A Multi Visual Feature Extraction model for X-Ray interpretation

Patel Ketul(ketvpate@iu.edu), and Patil Harshwardhan(hrpatil@iu.edu)
Indiana University, Bloomington

***Abstract*—This paper presents a deep learning approach for extracting visual features to generate report summaries from X-ray images. This is a comparative study of the models trained on the Indiana University - Chest X-Rays dataset and evaluated against BLEU, METOR, and ROUGE scores. In this project, we have conducted in-depth literature reviews and proposed a Multi Visual Feature Extraction model, geared towards optimizing the interpretation of Chest X-ray images.**

I. Introduction

Radiology is a branch of medicine that uses imaging technology to diagnose and treat diseases. Doctors specializing in this field, known as radiologists, are equipped with years of training and expertise to interpret X-rays, CT scans, MRIs, nuclear medicine scans, PET scans, ultrasounds, etc. As the number of medical images continues to grow, so too does the workload of radiologists responsible for interpreting them. The sheer volume can overwhelm even the most skilled radiologists, potentially impacting their ability to analyze and interpret them effectively. One possible option is to automate this process.

Machine learning stands out as the most effective way to automate the analysis and diagnosis of medical images. Its diverse applications include:

- **Image segmentation:** Precisely segmenting anatomical structures like the brain, spine, lungs, liver, kidneys, and colon for accurate measurements and analysis.[6]
- **Computer-aided diagnosis and detection:** Detecting and highlighting suspicious lesions in CT or MRI scans, like mammograms and CT colonography, to assist radiologists in making faster and more accurate diagnoses.[10]
- **Brain function analysis:** Analyzing fMRI data to understand brain activity and diagnose neurological diseases, providing valuable insights into brain function and behavior.
- **Content-based image retrieval:** Searching large medical image databases efficiently based on specific features.
- **Text analysis of radiology reports**: Utilizing NLP and NLU to extract key information, identify trends, and even assist in generating reports, streamlining workflow and improving patient care efficiency.

These applications demonstrate machine learning's transformative potential in radiology.

Out of all theses applications, we caught our eye on the topic of medical image interpretation and reporting. Many factors can contribute to faulty report generation - lack of knowledge, staff shortage, excessive workload, etc. This process can be error-prone, even for experienced radiology specialists. To reduce the error occurrences, an automated system is needed. [2]

In this paper, we will present a novel approach for automated report generation. Our primary focus is on the combination of the distinct visual feature extraction models, namely DenseNet121, InceptionV3, and Xception. The contributions of our work lie in evaluating and comparing the performance of these individual models and combinations of them, to discern their effectiveness in the automated generation of radiology reports. For comparing performance we will use BLEU [16], METEOR [17], and ROUGE[18] metrics.

II. Related Work

Existing literature on automated medical report generation based on radiological images centers around the recurrent neural network with attention mechanisms[5]. Notably, research concentrating on the Indiana University X-Ray dataset has sequentially addressed the classification of frontal and lateral views [9], the grading of cardiomegaly severity [10], and the computer-aided diagnosis (CAD) of lung diseases [11]. These studies commonly employed CNN and Variational Topic Inference for classification and exploratory analysis. Additionally, recent work in image camptioniong explores using GRU models [12], knowledge graphs [13], and Multimodal Large Language Models for automated report generation [14].

III. Methodology

A. Data

In this study, we employ the Indiana University Chest X-Ray collection [15] as our primary dataset for evaluation. This data set encompasses a total of 7,471 X-ray images and there are about 3955 patients text reports available in .XML format. Notably, some reports are associated with more than one image. Each report is uniquely linked to a pair of images representing the frontal and lateral views.

The content of these reports is structured in XML format, necessitating XML parsing to extract and convert the information into a more accessible CSV format. Each report consists of the following sections: impression, findings, tags, comparison, and indication.

For our model, we consider both image and text inputs. The image data consists of the X-ray images associated with each report. Additionally, we extract textual information from the abstract, comparison, indication, and findings sections of the reports to serve as text input features.

