

## YOLOv3 とドメイン知識を用いた CT 画像の病変部位検出

## Lesion Detection in Computed Tomography Images using YOLOv3 and Domain Knowledge

西郷 彰<sup>\*1</sup> 林 直輝<sup>\*2</sup> 伊藤 孝太郎<sup>\*2</sup>  
 Saigo Akira Hayashi Naoki Ito Kotaro

<sup>\*1</sup>株式会社 リクルートテクノロジーズ アドバンスドテクノロジーラボ  
 Recruit Technologies Co., Ltd. Advanced Technology Lab.

<sup>\*2</sup>株式会社 NTT データ数理システム シミュレーション&マイニング部  
 NTTDATA Mathematical Systems Inc. Simulation & Mining Division

Lesion detection in computer tomography (CT) images using deep neural networks (DNN) have been researched in computer-aided detection area. A dataset of large-scale annotated CT images, called DeepLesion, has also been published. However, the conventional lesion detection method needs many false positives per a image (FPI) to realize high sensitivity. Besides, it uses a constant CT value for all images in DeepLesion, thus there is a divergence from the medical site. On the other hand, one after another object detection methods have been proposed in the DNN community. In this study, we carried out experiments for FPI to decrease using You Only Look Once version 3; a novel object detection method. We also use medical setting CT value each image to bring it closer to the site. The experimental results show that our method is more accurate than conventional one in the sense of the sensitivity given same average FPI.

## 1. はじめに

画像解析分野において、畳み込みニューラルネットワーク (convolutional neural network, CNN) [2] が成功を収めている。CNN は多層のニューラルネットワークと畳み込み演算を用いて、入力画像から特徴量を自動抽出することが可能で、従来の人手による特徴量設計を用いた統計的画像認識技術を上回る認識性能を持つ。医療画像解析においても応用されており、皮膚可視画像 [1]、レントゲン画像 [3]、計算機断層撮影 (computer tomography, CT) 画像 [6]、磁気共鳴画像 (Magnetic Resonance Imaging, MRI) [7] などが挙げられる。CNN の学習には大量の画像データが要求されるが、CT 画像については DeepLesion と呼ばれる大規模データセットが公開されており、これを用いた病変部位検出手法が提案されている [6]。しかし、検出対象の CT 画像 1 枚につき多くの偽陽性のバウンディングボックス (bounding box, BB) の出現を許容しなければ高い感度<sup>\*1</sup>を実現することはできなかった。

深層学習を用いた物体検出 (深層物体検出) の手法が今日盛んに研究されており、CADe にも応用されている。近年の深層物体検出手法の 1 つとして You Only Look Once version 3 (YOLOv3) が挙げられる [4]。本研究では、YOLOv3 を DeepLesion データセット内の胸部画像に適応する。実際の医療現場で用いられる CT 値ごとにモデルを作成し、先行研究を上回る FROC 曲線<sup>\*2</sup>が得られた、すなわち感度あたりの偽陽性数を削減したことを報告する (図 1 参照)。また、放射線医師が立体的な情報を利用して正常組織とよく似た病変を識別することを参考に、DeepLesion において病変部位がラベ

ル付けされている CT 画像の近傍のスライスも考慮するように YOLOv3 を改変しての実験も実施した。

本論文の構成は次の通りである。第 2 節で関連研究として、DeepLesion データセットとその病変部位検出モデルと、立体情報考慮について概説する。第 3 節で、本研究における実験方法を説明し、第 4 節でその結果を述べる。第 5 節では、実験結果に対する考察を行い、第 6 節で本論文を結ぶ。

## 2. 関連研究

本節で関連研究について述べる。まず、2.1 及び 2.2 小節で DeepLesion データセット及びそれを用いた物体検出手法の概要を説明し、次に 2.3 小節で先行研究の課題を述べる。2.4 小節では、偽陽性 BB 対策として提案された、3 次元コンテキストを考慮した場合の手法を概説する。

