Semnan University

Journal of Modeling in Engineering

Journal homepage: https://modelling.semnan.ac.ir/

ISSN: 2783-2538

Medical Report Generation for Chest X-rays Using Convolutional Recurrent and Attention-Based Architectures

Fardin Ghaderi[1], Mohammad Bagher Khodabakhshi[1,*], Shahriar Jamasb[1]

[1] *Biomedical Engineering Department, Hamedan University of Technology, Hamedan, Iran*

Received: 2024-11-05 Revised: 2025-05-30 Accepted: 2025-06-21

Abstract

Medical images are central to diagnosis, but writing clear, complete radiology reports can be challenging for trainees and time-consuming for experts. We present an automatic report generation system that converts an image into a coherent narrative, aiming to reduce errors and speed documentation. The model combines modern image analysis with attention-guided language generation to identify key findings and express them in standard report style. Trained on paired images and reports, it was evaluated using common text quality measures, including BLEU 1-4, ROUGE-L, and CIDEr-D. Compared with prior work, our approach produced longer, more informative reports with higher alignment to reference texts, improving CIDEr-D by 7.2% and ROUGE-L by 3.2%, while maintaining competitive BLEU scores. These gains suggest the system can support clinicians by providing accurate drafts and template-like structure that highlight salient details. In practice, such tools could enhance reporting consistency, assist less experienced readers, and help manage growing imaging workloads.

Keywords :

Medical image processing, Recurrent deep neural networks, Automatic image captioning, Encoder, Decoder, Attention mechanism

1. Introduction

Manual medical report writing is time-consuming and prone to human error due to physician fatigue. To address this, automated report generation combines Computer Vision and NLP. Previous methods have explored hierarchical LSTMs [1], dual LSTMs to reduce data bias [2], multimodal fusion [3-5], weakly supervised frameworks [6], and iterative sentence generation [7]. While biomedical language models like BioBERT [8] and BioGPT [9] have advanced text processing, they either focus on discriminative tasks or require better visual integration. To overcome the challenge of generating comprehensive reports without relying on expert-annotated image classes, the paper introduces Res-LSTM-Attn. This recurrent convolutional architecture uses an attention mechanism in the decoder to generate long paragraphs word-by-word via image-word matching and is evaluated on the IU X-Ray dataset as the gold standard.

2. Methodology

2.1. Dataset Introduction

This study utilizes the IU X-Ray dataset, which is one of the most widely used datasets for medical image report generation. It comprises 7,470 chest X-ray images, including frontal and lateral views for each patient. Typically, radiology reports consist of several standard sections: “Indication” (reason for the X-ray), “Findings” (radiologist’s observations), and “Impression” (summary of relevant findings for diagnosis). Each report is usually associated with one or more chest X-ray images.The dataset contains 3,955 radiology reports associated with 7,470 images (dimensions 2048×2496). In this study, cases lacking complete two-view images or missing the “Findings” and “Impression” sections were excluded, resulting in a smaller subset of 2,867 reports linked to frontal images. This exclusion was necessary to ensure complete data for training the Res-LSTM-Attn model, which is designed to generate both sections simultaneously. However, removing incomplete reports might introduce data bias. Images were resized to 224×224 pixels. During preprocessing, meaningful words in the “Findings” and “Impression” sections were tokenized, while personal information, numbers, punctuation marks, Greek letters, and abbreviations were removed. This yielded 1,645 unique words. Special [start] and [end] tokens were added so the network can recognize the beginning and end of sentences. Words were then converted into numerical tokens, and all sentences were padded to the length of the longest sentence.Vocabulary sizes for the separate “Findings” and “Impression” sections are 1,400and 1,070 words, respectively.

The maximum report length is 163 words for combined reports, and 145 and 107 words for individual "Findings" and "Impression" sections. Finally, a random split of 10%(280 reports) was used as the test set for all evaluations.

2.2. Proposed Res-LSTM-Attn Model

The proposed model is based on an encoder-decoder architecture. A ResNet50 convolutional neural network is used as the encoder to extract features. An LSTM network serves as the recurrent decoder to generate words and understand their semantic relationships. The decoder is equipped with an attention mechanism to achieve a higher-level understanding of semantic connections. The model is named Res-LSTM-Attn, reflecting the combination of convolutional and memory-based networks with an attention mechanism. The overall architecture of the proposed model is illustrated in Figure 1. The framework consists of four main components:

1. An image encoder to extract visual features from chest X-rays.
2. A multi-label model to predict report words from the extracted visual features.
3. An LSTM language model that processes previously predicted words.
4. A language model with Multi-Head Self-Attention utilizing pre-trained FastText word embeddings for the latest predicted word. The outputs from these sections are fused in a fully connected (FC) layer to generate the report word by word.

