視線データを活用した深層学習による胸部 X 線写真の診断的分類

Diagnostic Classification of Chest X-Rays Pictures with Deep Learning Using Eye Gaze Data

井上 大輝 [*1] Taiki Inoue
木村 仁星 [*2] Nisei Kimura
中山 浩太郎 [*2, 3] Kotaro Nakayama
作花 健也 [*4] Kenya Sakka
Abdul G. A. R. [*2] Abdul G. A. R.
中島 愛 [*5] Ai Nakajima
Radkohl P. [*2] Radkohl P.
岩井 聡 [*6] Satoshi Iwai
河添 悦昌 [*6] Yoshimasa Kawazoe
大江 和彦 [*6] Kazuhiko Ohe
松尾 豊 [*2] Yutaka Matsuo

[*1] 東京大学大学院薬学系研究科
Graduate School of Pharmaceutical Sciences, The University of Tokyo

[*2] 東京大学大学院工学系研究科
Graduate School of Engineering, The University of Tokyo

[*3] NABLAS 株式会社
NABLAS Inc.

[*4] 東京大学大学院新領域創成科学研究科
Graduate School of Frontier Sciences, The University of Tokyo

[*5] Aalto University
Aalto University

[*6] 東京大学大学院 医学系研究科
Graduate School of Medicine, The University of Tokyo

Automatic diagnosis of chest X-ray pictures with deep learning has been extensively studied in recent years. In order to improve the accuracy, it is important how to input small localized areas which are disease specific while at the same time using the information that can be obtained by the whole picture. We considered that human eye-gaze fixations can be a biomarker that indicates areas specific to disease. In this study, we propose a deep learning model utilizing eye-gaze data. We demonstrate that the classification shows the better accuracy on using eye-gaze data of experienced doctors than eye-gaze data of non doctor or non-use of eye-gaze information.

1. はじめに

今日，肺がんを始めとした病態の診断を実施する上で胸部 X 線写真は広く利用されているが，これには専門家である放射線医による適切な読影が必要である．しかし，放射線医の人材不足は世界中で共通の問題であり，放射線医にかかる負担を軽減する技術や仕組みの開発が望まれている．このような背景ならびに近年の深層学習技術の発展に後押しされ，胸部 X 線写真の自動診断は現在盛んに研究されている．

胸部 X 線写真の深層学習による診断の先駆けとしては 2017 年に発表された CheXNet [Rajpurkar 17] が挙げられる．これは既存の深層学習ネットワークである DenseNet [Huang 16] を基盤としたモデルであり，112,120 枚の胸部 X 線写真を使用して 14 種類の病態を分類した結果，放射線医を上回る診断精度を記録した．しかし翌年には，AG-CNN [Guan 18] と呼ばれるモデルが CheXNet の診断精度を更新している．CheXNet が胸部 X 線写真全体をダウンスケールした画像を入力としているのに対し，AG-CNN は Attention 機構 [Mnih 14] を採用することで，胸部 X 線写真全体に加えて，異常と疑われる領域の局所画像を入力としている．このように，胸部 X 線写真を診断する深層学習モデルの精度向上には異常と疑われる局所画像を正確に抽出し，入力とすることが重要であると言える．

AG-CNN は胸部 X 線写真から得られる情報を基に異常と疑われる局所画像を抽出しているのに対し，本研究では「診断時に医師が凝視している領域を異常と疑われる局所画像として抽出できるのではないか」という仮説を立てた上で，視線データを基に抽出された局所画像を入力とする深層学習モデルを構築した．

連絡先：井上大輝，東京大学，〒113-0033，東京都文京区本郷 7-3-1，taiki-inoue45@g.ecc.u-tokyo.ac.jp

2. 手法

2.1 データセット

本研究では，東京大学附属病院より正常 47 枚，異常 48 枚，合計 95 枚の胸部 X 線写真を提供を受けた．その 95 枚 の写真を用いて，医師 5 名および医師訓練を受けていない成人男女 5 名に対して表示された胸部 X 線写真が正常・異常のいずれであるかを分類する試験を各被験者ごとに 2 回ずつ行い，その際に各被験者の視線データを収集した (Figure 1).