## 2.1 DeepLesion データセット

DeepLesion[6] はアメリカ国立衛生研究所 (national institutes of health, NIH) において集められた CT 画像のデータセットである。4459 人の患者から検出した 32,735 の病変がアノテートされた CT 画像と、その CT 画像の近傍スライスで構成されている。総計 500GB 近い大規模なデータセットであり、頭部から脚部に至るまで様々な部位の病変画像が収録されている。学習データ以外の評価・検証データについては、各病変のおおよその部位が“just reference”程度の確度でアノテートされており、骨、腹部、縦隔、肝臓、肺、腎臓、軟部組織、骨盤の 8 種類がある。

各 CT 画像には DICOM\_windows というそれぞれ既定の CT 値が定められており、ある画像は肺野用の  $[-1500, 500]$ 、別の画像は縦隔用の  $[-175, 275]$ 、といったように既定 CT 値は画像ごとに一般に同じではない。CT 値の場合の数は 55 通りだが、全体の 97% は  $[-1500, 500]$  または  $[-175, 275]$  である。この CT 値は肺野や縦隔といったそれぞれの部位を撮影する際に実際に医療現場で用いられる数値がアノテートされている。各 CT 画像は容量削減のために低いコントラストに変

連絡先: 西郷 彰, 株式会社 リクルートテクノロジーズ アドバンスドテクノロジーラボ, 東京都千代田区丸の内 1-11-1, saigo@r.recruit.co.jp

<sup>\*1</sup> 感度 (sensitivity) は再現率 (recall) と同値である。医療分野では感度という表現がされる。

<sup>\*2</sup> Free-response Receiver Operating Characteristic 曲線。横軸に画像 1 枚当たりの偽陽性数の平均値を、縦軸に感度をプロットして描画する。

換された形で収録されており、DICOM\_windows のアノテーションを用いて再変換することで復元できる。

また、各病変部位の位置情報として、人体内における相対座標が与えられている。この相対座標は、直立しているヒトを正面から見た時に左から右の方向を  $x$ 、腹から背の方向を  $y$ 、頭から足の方向（頭尾方向）を  $z$  としている。

## 2.2 先行研究

DeepLesion データセット構築の論文 [6] において、それを学習した物体検出手法の提案もされている。以下、この DeepLesion に対する物体検出手法を本論文では DLDet と書く。この先行研究の主な貢献は、これまで限られた状況でしか入手できなかった CT 画像について、大規模なデータセット DeepLesion を整備・公開したこと、そして実際にそのデータセットに対する検出手法 DLDet を提案したことである。この DLDet は DeepLesion 全体という包括的な病変を対象としている。

DLDet では評価データのおおよその部位情報を用いて病変クラスの擬似ラベルを学習データに割り振り、人体内の相対的な頭尾方向座標を自己教師付き学習による回帰で推定してから物体検出の学習を行っている [6]。包括性を持つために、DeepLesion データセットについて、医療現場で用いる上述の既定 CT 値ではなく  $[-1024, 3071]$  という広範囲の値で統一して指定した CT 画像を用いて学習している。このように学習した DLDet の FROC 曲線はおおよその部位ごとに報告されており、感度 0.9 を達成するには肝臓で 10、肺で 15 以上の画像 1 枚当たりの偽陽性数 (false positives per a image, FPI) が必要になる [6]。

## 2.3 課題

先行研究には次の問題点がある：

- 高感度を実現するために必要な FPI が少なくない。

CADe のユーザは大量の CT 画像を読影する医療サイドであり、真陽性を見逃しを減らした際の大量の偽陽性 BB は予測結果の視認性を下げることになる。

- 医療現場とは異なる CT 値を学習している。

$[-1024, 3071]$  という固定された CT 値を用いて復元している [6, 5]。これは医療現場で DeepLesion の元画像にアノテーションが付けられたときの CT 値（肺野  $[-1500, 500]$  や縦隔  $[-175, 275]$  など）とは異なり、現場で使われている値ではない。そのため、CADe が直面する実際的な状況と乖離したデータを学習している。

