Noname manuscript No.
(will be inserted by the editor)

Integrating Convolutional and Recurrent Neural Networks for Enhanced Medical Image Captioning

Andreas Kanavos · Gerasimos Vonitsanos · Phivos Mylonas

the date of receipt and acceptance should be inserted later

Abstract The rapid expansion of digital medical imaging technologies demands advanced tools for efficient and accurate image analysis. This research introduces a novel approach to medical image captioning, integrating Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to enhance the automatic generation of descriptive text for medical images. Our proposed model exploits the robust feature extraction capabilities of CNNs alongside the advanced sequential data processing of RNNs. We incorporate an attention mechanism that selectively focuses on diagnostically significant areas within images, thereby improving the relevance and accuracy of the generated captions. The effectiveness of our model was validated using an extensive set of evaluation metrics, including BLEU scores for linguistic quality and traditional classification metrics for accuracy. Results indicate that our model significantly outperforms existing systems in syntactic coherence and semantic accuracy, making it a valuable tool for aiding clinical decision-making and enhancing medical documentation.

Keywords Medical Image Captioning · Convolutional Neural Networks · Recurrent Neural Networks · Deep Learning · Automatic Image Annotation

Andreas Kanavos
Department of Informatics
Ionian University, Corfu, Greece
E-mail: akanavos@ionio.gr

Gerasimos Vonitsanos
Computer Engineering and Informatics Department
University of Patras, Patras, Greece
E-mail: mvonitsanos@ceid.upatras.gr

Phivos Mylonas
Department of Informatics and Computer Engineering
University of West Attica, Athens, Greece
E-mail: mylonasf@uniwa.gr

1 Introduction

Medical imaging is fundamental to modern healthcare, providing essential insights into the human body's structure and function through non-invasive techniques such as X-rays, Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and histopathological slides [15, 16]. With the evolution of imaging technology, there has been a surge in the volume of digital medical images, necessitating sophisticated computational tools to support accurate diagnostic decisions. The integration of Artificial Intelligence (AI) and Deep Learning (DL) into medical imaging represents a transformative shift in diagnostic practices, enhancing the precision and efficiency of patient care [4, 10].

The complexity of medical images poses significant challenges in clinical settings, requiring expert interpretation to inform effective diagnosis and treatment plans. For instance, radiologists examine CT scans for subtle disease indicators such as early-stage cancers. At the same time, pathologists must detect cellular anomalies in histopathological slides crucial for diagnosing conditions like cancer and inflammatory diseases [13, 17].

While technological advancements have enhanced image resolution and detail, aiding in more accurate diagnoses, they have exponentially increased the data volume, potentially overwhelming healthcare professionals. For example, databases like the Digital Database for Screening Mammography (DDSM) and the International Skin Imaging Collaboration (ISIC) contain extensive collections of images that require meticulous analysis to identify critical features indicative of diseases such as breast cancer and melanoma [14, 28].

The manual interpretation of these images is labour-intensive and susceptible to human error. The growing demand for medical imaging services, compounded by a shortage of trained specialists, underscores the urgent need for automated tools to provide timely, accurate interpretations, mitigating the risks of diagnostic delays and errors. Recent AI breakthroughs, particularly in deep learning, have demonstrated significant potential in automating tasks such as classification, detection, and segmentation in medical image analysis [7]. Extending these advancements to the automatic generation of descriptive texts for medical images—a medical image subtitling process—can significantly enhance healthcare delivery. This approach merges image analysis with Natural Language Processing (NLP) to produce interpretable, content-rich descriptions crucial for clinical practice [20, 29].

Nonetheless, the variability and complexity of medical images pose substantial challenges to model generalizability and accuracy. Models effective on specific datasets, such as chest X-rays, often falter under different conditions or patient groups, leading to misdiagnoses and inappropriate interventions [6, 10].

This study introduces a novel, robust model that combines Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to generate medical image subtitles. Our model utilizes CNNs for extracting detailed visual features and integrates bidirectional LSTM layers within the RNN to formulate coherent, contextually relevant captions [8, 25]. An attention mechanism is also employed to focus dynamically on diagnostically significant image regions, improving the relevance and accuracy of the output [18, 31]. The efficacy of our model is thoroughly evaluated through quantitative metrics like BLEU scores and qualitative assessments, showcasing its superior performance in generating precise, consistent, and clinically pertinent medical image subtitles over existing models.

The remainder of this paper is organized as follows: Section 2 reviews the related work and the development of techniques in medical image processing and the integration of AI technologies. Section 3 details the methodology employed, describing the architectural design and functionalities of the proposed models. In Section 4, we elaborate on the implementation specifics, including the dataset overview, training procedures, and evaluation metrics. Section 5 presents a thorough experimental evaluation of our models, showcasing their performance against established benchmarks. Finally, Section 6 concludes the paper with a summary of findings and an outline of future research directions.

