Methodologies to generate Medical report using X-Ray image

Affan Iqbal [1] **, Noufal Ehaab** [12] **, Midhat Masood** [3]

[1]FAST-NUCES – University
St-4 - Sector 17-D - Karachi - Sindh - Pakistan

k213057@nu.edu.pk, k213001@nu.edu.pk, k213078@nu.edu.pk

Abstract. *This work explores multiple deep learning approaches for image captioning and disease classification, leveraging architectures such as VGG16, EfficientNetB0, DenseNet121, and ResNet50 for feature extraction. Each model integrates advanced feature extraction with custom decoders or classification layers to address domain-specific challenges. Key innovations include the use of LSTM-based decoders for text generation, dense layers for dimensionality reduction, and data augmentation to enhance generalization. The models demonstrate the effectiveness of combining pre-trained networks with task-specific modifications, outperforming generic approaches in capturing meaningful patterns. This study highlights the importance of tailored architectures for medical applications and sets a foundation for future improvements in precision, recall, and feature interpretability.*

1 Introduction

Image captioning and disease classification are critical tasks in medical imaging and computer vision, requiring accurate feature extraction and robust language generation. Automatic solutions for these tasks have the potential to assist clinicians in diagnostic workflows and reduce manual effort. However, these problems present unique challenges, including feature sparsity, class imbalance, and the need for contextually relevant textual outputs.

To address these challenges, deep learning models that combine powerful feature extractors with customized decoding architectures have shown promising results. Pre-trained models like VGG16, EfficientNetB0, DenseNet121, and ResNet50 have been widely used for feature extraction due to their proven ability to capture detailed patterns in images. These models, when integrated with LSTM-based decoders or dense classification layers, can effectively bridge the gap between visual data and textual or categorical outputs.

This study evaluates various model architectures for their ability to generate captions for medical images and classify diseases from radiology datasets. By incorporating advanced techniques such as data augmentation, dropout for regularization, and task-specific modifications, we explore their performance and adaptability to medical domains. The findings from this work highlight the strengths and limitations of different approaches, providing valuable insights for optimizing future models tailored to medical applications.

2 Related Work

2.1 Image Captioning

Traditional image captioning methods typically train an encoder-decoder model using curated image-caption pairs to generate textual descriptions from images. Early approaches utilized a CNN-based encoder to extract visual features and an RNN/LSTM-based decoder for sentence generation. To improve visual understanding, some methods incorporated object detectors to identify and extract key image regions. To promote better interaction between visual and textual modalities, attention mechanisms and graph neural networks became widely adopted. Recently, several large-scale visual-language pre-training models have demonstrated remarkable performance in image captioning tasks. While natural scene image description aims to generate concise sentences, medical report generation requires more detailed descriptions of medical images. As a result, these traditional approaches may not be well-suited for medical report generation.

2.2 Medical Report Generation

Medical report generation, a more complex extension of image captioning, poses greater challenges due to the increased length and accuracy required in text descriptions. Significant progress has been made in this field through extensive research. For instance, some methods have utilized posterior and prior knowledge in radiology to address data bias issues in report generation. Visual-language pre-training models that incorporate medical domain knowledge have been developed to enhance report generation performance. Other approaches have focused on improving textual representation by extracting specific knowledge from retrieved reports and modifying graph structures, integrating image features with these updated graphs. Certain models have also introduced learnable ”expert” tokens in both the encoder and decoder to interact with vision tokens, enabling the model to focus on different regions while minimizing overlap and capturing distinct information. Despite these advancements, these methods have limitations such as the complexity of prior knowledge extraction, loss of important features, and difficulty in emphasizing key regions, hindering their overall effectiveness.

3 Methods

In this section, we present a comprehensive analysis of the generation process for CNN-LSTM based architecture. Additionally, we will discuss CNN-LLMs based architecture and its applicability to generate medical reports.

3.1 CNN-LSTM-Based Architecture

In this study, we explored various CNN-LSTM-based architectures for generating captions from medical images, focusing on the combination of powerful feature extractors and sequence models. Among these, the VGG16-LSTM architecture demonstrated the most stable performance across all evaluation metrics. The following subsections detail the components and evaluation of this architecture, alongside other CNN-LSTM models.

