Automated Radiology Report Generation Using Transformers

Wimukthi Nimalsiri
Department of Computer Science and Engineering
University of Moratuwa
Moratuwa, Sri Lanka
wimukthi.18@cse.mrt.ac.lk

Mahela Hennayake
Department of Computer Science and Engineering
University of Moratuwa
Moratuwa, Sri Lanka
mahela.18@cse.mrt.ac.lk

Kasun Rathnayake
Department of Computer Science and Engineering
University of Moratuwa
Moratuwa, Sri Lanka
kasunr.18@cse.mrt.ac.lk

Thanuja D. Ambegoda
Department of Computer Science and Engineering
University of Moratuwa
Moratuwa, Sri Lanka
thanujaa@uom.lk

Dulani Meedeniya
Department of Computer Science and Engineering
University of Moratuwa
Moratuwa, Sri Lanka
dulanim@cse.mrt.ac.lk

2023 3rd International Conference on Advanced Research in Computing (ICARC) | 979-8-3503-4737-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICARC57651.2023.10145699

***Abstract*—Given the rapid increase of respiratory illnesses in recent times, the demand for medical report writing for chest X-Rays (CXR) has significantly increased. In practice, a specialized medical expert has to go through an X-Ray image to compile the accompanying report, which is tedious, not scalable, and potentially prone to human error. Therefore, automatic medical report generation (AMRG) solutions for CXR as a diagnostic assistance tool could play an important role in lowering the burden on radiologists, making them more productive. However, current AMRG solutions are still lagging far behind the performance of human experts due to the reasons such as the inability to extract the most relevant features to be used for the compilation of the report. We address this by proposing MERGIS: MEdical Report Generation using the Image Segmentation approach. MERGIS is a modern transformer-based encoder-decoder model that leverages image segmentation to improve the accuracy of automatic report generation. In this approach, the CXR images are segmented before feeding into the model, enabling the encoder to extract relevant visual features of the medical image resulting in more accurate radiography reports. The proposed model outperforms the current state-of-the-art model for report generation on the MIMIC-CXR dataset with performance scores: BLUE-1 = 0.296, METEOR = 0.128, ROUGE L = 0.335, and CIDEr = 1.150.**

***Keywords*—medical report generation, Chest X-Ray, transformer, image segmentation, self-attention**

I. Introduction

Medical imaging plays an important role in medical diagnosis, providing insights to the state of organs and tissue in a non-invasive manner. Radiologists analyze radiographs such as X-Ray images to produce detailed reports to accompany the images so that other specialists can extract the information more effectively. The requirement for generating reports of chest X-Rays (CXR) has increased given the recent outbreak of COVID-19. However, writing an accurate report is time-consuming and results in increasing the waiting time of patients, invoking dissatisfaction, and inconvenience [1]. Thus, machine learning-based automation could be used to make the medical report generation workflow much more efficient. However, it is challenging given the following requirements:

(i) Generated reports should include coherent content with accurate medical terminology of the findings.
(ii) Medical reports should contain relevant and precise information about the diagnosis to get insights into a patient's condition.

This research intends to increase the efficiency of CXR report creation by providing an accurate report for radiologists to use as a reference. Medical report generation is an exciting area of research with the advancement of robust deep-learning techniques for image analysis and text generation. Usually, a formal medical report consists of impressions and findings. These two sections are concatenated to create a detailed report which will be used as the input for the proposed model. The generated reports need standard medical terms describing any anomaly since these CXR reports are observed by other specialists in the medical diagnosis process. In addition, generating a more accurate, informative, and coherent report without clinician intervention is challenging. Yet a robust deep learning model can address it by generating an accurate medical report with medical terminologies.

