The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

文字単位のアテンション機構を用いた胸部 X 線写真の所見テキスト生成手法

Character-level Text Generations with Attention for Chest X-ray Diagnosis

作花 健也[*1] Sakka Kenya
中山 浩太郎[*2, 3] Kotaro Nakayama
木村 仁星[*2] Nisei Kimura
井上 大輝[*4] Taiki Inoue
山口 亮平[*5] Ryohei Yamaguchi
河添 悦昌[*5] Yoshimasa Kawazoea
大江和彦[*5] Kazuhiko Ohe
松尾豊[*2] Yutaka Matsuo

[*1] 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
[*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] 東京大学大学院薬学系研究科 Graduate School pf Pharmaceutical Sciences, The University of Tokyo
[*5] 東京大学大学院医学系研究科 Graduate School of Medicine, The University of Tokyo
[*6] 東京大学附属病院 The University of Tokyo Hospital

Medical images are widely used in clinical practice for diagnosis and treatment, and much time is spent on diagnosis. Therefore, research to automatically generate cases from medical images has been actively conducted in recent years. However, it is difficult to judge the case as a classification problem because there are orthographic variants in the case written in the medical certificate.

In this paper, we aimed to automatically generate character-level cases in order to cope with orthographic variants on chest X-ray images. In addition, the interpretability of the result was improved by introducing an attention mechanism. As a result, it was confirmed cases with features such as position information were generated, and the effectiveness of character-level approach was shown in text generation of medical images.

1. はじめに

医療画像は，診断および治療のために広く利用されている．しかし，医療画像を読影するためには高い専門性が必要であるため，放射線科医師などの専門家でも経験によって読影結果に差が出てしまう．具体的には，胸部 X 線写真を正しく読影するためには以下の 5 つのスキルが求められている [Delrue 11]. (1) 通常時の胸部の解剖学的構造，および胸部疾患の基本的な生理学についての知識を持っていること．(2) 経験的に得られた固定的なパターンを通して胸部 X 線写真を分析すること．(3) 経時的な変化を考慮して評価を行うこと．(4) 患者の臨床症状および病歴に関する知識を持っていること．(5) 他の診断結果(臨床検査の結果[血液，痰]，心電図，呼吸機能調査)との相関に関する知識を持っていること．このような高い専門性が必要なことに加えて，胸部 X 線写真は患者の状況や重要な情報を把握するための方法として最も普及している方法の一つであり，救急医療や健康診断など様々な場面で日々大量の撮影が行われている．この結果，放射線科医師を含む医療従事者へ大きな負担が発生しており，その解決が求められていた．そのため，クラス分類問題として所見を自動で検出する研究が盛んに行われている [Rajpurkar 17]. さらに，ここ1～2年では画像から直接，所見のテキスト情報を出力する end-to-end のモデルなども提案されてはじめている [Jing 18].

所見は，専門家により表記方法に揺らぎが生じる．例えば，異常領域が肺の両側にある際に，「両肺」や「両側肺」のように複数の表現方法が存在する．そのため，クラス分類問題や単語単位での所見生成には複雑な前処理が必要なことに加え，汎用性が低くなるといった問題がある．これらの問題に対応するために，本稿では表記方法の揺らぎがある場合にも有効な手法である，文字単位 (character-level) [Zhang 16]での所見生成を行なった．

本稿では，胸部 X 線写真を入力として受け取り，その画像に対する所見を文字単位で生成する Encoder-Decoder モデルを提案した． Encoder 側では入力画像に対して特徴マップを抽出する．Decoder 側ではアテンション機構 (Attention mechanism)によって重み付けされた特徴マップと 1 ステップ前の状態をもとに文字を出力する．

本稿の主な貢献は次の 2 つである．(1) 医療画像において，文字単位での所見生成を行なった．(2) アテンション機構を導入することで，精度向上だけでなく，生成された各文字が画像のどの部位を参照して生成されたかを視覚的に示し，結果の解釈性を高めた．