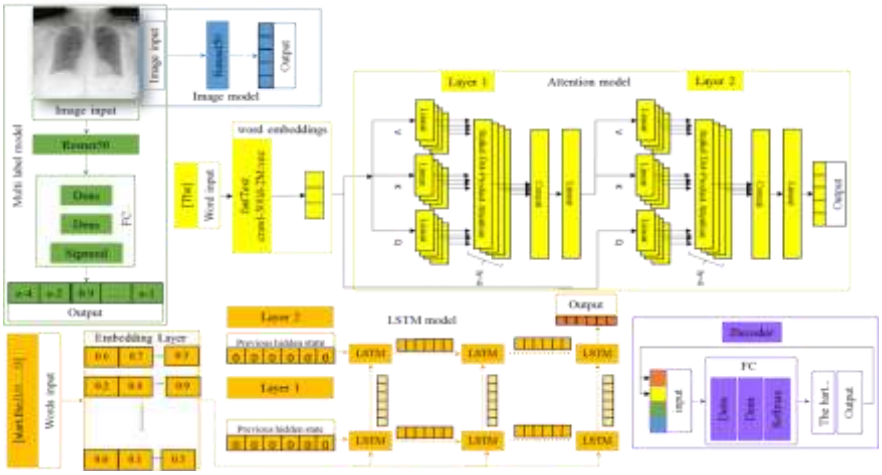

Figure 1. Overall Structure of the Res-LSTM-Attn Model

2.2.1. Image Encoder Model

A ResNet50 network is employed as the image encoder. It is one of the most efficient CNN feature extractors, utilizing skip connections to overcome the vanishing gradient problem common in deep networks. The visual features extracted from ResNet50 are fed into both the multi-label model and the report generation module.

2.2.2. Multi-Label Model

In multi-label classification, each data sample may be associated with multiple labels. Extracted ResNet50 features are passed through two fully connected layers. The final outputs represent the probability of the input image belonging to each word in the dictionary. This module is trained independently using the binary cross-entropy loss function:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}[\, y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \tag{1}$$

where $y_i$is the ground-truth label for each word, $p_i$is the predicted probability, and $N$is the dictionary size.

2.2.3. LSTM Model

A two-layer LSTM structure is utilized. The hidden states of the first layer serve as inputs to the second. The process of the first LSTM layer is defined as:

$$h_{1,t}, c_{1,t} = LSTM_1(x_t, h_{1,t-1}, c_{1,t-1}) \tag{2}$$

where $x_t$is the word embedding matrix. An attention module generates the context vector z_t:

$$z_t = f_{att}(h_{1,t}, v) \tag{3}$$

Then, $z_t$and $h_{1,t}$are fed into the second LSTM:

$$h_{2,t}, c_{2,t} = LSTM_2([z_t, h_{1,t}], h_{2,t-1}, c_{2,t-1}) \tag{4}$$

Finally, the output of the second LSTM is passed to a softmax layer for probability distribution prediction:

$$p_t = softmax(Wh_{2,t} + b) \tag{5}$$

2.2.4. Attention Model

A multi-head self-attention mechanism is employed to capture long-range dependencies and semantic information from the FastText word embeddings. The scaled dot-product attention is calculated as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{6}$$

where$Q, K, V$are the query, key, and value matrices, and $d_k$is the input dimension. The multi-head attention concatenates the results of Nparallel attention heads:

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_N)W^O \tag{7}$$

This computation allows the model to learn additional semantic information from different representation subspaces.

2.2.5. Fully Connected Model

The outputs from the preceding modules are concatenated and fed into a fully connected layer with 2,048neurons, followed by a softmax layer sized to match the vocabulary. During generation, the [start] token initiates the process alongside the image embeddings and multi-label outputs. Words are generated greedily:

$$w_t = \arg\max\ p(w_t \mid I, M, w_{0:t-1}; \theta) \tag{8}$$

where $I$is the image encoding, $M$is the multi-label output, and θrepresents the model parameters. The report generation model is trained using categorical batch cross-entropy loss:

$$L_{CE} = -\sum_{i=1}^{V} y_i \log(p_i) \tag{9}$$

where $y_i$is the one-hot ground truth for the correct word, $p_i$is the predicted probability distribution, and $V$is the vocabulary size.

Table 1. Comparison of the Proposed Model's Results with other Studies

Paper	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr-D
Findings + Impression						
Huang et al. [10]	0.476	0.340	0.238	0.169	0.347	0.297
Singh et al. [11]	0.374	0.224	0.153	0.110	0.308	0.360
Res-LSTM-Attn (Ours)	0.4558	0.3080	0.2191	0.1558	0.3900	0.4010
Findings						
KERP [12]	0.482	0.325	0.226	0.162	0.339	0.280
RTMIC [13]	0.350	0.224	0.143	0.096	-	0.323
Res-LSTM-Attn (Ours)	0.4275	0.2870	0.2038	0.1478	0.3534	0.4170
Impression						
CMAS [14]	0.401	0.290	0.220	0.166	0.521	1.457
Res-LSTM-Attn (Ours)	0.5624	0.4964	0.4211	0.3271	0.6044	2.3419

3. Results and Discussion

This section details the implementation and evaluation of the Res-LSTM-Attn model, trained using TensorFlow 2.16.2 on a 16 GB GPU. The optimal decoder uses a 2-layer LSTM with 256 neurons and a 4-head attention mechanism, offering an effective balance between computational efficiency and semantic comprehension without the heavy overhead of Transformer models. As shown in Table 1, the proposed model demonstrates superior performance compared to previous baselines. For combined "Findings + Impression" reports, it achieves the highest CIDEr (0.4010) and ROUGE-L (0.3900) scores. For "Findings" alone, it secures the top CIDEr score (0.4170), and it significantly outperforms the CMAS model across all metrics when generating the "Impression" section.