This study proposes a transformer-based encoder-decoder architecture to generate medical reports. The encoder consists of a Convolutional Neural Network (CNN) as the backbone to extract the visual features of the CXR image. The decoder is similar to the original transformer used in Natural Language Processing [2], which uses a multi-head attention mechanism. The words are embedded using Byte-Pair Encoding (BPE),

979-8-3503-4737-1/23/$31.00 ©2023 IEEE

which is a popular subword-based tokenization algorithm used by state-of-the-art NLP models [3]. Although similar state-of-the-art models output better results, most of them fail to optimize the visual feature extraction head, which helps to invoke better linguistic expressions in generated reports.

Subsequently, in this paper, we strive to leverage the transformer-based encoder-decoder model by experimenting with different CNN architectures as the backbone of the encoder. Additionally, we use segmented medical images to train the final model to take maximum advantage of the self-attention mechanism, which is the core concept behind transformers. The high-level architecture of the entire process can be observed in Fig. 1. Several experiments were carried out to select and validate the proposed approach.

First, we used a DenseNet-121 as the encoder backbone for feature extraction of the base model. The base model has given better results compared to the scores of recent transformer-based deep learning models. However, a comparative analysis of different CNN models should be performed to improve feature extraction further. The model recorded even better results with ResNet101, which was the most efficient CNN obtained as the encoder backbone from the comparative analysis. Ultimately, segmented images were used to amplify the model's ability to focus on the lung area. The result of this model surpasses all other approaches in image feature extraction and overall text generation. The novel contributions of this paper are as follows:

1) Phase I: Analysis of the effect of different CNN architectures for visual feature extraction to improve the performance of medical image report generation, hence improving the quality of reports generated.
2) Phase II: Introducing image segmentation and morphological post-processing as precursors to the report generation workflow resulting in an increase in accuracy.

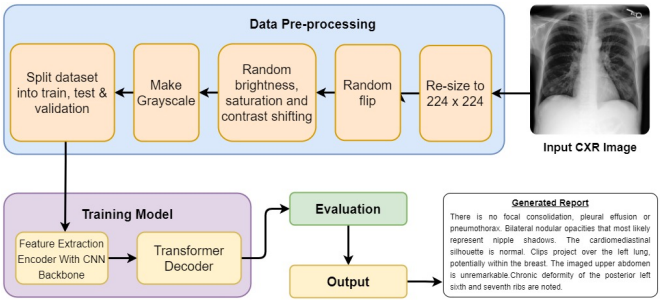


Fig. 1: Proposed baseline model with data pre-processing

The paper is structured as follows. Section II depicts the work related to transformer-based medical image models. A comprehensive description of the proposed model is presented in Section III. The experimental result analysis with the state-of-the-art methods is bestowed in Section IV.

II. Background

The first self-attention-based transformer architecture was introduced by Vaswani et al. [2]. They proposed a scaled dot-product attention mechanism and a multi-head attention mechanism that can run several attention layers in parallel. This method was a solution for the vanishing gradient issue of recurrent neural networks (RNNs). With the attention mechanism, transformers can learn contextual relations between words. Therefore, it can provide more accurate predictions.

Among several studies on medical image report generation, Jing et al. [4] have utilized a hierarchical LSTM decoder to generate medical reports. Their multi-task learning with a co-attention mechanism got 0.517, 0.386, 0.306, and 0.247 scores for BLEU-1, BLEU-2, BLEU-3, and BLEU-4, respectively, for the IU X-Ray dataset. They have made a tag prediction for the input image, and each tag is represented using a word-embedding vector, which acts as a semantic feature for the respective image. A CNN has been used as the visual feature extractor, and both the visual features and semantic features fed into the co-attention mechanism for further capturing of features. In [5], Li et al. have presented a different approach based on extracted disease graph. The encoder module has been used to transform visual features into a structured abnormality graph. After that, the retrieve module is used to load the text template, which matches the abnormalities, and another module called Paraphrase has used to re-write the generated report according to the specific case. Using this model, they achieved a BLEU-1 score of 0.482 from the IU X-Ray dataset.