- DeepLesion のアノテーションは完全ではない。

DeepLesion データセットのアノテーションはすべての病変と疑われる部位を網羅していない [6]。非専門家には正常組織との区別が困難な病変へのアノテートの有無が混在しており、FPI の多さの原因になっていると考えられる。

以上の点が、DLDet を検出モデルとして医療現場で実践するにあたって課題となる。このうち、偽陽性 BB を削減する手法として、次小節で挙げる立体情報の利用という方法がある。

## 2.4 立体情報を考慮した先行研究

CT 画像での病変検出は、専門家は立体的な情報を利用して正常組織か異常かを識別している。例えば、血管や気管は連続している管上組織であるので、近傍スライスにも似た位置に同じ組織があることが多く、近傍スライスに写っていないかどうかで管上の組織ではなく病変と判断される。単体の CT 画像

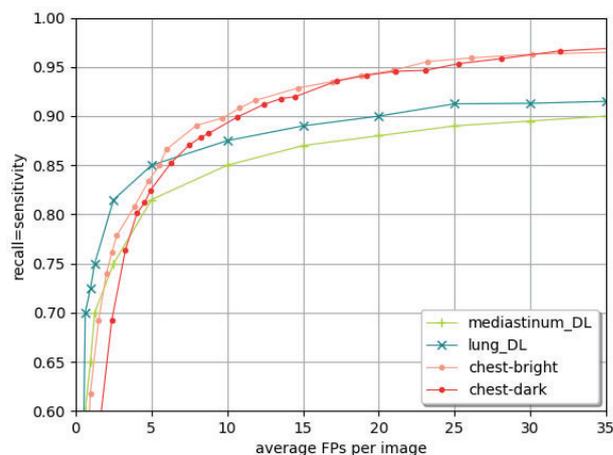


図 1: 平面情報のみを用いるとき、検証データについての各条件下での FROC 曲線。縦軸が再現率すなわち感度で、横軸が FPI の検証データセット内での平均値である。各曲線は次の通りである；+記号でプロットされた黄緑の曲線：先行研究 DLDet の縦隔病変に対するもの、x 記号でプロットされた青色の曲線：先行研究 DLDet の肺病変に対するもの、・記号でプロットされた濃赤色の曲線：chest-dark を学習したモデルのもの、・記号でプロットされた淡赤色の曲線：chest-bright を学習したモデルのもの。

だけでは、正常組織と病変の画像としての見た目は似通っており、FPI が多くなりやすい。そのため、立体情報も学習させることで同じ感度を達成する偽陽性数を減少させる研究がされている。病変部位も本来立体的であるため、すべてのスライスにアノテーションを行いすべてのスライスを学習することが理想的であるが、このナイーヴな方法はメモリやストレージの消費量が激しいだけでなく、アノテーションの手間が多大なものになるという欠点がある [5]。例えば、1 回の診察結果が 30 枚のスライスからなるとすると、1 枚のスライスのアノテーション作業の 30 倍の手間がかかり、この 30 枚の画像組を 1 データとすると単純計算としてメモリ消費量も 30 倍に増えることになる。特にアノテーションは専門の放射線医が行わなければならないため、十分なサンプルサイズを確保することが困難になる。

この問題を解決するために、単体の CT 画像に加えて前後数枚程度の画像組を 1 データとし、またアノテーションはその中心画像にのみに付けるという手法が提案されている [5]。CT 値は DLDet と同様の値に固定し、またスライス幅は 2mm になるように補正している \*3。

立体情報を用いたこの手法においては、偽陽性数 4 のときのおおよその部位ごとの感度と、全部位をまとめた FROC 曲線が報告されており、アノテーション付けられているスライスと前後 13 枚ずつの合計 27 枚を 1 データとした場合では DLDet における肝臓と肺の中間程度の性能が全部位まとめてであっても実現できている。前後 1 枚ずつの 3 枚の画像組を 1 データとした場合では、感度 0.9 を達成する偽陽性数はおよそ 20 である。ただし、こちらでも学習した CT 値は現場と異なる上述の固定値である。

