MDLab GM 資料
21 年 10 月 26 日 (火)

腹部超音波画像からの腫瘍検出

B3 原 英吾

1　研究背景および目的

- 背景
  - 検査実施者は超音波器具の操作と同時に診断を行わなければならず高難易度
  - 肝臓は沈黙の臓器と呼ばれ，炎症やガンがあっても初期には自覚症状がほとんどない
    - 症状を自覚しているときには重症化しているケースが多い
  - 機械学習による診断のサポート
    - 提供されているデータセットには，図 1 の様に明らかなアノテーション不足のある画像が存在する
- 目的
  - 既存の研究を踏まえたモデルの精度向上
    - noisy label[1]による精度低下の改善
  - 超音波支援システムの開発
    - 早期発見につながると良い

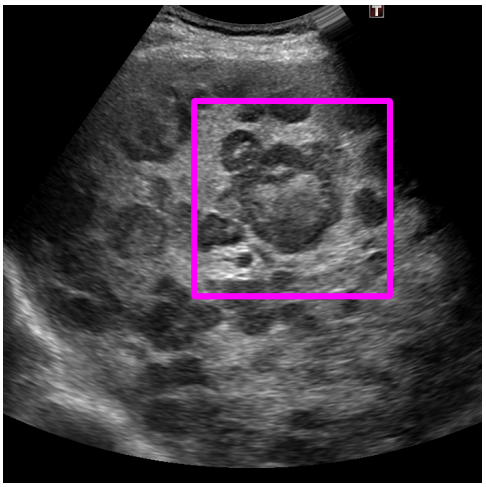
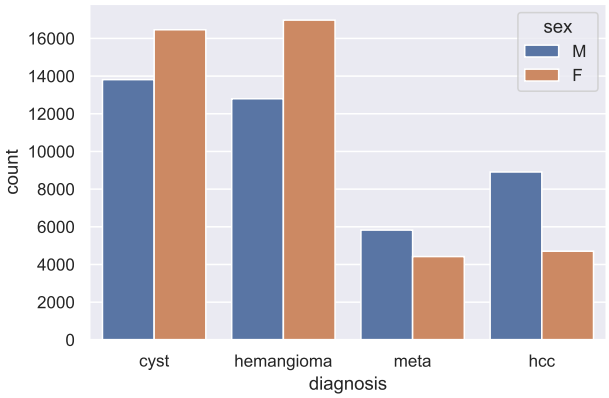


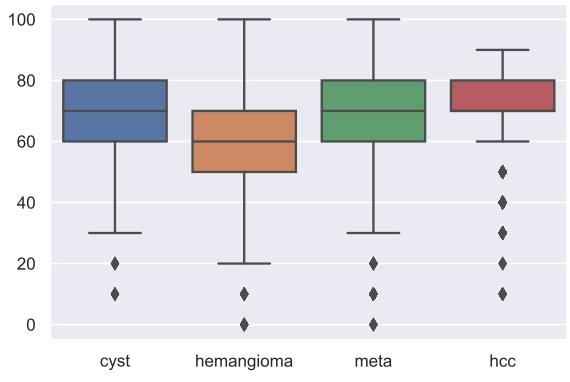
図 1: アノテーション不足のある診断画像例

2　これまでの研究のまとめ

- データセット
  - 国立研究開発法人日本医療研究開発機構 (AMED)[2]が提供している延べ 8 万人に及ぶ以下のデータが付随している
    - 腹部超音波画像，ROI
    - 年齢，性別



