに対するモデルの予測確率の変化を計測することで，ラベルノイズの抑制が予測の確実性に与える影響を評価した. さらに，モデルが判定の根拠とした画像領域を特定するため，Grad-CAM を用いたヒートマップによる可視化解析を行った.

4　実験結果と考察

表 1 により GPT-4o mini が医療報告文の否定表現や文脈を正確に理解できていることが示唆された. また, GPT-4o mini による再ラベリングの結果，全ラベル数 6429 件に対し，ノイズ除去数が 2286 件であり，36%の

表 1: MeSH と LLM の判定乖離における具体例

放射線読影レポートの抜粋
No stable cardiomegaly, ... Stable right basilar calcified granuloma. ... Stable cardiomegaly without acute cardiopulmonary abnormality.
MeSH ラベル（既存）: 1 (Positive) 判定理由（推測）: 文末の「Stable cardiomegaly」という単語に反応。
LLM ラベル（提案）: 0 (Negative) 判定理由: 文頭の「No stable ...」および全体の文脈から否定と判断。

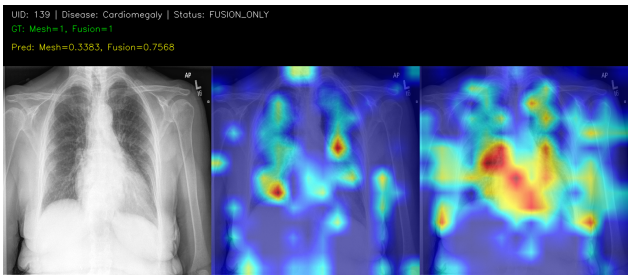

図 2: Cardiomegaly 症例における Grad-CAM の比較（左：元画像 中：MeSH 学習モデル 右：提案手法学習モデル）

ラベルノイズを削減した．Other Finding では，2323 件のうち，1816 件が適切なラベル付として改善された. Normal では，1348 件のうち，385 件が適切なラベル付として改善された．cardiomegary では，319 件のうち 9 件が適切なラベル付として改善された. 他のラベルに関しても適切にラベル付を行えた.

4.1　可視化による識別根拠の妥当性評価

提案手法によるラベル浄化が, 画像分類モデル（ViT）の識別根拠に与える影響を評価するため，Grad-CAM を用いた可視化を行った. 図 2 に，心肥大（Cardiomegaly）症例における比較結果を示す. 既存の MeSH ラベルで学習したモデル（中央）では，予測確信度が 0.3383 と低く，注視領域も肺野全域に分散しており，疾患特有の解剖学的特徴を捉えられていない. 対して，提案手法による浄化ラベルで学習したモデル（右）では，予測確信度が 0.7568 へと大幅に向上した. さらに，注視領域が心臓境界部へ集約されており，医学的にも妥当な根拠に基づいて判定が行われていることが確認された. 全ての心肥大テストデータにおいては 22 個中 7 個で改善，8 個が現状維持，7 個が悪化した.

5　まとめ

本研究では，LLM を活用することで IU-Xray データセットの MeSH ラベルに含まれるノイズを効果的に除去できることを示した. ノイズ除去したラベルを用いることで，画像分類モデル（ViT 等）の学習阻害要因が取り除かれ，精度の良い診断モデルの構築が可能になると考えられる. 定性的な側面として，ノイズが除去されたラベルで学習したモデルは，既存モデルと比較して予測確信度が向上し，Grad-CAM による注視領域が分散した状態から解剖学的に妥当な疾患部位へと集約されることを確認した.

参考文献

[1] Y. Zhang, et al., AAAI,34(7), 12910–12917,2020.

[2] A. Alqutayfi et al., JAIT,16(2), 264–273, 2025.