正常・異常の分類は，ディスプレイ上に 95 枚の胸部 X 線画像の中から 1 枚ずつランダムに画像を表示させ,被験者がマウスの左右クリックによって正常・異常の分類を行えるソフトウェアを独自に開発し，それを利用した．

また視線データは，Leveque らによる研究 [Leveque 18] を参考に，赤外線型のアイトラッキングデバイス Tobii Eye Tracker 4C

を利用して収集することとした. 得られた視線データは, 約 11 ミリ秒に 1 回の間隔で視線の x, y 座標が記録された時系列データである.

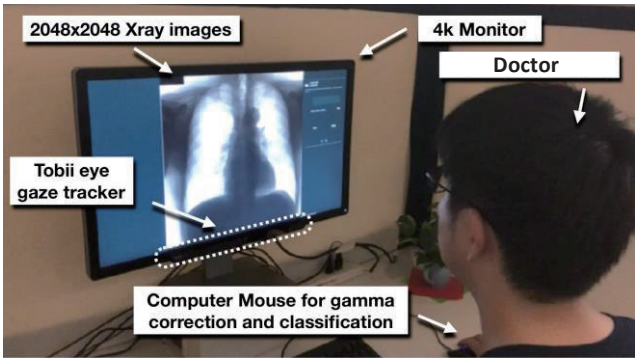

Figure 1 | 視線データ収集時の様子

2.2 胸部 X 線写真の前処理

入力データのもとになる胸部 X 線写真は, 既存の深層学習モデルである CheXNet および AG-CNN を参考にして, 平均を 0.485, 標準偏差を 0.229 として正規化を行った. また, データ拡張と頑健性向上を目的とした, 鏡映や回転・せん断変形による画像の水増し方法は, 本研究では実施しなかった. これは, 胸部 X 線写真は左右非対称であり, また読影では各部位の傾きなどが重要な情報になるためである.

2.3 モデルの構造

本研究で提案するモデル (Figure 2) では, まず Maximum Sampling を行うことで, 胸部 X 線写真の中で視線が最も集まっている領域 (局所画像) を抽出する. そして, 胸部 X 線写真を 2048×2048 から 224×224 にダウンスケールした画像 (全体画像) と, 局所画像を重ねて複数チャネルとしたものを DenseNet-121 の入力とし, 正常および異常の確率を出力させた.

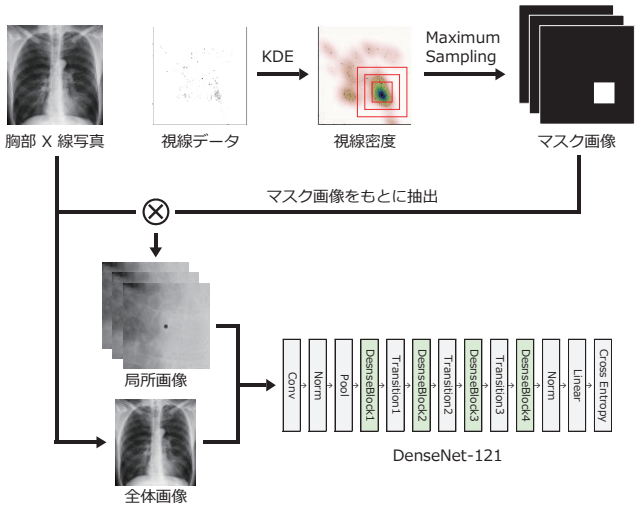

Figure 2 | 本研究で提案するモデルの構造

(1) Maximum Sampling

局所画像を抽出するため, 確率変数の確率密度関数を推定するノンパラメトリック手法の 1 つである KDE (Kernel Density Estimation) を用い, 視線データ (Figure 3-a) から視線密度 (Figure 3-b) を算出した. そして, 視線密度が最も高い点を中心として局所画像を抽出した (Figure 3-c).