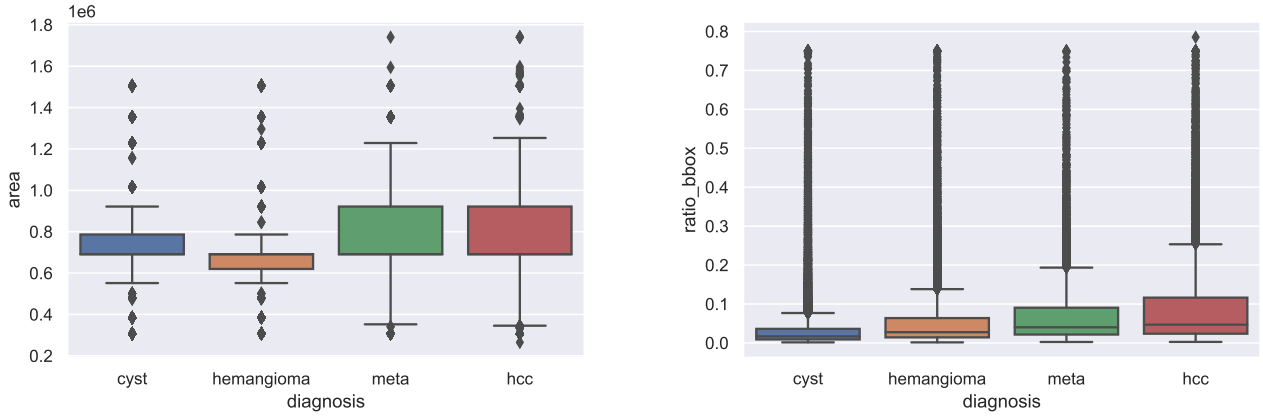
(a) 性別毎の画像枚数



(b) 診断名毎の年齢分布

[1]今回は図 1 の様なアノテーションが不足しているものを指す
[2]https://www.amed.go.jp/

- 性別 (図 2a)
  - ∗ hcc(肝細胞癌) は男性が罹患しやすい
  - ∗ hemangioma(血管腫) は女性が罹患しやすい
- 年齢 (図 2b)
  - ∗ hcc(肝細胞癌) は比較的高齢者が罹患しやすい
  - ∗ cyst(単純嚢胞)，hemangioma(血管腫) の分布にははあまり特徴がない
  - ∗ meta(転移性肝癌) における 0 歳はラベルミスである可能性が高い



(a) 診断名毎の画像サイズ ($h \times w$) の分布　　(b) 診断名毎の bbox の割合

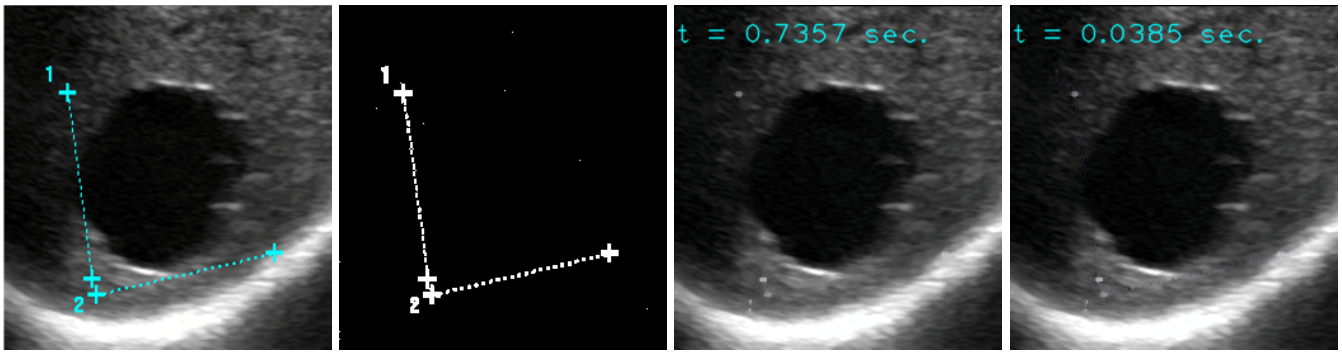
図 2: データセットに含まれているメタデータの分布

- 図 2a から hemangioma(血管腫) は比較的画像サイズが統一されていることが読み取れる
  - ∗ 血管腫においては腫瘍の大きさが血管に依存するためあまり偏りが生じていない？
- 図 2b から cyst(単純嚢胞) は他の診断と比べて bbox の割合が低い ($\frac{1}{2}$ 程度) であることが読み取れる

3 前回の GM からの進捗

- データクレンジングのプログラムの統合及び改善

1. 400 × 400 以下の画像の除外
2. Perceptual Hash を利用した類似画像の除外
   - 差が 2 以下であればスキップ
3. 青色や黄色のスケールの除去
   - 図 3b の様に HSV 空間で対象の色のある部分の mask を生成
4. 使用する画像パスの書き出し



(a) クレンジング前　(b) マスクの生成　(c) 改善前　(d) 改善後

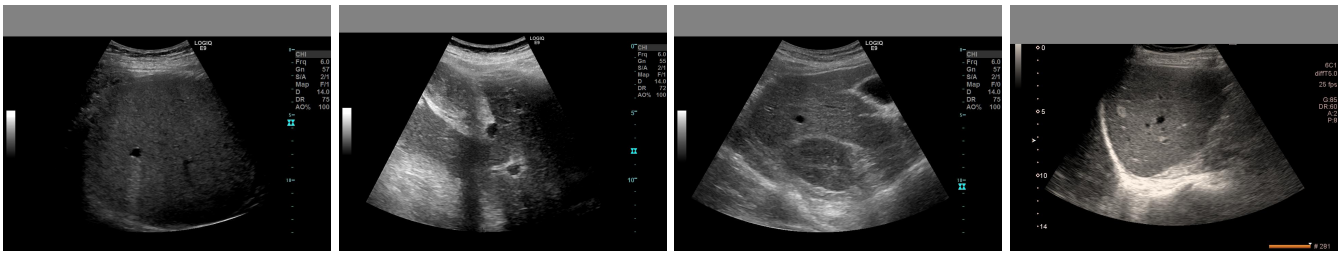
図 3: データクレンジングを行った結果

  - 以前まではデータセット構築として以下を個別に行っていたため時間がかかっていた
  - 特にデータクレンジングにおいて，図 3c の様に 1 枚あたり約 0.7357 秒も必要としていた
    * 改善後は約 0.0385 と図 3d の様に元の精度を保ったまま約 20 倍高速化