2 Related Work

Medical image analysis is crucial in healthcare, offering vital insights into the human body through various imaging modalities, each uniquely enhancing clinical practices and medical research. X-ray imaging, one of the earliest techniques, is essential for examining skeletal structures and detecting lesions. Magnetic Resonance Imaging (MRI) provides detailed views of soft tissues and the central nervous system, which is critical for diagnosing tumours and neurological disorders. Computed Tomography (CT) offers images of complex body structures, indispensable for diagnosing cerebral and vascular diseases. Meanwhile, histopathological slides provide detailed cellular and tissue information, crucial for diagnosing diverse pathological conditions [5].

Integrating AI into medical imaging marks a transformative era. Advanced machine learning models, intense CNNs and RNNs enhance the automatic analysis and interpretation of medical images, boosting diagnostic accuracy and efficiency. For instance, CNNs excel in pattern recognition within X-rays, facilitating rapid fracture identification, while RNNs effectively track disease progression through sequential MRI scans. Research continues to advance these models' accuracy and processing speeds, employing larger, more diverse datasets and techniques like transfer learning. This approach utilizes pre-trained models adapted with minimal training for specific applications, pivotal in personalizing diagnostic tools [4, 27].

However, each imaging modality comes with challenges. X-ray imaging, while crucial for diagnosing conditions from fractures to lung diseases, struggles with overlapping structure structures and demands high-resolution imaging to differentiate between similar tissues [1, 17]. MRI, known for its superior tissue contrast, requires high contrast resolution for accurate pathology identification, with techniques like functional MRI (fMRI) providing essential brain activity insights. CT imaging combines multiple X-ray views to create detailed cross-sectional images vital for diagnosing cancers and cardiovascular conditions. Innovations such as low-dose CT (LDCT) improve lung cancer screening by detecting early-stage diseases with reduced radiation risks [13].

The complexity and volume of digital medical images pose significant challenges. Traditional diagnostic processes rely heavily on the expertise of radiologists and clinicians trained to discern subtle abnormalities. Recent AI advancements aim to address issues such as image artefacts from metal implants, which can obscure critical details and lead to diagnostic errors. Moreover, medical image analysis extends beyond pathology classification, including object detection and segmentation. Techniques like YOLO (You Only Look Once) and Faster R-CNN have been adapted to medical imaging, identifying and delineating pathological conditions to assist radiologists in highlighting crucial areas for more accurate and efficient diagnoses [22, 32].

Advances in image segmentation through AI, particularly with models like U-Net, automate the division of complex images into meaningful components, which is essential for precise diagnostics and tailored treatments. The emerging field of medical image subtitling combines computer vision and natural language processing to enhance interpretability [23]. These models typically integrate CNNs for robust feature extraction with RNNs enhanced by attention mechanisms, producing descriptive textual summaries of medical images. This synthesis of visual and textual analysis not only aids in clinical decision-making but also streamlines the clinical workflow, illustrating the profound impact of AI on medical diagnostics [30].

3 Methodology

The methodology adopted in this research is geared towards developing an effective multi-modal deep learning system, Model-80 and Model-70, which integrate sophisticated image processing and natural language processing techniques. This section outlines the approach taken, beginning with the architectural design of our models, each tailored to specific training/validation split ratios to optimize performance and robustness in medical image captioning.

3.1 Model Architecture Overview

Our models integrate image features and sequence data to produce medically pertinent predictions. The architecture includes:

- **Encoder for Image Features:** Includes an input layer for 4096-dimensional vectors, a dropout layer to combat overfitting, and a dense layer with ReLU activation to compress features.
- **Sequence Feature Encoder:** Processes data through an input layer and an embedding layer that converts tokens into dense vectors, a dropout layer, and a bidirectional LSTM layer to capture contextual information.
- **Attention Mechanism:** Employed to enhance the model's focus by computing a context vector, averaged using GlobalAveragePooling1D.
- **Decoder:** Combines image and sequence features through an Add layer, followed by dense layers with ReLU activation and dropout for normalization. The final layer uses softmax activation for classification tasks.
- **Optimization with Adam:** The model employs the Adam optimizer for effective gradient descent and optimal convergence across training epochs.