The target variable for our analysis is the "Impression" section of the medical reports, which is treated as a text feature. The combination of image and text inputs allows the model to learn complex patterns and relationships essential

Comparison: None.

Indication: Positive TB test

Findings: The cardiac silhouette and mediastinum size are within normal limits. There is no pulmonary edema. There is no focal consolidation. There are no XXXX of a pleural effusion. There is no evidence of pneumothorax.

Impression: Normal chest x-XXXX.

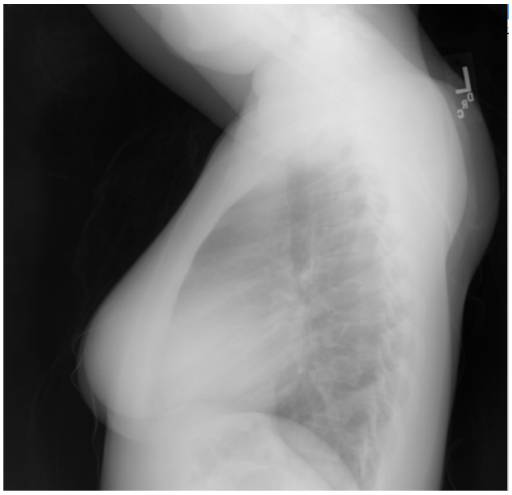

Fig. 1. A sample report for the chest x-ray where the comparision, indication, findings and impressions are mentioned. [15]

for automatic report generation in the field of medical image interpretation.

B. Data Pre-processing and Augmentation

The IU Chest X-ray dataset is organized into two primary directories: 'NLMCXR_png' and 'NLMCXR_reports'. The 'NLMCXR_png' directory houses 7,471 images, while 'NLMCXR_reports' contains 3,955 patient reports stored in XML format. The initial step involves scraping these XML files to extract and convert data. Key fields extracted include "image_id", "image_caption", "comparison", "indication", "findings", and "impression". Text processing involved removing extraneous symbols and replacing null values with appropriate sentences.

Regarding the chest X-ray images, we noted that 3,405 reports referenced more than two images. Our objective was to limit this to precisely two images per report, focusing solely on frontal and lateral views for simplicity and clarity.

For classifying these frontal and lateral views, we employed the K-means clustering algorithm in conjunction with the VGG16 model, which is pre-trained on ImageNet. This approach effectively segregated data, aligning frontal and lateral image IDs with corresponding findings from the reports. Consequently, the refined dataset comprises 3,302 entries, each including both a frontal and a lateral image.

C. Model Architecture

In the model architecture illustrated in Fig. 2, the framework is segmented into three distinct sections. Initially, the input consists of X-ray images which are processed through a Convolutional Neural Network (CNN) architecture for the extraction of visual features. This part of the model uses popular methods like DenseNet121, InceptionV3, and Xception for feature generation. This feature tensor serves as the precursor for the subsequent Recurrent Neural Network (RNN) processing stage.

Concurrently, textual data extracted from medical reports undergo a transformation into word embeddings. For data preparation, we tokenized all the words in the findings and impression in the dataset and obtained 2,218 unique words. Considering that the c size is already very small, we decided not to drop infrequent words with only once or twice appearances. We also added two special tokens, ¡start¿ and ¡end¿, into the vocabulary to indicate the start and the end of a sentence. To evaluate our models, we randomly picked 250 reports to form the testing set. All evaluations are done on the testing set.This process is crucial as it converts raw text into a format that is amenable to deep learning algorithms. The synergy between these word embeddings and the RNN module is pivotal for the accurate generation of outputs.

A substantial focus of the existing research has been on the refinement of text feature extraction. We propose a novel approach that combines three diverse visual feature generators (DenseNet121, InceptionV3, and Xception) shown by Fig. 3 By using these different methods together, our model can learn more details and pick up on things that might be missed by using just one method. This should make the model more effective at understanding both the images and the reports.

D. Evaluations

BLEU [16], METEOR [17], and ROUGE[18] are commonly used evaluation metrics for medical image report generation, which are adapted from machine translation and text summarization.

