FTT : Fourier Transform based Transformer for Brain CT Report Generation

*Note: Sub-titles are not captured in Xplore and should not be used

Jieun Kim
Graduate School of Information
Yonsei University
Seoul, Republic of Korea
lilly9928@yonsei.ac.kr

Byeong Su Kim
Dept. of Artificial Intelligence
Yonsei University
Seoul, Republic of Korea
kbs23@yonsei.ac.kr

Insung Choi
Dept. of Integrative Medicine
Yonsei University
Seoul, Republic of Korea
gamma@yonsei.ac.kr

Zepa Yang
Guro Hospital Biomedical Research Center
Korea University
Seoul, Republic of Korea
yangzepa@korea.ac.kr

Beakcheol Jang
Graduate School of Information
Yonsei University
Seoul, Republic of Korea
bjang@yonsei.ac.kr

2024 International Conference on Artificial Intelligence in Information and Communication (ICAIIC) | 979-8-3503-4434-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICAIIC60209.2024.10463282

***Abstract*—The interpretation of Brain Computed Tomography (CT) scans predominantly falls under the purview of specialized radiologists. However, given the challenges associated with excessive workloads, human resource limitations, urgency in emergency scenarios, and inconsistencies in outsourced interpretations, the margin for diagnostic errors is substantial. To ameliorate this issue, burgeoning research has been directed towards the automatic synthesis of various medical diagnostic reports. Contrary to conventional image captioning tasks, the domain of medical report generation is fraught with inherent biases, making it arduous to accurately extract features pertinent to specific pathological lesions. Moreover, redundant descriptions of normative areas further impede the precise delineation of anomalies. To address these challenges, this paper introduces a novel transformer architecture that synergizes lesion detection algorithms with Fourier Transform techniques. Experimental results indicate that our proposed model outperforms existing combined-embedding models and exhibits enhanced performance when applied to Fourier-transformed image data.**

***Index Terms*—Automated Medical Reporting, Brain CT Scans, Transformer Architecture, Fourier Transform, Lesion Detection**

I. Introduction

Stroke is a significant medical condition, accounting for the third leading cause of mortality in South Korea as of 2021. Specifically, cerebral hemorrhage is implicated in approximately 20% of these instances. The case fatality rate of cerebral hemorrhage varies globally and regionally, ranging from 30% to 50%, underscoring the urgency for accurate diagnosis and timely intervention. Typically, cerebral hemorrhages are swiftly detected via non-invasive cranial computed tomography (CT) scans. However, accurate interpretation of these images mandates the expertise of specialized radiologists. The propensity for diagnostic errors exists, exacerbated by factors such as high workloads, personnel deficits, emergent cases, and issues related to outsourced image interpretation. To address these challenges, ongoing research in the domain of medical artificial intelligence is aimed at automating medical report generation, thus expediting its integration into clinical workflows. Unlike conventional image captioning tasks in artificial intelligence [1], medical report generation is complicated by variances in imaging data. A pivotal challenge involves the presence of data biases, both in imaging and textual data [2], [3]. In the realm of CT scans, the fraction of slides depicting lesions is minuscule compared to the total image set. Additionally, these lesions are generally small in dimension. Furthermore, brain structures are ontogenetically similar across individual patients, often leading to model training that is skewed towards these common patterns. As a consequence, there is a significant challenge in feature learning pertaining to the lesion itself. Textual data often includes redundant descriptions of normal anatomical regions, thereby constraining the model's capacity to identify unique pathological features. These intricacies compound the complexities involved in processing unstructured data, accurate lesion identification, and the generation of clinically relevant sentences in medical reports. To tackle these issues, this preliminary study introduces a novel computational approach. The method involves cropping the lesion area in a CT scan within a bounding box and then applying Fourier Transform to this cropped region to re-domainize the image. This approach isolates the lesion characteristics, minimizes noise, and reduces computational overhead during model training [4], [5]. Subsequently, the Fourier-transformed image is used as the input for a Vision Transformer (ViT) [6], integrated with a Transformer's decoder [7] to synthesize the radiological report. The ViT's proficiency in contextual feature capture

