2B1-3

深層学習による胸部X線写真からの診断補助

Diagnosis support from Chest X-ray pictures with Deep Network

黒滝 紘生 *1 Hiroki Kurotaki　中山 浩太郎 *1 Kotaro Nakayama　上原 雅俊 *2 Masatoshi Uehara　山口 亮平 *3 Ryohei Yamaguchi　河添 悦昌 *4 Yoshimasa Kawazoe
大江 和彦 *3 Kazuhiko Ohe　松尾 豊 *1 Yutaka Matsuo

*1東京大学工学系研究科技術経営戦略学専攻
The Department of Technology Management for Innovation, Graduate School of Engineering, The University of Tokyo
*2東京大学工学部計数工学科
Department of Mathematical Engineering and Information Physics, School of Engineering, The University of Tokyo
*3東京大学大学院医学系研究科
Graduate School of Medicine, The University of Tokyo
*4東京大学医学部附属病院
The University of Tokyo Hospital

X-ray pictures of the chest are used to detect abnormalities or diseases (e.g. rib fracture, lung cancer, pneumonia etc). It is beneficial if we managed to support clinical judgement by doctor automatically with machine learning models. We propose a method to detect abnormal images from chest X-ray images for both unsupervised and supervised settings. For unsupervised task we used variational autoencoder (VAE) and for supervised we used convolutional network (CNN). We verified our method with a chest X-ray image dataset provided from The University of Tokyo Hospital. Our method successfully discriminated the abnormal images from the normals with high accuracy.

1. はじめに

2 次元胸部 X 線写真は，肺がんを始めとする病態の診断のために有用な，多種の情報を含んでいる．しかし，診断を適切に下すことは難しく，専門医による判断が必要となる．機械学習によって，専門医による診断の自動補助ができれば，医療の現場において有用だと考えられる．例えば，健康診断にて撮影された多数の 2 次元胸部 X 線写真を，自動で診断したり，簡単な所見を生成できれば，どの患者が精密検査を受けるべきか，より素早く判断できる．

一方で，畳み込みネットワーク (以下 CNN) や変分オートエンコーダ (以下 VAE) などの深層学習モデルは，画像処理を主とする様々なタスクにて高い性能を達成しており，胸部 X 線写真の処理にも有効であると考えられる．

そこで本研究では，深層学習の高い画像処理性能を，診断補助の問題に応用して，医師の所見作成を補助するため，肺の X 線写真を中心とするデータより，肺の異常検知を行う方法を提案した．医用画像のデータセットに症状のラベルや所見をつけるためには，医師による判断が必要となるため，多量の教師ラベルつきデータセットを得ることは難しい．そこで我々は，教師なしと教師ありのそれぞれのタスク設定に対する，異常検出の手法を提案した．具体的に述べると，教師なし学習では，生成モデルの一種である VAE を用い，正常データでモデルを学習させた後，テスト用データの周辺対数尤度を，正常・異常データ間で比較した．教師あり学習では，CNN によって正常ラベルと異常ラベルの識別問題として学習させた．2 つの手法は異なる状況を互いにカバーしており，将来的には双方を組み合わせることも考えられる．

連絡先: 黒滝 紘生, 東京大学工学系研究科, 〒 113-0033 東京都文京区本郷 7-3-1 工学部 2 号館, kurotaki@weblab.t.u-tokyo.ac.jp

深層学習による 2 次元胸部 X 線画像からの異常検出の先行研究としては，[Shin 16] や [Ypsilantis 16] がある．[Shin 16] は CNN とリカレントネットワーク (RNN) の組み合わせにより，肺の異常箇所や種類を表すラベルを識別させた．[Ypsilantis 16] はアテンションモデルによって，心拡大および医療機器の埋め込みを検出させた．このとき，画像のどこに着目すれば，所見に必要な情報が得られるかを，モデルに学習させることができた．また，[Kim 16][Hwang 16] は，画像全体についたラベルから，ピクセル単位の異常箇所推定を行う手法を提案した．しかし，我々の知る限りでは，生成モデルを用いて教師なし学習による異常検知を行った研究はなく，教師あり学習のタスクでも性能の発展向上に余地がある．我々の研究は，これらの先行研究を補完すると共に，新たな知見を付け加えることを目標としたものである．