また, 人間の視野が複数階層により成り立っている [Pickrell 03] ことを考慮し (Figure 4), 最大で 4 階層とする局所画像の階層化を行った. 階層 r_i の局所画像は, 2048×2048 の胸部 X 線写真の中で最も視線密度が高い点を中心として, 一辺が s_{r_i} の画像となるよう抽出した後, 224×224 にダウンスケールすることで得られる. s_{r_i} は以下の数式に従う.

$$s_{r_i} = 224 \div (1 - 0.25 \times (i - 1)), \qquad i = \{1, 2, 3, 4\}$$

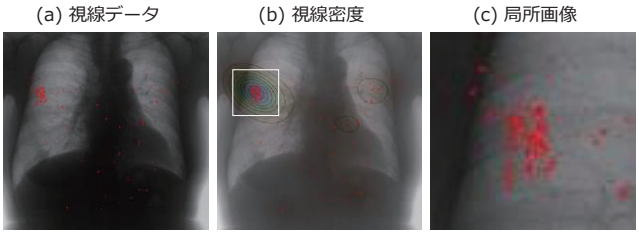

Figure 3 | Maximum Sampling の過程

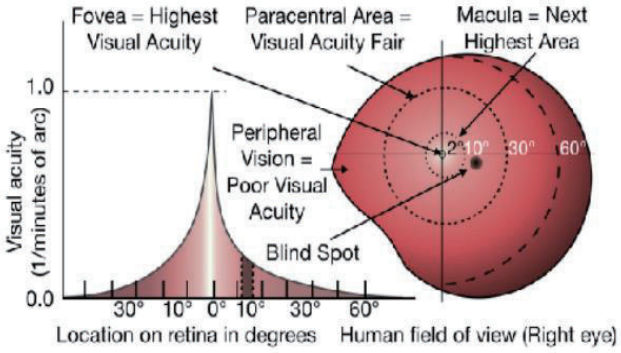

Figure 4 | 人間の視野に関するモデル

(2) 畳み込みニューラルネットワーク

畳み込みニューラルネットワーク (以下 CNN) は深層学習モデルの 1 つで, 画像処理などのタスクにおいて高い性能を示している [Krizhevsky 12]. 特に CNN の派生である DenseNet [Huang 16] は, CheXNet および AG-CNN の基盤となっているネットワークであるため, 本研究でも同様に DenseNet を基盤としてモデルを構築した.

なお, 入力次元におけるチャネル方向の扱い方に関しては本来の DenseNet と異なる方式を採用している. DenseNet では RGB の色情報チャネル方向に 3 次元として組み込んでいるが, 本研究において色情報はグレースケールのため 1 次元のみであり, チャネル方向では色の代わりに視野の階層数を表している. 例えば階層数が 2 の場合, 入力次元は 2 となる.

出力層は Softmax (2 クラス), 活性化関数は ReLU, 目的関数は CrossEntropyLoss, 最適化手法は Adam ($\beta1$=0.9, $\beta2$=0.999, ε=10^{-8}), 学習率は 10^{-5}, 事前学習は ImageNet, エポック数は 100 とした.

2.4 評価方法

本研究で使用したデータセットは 190 枚 (ユニークな 95 枚の胸部 X 線写真に対して n = 2 で視線データを収集) となっており, これらを訓練データ 152 枚, テストデータ 38 枚に分割した.