In another study, Wijerathna et al. [6] have used the LXMERT (Learning Cross-Modality Encoder Representations from Transformers) model, which is mainly designed for question-answering tasks using images. The authors have modified the LXMERT model and used it for medical report generation. They have used the ChexNet model as the primary feature extraction model and integrated a memory into the decoder. With these improvements, they have achieved a BLEU-1 score of 0.498 by using the IU X-Ray dataset. A similar study has been done by Amjoud et al. [1], using a transformer-based approach. The model consists of a feature extractor, a separate encoder, and a decoder. DenseNet-121 is used to extract features, and it is trained on the ImageNet database using an Nvidia T4 GPU. The model achieved 0.479, 0.205, and 0.380 scores for BLEU-1, METEOR, and ROUGE metrics, respectively.

Chen et al. have followed a similar approach for radiology report generation [7], with three major components similar to the previously mentioned study. They have changed the feature extractor to a ResNet and integrated a memory module into the decoder. The purpose of the extra memory module is to improve the original layer normalization with MCLN (Memory-driven Conditional Layer Normalization) for each decoding layer. They achieved a BLEU-1 score of about 0.353 using the MIMIC-CXR dataset [8], which they used for training and evaluation. They evaluated the model using the IU X-RAY dataset and recorded a score of 0.470 for BLEU-1.

"RATCHET" is another study that followed a similar approach to the medical image report generation using transformers [3]. In this model, the encoder has been replaced with the pre-trained DenseNet-121, and it works as the primary image

feature extractor. The decoder architecture of the transformer remains the same as the base model.

Although several studies have addressed medical image report generation, most have not focused much on improving visual feature extraction. Therefore, in this study, we apply image segmentation to improve the feature extraction process and use a transformer-based model to generate radiology reports. Furthermore, several studies have used different CNN architectures to classify CXR images [9]. In our study, we also perform a comparative study to identify the most suitable CNN model for visual feature extraction as the backbone of the encoder.

III. Methodology

A. Datasets

We used the MIMIC Chest X-ray (MIMIC-CXR) Database v2.0.0 [8], a large publicly available dataset of chest radiographs as the main dataset. It contains 377,110 CXR images from 227,827 imaging studies, where each study has frontal and/ or lateral views. This multi-class dataset has 14 derived labels. The original dataset contains rich DICOM (.dcm) images of high-resolution with free-text radiology reports. The uncompressed size of this dataset is approximately 4.6 TB, which is much larger than other publicly available CXR datasets. Since we use a transformer-based approach for medical report generation, a larger dataset similar to MIMIC-CXR tends to give better results.

In this study, we used the MIMIC-CXR-JPG dataset, which contains CXR images in JPG format with structured labels derived from free-text radiology reports due to computational resource constraints. The corresponding CXR report states the observations such as "Heart size is likely normal. Lungs are clear taking into account low lung volumes." and impressions such as "Mild volume overload in the background of low lung volumes.". They have converted the DICOM images in the original dataset to a compressed JPG format using a lossy compression algorithm. The corresponding JPG dataset is approximately 557.6 GB in size, which is significantly lower than the original dataset. We split the dataset into a ratio of 80:10:10 for training, testing, and validation, with 194639, 24344, and 24342 images, respectively.

Moreover, we used two CXR datasets, namely Montgomery County (MC) [10] and Shenzhen Hospital (SH) [11], to train the segmentation model. The MC dataset includes manually segmented lung masks containing 138 posterior-anterior x-rays. The left and right binary lung masks are available separately on different directories and must be combined when training the model. The SH dataset contains 662 frontal chest X-rays, of which 326 are normal cases, and 336 are cases with manifestations of Tuberculosis, including pediatric X-rays. The combined dataset was split into a ratio of 80:10:10 to obtain the train, test, and validation datasets.

B. Process View

The transformer-based proposed model with image segmentation is shown in Fig. 2. Initially, the CXR image is segmented and saved it locally. We have segmented the images from the model trained using MC and SH datasets. Then the segmented images were passed into the feature extraction module.