2. 関連研究

2.1 文字単位の言語モデル

自然言語処理やキャプション生成の分野では，一般的に単語単位での分散表現を学習した言語モデルが用いられている [Vinyals 15]. しかし，単語単位での言語モデルは前処理が複雑であることや表記方法の揺らぎに弱いといった問題点がある．そこで近年，ユーザー生成コンテンツなどのようにテキストが一定の規則に基づいて整えられていないような場合においても有効である，文字単位での言語モデルが提案されている [Zhang 16]. 文字単位の言語モデルの他の利点として，日本語で特に困難な単語分割の必要がないという点が挙げられる．しかし，長

連絡先：作花健也，東京大学，〒113-8656 東京都文京区本郷 7-3-1，sakka_kenya_17@stu-cbms.k.u-tokyo.ac.jp

期的な依存関係を学習する必要があるため活用が困難であった. 医療画像の所見は, 文字列長が他のキャプション生成の問題と比べて短い. そのため, 専門家により記述方法が異なる医療画像の所見テキスト生成のための手法として, 文字単位の言語モデルは有効な手法であると考えられる.

2.2 アテンション機構

アテンション機構を用いた画像のキャプション生成の方法として, ハード・アテンション (Hard-Attention)とソフト・アテンション (Soft-Attention)の 2 つの方法が考案されている [Xu 16]. ハード・アテンションは, アテンションの重みを特定の領域に絞り込んで適用する特徴を持つ. 一方, ソフト・アテンションは, アテンションの重みを広範囲に適用する特徴を持つ. 胸部 X 線写真の読影は, 広範囲の特徴を捉える必要があることから, 本稿ではソフト・アテンションを用いて所見生成を行なった.

2.3 医療画像における所見生成

医療画像の所見生成にアテンション機構を用いた研究も行われている [Zhang 17, Jing 18]. しかし, 英語の文章を対象としており, 単語単位での所見生成を行なってる. 特に Jing らの研究 [Jing 18]では, 精度向上のために所見に関するタグが用いられているため, 通常の読影では付与されない情報を専門家が付与する必要がある.

他の言語と異なり, 日本語の文章に含まれる文字は位置情報や前後関係などの情報を豊富に持つ. そのため, 日本語の所見生成においては, アテンション機構を用いた文字単位の手法が有用であると考えられる.

3. 提案手法

本稿では, 胸部 X 線写真を入力として受け取り, ソフト・アテンションを用いて所見を文字単位で出力する Encoder-Decoder モデルを提案した (図 1). モデルは, 正解の所見と生成された所見の交差エントロピー誤差とアテンション機構に関するソフトマックス関数の和で定義された損失関数を最小化するように, Encoder, Decoder を End-to-End で学習した. Early-Stopping を用いて, 2epoch 連続して検証データで精度の向上がないとき学習を打ち切った.

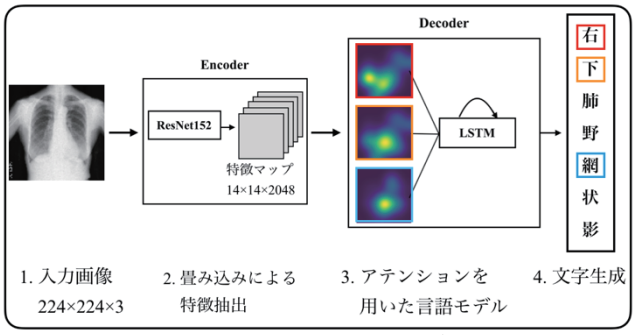

図 1 Encoder-Decoder モデル

3.1 Encoder

ResNet152 [He 15]の事前学習済みモデルを用いて, 入力画像に対する特徴マップを取得した. しかし, 事前学習は一般物体で行われているため, 胸部 X 線写真ではうまく特徴が抽出できないと考えられる. そのため, 胸部 X 線写真を用いて所見の頻度が高い上位 11 クラスの分類問題として, 再度事前学習を行なった.

胸部 X 線写真をモデルに入力する際には, 以下の前処理を行なった. まず, 画像のサイズが 2048×2048 と大きいため, 224×224 にリサイズした. その後, 3 チャネル画像の各チャネルの値を平均 (0.485, 0.456, 0.406), 分散 (0.229, 0.224, 0.225)に正規化した.

学習時には, モデルの汎用性を高めるためにデータ拡張 (Data Augmentation)を用いた. 入力画像を 256×256 にリサイズした後, Random Crop を行い 224×224 の領域を抽出し, 50%の確率で Horizontal flip を行なった. 最適化関数は, Adam を利用し, 学習率は 0.0001 に設定した.