\*3 補正手法は報告されていない。

### 3. 方法

前述の課題を踏まえて、次に述べる方法で数値実験を行った。使用したデータと前処理、ネットワークアーキテクチャ、実験条件の順で述べる。

#### 3.1 データ

学習・評価・検証それぞれに対応するデータの分割は、DeepLesion の設定を継承した。後述のロジスティック回帰による胸部、上腹部、下腹部の 3 クラス分類器の学習に用いたデータはこの分割に従った際の評価用データとした。

DeepLesion の CT 画像を用いたが、先行研究とは異なり各画像に対応付けられた CT 値で復元した。また、胸部の画像のみを用いた。実際の読影にあたっては、病変が存在する組織はわからなくとも、CT スキャンを行う人体の範囲は既知であるので、これを利用して胸部、上腹部、下腹部ごとに検出モデルを作成する方が高い精度が期待できるためである。本論文では胸部について実施した実験を報告する。

胸部、上腹部、下腹部の分類は次の方法を用いた。軟部組織や骨は全身に存在するため、これら以外の部位の大まかな分類として、縦隔と肺は胸部に、肝臓と腎臓と腹部は上腹部、骨盤は下腹部と定義した。縦隔などのアノテーションは評価データと検証データについているが、評価データの 6 割を学習させて 3 クラス分類器をロジスティック回帰により作成し、その Ridge 正則化パラメータを評価データの 2 割で調整した。このロジスティック回帰においては、病変の  $x$  及び  $z$  座標を説明変数とした。残った 2 割の評価データに対する精度 (accuracy) は約 94%であった。この 3 クラス分類器を元の学習データに適応した結果を以て、胸部、上腹部、下腹部の擬似ラベルとした。このうち、胸部・下腹部はそれよりも上・下であるような  $z$  座標のデータ (頭部や頸部・脚部) も含まれるクラスになっていたため、本研究で実際に胸部画像として用いたのは上記の胸部クラスの画像のうち、 $z \geq 0.39$  を満たすものとした。

画像のアノテーションに対して以上の前処理を実施した。画像そのものへの前処理として、各画像に DICOM\_windows としてアノテーションされている CT 値をそれぞれ対応付けてコントラストを復元した。ドメイン知識として、この CT 値は DeepLesion 作成にあたって対応する病変を専門の放射線医師が読影・検出した際の数値であるため、実際的な状況で用いられる値である。こうして復元した胸部画像のうち、 $[-1500, 500]$  が約 41%を、 $[-175, 275]$  が約 54%をそれぞれ占めていたため、それぞれの CT 値ごとに検出モデルを作成した。以下、CT 値が  $[-1500, 500]$  である胸部画像のデータセットを chest-bright、 $[-175, 275]$  については chest-dark とそれぞれ呼称する。ニューラルネットワークに入力する際には  $[0, 255]$  なる画素値から 127.5 を引いて 255.0 で割るという正規化を行った。また、後述するネットワークアーキテクチャが RGB 画像に対するものであったため、モノクロ画像から RGB への変換を施した。

#### 3.2 ネットワークアーキテクチャ

ネットワークアーキテクチャは YOLOv3 を用いた。YOLOv3 には物体検出を行う層が 3 個あり、それぞれにアンカーボックスが対応付けられており [4]。そのサイズはオリジナルの YOLOv3 で用いられているものを流用した。また、オリジナルでは 80 クラスの物体を検出する [4] が、CADe においては診断を自動化するのではなく病変の疑いのある箇所を自動的に検出することが重要なため、本研究では明らかな悪性腫瘍から良性の炎症・画像だけでは確認できないが病変の疑いのある箇所のすべてを「病変 (lesion)」という 1 クラスにまとめ

た。DeepLesion データセットにおいても、病変の種類はアノテーションされていない [6]。一般に、クラス数  $c$  の YOLOv3 において物体検出を行う層の直前の畳み込み層のフィルター数は  $3(c+5)$  で与えられるため [4]、本実験では 18 とした。