The result of the feature extraction head is passed as the Key(K) and the Query(Q) inputs to the second multi-head attention (MHA) layer of the decoder. The Value(V) input of the second MHA layer comes from the first masked multi-head attention (M-MHA) layer, and it contains the embedding information of vocabulary we created using the medical reports. Then the attention matrix is calculated by matching the Q and the K against each other and expedited using a scaled dot product and a softmax operation. $softmax(QK^T)$ results in a probability matrix which will be further multiplied by V as shown in (1) to obtain the localized values that the model should focus on. Here, $Q : Query$, and $K : Key$ are the same matrices from the output of the feature extraction layer. The value V is the matrix of word tokens with positional embedding, and d_k is the dimension of the key matrix.The decoder uses these values to generate the next predicted token.

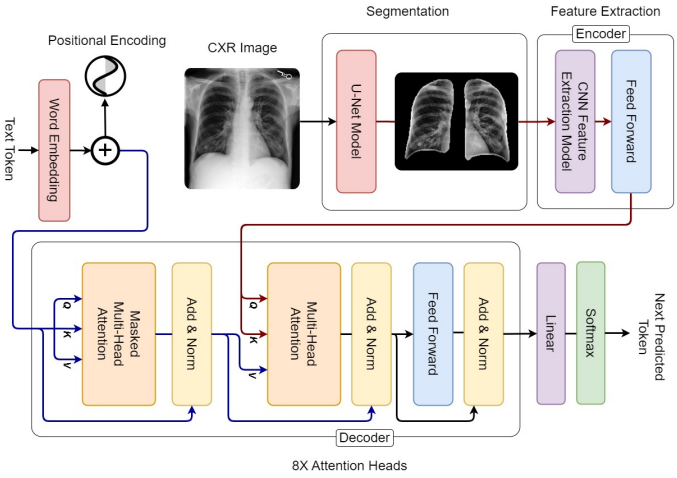


Fig. 2: Proposed model with CXR image segmentation

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

C. Transformer-based encoder-decoder

We used DenseNet-121 model as the primary feature extractor of the baseline model, as inspired by the results of RATCHET [3]. It is a densely connected convolutional neural network, including multiple average pooling layers and a fully connected layer. The model is trained using the Adam optimizer with a learning rate of 1×10^{-5} and a batch size of 32 for 10 iterations. Initially, the medical images are resized to 224 x 224, and applied data augmentation techniques such as include random flipping, random brightness, saturation, and contrast shifting. Next, the images were turned into grayscale, and the overall process is shown in Fig. 1. Visual feature extraction is the primarily focused module of the proposed system, as it significantly impacts the model performance by identifying the abnormalities and defects in CXR images.

Word embedding is also an important part of report generation, and the model needs to possess syntactic and semantic knowledge of the language to generate an accurate report of the medical image. Moreover, medical terminology strongly influences the tokenization process and subsequently leads to a better learning curve of the model. We have used the Byte Pair Encoding (BPE) tokenizer, which is provided by the Huggingface library. The BPE breaks the word into one or more sub-words. The primary idea of these tokenizers is that more frequent words should be given unique IDs, and infrequent words should break into more sub-words while preserving the meaning.

A vocabulary of 20000 words is generated using the BPE tokenizer and the medical reports in the MIMIC-CXR dataset. We only took the words that occurred more than two times to create the vocabulary. In addition, the medical reports are created using the combination of both the Findings and Impression sections of the original report to obtain a detailed report. The $Findings$ section contains the observations of the radiologist about the medical image, and the Impressions section includes a summary of the findings, symptoms, and clinical history. Some words are randomly changed to $< mask >$ tokens to make the process more robust.