3.2 Decoder

ソフト・アテンションによって重み付けされた 14×14×2048 の特徴マップと 1 ステップ前の状態を入力として受け取り, 1 層の LSTM [Hochreiter 97]を用いて所見の文字を 1 文字づつ出力するモデルを構築した. 一文字ずつアテンション機構を用いて出力することで, より詳細に前の状態を反映した所見生成を行い, 結果の解釈性を高めることもできる.

Decoder のハイパーパラメータはそれぞれ以下のように設定した. 隠れ状態を 256 次元, 最適化関数は Adam を利用し, 学習率は 0.0004 に設定した. 今回のデータセットでは, 文字の種類が 300 以下と比較的少ないため, 分散表現は用いていない.

4. 評価実験

4.1 データセット

東京大学附属病院より提供していただいた 18,004 枚の胸部 X 線写真を用いて評価実験を行なった. これらの各写真には, 専門家による所見情報が付与されている. データセット内の所見情報は, 全て同一の専門家によってラベル付けされているのではなく, 複数人の専門家によってラベル付けされているため, 同じ所見でも異なる記述方法で書かれていることもある.

上記のデータセットの所見情報から, ”前回と変化なし”などのように時系列情報が必要なデータに関しては, データセットから除いた. 加えて, ”手入力 部位”などのように所見情報として, 不必要なものも除き, データセットの整形を行なった.

4.2 実験設定

提案手法の有効性を BLEU スコア [Papineni 02]を用いて検証した. データ量を保証するために, 所見の出現回数に閾値を設定した. 本稿では, 閾値を 5 と 30 に設定した 2 つのデータセットを用いて検証を行なった (表 1). 閾値を 5 に設定した場合は全 18,004 件の所見のうち 91.5%, 閾値を 30 に設定した場合は 84.3%の所見を含んでいる.

上記で得られた 2 つのデータセットについて, ラベルの分布を保持するように各ラベルの出現回数の 80%を訓練データ,残り 20%の内それぞれ 10%を検証データとテストデータとして抽出した. ラベルによっては, データ数が 3 枚程度のものもあり学習データとして不十分である. 加えて, ラベルごとの頻度のばらつきが大きい. そのため, 訓練データに関して, 異常なし以外のラベルのデータ数を 100 枚にし, 異常なしとその他のラベル (異常あり)の比率が元のデータセットの分布に従うようにアップサンプリングを行なった.

表 1 データセットの詳細

	症例数	文字の種類	データ数 (訓練)	データ数 (検証)	データ数 (テスト)
症例の出現頻度の閾値：5	158	128	12,893	1,616	1,616
症例の出現頻度の閾値：30	25	61	11,887	1,472	1,472

4.3 実験結果

各文字毎のアテンション機構の結果と生成された文字列の結果を示す (図 2). デバイスが埋め込まれている患者の胸部 X 線写真では, 所見生成時にデバイスの周辺にアテンション機構が機能していることが確認できる (図 2 A). 「両側肺尖胸肥厚」のように, 所見に位置情報が含まれる例では, アテンション機構が位置情報を反映するように機能していることが確認できる (図 2 B). モデルによって生成された所見の中では, 「肥厚」と「癒着」のように素人目では判断が難しいような誤りが見られた (図 2 C). これらの結果は, 専門家に確認してもらい, どの程度診断に影響のある誤りであるかを判断する必要がある. 胸部 X 線写真のみの情報から, 「異常なし」と判断するためには, 画像を網羅的に読影する必要がある. モデルが「異常なし」と所見生成した際のアテンション機構の推移を確認すると, 他の所見の場合とは異なり広範囲にアテンション機構が機能していることが確認できる (図 2 D).

各データセットのテストデータに対する精度評価の結果を示す (表 2). どちらのデータセットでも, 高いスコアが出ていることがわかる. テストデータは, 実際の分布をもとに作成されているため, 「異常なし」のラベルが全体の 8 割含まれている. そのため, 「異常なし」の所見を正しく判断できるモデルであれば, BLEU スコアが高くなる. そこで, 「異常なし」以外のラベルについてのみ BLEU スコアを計算した結果も示した.