BLEU (Bilingual Evaluation Understudy): This metric is widely used in evaluating medical image report generation. It works by comparing the generated report with a standard report, focusing on how many words or phrases (n-grams) they have in common. BLEU comes in various forms, such as BLEU-1 for single words and BLEU-2, -3, and -4 for larger word groups. Despite its simplicity and correlation with human judgment, BLEU has limitations. It mainly assesses superficial text similarities and doesn't fully capture the semantic depth or coherency of the content.

METEOR (Metric for Evaluation of Translation with Explicit ORdering): This metric builds upon BLEU by incor-

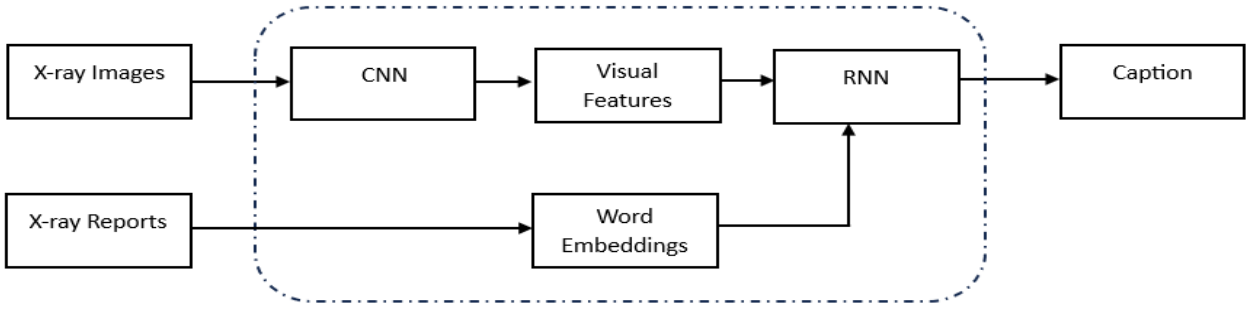

Fig. 2. A basic architecture of the ENCODER-DECODER model used for image captioning.

porating a balance of precision and recall, with more emphasis on recall. It uses the Porter stemmer and WordNet to enhance its evaluation, especially in recognizing longer matching word sequences. METEOR stands out for considering both surface-level and deeper semantic similarities, including handling synonyms and paraphrasing. However, like BLEU, METEOR might not fully assess the overall coherence of the text.

ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation—Longest Common Subsequence): ROUGE-L is distinct in focusing on the longest common subsequence shared between the machine-generated and reference texts. It assesses this in terms of recall, precision, or a combination (F-measure). This metric is more oriented towards evaluating semantic content and coherence, aligning well with human judgment. However, its focus on the longest subsequence means it might miss other quality aspects of the text.

E. Results

In our comparative analysis, distinct performance trends emerged among the evaluated methods as shown in TABLE. I. The Vanilla CNN-RNN method serves as a baseline model in the context of our study. Surprisingly, our proposed method, the Multi-Encoder-Decoder, recorded uniform scores of 0.1 across all metrics. This uniformity and significantly lower performance as compared to other methods raise concerns about the efficacy of this approach or possible limitations in our implementation. Further investigation and refinement are necessary to understand and address these shortcomings.

IV. Conclusion

The baseline model, Vanilla CNN-RNN, provided a foundational benchmark for our comparisons. Intriguingly, our proposed Multi-Encoder-Decoder method demonstrated a unique trend, yielding uniform scores of 0.1 across all evaluation metrics. This unexpected result significantly deviates from the performance of other evaluated methods, prompting a re-evaluation of the approach's effectiveness or potential limitations in our implementation.

Moving forward, our research suggests several potential avenues for further exploration. Refinement and optimization of the Multi-Encoder-Decoder architecture could be a starting point. Additionally, a deeper analysis into the model's components and training methodology may provide insights into the observed performance patterns. Future studies could also consider alternative architectures or hybrid models that might overcome the limitations identified in this research.