3.2 CNN Feature Extraction

The CNN architecture employs multiple convolutional layers to capture various optical features, from low-level features like edges and textures to high-level features such as anatomical structures and pathological patterns. Maximum pooling layers reduce the spatial dimensions of these feature maps, while a dropout layer enhances model generalizability by randomly discarding units during training. Finally, a fully connected (dense) layer transforms the extracted features into a fixed-size vector representation, summarizing the critical information from the image for input into the RNN to generate captions [7, 13, 24].

3.3 RNN for Caption Generation

The RNN architecture begins with an embedding layer that converts each word in the vocabulary into a dense vector representation, capturing the semantic meaning necessary for coherent text generation. Bidirectional LSTM layers retain important information across extended sequences and enhance predictive capabilities by processing forward and backwards input sequences. An attention mechanism allows the model to focus on relevant areas of the image during caption generation, ensuring accurate and clinically significant outputs. Additional dense layers with ReLU activation functions introduce nonlinearity, enhancing feature representation and capturing complex patterns within the data. The final output layer uses a softmax activation function to predict the probability distribution over the vocabulary for the next word in the caption sequence, guiding the selection of the most appropriate word at each step [7, 13].

4 Implementation

This section translates the theoretical concepts outlined in the Methodology into actionable steps, detailing the practical implementations and the evaluation strategies employed. These steps ensure that each component supports the robust operation of our multi-modal deep learning system, Model-80 and Model-70, and validates the efficacy of our model in generating clinically relevant captions for medical images. By leveraging cutting-edge computational resources and advanced machine-learning techniques, we aim to demonstrate that our findings are theoretically sound, practically viable, and effective in real-world scenarios.

4.1 MediCat Dataset Overview

The dataset, sourced from the MediCat repository [26], includes 50,000 images, specifically selected for their maximum description size of 20 words to ensure homogeneity and the production of informative captions. Before training, images were preprocessed to normalize and augment data, addressing variations in image quality and format, which are common challenges in medical datasets.

4.2 Training Process

The training process is designed to optimize the model's performance and ensure the production of accurate and clinically relevant captions. We focus on selecting appropriate loss functions, optimization strategies, and regularization techniques, which are crucial for handling the high variability and imbalance in medical datasets.

Table 1 Overview of Image Sources and Annotations in the MediCat Dataset

Description	**Details**
Images from open access articles	217,060 images from 131,410 open access articles
Figure and subfigure captions annotations	7,507 annotations of captions and subfigures for 2,069 images
Embedded references	Embedded references for around 25,000 images in the ROCO dataset

- **Loss Function:** Categorical cross-entropy measures the discrepancy between predicted and actual word sequences, optimizing model predictions by minimizing the negative log-likelihood of correct words [3].
- **Optimization:** The Adam optimizer facilitates efficient gradient descent and faster convergence, with dynamic adjustments of the learning rate for each parameter [12].
- **Batch Size and Epochs:** A batch size of 32 and 100 epochs balance computational efficiency with robust learning tailored to the complexity and size of the medical image datasets [2].
- **Learning Rate and Regularization:** Adaptive learning rate and techniques like dropout and L2 regularization are implemented to enhance model generalizability and prevent overfitting [2].

4.3 Evaluation Metrics

A robust set of evaluation metrics, including accuracy, precision, recall, and F1 score, are used to assess model performance comprehensively. These metrics are complemented by the BLEU score, which evaluates the linguistic quality of generated captions by measuring the accuracy of n-grams between the produced captions and reference captions [1, 7].

4.4 Experimental Setup

The models are trained and evaluated in a controlled environment using powerful GPUs, significantly reducing the computational time required for training. We use TensorFlow for its robust neural network capabilities and flexibility. For Model-80, the dataset is split into training (80%), validation (10%), and test (10%) sets. In contrast, for Model-70, the dataset is divided into training (70%), validation (15%), and test (15%) to assess the impact of different training/validation splits on model performance. These setups ensure that each model is rigorously tested on unseen data, providing a reliable measure of effectiveness. Systematic tuning of hyperparameters, including learning rate, beam size, and number of layers, is conducted using grid search and other optimization techniques to ensure the best performance of our models under varied conditions.

5 Experimental Evaluation

The performance of our models, Model-80 and Model-70, was rigorously evaluated using a combination of traditional classification metrics and the BLEU score to assess their natural language generation capabilities. This detailed evaluation strategy ensures a balanced assessment of the models' accuracy and linguistic quality in medical image captioning within a medical context.

5.1 Model Performance Metrics

We utilized a set of metrics to evaluate different aspects of model performance, illustrating the trade-offs between model configurations. The results, presented in Table 2, underscore the importance of parameter tuning in achieving optimal model performance, particularly in applications like medical image captioning, where accuracy and reliability are crucial.