我々は，東大医学部附属病院より提供された，医師による所見ラベルつきの，2 次元胸部 X 線写真を中心とした画像データセットを用いて，提案手法を検証した．提案手法によって，データセットのうちから，部位や撮影条件の異なる画像および，診断の上で異常とされた胸部 X 線画像を，通常の胸部 X 線画像から高い精度で識別できることを確かめた．

2. 手法

本研究では，深層学習モデルのうち，変分オートエンコーダと畳み込みニューラルネットワークを用いた．

2.1 変分オートエンコーダ

変分オートエンコーダ (以下 VAE)[Kingma 13] は，深層学習による生成モデルの一種で，入力されるデータベクトルと，データの説明変数を表す隠れ状態ベクトルの，同時分布をモデル化している．学習はデータセット $\mathbf{X}$ の各点 $\mathbf{x}$ の対数周辺尤

度 $\log p_{\boldsymbol{\theta}}(\mathbf{x})$ の変分下界

$$\mathcal{L}(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\phi}) = \log p_{\boldsymbol{\theta}}(\mathbf{x}) - D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})) \quad (1)$$

の最大化によって行われる．ただし $\boldsymbol{\theta}$ は分布 $p_{\boldsymbol{\theta}}$ の，$\boldsymbol{\phi}$ は事後近似分布 $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ のパラメータである．学習の後，モデルにテスト画像データを入力すると，画像の $\log p_{\boldsymbol{\theta}}(\mathbf{x})$ の変分下界 (以下 VLB) を評価できる．VLB の値がゼロに近いほど，訓練データで学習されたモデルと，入力したテストデータの分布が近いと考えられる．これを我々の課題に即していうと，正常な肺の画像を訓練データとして学習を行った後，テストデータとして，正常または異常な画像を入力したとき，正常な画像における VLB の方が，異常な画像に比べて，よりゼロに近く出ると予想される．よって VLB を比較することで，異常な肺の画像を検知できると考えられる．

我々のモデル構造では，計算時間の都合上，入力画像を後述する方法で 100×100 に縮小した．これを 10,000 次元の入力ベクトルに変換の後，VLB の観測データとして入力した．VLB の隠れ変数の次元は 256，バッチサイズは 100，ϵ は 1.0，最適化法は RMSProp とした．学習は 50 エポック行った．全ての VLB は，3 回の実験による平均値を取った．

2.2 畳み込みネットワーク

畳み込みネットワーク (以下 CNN)[Krizhevsky 12] は深層学習モデルの一つで，画像処理などのタスクにて高い性能を挙げている．我々は CNN を用いて，与えられた画像が正常か異常かを識別させる，2 クラス識別問題を行った．入力画像は，VAE と同様 100×100 にサイズ変換したものを，2 次元のまま用いた．ネットワーク構造は，3×3 畳み込み層 (フィルタ数 32) を Conv32，2×2 最大プーリングを MP，全結合層を FC として，入力層 - Conv32 - Conv32 - MP - (Dropout0.25) - FC128 - (Dropout0.5) - Softmax(2 クラス) とした．バッチサイズは 128，活性化関数は ReLU，目的関数は多クラスクロスエントロピー，最適化手法は Adam とした．学習は 50 エポック行った．

2.3 データ画像の前処理

データの胸部 X 線写真は，計算時間の都合上，100×100 に縮小して用いた．このとき，データの性質上，オリジナルの写真が正方形になっておらず，単純な縮小では 100×100 にならない場合が多くあった．その場合，まず，長辺が 100 ピクセルになるまで比を維持して縮小した後，画像中の最小の値で短辺を 100 ピクセルまで埋めたてた．ここで最小の値を選んだのは，X 線写真の本来の外周部を模すためで，外周は背景のため値が最も小さくなる傾向が見られたからである．

また，数字認識や物体認識のデータセットで用いられる，鏡映や回転・せん断変形によるデータ拡張と頑健性向上の手法は，本研究では行わなかった．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．

3. 実験

我々は，前章で述べた VAE と CNN の各手法によって，2 次元肺 X 線画像の識別タスクを行った．

3.1 データセットとその分割方法

実験に用いたデータセットの概要を，表 1，2 に示した．データセットは，東京大学医学部附属病院に拠るもので，モノクロ X 線写真データ 17,848 枚のそれぞれに，撮影条件などを表すメタデータと，医師による診断データが紐付けられている．メタデータは DICOM という形式で記述されている．診断データは医師が作成したもので,「所見」と「最終判断」の 2 つから成る.「所見」は，画像に見られる陰影パターンや診断を，5〜30 字ほどで自由に記した文字列である.「最終判断」は，病的意義の有無を示している．