After the positional embedding of the medical reports, we used M-MHA to understand the meaning of each word in the medical report using the self-attention mechanism. We have used eight multi-head attention modules to understand each word's context in the medical report. The maximum length of the generated sequence is set to 128 tokens considering the longest radiology report found in the MIMIC-CXR dataset. Moreover, an iterative process generates each token after the initial $< start >$ token and concatenates with the previously generated sequence, which will be used as the input to generate the next token. Subsequently, the process terminates when it predicts the $< end >$ token or exceeds the maximum limit.

D. Comparative analysis of CNN models

The baseline model we proposed in phase I uses a DenseNet-121, as inspired by the model presented in RATCHET [3]. Moreover, motivated by the existing comparative studies on identifying the best performing CNN models in CXR classification [12], we experimented with several CNN models, namely DenseNet, ResNet, MobileNet, Inception, and Xception, to select the best performing CNN model to be used as the backbone of the encoder in this study. In order to assess the performance of each CNN model, we used random pathological classes and evaluated the performance of each model to move forward based on the results stated in TABLE I.

E. Medical image segmentation using U-Net

During phase II, the medical images are segmented to enhance the process of extracting image features from CXR images. Inspired by [13], an efficient U-Net architecture is used to conduct segmentation on CXR images. Both the images and masks from the combined dataset of SH dataset and MC dataset were resized to 224 x 224 and normalized to a range within [0,1]. Data augmentation steps were also performed to reduce the class imbalance. The Adam optimizer with a learning rate of 1×10^{-5}.The model was trained for 20 epochs along with a batch size of 5. Since the attention-based mechanism amplifies necessary details to focus more on the essential aspects of data, a segmented image would give even more promising results when combined with attention. The U-Net architecture comprises a convolution neural network for both downsampling and upsampling, whose training strategy relies on the strong use of data augmentation to improve the efficiency of available annotated samples. The segmented images were fed into the encoder of the model, which contains a ResNet101 as the encoder backbone since we obtained the best results from it using the comparative analysis.

The lung segmentation masks were dilated to load lung boundary information within the training set, and the images were resized to 512x512 pixels. Then, the entire MIMIC-CXR dataset is segmented using the trained U-Net model and saved with the exact folder structure as the original dataset. The images were resized to 512x512 pixels and saved in a reduced resolution locally. The segmented MIMIC-CXR images have shown better performance compared to the base model. The image segmentation module that extracts the lung area is used to comply with the computational resource constraints and less mobility due to the large size of the MIMIC-CXR dataset. This approach reduced the size of the dataset, making it easier to train the model in multiple instances.

IV. Results and Discussion

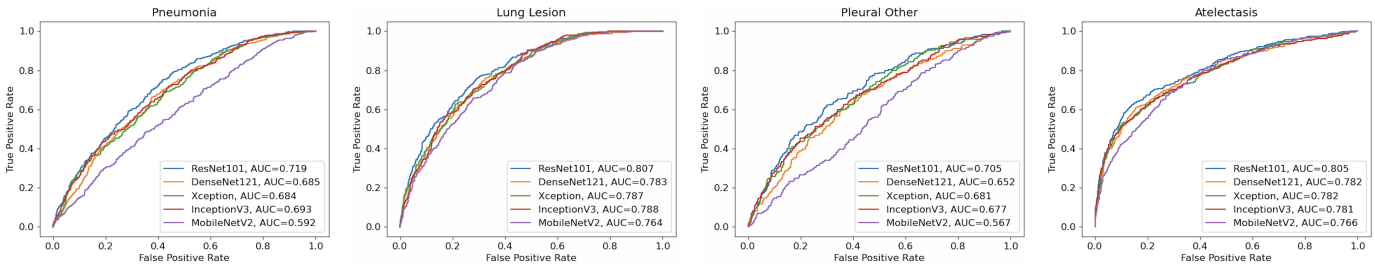
Most of the state-of-the-art transformer-based encoder-decoder models primarily focused on improving the attention mechanism of the transformer decoder [6] [7]. All of those mechanisms prioritize the learning focus of the overall transformer-based model. Thus, many models do not give much attention to improving the encoder for better visual feature extraction. A systematic evaluation of different CNN models for visual feature extraction and usage of image segmentation on attention-based transformer models is much needed to get the maximum out of the encoder.