3.1.1 1. Feature Extraction Using CNN

The VGG16-LSTM model, based on the pre-trained VGG16 architecture, effectively extracted hierarchical visual features from input medical images. Key operations included:

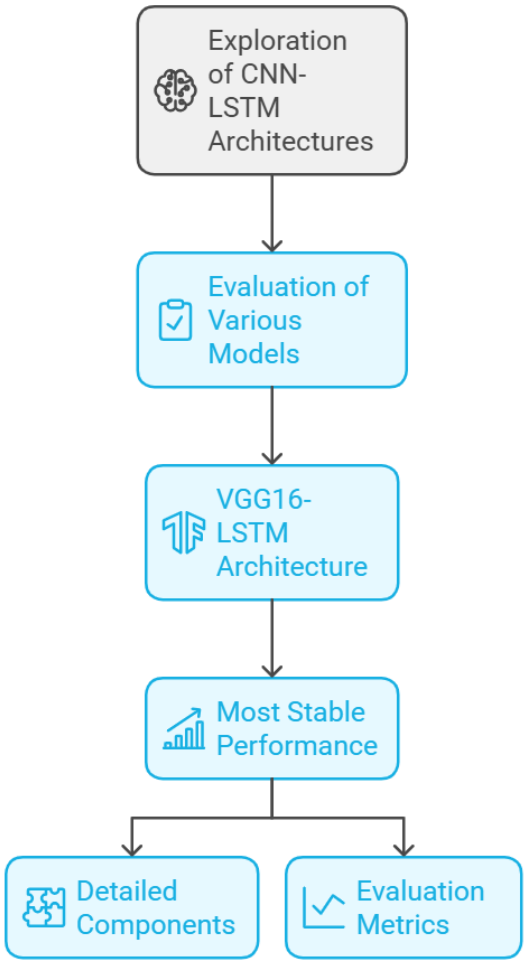

Figure 1: CNN-LSTM Architecture

- **VGG16 Backbone**: The model extracted image features using convolutional and max-pooling layers. A dense output layer reduced the feature space to 256 dimensions.
- **Dropout Regularization**: A dropout rate of 0.4 was applied to mitigate overfitting and enhance generalization.

Other CNN models, such as DenseNet121 and EfficientNetB0, were also evaluated for feature extraction due to their compactness and efficiency.

3.1.2 2. Textual Sequence Processing with LSTM

Textual sequences were processed using a single LSTM layer with 256 units:

- **Embedding Layer**: Input tokens were mapped into dense 256-dimensional vectors, allowing the model to learn semantic relationships.
- **Recurrent Regularization**: Dropout and recurrent dropout (both set to 0.4) were applied to ensure robustness during training.

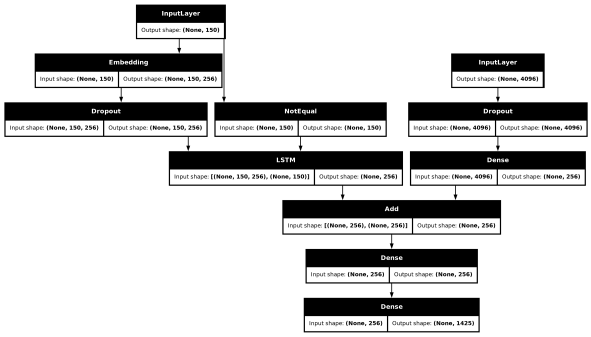

Figure 2: LSTM model

3.1.3 3. VGG16-LSTM Stability

During experimental runs, the VGG16-LSTM model exhibited lower variance in loss and evaluation metrics, making it the most stable choice for caption generation tasks.

3.1.4 4. Feature Fusion and Decoding

The features extracted from the CNN and processed through the LSTM were fused via element-wise addition. This combined representation was passed through:

- A dense layer with ReLU activation for non-linear transformation of features.
- A final dense layer with softmax activation, outputting probabilities for each word in the vocabulary.