The authors emphasize ROUGE and CIDEr over BLEU, as they better evaluate both clinical accuracy and essential content coverage. Despite the strong results, the model faces limitations such as occasional grammatical issues due to relying on the cross-entropy loss function instead of reinforcement optimization, lower BLEU-4 scores stemming from the limited vocabulary of the dataset, and reduced sentence diversity caused by prioritizing fast greedy decoding over beam search.

4. Conclusions

This study introduces the advanced Res-LSTM-Attn model for the automated, word-by-word generation of detailed chest X-ray reports. The proposed method extracts semantic features without requiring image labels, thereby aiding physicians in clinical decision-making. However, current limitations include a lack of disease diversity in the dataset and high hardware requirements for training and testing. For future research, the authors suggest evaluating the model on more diverse datasets, replacing convolutional networks with Vision Transformers, and upgrading the recurrent LSTM to a Transformer decoder. Furthermore, integrating specialized biomedical language models like BioBERT for text feature extraction and BioGPT for text generation is recommended to significantly enhance the accuracy, fluency, and coherence of the generated reports.

5. References

[1] Jing, Baoyu, Pengtao Xie, and Eric Xing. “On the Automatic Generation of Medical Imaging Reports.” *ArXiv Preprint ArXiv:1711.08195*, 2017.

[2] Harzig, Philipp, Yan-Ying Chen, Francine Chen, and Rainer Lienhart. “Addressing Data Bias Problems for Chest X-Ray Image Report Generation.” ArXiv Preprint ArXiv:1908.02123, 2019.

[3] Yang, Shaokang, Jianwei Niu, Jiyan Wu, Yong Wang, Xuefeng Liu, and Qingfeng Li. “Automatic Ultrasound Image Report Generation with Adaptive Multimodal Attention Mechanism.” Neurocomputing 427 (2021): 40–49.

[4] Li, Yuan, Xiaodan Liang, Zhiting Hu, and Eric P Xing. “Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation.” Advances in Neural Information Processing Systems 31 (2018).

[5] Shi, Jijun, Shanshe Wang, Ronggang Wang, and Siwei Ma. “AIMNet: Adaptive Image-Tag Merging Network For Automatic Medical Report Generation.” In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 7737–41. IEEE, 2022.

[6] Han, Zhongyi, Benzheng Wei, Stephanie Leung, Jonathan Chung, and Shuo Li. “Towards Automatic Report Generation in Spine Radiology Using Weakly Supervised Framework.” In Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11, 185–93. Springer, 2018.

[7] Xue, Yuan, Tao Xu, L Rodney Long, Zhiyun Xue, Sameer Antani, George R Thoma, and Xiaolei Huang. “Multimodal Recurrent Model with Attention for Automated Radiology Report Generation.” In Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I, 457–66. Springer, 2018.

[8] Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. “BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining.” Bioinformatics 36, no. 4 (2020): 1234–40.

[9] Luo, Renqian, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. “BioGPT: Generative Pre-Trained Transformer for Biomedical Text Generation and Mining.” Briefings in Bioinformatics 23, no. 6 (2022): bbac409.

[10] Huang, Xin, Fengqi Yan, Wei Xu, and Maozhen Li. “Multi-Attention and Incorporating Background Information Model for Chest x-Ray Image Report Generation.” IEEE Access 7 (2019): 154808–17.

[11] Singh, Sonit, Sarvnaz Karimi, Kevin Ho-Shon, and Len Hamey. "From chest x-rays to radiology reports: a multimodal machine learning approach." In 2019 Digital Image Computing: Techniques and Applications (DICTA), pp. 1-8. IEEE, 2019.

[12] Li, Christy Y., Xiaodan Liang, Zhiting Hu, and Eric P. Xing. "Knowledge-driven encode, retrieve, paraphrase for medical image report generation." In Proceedings of the AAAI conference on artificial intelligence, vol. 33, no. 01, pp. 6666-6673. 2019.

[13] Xiong, Yuxuan, Bo Du, and Pingkun Yan. "Reinforced transformer for medical image captioning." In International workshop on machine learning in medical imaging, pp. 673-680. Cham: Springer International Publishing, 2019.

[14] Jing, Baoyu, Zeya Wang, and Eric Xing. “Show, Describe and Conclude: On Exploiting the Structure Information of Chest x-Ray Reports.” ArXiv Preprint ArXiv:2004.12274, 2020.