Generating Chest X-Ray Reports using an Encoder-Decoder Model with Attention

Adriana Rupérez Arellano
Data Science and Engineering Department
Case Western Reserve University
axr1429@case.edu

Austin Angel
Computer Science Department
Case Western Reserve University
axa2074@case.edu

Abstract

This project explores the generation of medical reports from chest X-rays images using an Encoder-Decoder model with an attention mechanism. The built model uses a visual pretrained encoder and a GRU based recursive decoder. The goal is to generate coherent and relevant findings. It is evaluated using different metrics such as BLEU, ROUGE or METEOR scores. This model is able to produce structured reports, however it lacks of specific clinical details. Limitations and possible future work is also discussed.

1 Introduction

The generation of medical reports could potentially reduce workload for radiologists and speed up clinical processes. Therefore the goal is to generate these medical reports conditioned to an image, a chest x-ray. The architecture of the proposed model follows an encoder-decoder with attention mechanism. The model will be developed, discussed and analyzed.

2 Methodology

2.1 Dataset

The dataset used for this project has been extracted from OpenI and it contains 7,500 chest x-rays images and 4,000 medical reports. From the reports, relevant sections have been extracted (Findings and Impressions). Additionally, the reports have been paired with the images IDs. After processing the dataset, a vast amount of reports were paired with 2 or more images, some content was anonymized ("XXXX" or "xxxx") and hence eliminated, and inconsistent format made the pairing of image-report and extraction of relevant information difficult.

2.2 Model Architecture

The model architecture uses an Encoder-Decoder framework with attention, using a DenseNet121 pre-trained on RadImageNet as the encoder. This encoder processes the 448x448 input X-ray images to extract both global context features used for decoder initialization and a grid of spatial features that are fed into the attention mechanism. A Bahdanau attention module calculates a context vector at each decoding step by weighting these spatial features based on the decoder's previous hidden state. The decoder itself is a single layer GRU which takes the combination of the attention context vector and the embedding of the previously generated token as input. A final linear layer followed by a softmax function predicts the probability distribution over the vocabulary for the next token, with greedy search used during report generation and dropout applied during training for regularization.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

2.3 Training

The model was trained on an 80/20 split of the 3,331 filtered image report pairs, using the AdamW optimizer and minimizing CrossEntropyLoss. An attention entropy penalty (λ=0.1) was added to the loss for stability. Training utilized batches, teacher forcing, and gradient clipping. Early stopping based on validation loss on training around epoch 39, with the best performing weights saved. During this process, the DenseNet121 encoder was fine tuned, where the GRU decoder and attention components were trained from scratch.

3 Results

3.1 Quantitative Results

The model trained using the setup described above, was evaluated on the validation set of 666 image report pairs. Reports were generated using greedy decoding for evaluation, although sampling methods were used on the qualitative side. The scores across the validation set are presented below:

Metric	Score
BLEU-4	2.94
ROUGE-1	28.67
ROUGE-2	7.43
ROUGE-L	26.67
METEOR	20.24

Table 1: Evaluation scores of the model on the validation set.

Interpreting these scores, the relatively high ROUGE-1 and ROUGE-L values show that the model is capable of generating reports that capture some of the overall sentence structure present in the reference reports. This indicates a basic understanding of common terms and report organization. However, the very low BLEU score and the low ROUGE-2 score show a significant limitation. The model struggles to generate precise phrasing and reproduce longer sequences of words accurately. This shows that generated text that might be grammatically accurate and contain relevant keywords but lacks the exact clinical nuance and detailed descriptions found in the ground truth. The METEOR score of 20.24 suggests a moderate level of semantic overlap, showing that the model captures some meaning but requires a lot of improvement, especially medical accuracy where synonyms or stemmed words might not work.

3.2 Qualitative Results

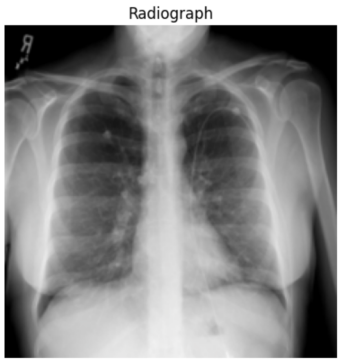
Inspection of generated reports by their ground truth references provides insights about the quantitative metrics. Several patterns shown, as illustrated in Figure 1 and in Figure 2.

Strengths: The model consistently generates sentences and correctly identifies major anatomical structures like the heart, lungs, and mediastinum. Phrases that are accurate, such as “Heart size is normal,” “The lungs are clear,” or “No pleural effusion or pneumothorax,” are produced, showing the high prevalence of normal findings in many radiology datasets. The reports follow a structure, resembling human written findings or impressions.

Weaknesses: The most significant limitation is the frequent misses of specific clinical behaviors mentioned in the reference reports. The model fails to mention findings like "scattered calcifications, subsegmental atelectasis, displaced rib fractures, or specific types of opacities", even when visible in the image. Instead, it defaults to more generic statements. Repetitive phrasing within a single report or across different reports was also found. While using BioClinicalBERT helps with medical terms, the model sometimes generates awkward phrasing compared to radiological terminology.