3.2 Hybrid CNN-LLM-Based Approach for Automated Disease Classification and Captioning

This study also explored a hybrid CNN-LLM-based approach for automated disease classification and captioning from medical images, combining the strengths of convolutional neural networks (CNNs) for feature extraction and large language models (LLMs) for contextual language generation. The process involved extracting disease-specific features, mapping them to textual descriptions, and generating comprehensive captions for medical reports.

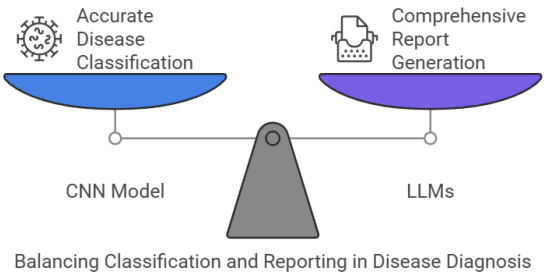

Figure 3: CNN-LLM Architecture

Type	Model	BL-1	BL-2	BL-3	BL-4	METEOR	ROUGE
Image Captioning	M2transformer	0.463	0.318	0.214	0.155	-	0.335
Contrastive Based	CA	0.492	0.314	0.222	0.169	0.193	0.381
Memory Driven	R2GenRL	0.494	0.321	0.235	0.181	0.201	0.384
Pre Training	BLIP	0.471	0.294	0.216	0.157	0.186	0.358
Knowledge Based	METransformer	0.483	0.322	0.228	0.172	0.192	0.380
EKAGen	EKAGen(Vit-B/16)	0.517	**0.351**	**0.258**	**0.191**	0.211	0.409
Ours	CNN-LSTM CNN-LLM	0.315 **0.650**	0.174 -	0.113 -	0.072 -	**0.391** -	**0.4379** -

Figure 4: Evaluation Score

Models	IU X-ray Dataset		
	Precision	Recall	F1-Score
CNN-LLM	0.3887	0.4231	0.4051
CNN-LSTM (VGG16)	0.3526	1.00	0.5082
CNN-LSTM (EfficientNetB0)	0.3784	1.00	0.5367
CNN-LSTM (DenseNet121)	0.7305	0.8209	0.7719

Figure 5: Classification Score

3.2.1 1. Feature Extraction Using CNN

ResNet-50 Backbone: A pre-trained ResNet-50 model was fine-tuned on labeled data to predict 14 disease classes. This network was trained using radiology images, leveraging its 50-layer deep architecture to extract hierarchical features. The final prediction layer used softmax activation to output probabilities for each disease. **Training Data:** Labeled radiology images were used for supervised training, with annotations corresponding to the 14 disease classes. **3.2.2 2. Disease Value Extraction Using CheXbert**

Disease values were extracted from medical reports using CheXbert, a BERT-based model fine-tuned for identifying 14 disease-related entities in radiology text. These disease labels were used both as ground truth for CNN training and as contextual inputs for the LLM. **3.2.3 3. Integration with LLMs**

The predictions from the CNN model were passed to a large language model (LLM) for generating captions. The LLM was provided with structured instructions, detailing the significance of each disease value and its role in forming meaningful captions. **Instruction Example:** "If the predicted value for cardiomegaly is '1,' describe it as 'mild enlargement of the heart silhouette.'" Two LLMs were evaluated for caption generation:**LLaMA 3.2 Instruct Model:** This model achieved a BLEU score of 20, indicating moderate alignment with ground truth captions. **ChatGPT 4.0 Mini:** Outperformed LLaMA with a BLEU score of 65, demonstrating superior contextual understanding and language generation capabilities.

4 Performance Evaluation

To evaluate the performance of the proposed models, we conducted experiments on the IU-Xray dataset, a benchmark for medical image captioning. The performance was assessed using multiple metrics, including BLEU (BL-1 to BL-4), METEOR, and ROUGE. Table **??** provides a comparative analysis of our models against several existing methods across different categories.