Table 2 Performance Metrics of Models with Different Parameter Sets

Model	Batch Size	Train/Split Set	Accuracy	Precision	Recall	F1
Model-80	64	80/20	0.4941	0.7095	0.3031	0.4244
Model-70	128	70/30	0.3936	0.6336	0.1849	0.2862

These results indicate that Model-80, with a more favourable training-to-validation split and smaller batch size, achieves a better balance across all metrics, particularly in precision and F1 score, compared to Model-70. This demonstrates the challenges of model training on limited data with its lower overall performance.

5.2 BLEU Metric Evaluation

The BLEU (Bilingual Evaluation Understudy) score was employed to quantitatively measure the linguistic quality of text generated by our models compared to human-crafted reference texts. This metric is particularly relevant in assessing AI models' natural language generation capabilities [19].

Table 3 indicates that both models perform competitively, with slight variations in BLEU-1 and BLEU-2 scores reflecting differences in lexical selection and syntactic structuring. Model Model-70 demonstrates marginally higher scores, possibly due to a broader dataset or different parameter tuning, which might have influenced its ability to match the reference texts closely. The robust syntactic structuring of Model-80, despite a slightly narrower lexical diversity, is notable.

Table 3 BLEU Score Evaluation for Medical Image Captioning Models

Model	BLEU-1	BLEU-2
Model-80	0.309	0.205
Model-70	0.310	0.207

5.3 Comparative Analysis

This subsection benchmarks the performance of our models against contemporary methodologies in medical image captioning, providing a crucial assessment of their relative effectiveness.

Table 4 confirms that our models achieve the highest scores among the compared models in both metrics, underscoring the efficacy of their architecture in creating descriptive and clinically relevant captions. Including attention mechanisms allows our models to focus on pertinent parts of images, ensuring that the generated text accurately reflects critical aspects of the medical images.

Table 4 Comparative Analysis of Algorithms with BLEU-1 & BLEU-2 Metrics

Model Name	BLEU-1	BLEU-2	Description
ImageCLEF 2017 Model [17]	0.142	0.070	Neural Captioning for the ImageCLEF 2017 Medical Image Challenges
Retinal Image Captioning [9]	0.158	0.076	Contextualized Keyword Representations for Multi-modal Retinal Image Captioning
Global-Local Visual Extractor [13]	0.148	0.074	Cross Encoder-Decoder Transformer with Global-Local Visual Extractor for Medical Image Captioning
Proposed Model	0.309	0.205	Hybrid CNN-RNN model with integrated attention mechanisms

6 Conclusions and Future Work

This research introduced an advanced approach to medical image captioning that integrates CNNs and RNNs, enhancing the generated text's accuracy and richness. The deployment of Model-80 and Model-70 demonstrated that a careful balance of training and validation and sophisticated model architecture can significantly improve performance in medical image analysis. Our models have proven highly effective in producing clinically relevant captions through rigorous evaluation using traditional classification metrics and BLEU scores.

Looking forward, we aim to enrich the dataset diversity by incorporating a more detailed range of medical images and annotations. This expansion will help refine the models' ability to generalize across different medical scenarios, which is crucial for real-world applications [21]. Collaborations with medical professionals will play a vital role in this phase, ensuring that the generated captions meet the practical needs of clinical practice.

Additionally, exploring real-time captioning systems and adopting more advanced neural network architectures, such as Transformers, could significantly push the boundaries of what is currently achievable in medical image analysis [11].