- 付録 A におけるソースコード 1 の様に付随していたメタデータを COCO-Dataset の規則に則った json 形式に
  - train，validation，test に分割

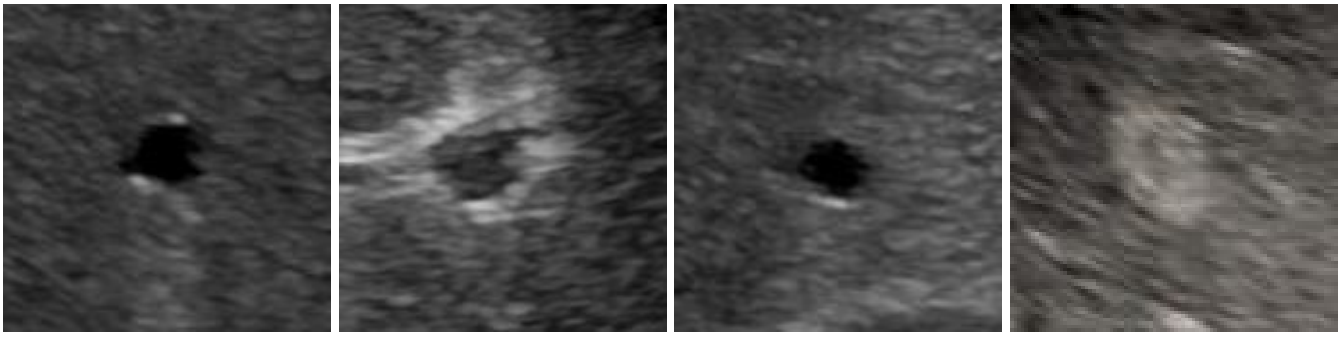
filename	データ数
train.json	67122
validation.json	8390
test.json	8391

- 付録 A におけるソースコード 1 の様に付随していたメタデータを COCODataset の規則に則った json 形式に
  - COCO 形式のデータを扱うために以下を gpgpu の環境にインストール
    * pycocotools
    * crowdposetools
- 元の画像から腫瘍の部分の切り出しを行った



(a) 000002.jpg (b) 000019.jpg (c) 000027.jpg (d) 000040.jpg

図 4: 元の画像



(a) 000002.jpg (b) 000019.jpg (c) 000027.jpg (d) 000040.jpg

図 5: 腫瘍の切り出しを行った結果

4 今後の課題&スケジュール

- 11/30 まで
  - 新規に作成した json 形式のデータでモデルの学習コードを完成させる
- できるだけ早めに
  - 研究の方向性を決める
  - Confident Learning [2] を利用してみる
    * ラベルにノイズが含まれていると予想されるデータセットに対して精度を向上させることができる
    * pip でインストールできる cleanlab[3]というライブラリを用いることで簡単に使える
      · 調べてみたら元は Keras？

[3] https://github.com/cleanlab/cleanlab

付録 A　json ファイル

```
{
    "info": {
        "description": "AMED Dataset",
        "url": "https://www.amed.go.jp",
        "version": "1.0",
        "year": 2021,
        "contributor": "Japan Agency for Medical Research and Development",
        "date_created": "2021/10/22"
    },
    "licenses": [{
        "url": "hhttps://www.amed.go.jp",
        "id": 1,
        "name": "Japan Agency for Medical Research and Development"
    }],
    "images": [
        {
            "license": 1,
            "file_name": "000000.jpg",
            "height": 873,
            "width": 1164,
            "id": 0
        },
        ...
    ],
    "annotations": [
        {
            "iscrowd": 0,
            "image_id": 0,
            "bbox": [317.5,417.5,164.0,164.0],
            "category_id": 1,
            "id": 0,
            "age": 80,
            "sex": 1
        },
        ...
    ],
    "categories": [
        {
            "supercategory": "cancer",
            "id": 1,
            "name": "cyst"
        },
        ...
    ]
}
```

ソースコード 1: 変換後の json ファイル

参考文献

[1] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep High-Resolution Representation Learning for Visual Recognition, 2020.

[2] Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. Confident Learning: Estimating Uncertainty in Dataset Labels, 2021.