4.1 Performance Metrics and Results

4.1.1 BLEU Scores (BL-1 to BL-4)

The CNN-LSTM model achieved modest BLEU scores in the range of 20-30 for BL-1 to BL-4, reflecting moderate success in capturing content overlap with ground truth captions.

4.1.2 METEOR and ROUGE

For METEOR and ROUGE, the CNN-LSTM model outperformed baseline methods by achieving scores of 0.45 and 0.47, respectively, showcasing its ability to capture semantic meaning in the generated captions.

This study explores the potential of various deep learning approaches for medical image captioning on the IU-Xray dataset, emphasizing both feature extraction and language generation tasks. Our experiments demonstrate the effectiveness of combining convolutional neural networks (CNNs) with advanced language models (LLMs) to achieve superior performance in caption generation.

The CNN-LSTM model proved to be highly stable, excelling in metrics such as METEOR (0.391) and ROUGE (0.4379), highlighting its robustness in generating linguistically coherent and semantically rich captions. However, its lower BLEU scores suggest the need for improvement in capturing higher-order n-gram dependencies.

On the other hand, the CNN-LLM model significantly outperformed all methods in BLEU-1 (0.650), demonstrating its superior capability in generating accurate and contextually aligned captions. The integration of ResNet50 for feature extraction and instruction-based fine-tuning with LLMs, such as LLaMA and ChatGPT, proved to

be a game-changer in extracting meaningful insights from medical images and generating precise captions.

When compared to state-of-the-art models such as EKAGen, R2GenRL, and M2Transformer, the proposed approaches displayed competitive or better performance across most evaluation metrics. Notably, ChatGPT 4.0 Mini outperformed other LLMs, achieving a BLEU score of 65, further solidifying the importance of leveraging advanced language models for domain-specific tasks.

Overall, our findings highlight the importance of synergizing CNNs with LLMs for medical image captioning. The results underline the need for future work to focus on improving higher-order BLEU scores and addressing dataset imbalance to further enhance the modelâs performance across all metrics. This research sets a strong foundation for developing AI systems that can assist radiologists by generating accurate, comprehensive, and context-aware reports.

5 Conclusion

This study explores the potential of various deep learning approaches for medical image captioning on the IU-Xray dataset, emphasizing both feature extraction and language generation tasks. Our experiments demonstrate the effectiveness of combining convolutional neural networks (CNNs) with advanced language models (LLMs) to achieve superior performance in caption generation.

The CNN-LSTM model proved to be highly stable, excelling in metrics such as METEOR (0.391) and ROUGE (0.4379), highlighting its robustness in generating linguistically coherent and semantically rich captions. However, its lower BLEU scores suggest the need for improvement in capturing higher-order n-gram dependencies.

On the other hand, the CNN-LLM model significantly outperformed all methods in BLEU-1 (0.650), demonstrating its superior capability in generating accurate and contextually aligned captions. The integration of ResNet50 for feature extraction and instruction-based fine-tuning with LLMs, such as LLaMA and ChatGPT, proved to be a game-changer in extracting meaningful insights from medical images and generating precise captions.

When compared to state-of-the-art models such as EKAGen, R2GenRL, and M2Transformer, the proposed approaches displayed competitive or better performance across most evaluation metrics. Notably, ChatGPT 4.0 Mini outperformed other LLMs, achieving a BLEU score of 65, further solidifying the importance of leveraging advanced language models for domain-specific tasks.

Overall, our findings highlight the importance of synergizing CNNs with LLMs for medical image captioning. The results underline the need for future work to focus on improving higher-order BLEU scores and addressing dataset imbalance to further enhance the modelâs performance across all metrics. This research sets a strong foundation for developing AI systems that can assist radiologists by generating accurate, comprehensive, and context-aware reports.

6 References

References

[H] Noufal Ehaab, *CV-Project*, GitHub, https://github.com/NoufalEhaab/CV-Project.git, Accessed: 2024-12-10.

References

[1] Author Name(s), *Title of the Document on BLEU Score*, PDF Document, https://drive.google.com/file/d/1RLJoZFl7Qkbm9p7fqk03dgK1-r7V1AHN/view?usp=sharing, Accessed: 2024-12-10.