#### 3.3 実験

上述の方法で得られた学習・評価・検証データそれぞれの枚数は chest-bright では (3681, 937, 785)、chest-dark では (5074, 994, 1010) となった。この分割に従って chest-bright 及び chest-dark それぞれを上記の YOLOv3 で学習して病変検出モデルを作成した。活性化関数 leaky ReLU のパラメータ、学習率、勾配法は YOLOv3 の元論文 [4] と同様とした。評価指標としては評価・検証データに対して先行研究 [6] と同様に FROC 曲線を用いた。120 エポックの学習を行い、各エポックごとで評価データに対する YOLOv3 の損失関数値を取得し、120 エポック目に最も近い損失関数極小化エポックのモデルの FROC 曲線を描画した。立体情報を利用する場合は、スライス数を 3 とし、それに合わせてネットワークの入力層のチャンネル数を 3 倍した [5]。

FROC 曲線を描画にあたっては、non-maximum suppression 法の閾値を 0.1 で固定し、検出物体の confidence 閾値を変化させたときの FPI 及び感度をプロットした。比較のために、DLDet の肺及び縦隔の FROC 曲線を併記した。chest-bright は肺の、chest-dark は縦隔の病変を読影する際の CT 値を学習したモデルであるので、それぞれ DLDet の対応部位の FROC 曲線と比較した。

### 4. 結果

本論文では平面情報のみを利用する場合について報告する。検証データに対する FROC 曲線は図 1 のようになった。DLDet では達成できない感度 0.95 を chest-bright、chest-dark のどちらも 22,24FPI 程で実現できた。DLDet では比較的感度が高い肺についてであっても、感度 0.9 の達成には 20FPI 必要とされるのに対して、実験結果では chest-bright、chest-dark のどちらも 10FPI 前後で実現できるようになった。感度 0.85 については、chest-bright は DLDet と遜色なく 5.5FPI 程度で実現することができ、chest-dark では DLDet の 10FPI よりも少なく約 6FPI で実現された。感度 0.8 については、chest-bright では DLDet の 2FPI にやや劣る約 3.5FPI であるものの、chest-dark は DLDet の約 4.5FPI よりわずかに少なくおよそ 4FPI で実現可能であった。感度ごとの実画像への検出例を図 2 及び 3 に示す。

### 5. 考察

感度は真陽性率であるため、この値が高いほど (偽陽性の増加を犠牲にしつつ) 病変の見逃しを減らすことができる。病変の検出を目的としているため、高い感度における偽陽性数の少なさが重要であると考えられる。感度 0.85 以上の領域 (高感度領域と呼ぶ) において、実現可能な最大感度を引き上げただけでなく特に感度 0.9 以上では先行研究と同程度の感度を實現する FPI を著しく削減することができた。感度が 0.85 以上 0.9 未満においても先行研究と同程度かそれより少ない FPI であった。また、感度が 0.8 以上 0.85 未満の領域 (中感度領域と呼ぶ) では、先行研究よりわずかに多い FPI で同じ感度を達成できているため、FPI の少なさを重視しても中感度領域では検出性能が大きく劣化することはないと考えられる。以上から、本研究では中感度領域における検出性能をほぼ落とすこと

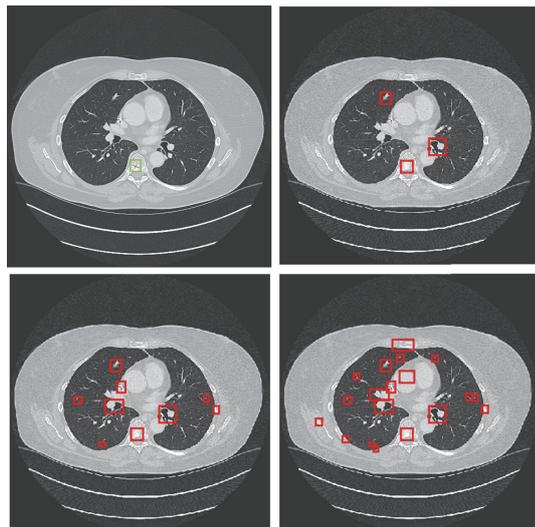


図 2: chest-bright の検出例. 左上がアノテーションで, 右上, 左下, 右下の順に感度が 0.8, 0.85, 0.9 となっている.