表 2 各データセットの精度評価の結果

	BLEU 1	BLEU 2	BLEU 3	BLEU 4
閾値：5 (元の分布)	0.7105	0.7053	0.7021	0.6966
閾値：30 (元の分布)	0.8366	0.8355	0.8347	0.8262
閾値：5 (異常ありのみ)	0.1362	0.1135	0.1002	0.0760
閾値：30 (異常ありのみ)	0.3621	0.3547	0.3493	0.2926

5. 考察

所見生成において, アテンション機構がデバイスの有無や位置情報を捉えることができていることが確認された (図 2 A, B, C). 位置情報が出力された時とアテンション機構がその位置に働いている時が一致していない例もある. これは, 読影に必要な異常部位がその位置にあるため, 時系列的に前後して関係のある文字を出力する時に参照されるためであると考えられる.

医療画像の所見には, 表記の揺らぎが存在するため, 既存の BLEU などの評価指標では正しく評価することが難しい. 例えば, 「両側肺尖胸膜肥厚」と「両肺尖部胸膜肥厚」は意味的に同じ所見を表している (図 2 B). しかし, 既存の評価指標では各文字の位置が考慮されていることや単語のオントロジーが考量されていないため, これら二つの所見の類似度は低くなる.

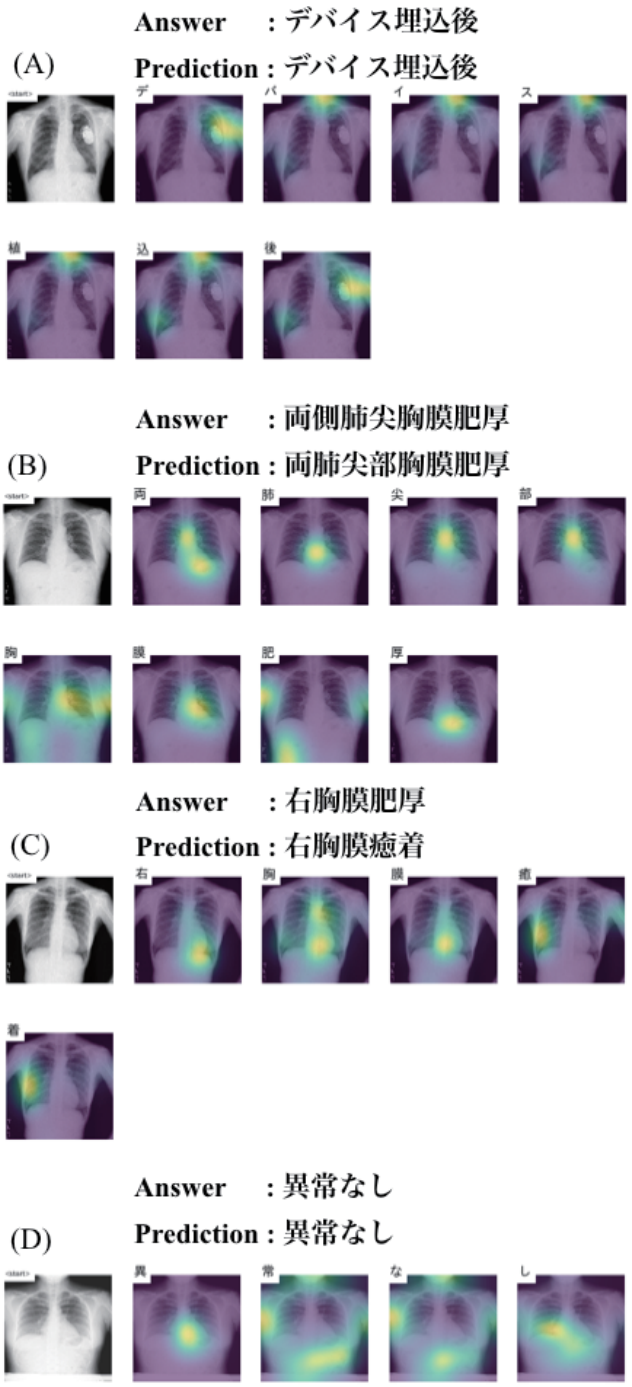

図 2 所見生成の結果とアテンション機構の推移

医療画像の所見の実際の分布を反映したテストデータでは, 高い BLEU スコアが得られた (表 2). 実際の分布では, 「異常なし」が全体の所見の 8 割を占める. そのため, 常に「異常なし」を出力することで精度が上がってしまう可能性がある. しかし, 本稿で提案した手法で作成されたモデルでは, 81 種類の所見が生成されており, 「異常なし」のデータに対して正しく「異常なし」と出された結果は 68%程度, 異常所見に対して誤って「異常なし」と出した結果は 11%程度であった. 「異常なし」の精度が低い原因として, 「異常なし」と判断する際には, 画像全体を見る必要があり問題設定として難しいことが考えられる. 結果からも, 他の例では異常部位に重点的にアテンション機構が働いているのに対して, 「異常なし」ではアテンション機構が広範囲に働

いており，画像全体を見て判断していることが確認できる (図 2 D). テストデータから「異常なし」を除いた場合，BLEU スコアが下がった. この原因として，実験結果で例を挙げた精度評価指標の問題も考えられる.「異常なし」以外の所見では，特に表記の揺らぎが大きく BLEU で正しく評価を行うことが困難である.

本稿で提案した手法は，所見と重点的に見るべき領域を同時に確認できることから，仮に生成された所見に誤りがある場合でも専門家がアテンション機構が働いている領域を重点的に確認することで正しい読影が可能である. また，読影のためのスクリーニングとして利用することで，専門家の労力を下げることに貢献できると考えられる.

6. まとめ

本稿では，胸部 X 線写真におけるアテンション機構を用いた文字単位での所見生成を行う手法を提案した. さらに，アテンション機構を用いることで結果の解釈性を高めた. 提案した手法について，BLEU スコアを用いて精度評価することで，手法の有用性を確認した. 今後の応用として，アテンション機構の結果を用いることで，経験の少ない専門家が胸部 X 線写真の読影方法を学ぶ際の手助けも可能であると考えられる. 文字単位でのキャプション生成は，一般的に困難であるため実用例が少ないが，今後医療画像のみならず多くの分野で用いられることが期待できる. 医療画像は大量のデータを集めることが困難であるため，どのようなデータ拡張が有効か検証する必要がある. 特に本稿でも使用した Horizontal Flip が，左右で似た構造を持つ胸部 X 線写真においても有効であるか検証を行いたい.

本稿で提案した手法は，日本語のみならず他の言語でも適用が可能であると考えられる. 特に日本語と同様に文字の持つ情報量が多く，よりデータ量が豊富な中国での適用が期待される.

今後は，医療画像の所見のように単語の位置情報が重要でなく，類似語が多いような文章に対しても適切に評価できる指標を専門家と作っていく必要がある.