Initially, we implemented a transformer-based encoder-decoder model with a DenseNet-121 [3] for feature extraction. In phase I, a comparative analysis was done using different CNN models to find the best suitable feature extractor. The results are mentioned only for selected classes due to the database bias, which is a potential limitation of the study and can be minimized with proper data augmentation, although it is a tedious task.

To assess the performance, we have performed direct multi-label classification for the 14 classes available in the CheXpert, a large chest radiograph dataset with uncertainty labels [6]. The accuracy of a test yields its ability to differentiate a diseased person and a healthy person correctly. Similarly, the sensitivity of a test can identify a diseased individual correctly, while the specificity of a test can correctly classify an individual as 'healthy'. These values are more important

TABLE I: Evaluation Results of MIMIC-CXR Image Classification for Selected Classes. (Pn: Pneumonia, LL: Lung Lesion, PO: Pleural Other, At: Atelectasis)

CNN Architecture	Accuracy(%)				Sensitivity(%)				Specificity(%)			
	Pn	*LL*	*PO*	*At*	*Pn*	*LL*	*PO*	*At*	*Pn*	*LL*	*PO*	*At*
ResNet101	**63.5**	**75.6**	**68.7**	**78.4**	66.0	62.4	61.7	62.1	62.4	**78.8**	**69.9**	**86.0**
DenseNet-121	62.3	69.0	63.8	76.9	**75.0**	75.7	59.5	60.9	58.1	67.4	64.6	84.3
Xception	60.3	71.0	54.5	74.3	68.4	76.7	73.9	63.2	57.7	69.6	51.0	79.5
InceptionV3	59.9	71.4	60.1	74.8	71.6	70.4	66.9	61.9	56.0	71.7	58.9	80.8
MobileNetV2	60.7	61.2	40.8	70.3	48.2	**83.0**	**78.2**	**68.5**	**64.8**	55.8	34.1	71.1



(a) (b) (c) (d)

Fig. 3: ROC curves of ResNet101, DenseNet-121, Xception, InceptionV3, and MobileNetV2 related to pathological classes Pneumonia, Lung Lesion, Pleural Other, Atelectasis

in clinical settings for screening tests and the final confirmation of disease. Considering the results shown in TABLE I, sensitivities of most of the pathological classes are higher for MobileNet. Thus, it is more suitable for screening purposes. Testing positive from MobileNet will help the person to identify the diseases beforehand to take necessary premedications. In contrast, the specificities of LL, PO, and At are highest for ResNet, implying that it is a better model for the final confirmation of the disease for those classes since testing positive will guarantee that the patient is diseased. Additionally, observing the results of TABLE I, accuracies of all pathological classes among the selected CNN architectures are highest for ResNet. This ensures that it is better for both screening and the final confirmation of disease. Considering all the results, ResNet is more powerful compared to other CNNs, since it yields better results for both accuracy and specificity.

Furthermore, we have included the receiver operating characteristic (ROC) curves for selected classes that have comparatively large amounts of studies to evaluate the performance of CNNs to some extent. The ROC curves with the area under the curve (AUC) values are shown in Fig. 3. It can be observed that ResNet101 has shown comparatively better results among the considered CNN models for most of the cases. We have selected four pathological classes, namely Pneumonia, Lung Lesion, Pleural Other, and Atelectasis, to visualize the results, considering the higher number of image availability, that result in smooth ROC curves. Generally, classes with imbalanced datasets show less performance, and the model tends to provide ROC curves with discrete behavior.