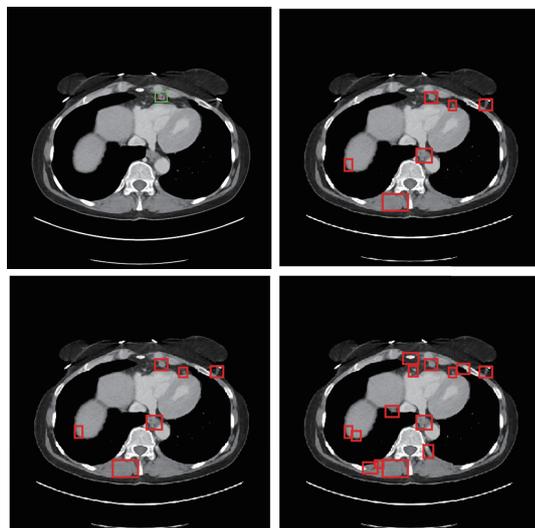


図 3: chest-dark の検出例. 左上がアノテーションで, 右上, 左下, 右下の順に感度が 0.8, 0.85, 0.9 となっている.

なく, 高感度領域において先行研究よりも十分少ない FPI を達成するという貢献ができたと考えられる.

本研究における胸部の定義として肺及び縦隔の病変の位置座標を使った分類モデルの出力を用いた. この分類モデルは病変の位置座標のみから胸部か上腹部か下腹部かを判定する. 実際の病変検出においては, CT スキャンを走査する範囲は既知でもどの部位に病変が存在するかは事前にはわからず, かつ胸部の骨や軟部組織の病変も存在する可能性がある. chest-bright 及び chest-dark はどちらもそれぞれ肺と縦隔の病変に限るわけではなく, 軟部組織や骨や一部肝臓も含む病変が混在しており, 胸部範囲の CT スキャン結果のデータセットとして現実的であると考えられる. また, 先行研究において軟部組織や骨の病変の検出の困難さが報告されている [6]. このように, 実際的かつ検出困難な病変も含んだデータセットで, 肺や縦隔のみで描画した先行研究の FROC 曲線を上回ったことは, 現場での応用に近づくことができたと考えられる.

## 6. 結論

DeepLesion データセットの胸部画像に対し, 現場で用いられる CT 値を YOLOv3 に学習させた結果, 先行研究よりも実際の設定で同じ平均 FPI に対する感度を向上させることができた. 今後の課題として, 特に感度と FPI がともに小さい領域での性能向上のためには, モデルが検出した偽陽性が本当に偽陽性であるかを再検討する, すなわち DeepLesion のアノテーション不備を改善するという施策が考えられ, 本稿投稿時点 (2019 年 2 月 8 日現在) で検討中である.

## 謝辞

本研究にあたり, 医療現場での読影業務と用いる CT 値について慶応義塾大学病院の荒井先生による指導を受けた.

## 参考文献

- [1] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.
- [2] Kunihiro Fukushima and Sei Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer, 1982.
- [3] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.
- [4] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.
- [5] Ke Yan, Mohammadhadi Bagheri, and Ronald M Summers. 3d context enhanced region-based convolutional neural network for end-to-end lesion detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 511–519. Springer, 2018.
- [6] Ke Yan, Xiaosong Wang, Le Lu, and Ronald M Summers. Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of Medical Imaging*, 5(3):036501, 2018.
- [7] Zhijie Zhu, Ghazaleh Haghiashtiani, and Michael C McAlpine. Biophysical sensing in deep tissue via mri. *Nature Biomedical Engineering*, 3(1):11, 2019.