V. Future Work

The findings of our current study open several promising avenues for future research, particularly in the application

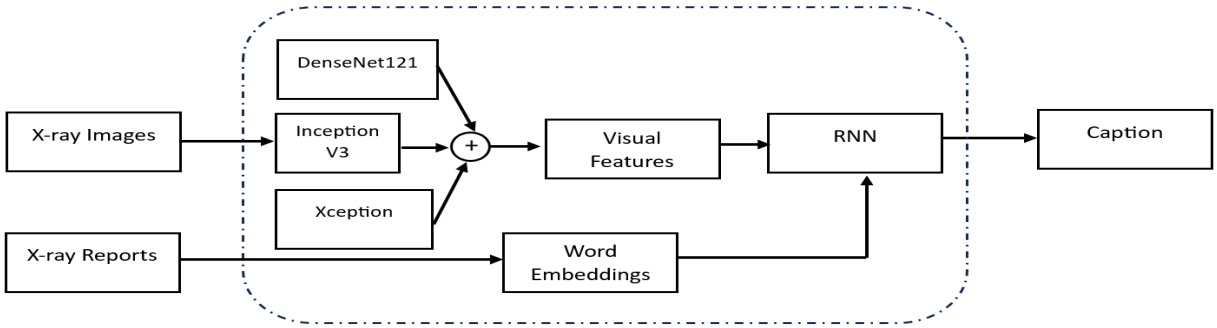

Fig. 3. Our proposed approach for combining encodings for visual features.

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE
Vanilla CNN-RNN [20]	0.273	0.144	0.116	0.082	0.125	0.226
Hierarchical generation [19]	0.437	0.323	0.221	0.172	0.244	0.325
recurrent-attention [2]	0.464	0.358	0.270	0.195	0.274	0.366
our-multi-encoder-decoder	0.1	0.1	0.1	0.1	0.1	0.1

TABLE I

COMPARISON OF TRADITIONAL METHODS AND OUR PROPOSED MODEL ON BLEU, METOR AND ROUGE METRICS

and advancement of deep learning models for X-ray image analysis and report generation:

- **Expanding Dataset Trials:** A crucial next step involves testing our models across a diverse range of X-ray datasets. This will not only help in assessing the models' adaptability and robustness but also provide insights into their performance across different clinical scenarios and imaging conditions.
- **Implementing Attention Transformers:** With the emerging significance of attention mechanisms in deep learning, especially for image and language processing tasks, integrating attention transformers into our model architecture could enhance its capability to focus on critical features in X-ray images. This enhancement is expected to yield more accurate and clinically relevant interpretations.
- **Utilizing Large Language Models:** The advancement in large language models presents an exciting opportunity to improve the natural language processing component of our system. Adapting these models for medical report generation could lead to significant improvements in the coherence, accuracy, and clinical relevance of the generated reports.
- **Exploring Curriculum Learning:** To refine our model's learning process, implementing curriculum learning could prove beneficial. By progressively increasing the complexity of the training data, we anticipate improvements in the model's learning dynamics, especially in handling complex or rare cases in radiographic analyses.
- **Cross-Modality Analysis Integration:** Investigating the integration of insights from various imaging modalities or related clinical data can offer a more holistic diagnostic approach. This cross-modality analysis has the potential to enhance the diagnostic accuracy and utility of our model.

Through these future initiatives, we aim to address the current limitations and explore new dimensions in the application of AI for medical imaging and diagnostics

REFERENCES

[1] Wang, Shijun, and Ronald M. Summers. Machine Learning and Radiology. Medical Image Analysis 16, no. 5 (2012): 933-951. https://doi.org/10.1016/j.media.2012.02.005.