我々はデータセットを以下の方法で分割して，正常と異常のいずれかにデータ群を振り分けた．ここで，DICOM 形式における”Modality”は「撮画手段」，”CR”は「コンピュータ X 線撮影 (Computed radiography)」を表している．

1. **撮画手段識別** ”Modality”が”CR”なら正常，他なら異常として分別
2. **診断識別** ”Modality”が”CR”のデータの中で，さらに診断が正常か異常かで分別

まず，メタデータに含まれる撮影条件のうち，項目”Modality”が”CR”のものと，それ以外に 2 分割した．この”CR”は，2 次元胸部 X 線写真における一般的な撮画手段である．データには，脳の断面など，肺以外のデータが含まれており，そのため”CR”以外のデータが混入している．また肺の画像でも，数は少ないものの，”CR”以外の条件で撮られたものが含まれている．肺以外の部位のデータや，撮画手段が異なるデータを検出することは，モデル適用の前提条件を揃えるための，重要なステップであると考えられる．この検出タスクを行うため，”CR”を肺のデータ，”CR”以外を肺以外のデータと見なして，データセットを二分した．

その上で，”CR”であるデータのみを更に，医師による診断が正常か異常かで二分した．具体的には，

- 診断データの「所見」欄が「異常所見なし」「異常なし」のいずれか
- 「最終判断」欄が「異常なし」

の 2 条件を同時に満たす画像のみを，正常データとした．一つも満たさないものは異常データに振り分け，片方のみを満たすデータは，カテゴリが曖昧なため，取り除いた．

なお，次に述べる VAE 実験と CNN 実験とで，データ総数が異なっているが，これはバッチサイズの処理の都合による．また，データ数合計が 17,848 枚に満たないのは，先述のように，ラベル付けが曖昧なデータを取り除いたためでもある．

3.2 VAE による実験

VAE では，正常なデータのみを 4:1 に分割して，順に訓練およびテストセットとした (表 1)．異常データは，全てテストセットとした．正常な訓練セットで学習させたモデルに，正常と異常のテストセットを独立に入力し，3 回の VLB 値の平均を比較して，仮説通り正常なデータに対する VLB の方が 0 に近くなっているか検証した．

3.3 CNN による実験

CNN では，正常と異常のデータを混ぜた後，4:1 に分割して，順に訓練およびテストセットとした (表 2)．訓練セットで学習させたモデルを用いて，テストセットにおける識別精度によって評価を行った．

4. 結果・考察

表 3 は，VAE 実験における，各テストセットにおける VLB の 3 回の平均値を示している．正常テストデータは共通のた

め，一マスに書いている．撮画手段が CR でない異常テストデータでは，CR である正常テストデータに比べて，VLB 値が負に大きな値をとっている．これは，異常テストデータよりも正常テストデータの方が，正常訓練データの分布に近いことを示しており，提案手法である，VLB の比較による異常検出を確認できたと考えられる．一方，診断において異常とされたテストデータの VLB は，正常テストデータとほとんど違いがなく，むしろ僅かではあるが，より 0 に近くなっている．予想に反した理由として最も大きなものは，画像の解像度を 100×100 に縮小したためだと考えている．なぜなら，医師が X 線画像から診断を下すときは，微細な陰影の変化を追うため，2,000×2,000 程度の画像を使うことが多いからである．

図 1〜6 は，各テストセットについて，最も VLB が低かった画像および高かった画像 10 枚を，VLB 値と共に示している．図 1 と図 2 を比較すると，非 CR 画像の中でも，肺でない画像は VLB が大きな負の値を取っており，一方で肺の画像は小さな負の値を取り，ほとんど正常のものに近づいていることがわかる．また，図 3〜6 を比較すると，正常画像と異常画像に共通して，撮影体位が崩れているときや，元データにおいて外周の枠が乱れているときに，VLB が負に大きくなっている．これらも予想の正しさを裏付ける結果と言える．なお，図 6 において，肺ではないデータが最も負の小さい VLB となっている．この原因の調査は今後の課題であるが，全体的な階調が肺に近かったためではないかと考えている．

また表 4 は，CNN 実験における，各テストセットでの精度を示している．精度は正しく識別できたデータの個数で算出している．Modality が”CR”(=正常な撮画手段) か否かを，精度良く識別できていることがわかる．また，診断所見の識別においても，ある程度までの予測が可能になっていると考えられる．