When the ROC curves for a model are more toward the main diagonal, the false positive rate is larger and the true positive rate is smaller, making the respective test less accurate. In Fig. 3, DenseNet, Xception, and Inception models are on par with each other and produced averagely better results for most of the pathological classes. On the other hand, MobileNet shows a comparatively low AUC score for each of the classes, and ROC curves related to MobileNet are more toward the main diagonal as shown in Fig. 3(a) and Fig. 3(c), making MobileNet the least accurate model from the selected models.

In comparison, DenseNet-121 performs better than MobileNet in all the classes and has better AUC scores as portrayed in Fig. 3(a), Fig. 3(b), and Fig. 3(d). Thus, it is evident that our baseline model, which uses a DenseNet-121 as the primary feature extractor, performs better than most of the state-of-the-art models. Generally, the test tends to be more accurate when the ROC curve is closer to the upper left corner. Singularly, ResNet101 performs better than all the selected CNN models, and it occupies the best ROC curves since most of those curves tend to deviate more toward the upper left corner, resulting in higher AUC scores. Thus, considering the results in TABLE I and ROC Curves in Fig. 3, a CNN encoder with ResNet101 is used as the primary feature extractor since ResNet101 performed finer in many cases.

The overall accuracy of using ResNet101 backbone as the visual feature extractor is stated in TABLE II. The NLG metrics include BLEU-1 (B-1), METEOR (MET.), ROUGE L (R L), and best language quality performance is highlighted in bold. MERGIS contains the results obtained using the segmented images, and we selected the encoder with the Resnet101 backbone as the most performing model using the comparative analysis.

Finally, we used the segmented images of MIMIC-CXR images as the input and evaluated the baseline model using the ResNet101 encoder backbone as the primary feature extractor (ResNet101 + Segmentation). The baseline model with a DenseNet-121 feature extractor and transformer decoder seems

TABLE II: NLP Evaluation on MIMIC-CXR Report Generation.

Model	**NLP Metrics**			
	B-1	*MET.*	*R L*	*CIDEr*
TieNet [14]	0.190	0.069	0.200	0.411
RATCHET [3]	0.232	0.101	0.240	0.493
LXMERT as Caption Decoder [6]	0.165	-	-	-
Our Experiments				
Baseline (DenseNet-121)	0.224	0.108	0.233	0.648
ResNet + without Segmentation	0.251	0.113	0.265	0.767
MERGIS (ResNet + Segmentation)	**0.296**	**0.128**	**0.335**	**1.150**

to have a better performance than the TieNet [14], which is a CNN-RNN-based model that uses ResNet for feature extraction. Thus, we can observe that Transformer based models outperform RNN-based models. The results of TieNet were obtained from [3] since they have reimplemented the TieNet model and evaluated using the MIMIC-CXR dataset.

The RATCHET model [3], which is a CNN-RNN-based medical transformer that uses a DenseNet-121 as the feature extractor, outperformed both our baseline model with DenseNet-121 and TieNet. Additionally, LXMERT, as caption decoder [6], has used ChexNet as the feature extractor. It is similar to DenseNet-121 since both have 121 layers and perform slightly lower than TieNet. In contrast, our model, which only uses ResNet101 as the encoder backbone, performed better than RATCHET on all the BLUE-1, METEOR, ROUGE L, and CIDEr scores. Further, studies have shown that ResNet101 provides better accuracy in general CXR classification [12]. Importantly, the proposed MERGIS model that uses segmented images outperformed all the other models by a significant margin. This clearly shows that enhancing the encoder, which is responsible for visual feature extraction, can significantly influence the overall performance of the transformer-based deep learning models.