その際，訓練・テストデータ間で，同一の胸部 X 線写真を出現させないようにし，正常・異常の割合を一定にしている．そして，訓練データを用いて学習させ，各エポックごとにテストデータを用いて AUROC を算出し，これを比較することでモデルの良し悪しを評価することにした．

この際，視線データの有用性を精査するため，胸部 X 線写真全体をダウンスケールした画像のみを入力とした場合の結果を Baseline として置き，本研究で提案するモデルと比較した．

また局所画像の階層化の有用性を精査するため，視線データの局所画像部分の階層数は，1 階層 (全体画像と合わせて 2 次元の入力) から 4 階層 (全体画像と合わせて 5 次元の入力) の場合までの AUROC をそれぞれ算出し，比較を行った．

3. 結果・考察

本研究におけるモデル評価の結果は Figure 5，Table 1 に示す通りである．A～E はそれぞれ医師訓練を受けていない 5 名，F ～ J は医師 5 名を表している．Table 1 の Actual Accuracy は各被験者が正常・異常の判定を行った結果の正答率であり，n Hierarchies は，全体画像＋1 ～ n-1 階層の局所画像を入力とした場合の AUROC を表している．

医師訓練を受けていない A ～ E の被験者の視線データを使用してテストを行った場合，全ての階層数において Baseline を下回る結果となった．一方で，医師である F ～ J の視線データを利用した場合，F および J が Baseline を上回る結果が得られた．なおこの際，2 ～ 5 Hierarchies としてテストを行ったが，各階層間で AUROC に有意な差は認められなかった．

これらの結果より，視線データはモデルの AUROC に影響を及ぼし，特に F や J のような Actual Accuracy の高い医師の視線データを採用することによって胸部 X 線写真における正常・異常の分類において効果的な作用をもたらすことが示唆された．

これはすなわち，より Actual Accuracy の高い被験者の視線データを収集・採用することによって更にモデルの精度が改善する余地があることを表している．

一方で，Actual Accuracy が最も高かった医師 G においては，モデル精度が全ての階層において Baseline を下回る結果であり，Actual Accuracy の高い視線データが必ずしもモデル精度向上に貢献しないことが分かる．この点に関しては診断における熟練度の違いが画像内の各部位への視線の滞留時間等の違いを生む可能性等も視野に入れた上で今後更なる検証が必要である．

4. 今後の展望

本研究を通じて，医師の視線データの利用が胸部 X 線写真に対する深層学習を利用した診断的分類において有効性を持つことが分かった．同時に，Actual Accuracy の高い医師の視線データ使うことが必ずしもモデルの改善を保証しないことも見た．これはすなわち本研究により開発したモデルにおいて捉え切れていない特徴の存在が示唆されているということであり，今後そうした特徴の検証を視野に入れた更なるモデル改善の余地が期待できる．

そうした特徴の一例としては，例えば視線データが持つ時系列的な情報の存在が挙げられる．今回の研究では Maximum Sampling により視線データを視線密度として取り扱ったため時系列的な情報が失われていたが，今後モデルに時系列的な情報を取り入れることにより今回の研究では表現し切れなかった情報を捉え，更なるモデルの改善に繋がる可能性が考えられるため，そうした観点を踏まえて引き続き検証を続けていきたい．

5. 謝辞

本研究は JSPS 科研費 JP25700032，JP15H05327，JP16H06562，国立研究開発法人日本医療研究開発機構 (AMED) の平成 28 年度「臨床研究等 ICT 基盤構築研究事業」の助成を受けたものです．

6. 参考文献

[Rajpurkar 17] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, Andrew Y. Ng: “CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning”, arXiv: 1711.05225, 2017.

[Huang 16] Gao Huang, Zhuang Liu, Laurens van der Maaten, Kilian Q. Weinberger: “Densely connected convolutional networks”, arXiv: 1608.06993, 2016.

[Guan 18] Qingji Guan, Yaping Huang, Zhun Zhong, Zhedong Zheng, Liang Zheng, Yi Yang: “Diagnose like a Radiologist: Attention Guided Convolutional Neural Network for Thorax Disease Classification”, arXiv: 1801.09927, 2018.

[Mnih 14] V. Mnih, N. Heess, A. Graves et al., “Recurrent Models of Visual Attention”, in Advances in neural information processing systems, 2014.