図 7〜10 に，CNN による識別結果と画像を，誤差関数の値でソートして，低い順および高い順に抽出した．図 7 の先頭 3 枚のみが，CNN によるラベル識別に失敗した画像であり，他の画像はどれも正しく識別できている．肺同士でも，撮画手段が異なってさえいれば，高い精度で識別できていることがわかる．図 9，10 では，診断で異常とされた肺の識別を行っている．この識別問題の精度を向上させるためには，解像度の向上や，異なる解像度サイズを階層的に扱うネットワーク構造の採用などが考えられる．これらの検証は，今後の課題である．

5. まとめ

本研究では，2 次元胸部 X 線写真における，医師の所見作成を補助するため，深層学習モデルを用いて肺の異常画像を検出する方法を提案した．教師なし学習の設定では，変分オートエンコーダ (VAE) を，教師あり学習では畳み込みネットワーク (CNN) をそれぞれ用いた．実際の臨床データを用いて実験を行い，提案手法によって，撮画手段の異常や，診断結果としての異常を，高い精度で識別できることがわかった．この結果は，医師による診断所見を，予測および生成するための基礎になると考えられる．

今後の課題としては，より精度の高い識別のため，モデルを改善していくことが必要である．例えば，入力画像の解像度を上げたり，肺画像の各領域を分割して，専用モデルを適用する方法が考えられる．また，アテンションモデルや階層型 CNN，レイヤースキップといったより高度なネットワーク構造の利用，正規化や勾配クリップ，重要度重みといった深層学習技術の適用による性能向上の検証も，重要な課題である．さらに，将来的には二つの手法を組み合わせることも考えられる．

表 1: VAE 実験のデータセット詳細

	CR-異常無し	CR-異常あり	非 CR
訓練	9,400	-	-
テスト	2,300	5,000	600

表 2: CNN 実験のデータセット詳細

データセット		総数	正常	異常
撮画手段識別 (CR)	訓練	9,964	9,436	528
	テスト	2,492	2,359	133
診断識別	訓練	13,515	9,964	4,079
	テスト	3,379	2,359	1,020

表 3: VAE 実験における，各テストセットでの VLB

タスク	正常	異常
撮画手段識別	-4813.6	-8758.1
診断識別		-4791.0

表 4: CNN 実験における，各データセットタスクの精度

タスク	テストデータ数	誤り数	識別精度
撮画手段識別	2,492	3	98.9%
診断識別	3,379	1,020	69.8%

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．

参考文献

[Hwang 16] S.Hwang, H.E.Kim: Self-Transfer Learning for Fully Weakly Supervised Object Localization. arXiv preprint arXiv:1602.01625,2016.

[Kim 16] H.E.Kim, S.Hwang: Deconvolutional Feature Stacking for Weakly-Supervised Semantic Segmentation. arXiv preprint arXiv:1602.04984,2016.

[Kingma 13] D.P.Kingma, M.Welling: Auto-Encoding Variational Bayes. arXiv preprint arXiv:1312.6114,2013.

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

[Shin 16] H.C.Shin, K.Roberts, L.Lu, D.D.Fushman, J.Yao, R.M.Summers: Learning to Read Chest X-Rays. CVPR 2016, pp. 2497-2506. 2016.

[Ypsilantis 16] P.Ypsilantis, G.Montana: Learning what to look in chest X-rays with a recurrent visual attention model. NIPS 2016 Workshop on Machine Learning for Health. 2016.

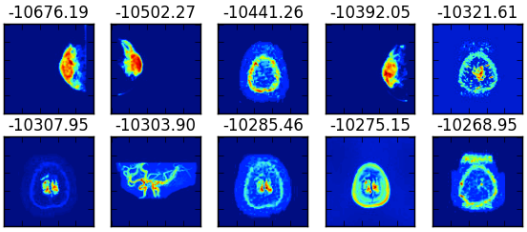

図 1: 非 CR 画像での VAE 変分下界 (ワースト 10)

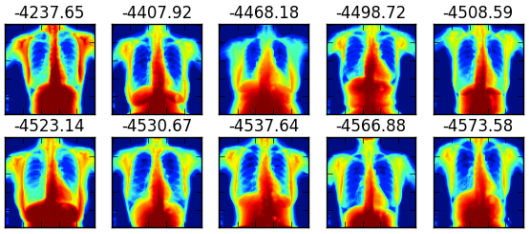

図 2: 非 CR 画像での VAE 変分下界 (ベスト 10)