[2] Y. Xue, T. Xu, L.R. Long, Z. Xue, S. Antani, G.R. Thoma, X. Huang Multimodal recurrent model with attention for automated radiology report generation Proceedings of the international conference on medical image computing and computer-assisted intervention (MICCAI), Springer, Granada, Spain (2018), pp. 457-466

[3] Mazurowski, M.A., Buda, M., Saha, A. and Bashir, M.R. (2019), Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI. J. Magn. Reson. Imaging, 49: 939-954. https://doi.org/10.1002/jmri.26534

[4] Sahiner, B., Pezeshk, A., Hadjiiski, L.M., Wang, X., Drukker, K., Cha, K.H., Summers, R.M. and Giger, M.L. (2019), Deep learning in medical imaging and radiation therapy. Med. Phys., 46: e1-e36. https://doi.org/10.1002/mp.13264

[5] Cho, Kyunghyun, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation." 2014. arXiv preprint, 1406.1078, cs.CL.

[6] Hesamian, Mohammad Hesam, Wenjing Jia, Xiangjian He, and Paul Kennedy. "Deep learning techniques for medical image segmentation: achievements and challenges." Journal of digital imaging 32 (2019): 582-596.

[7] Doi, Kunio. "Computer-aided diagnosis in medical imaging: historical review, current status and future potential." Computerized medical imaging and graphics 31, no. 4-5 (2007): 198-211.

[8] Chen, Chung-Ming, Yi-Hong Chou, Norio Tagawa, and Younghae Do. "Computer-aided detection and diagnosis in medical imaging." Computational and mathematical methods in medicine 2013 (2013).

[9] Z. Xue et al., "Chest X-ray Image View Classification," 2015 IEEE 28th International Symposium on Computer-Based Medical Systems, Sao Carlos, Brazil, 2015, pp. 66-71, doi: 10.1109/CBMS.2015.49.

[10] S. Candemir, S. Rajaraman, G. Thoma and S. Antani, "Deep Learning for Grading Cardiomegaly Severity in Chest X-Rays: An Investigation," 2018 IEEE Life Sciences Conference (LSC), Montreal, QC, Canada, 2018, pp. 109-113, doi: 10.1109/LSC.2018.8572113.

[11] Akbar, Wajahat, Muhammad Inam Ul Haq, Abdullah Soomro, Sher Muhammad Daudpota, Ali Shariq Imran, and Mohib Ullah. "Automated Report Generation: A GRU Based Method for Chest X-Rays." In 2023 4th International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), pp. 1-6. IEEE, 2023.

[12] Lee, Seowoo, Jiwon Youn, Mansu Kim, and Soon Ho Yoon. "CXR-LLaVA: Multimodal Large Language Model for Interpreting Chest X-ray Images." arXiv preprint arXiv:2310.18341 (2023).

[13] Liu, Weihua, Youyuan Xue, Chaochao Lin, and Said Boumaraf. "Dynamic Multi-Domain Knowledge Networks for Chest X-ray Report Generation." arXiv preprint arXiv:2310.05119 (2023).

[14] Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. J. Am. Med. Inform. Assoc. 23(2), 304–310 (2015)

[15] Indiana University. "Indiana University - Chest X-Rays (XML Reports)." 2023. Radiology Reports Collection. Accessed [Date]. Available at: https://i.imgur.com/PWo3x47.png.

[16] Papineni, K., Roukos, S., Ward, T., and Zhu, W. "Bleu: A Method for Automatic Evaluation of Machine Translation." In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA. ACL (2002), pp. 311-318.

[17] Banerjee, S., and Lavie, A. "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments." In Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation andor Summarization at ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005. Association for Computational Linguistics (2005), pp. 65-72.

[18] Lin, C.-Y. "ROUGE: A Package for Automatic Evaluation of Summaries." In Text Summarization Branches Out, Barcelona, Spain. Association for Computational Linguistics (July 2004), pp. 74-81.

[19] Krause, J., Johnson, J., Krishna, R., Fei-Fei, L.: A hierarchical approach for generating descriptive image paragraphs. In: CVPR, pp. 3337–3345 (2017)

[20] Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: CVPR, pp. 3156–3164 (2015)

[21] Bustos, Aurelia, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. "PadChest: A large chest x-ray image dataset with multi-label annotated reports." Medical Image Analysis 66 (2020): 101797. https://doi.org/10.1016/j.media.2020.101797.