4 Analysis and Discussion

This demonstrates that an attention based Encoder-Decoder model can learn the basic task of generating text descriptions from chest X-ray images. The model achieves good structure as indicated




```
--- Ground Truth Report ---

The heart is normal in size and contour. Scattered calcifications are noted, compatible with prior granulomatous disease. The lungs are clea
r, without evidence of infiltrate. There is no pneumothorax or effusion. No acute cardiopulmonary disease.

--- Generated Report ---

heart size, mediastinal contours are within normal limits. pulmonary vascularity is within normal limits. no focal consolidation, pneumothora
x, or pleural effusion. no acute pulmonary abnormality. no acute cardio
```

Figure 1: Example 1: The model captures general findings for normal anatomic structures such as heart and lungs but misses calcifications.


```
--- Ground Truth Report ---

Heart size is normal. Right lung is clear. Granulomatous disease in the bilateral. Subsegmental atelectasis in the left lower lung. No pneumo
thorax. No pleural effusion. Subsegmental atelectasis in the left lower lobe.

--- Generated Report ---

stable cardiomegaly. no pneumothorax. the heart size is normal. relative hypoin left pulmonary or pleural effusion or airspace disease. 2. no
adjacent infiltrate. please anterior
```

Figure 2: Example 2: Comparison between ground truth report and generated report. The model captures general findings and captures something on left lung, however it misses the specific abnormality.

by ROUGE-L and METEOR scores, and the attention mechanism successfully helps connect the generated terms to image regions. The use of a pretrained encoder contributes by providing relevant features adapted to the medical field.

Despite these successes, the model's performance falls short of clinical applicability primarily due to its lack of specificity and accuracy in identifying and describing abnormalities. The low BLEU and ROUGE-2 scores reflect this hardship in matching precise clinical wording.

First, the NLMCXR dataset has challenges including noise from multiimage reports being paired with single images, anonymization artifacts, and variability in reporting. The dataset size after filtering (3,331 pairs) is also small for training complex sequence generation models, limiting the model's exposure to different findings. Second, the model architecture itself has limitations. The single layer GRU decoder may struggle with capturing the dependencies and relationships often present in detailed medical narratives. More powerful architectures like Transformers have shown better performance in sequence modeling tasks and might be better suited. Third, the 7×7 spatial grid for

the attention mechanism might offer insufficient resolution for detecting small characteristics of chest X-rays. Finally, the standard NLP evaluation metrics used do not directly measure clinical accuracy. A report could achieve a high ROUGE score by mentioning common terms but still miss a critical finding.

5 Conclusions and Future Work

This project successfully implemented and evaluated an Encoder-Decoder model with Bahdanau attention for generating chest X-ray reports. The results are encouraging, showing the model's ability to produce fluent, structurally relevant text grounded in visual features through the attention mechanism. However, the generated reports currently lack the necessary clinical detail and accuracy for deployment, often producing generic descriptions and missing specific findings.

Significant improvements are needed to bridge the gap towards clinical utility. Key directions for future work include:

- **Architectural Enhancements:** Transitioning to Transformer based encoder-decoder architectures could improve the modeling of long range dependencies and complex language structures. Exploring attention mechanisms or using higher resolution feature maps for attention might enable better findings.
- **Data Augmentation and Refinement:** Utilizing larger, and cleaner datasets like MIMIC-CXR could expose the model to a wider range of reporting styles. Advanced data cleaning and preprocessing techniques are also needed.
- **Improved Generation Strategies:** Replacing greedy decoding with beam search could potentially help with more coherent and optimal report combinations. Using methods to reduce factual hallucination or improve grounding is also needed.
- **Clinically-Oriented Evaluation:** Developing and integrating evaluation metrics focused on clinical accuracy is needed. This could be by using systems like the CheXpert labeler to compare the presence or absence of specific findings in generated vs. reference reports, or calculating precision, recall, and F1 scores for key clinical terms.

In conclusion, while automated radiology report generation using attention-based models shows promise as an assistive technology, the current state requires further research and development, particularly focusing on enhancing clinical accuracy and incorporating more sophisticated evaluation methods, before it can be reliably integrated into clinical practice.

References

[1] Loshchilov, I. & Hutter, F. (2019). Decoupled Weight Decay Regularization. *International Conference on Learning Representations (ICLR)*.

[2] Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311–318).

[3] Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop* (pp. 74–81).

[4] Banerjee, S. & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization* (pp. 65–72).

[5] Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.-y., Mark, R.G., & Horng, S. (2019). MIMIC-CXR: A large publicly available database of labeled chest radiographs. *Scientific Data*, 6(1), 317.

[6] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D., Halabi, S., Sandberg, J., Jones, R., Larson, D., Langlotz, C., Patel, B., Lungren, M.P., & Ng, A.Y. (2019). CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), 590–597.