[Leveque 18] Lucie Leveque, Hilde Bosmans, Lesley Cockmartin, Hantao Liu: “State of the Art: Eye-Tracking Studies in Medical Imaging”, IEEE Access, 2018.

[Keane 03] Miller-Keane, Marie T. O'Toole, EdD, RN, FAAN: “Miller-Keane Encyclopedia & Dictionary of Medicine, Nursing & Allied Health”, Saunders, 2003.

[Krizhevsky 12] A.Krizhevsky, I.Sutskever, G.E.Hinton: “ImageNet Classification with Deep Convolutional Neural Networks”, NeurIPS, 2012.

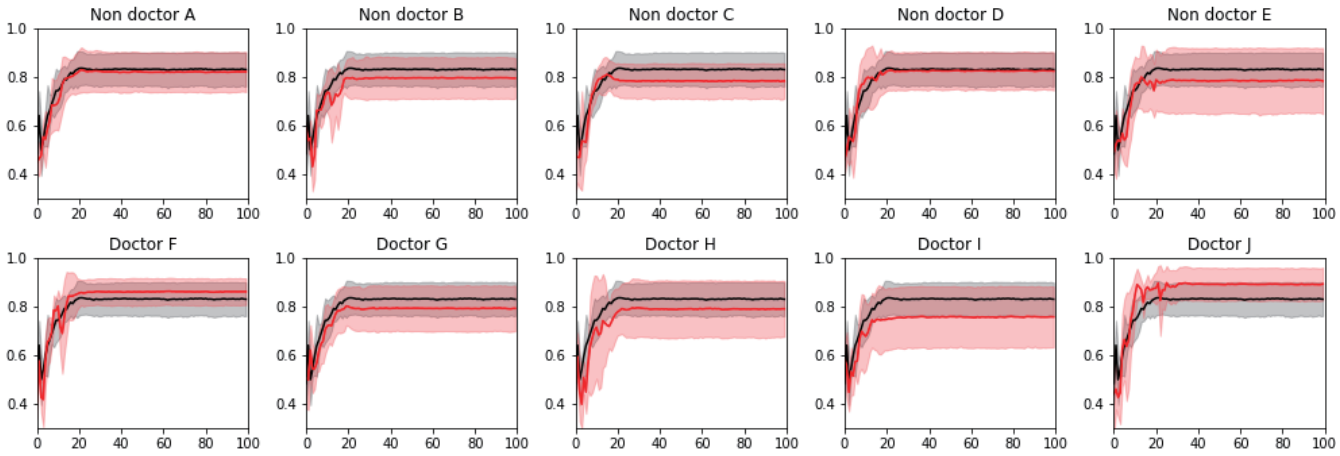

Figure 5 | 3 Hierarchies における各被験者のテストデータに対する精度の推移. 赤線:各被験者のテストデータに対する精度, 黒線:Baseline の精度, 薄い赤で囲まれた領域:各被験者のテストデータに対する精度±SD, 薄い黒で囲まれた領域:Baseline の精度±SD.

Table 1 | 各被験者の視線データおよびモデルごとの比較.

	Non doctor					Doctor				
Model	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
Actual Accuracy※1	0.770	0.750	0.530	0.620	0.540	0.830	0.860	0.740	0.780	0.850
2 Hierarchies※2	0.818	0.814	0.814	0.824	0.829	**0.856**	0.827	0.837	0.820	**0.856**
3 Hierarchies※2	0.829	0.800	0.810	0.827	0.800	0.863	0.802	0.795	0.760	**0.898**
4 Hierarchies※2	0.789	0.784	0.780	0.801	0.687	**0.882**	0.834	0.760	0.769	0.859
5 Hierarchies※2	0.770	0.751	0.758	0.832	0.658	0.866	0.798	0.789	0.776	**0.870**
Baseline※2	0.836									

※1:各被験者が正常・異常の判定を行った結果の正答率

※2:AUROC の値