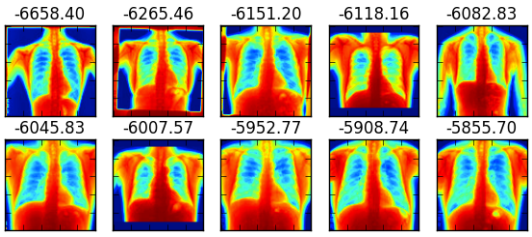

図 3: 診断正常画像での VAE 変分下界 (ワースト 10)

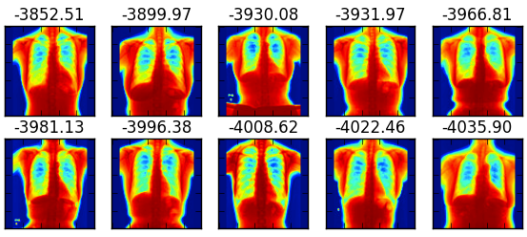

図 4: 診断正常画像での VAE 変分下界 (ベスト 10)

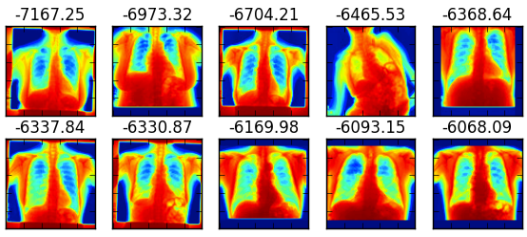

図 5: 診断異常画像での VAE 変分下界 (ワースト 10)

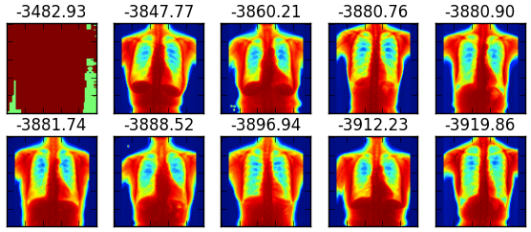

図 6: 診断異常画像での VAE 変分下界 (ベスト 10)

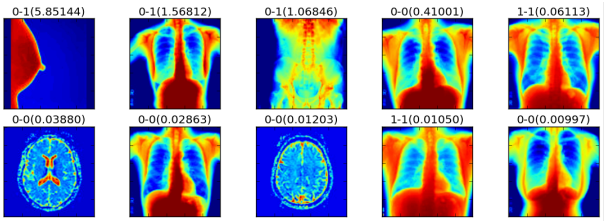

図 7: CNN 結果 (撮画手段識別セット，誤差ワースト 10) 数値は順に，真のラベル，予測ラベル，多クラスクロスエントロピー誤差関数の値．ラベルは 0 が異常，1 が正常を表す．図 8〜10 も同様.

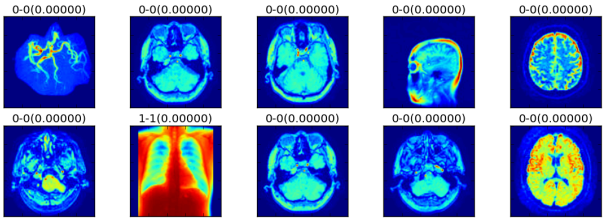

図 8: CNN 結果 (撮画手段識別セット，誤差ベスト 10)

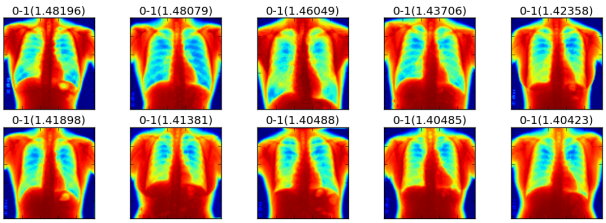

図 9: CNN 結果 (診断識別セット，誤差ワースト 10)

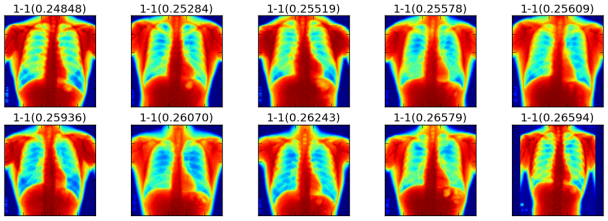

図 10: CNN 結果 (診断識別セット，誤差ベスト 10)