7. 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

[Delrue 11] Louke Delrue, Robert Gossenlin, Bart Ilsen, An Van Landeghem, Johan de Mey, and Philippe Duyck: Difficulties in the Interpretation of Chest Radiography, In Comparative Interpretation of CT and Standard Radiography of the Chest, page 27-49 , Springer, 2011.

[He 15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun: Deep Residual Learning for Image Recognition, arXiv: 1512.03385v1, 2015.

[Hochreiter 97] Sepp Hochreiter, and Jurgen Schmidhuber: Long Short-Term Memory , Neural Computation 9 , page 1735-1780 , 1997.

[Jing 18] Baoyu Jing, Pengtao Xie, and Eric P. Xing: On the Automatic Generation of Medical Image Reports , arXiv: 1711.08195v3, 2018.

[Papineni 02] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu: BLEU: a Method for Automatic Evaluation of Machine Translation, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), page 311-318, 2002.

[Rajpurkar 17] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Robyn L. Ball, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, Andrew Y. Ng: ChestXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning, aeXiv:1711.05225v3, 2017.

[Vinyals 15] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan: Show and Tell: A Neural Image Caption Generator, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), page 3156-3164, 2015.

[Xu 16] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention , arXiv: 1502.03044v3, 2016.

[Zhang 16] Xiang Zhang, Junbo Zhao, and Yann LeCun: Character-level Convolutional Networks for Text Classification, arXiv:1509.01626v3, 2016.

[Zhang 17] Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang: MDNet: A Semantically and Visually Interpretable Medical Image Diagnosis Network, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6428-6436, 2017.