References

1. Ayesha H, Iqbal S, Tariq M, Abrar M, Sanaullah M, Abbas I, Rehman A, Niazi MFK, Hussain S (2021) Automatic medical image interpretation: State of the art and future directions. Pattern Recognition 114:107,856
2. Bengio Y, Goodfellow I, Courville A (2017) Deep Learning, vol 1. MIT Press
3. Bishop CM (2006) Pattern Recognition and Machine Learning, vol 4. Springer
4. Chen J, Guo H, Yi K, Li B, Elhoseiny M (2022) Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 18,009–18,019
5. Chen X, Zitnick CL (2015) Mind's eye: A recurrent visual representation for image caption generation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 2422–2431
6. Chohan M, Khan A, Mahar MS, Hassan S, Ghafoor A, Khan M (2020) Image captioning using deep learning: A systematic literature review. International Journal of Advanced Computer Science and Applications (IJACSA) 11(5):62
7. Codella NCF, Rotemberg V, Tschandl P, Celebi ME, Dusza SW, Gutman DA, Helba B, Kalloo A, Liopyris K, Marchetti MA, Kittler H, Halpern A (2019) Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). CoRR abs/1902.03368
8. Hou B, Kaissis G, Summers RM, Kainz B (2021) RATCHET: medical transformer for chest x-ray diagnosis and reporting. In: 24th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Springer, Lecture Notes in Computer Science, vol 12907, pp 293–303
9. Huang J, Wu T, Worring M (2021) Contextualized keyword representations for multi-modal retinal image captioning. In: International Conference on Multimedia Retrieval (ICMR), ACM, pp 645–652
10. Huang J, Wu T, Yang CH, Worring M (2021) Longer version for "deep context-encoding network for retinal image captioning". CoRR abs/2105.14538
11. Kanavos A, Theodoridis E, Tsakalidis AK (2014) A pubmed meta search engine based on biomedical entity mining. In: 25th International Workshop on Database and Expert Systems Applications (DEXA), pp 82–86
12. Kingma DP (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980
13. Lee H, Cho H, Park J, Chae J, Kim J (2022) Cross encoder-decoder transformer with global-local visual extractor for medical image captioning. Sensors 22(4):1429
14. Lee RS, Gimenez F, Hoogi A, Miyake KK, Gorovoy M, Rubin DL (2017) A curated mammography data set for use in computer-aided detection and diagnosis research. Scientific Data 4(1):1–9
15. Livieris IE, Kanavos A, Tampakas V, Pintelas PE (2018) An ensemble SSL algorithm for efficient chest x-ray image classification. Journal of Imaging 4(7):95
16. Livieris IE, Kanavos A, Tampakas V, Pintelas PE (2019) A weighted voting ensemble self-labeled algorithm for the detection of lung abnormalities from x-rays. Algorithms 12(3):64
17. Lyndon D, Kumar A, Kim J (2017) Neural captioning for the imageclef 2017 medical image challenges. In: Working Notes of CLEF - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, vol 1866

18. Ma Z, Yang Y, Wang G, Xu X, Shen HT, Zhang M (2022) Rethinking open-world object detection in autonomous driving scenarios. In: 30th ACM International Conference on Multimedia (MM), pp 1279–1288
19. Papineni K, Roukos S, Ward T, Zhu W (2002) Bleu: A method for automatic evaluation of machine translation. In: 40th Annual Meeting of the Association for Computational Linguistics, ACL, pp 311–318
20. Park H, Kim K, Yoon J, Park S, Choi J (2020) Feature difference makes sense: A medical image captioning model exploiting feature difference and tag information. In: 58th Annual Meeting of the Association for Computational Linguistics, ACL, pp 95–102
21. Piri J, Mohapatra P, Acharya B, Gharehchopogh FS, Gerogiannis VC, Kanavos A, Manika S (2022) Feature selection using artificial gorilla troop optimization for biomedical data: A case analysis with covid-19 data. Mathematics 10(15):2742
22. Ren S, He K, Girshick RB, Sun J (2017) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 39(6):1137–1149
23. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, Lecture Notes in Computer Science, vol 9351, pp 234–241
24. Savvopoulos A, Kanavos A, Mylonas P, Sioutas S (2018) LSTM accelerator for convolutional object identification. Algorithms 11(10):157
25. Selivanov A, Rogov OY, Chesakov D, Shelmanov A, Fedulova I, Dylov DV (2023) Medical image captioning via generative pretrained transformers. Scientific Reports 13(1):4171
26. Subramanian S, Wang LL, Bogin B, Mehta S, van Zuylen M, Parasa S, Singh S, Gardner M, Hajishirzi H (2020) Medicat: A dataset of medical images, captions, and textual references. In: Findings of the Association for Computational Linguistics (EMNLP), Findings of ACL, pp 2112–2120
27. Vinyals O, Toshev A, Bengio S, Erhan D (2015) Show and tell: A neural image caption generator. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 3156–3164
28. Wang R, Lei T, Cui R, Zhang B, Meng H, Nandi AK (2022) Medical image segmentation using deep learning: A survey. IET Image Processing 16(5):1243–1267
29. Wu T, Huang J, Lin J, Worring M (2023) Expert-defined keywords improve interpretability of retinal image captioning. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, pp 1859–1868
30. Xu K, Ba J, Kiros R, Cho K, Courville AC, Salakhutdinov R, Zemel RS, Bengio Y (2015) Show, attend and tell: Neural image caption generation with visual attention. CoRR abs/1502.03044
31. Zhang J, Nie Y, Chang J, Zhang J (2021) Surgical instruction generation with transformers. In: 24th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Springer, Lecture Notes in Computer Science, vol 12904, pp 290–299
32. Zhang X, Dong X, Wei Q, Zhou K (2019) Real-time object detection algorithm based on improved yolov3. Journal of Electronic Imaging 28(5):053,022