V. Conclusion

We proposed MERGIS model, which is a transformer-based approach for medical report generation improved with image segmentation as an input preprocessing step. The overall research consists of two main phases. In the first phase, a comparative study between different CNNs, namely DenseNet, ResNet, MobileNet, Inception, and Xception, was used as the encoder backbone to extract the visual features. Since ResNet101 outperformed all other CNNs in comparative analysis, we used ResNet as the encoder model for the second phase. The second phase uses a transformer-based model with CXR images segmented using a modified U-Net architecture. Given the visual features, the transformer decoder generates the medical report. To improve the feature extraction and reduce the attention to unnecessary data on CXR images, we segmented the MIMIC-CXR images, which significantly improved the accuracy of the model. We have shown that the proposed MERGIS model yields better scores for BLUE, METEOR, ROUGE, and CIDEr metrics compared to the existing state-of-the-art transformer-based models on MIMIC-CXR dataset and proves the fact that image segmentation can significantly enhance the performance of transformer-based models.

References

[1] A. B. Amjoud and M. Amrouch, "Automatic generation of chest x-ray reports using a transformer-based deep learning model," in *2021 Fifth International Conference On Intelligent Computing in Data Sciences (ICDS)*. IEEE, Oct. 2021. [Online]. Available: https://doi.org/10.1109/icds53782.2021.9626725

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017. [Online]. Available: https://arxiv.org/abs/1706.03762

[3] B. Hou, G. Kaissis, R. M. Summers, and B. Kainz, "Ratchet: Medical transformer for chest x-ray diagnosis and reporting," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 293–303.

[4] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2018. [Online]. Available: https://doi.org/10.18653/v1/p18-1240

[5] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Knowledge-driven encode, retrieve, paraphrase for medical image report generation," 2019. [Online]. Available: https://arxiv.org/abs/1903.10122

[6] V. Wijerathna, H. Raveen, S. Abeygunawardhana, and T. D. Ambegoda, "Chest x-ray caption generation with chexnet," in *2022 Moratuwa Engineering Research Conference (MERCon)*, 2022, pp. 1–6.

[7] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, "Generating radiology reports via memory-driven transformer," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1439–1449. [Online]. Available: https://aclanthology.org/2020.emnlp-main.112

[8] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C. ying Deng, R. G. Mark, and S. Horng, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific Data*, vol. 6, no. 1, Dec. 2019. [Online]. Available: https://doi.org/10.1038/s41597-019-0322-0

[9] D. Meedeniya, H. Kumarasinghe, S. Kolonne, C. Fernando, I. De la Torre Díez, and G. Marques, "Chest x-ray analysis empowered with deep learning: A systematic review," *Applied Soft Computing*, vol. 126, p. 109319, 2022, doi: https://doi.org/https://doi.org/10.1016/j.asoc.2022.109319.

[10] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, and G. Thoma, "Two public chest x-ray datasets for computer-aided screening of pulmonary diseases," *Quantitative Imaging in Medicine and Surgery*, vol. 4, no. 6, 2014. [Online]. Available: https://qims.amegroups.com/article/view/5132

[11] S. G. Stirenko, Y. Kochura, O. Alienin, O. Rokovyi, P. Gang, W. Zeng, and Y. G. Gordienko, "Chest x-ray analysis of tuberculosis by deep learning with segmentation and augmentation," *CoRR*, vol. abs/1803.01199, 2018. [Online]. Available: http://arxiv.org/abs/1803.01199

[12] C. Fernando, S. Kolonne, H. Kumarasinghe, and D. Meedeniya, "Chest radiographs classification using multi-model deep learning: A comparative study," in *Proceedings of the 2nd International Conference on Advanced Research in Computing (ICARC)*, Belihuloya, Sri Lanka, 2022, pp. 165–170, doi: https://doi.org/10.1109/ICARC54489.2022.9753811.

[13] H. Kumarasinghe, S. Kolonne, C. Fernando, and D. Meedeniya, "U-net based chest x-ray segmentation with ensemble classification for covid-19 and pneumonia," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 18, no. 07, p. 161–175, 2022, doi: https://doi.org/10.3991/ijoe.v18i07.30807.

[14] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, "Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays," 2018. [Online]. Available: https://arxiv.org/abs/1801.04334