ICONS(Institute of Convergence Science),Yonsei University

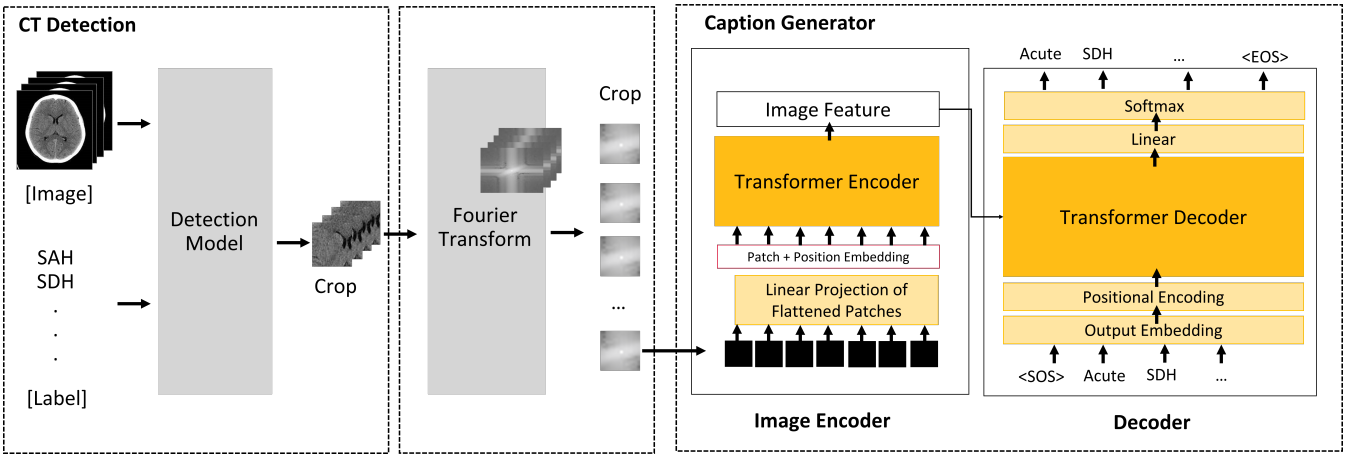

Fig. 1. Proposed Method

enhances the model's capacity to learn key imaging attributes [6]. Moreover, the Transformer algorithm, equipped with an attention mechanism, is ideally suited for the task of report generation, given its demonstrated efficacy in natural language processing [7]. The contributions of this work encompass the introduction of a new methodology for precise lesion identification, efficient feature extraction, and the flexible generation of medical text. This study initially focuses on employing Fourier-transformed cropped lesion images as training data. However, future research will expand to include other mathematical transformations, such as the Radon and Hough Transforms, as alternative domains for feature extraction and report synthesis.

II. Related Work

Advancements in deep learning technologies have spurred extensive research into their application for generating image captions and medical reports. Specifically within the healthcare sector, deciphering complex medical images and automatically generating descriptive reports pose formidable challenges. Among the myriad neural network architectures, the fusion of encoder-decoder frameworks and attention mechanisms has been particularly efficacious in advancing the field of image captioning research [8]–[10]. These architectures are extensively employed as robust techniques for extracting visual features from images and generating captions therefrom. Notably, these strategies are instrumental in the generation of medical reports. Numerous studies have advocated for the use of CNN-RNN architectures to predict image tags and structure reports accordingly [11]. Moreover, attempts to leverage pretrained language models like GPT-2 have been made to condition the generation of reports [12]. This is facilitated by introducing novel key and value weights that can be mapped onto the decoder's own attentional framework. However, conventional image captioning methodologies often fall short when applied to medical reporting, owing to distinct feature disparities between general image data and medical imagery. Consequently, specialized algorithms such as DeltaNet have been investigated to produce more reliable reports via iterative search-and-update mechanisms [13]. Building upon these technological strides, considerable focus has been directed towards medical image-to-verbal pretraining in recent years. In this study, a holistic framework termed Vision-Language Transformer (VLT) is introduced through diverse datasets [14], serving as the foundation for various medical vision-language pretraining (VLP) endeavors. To further enhance report quality, an innovative approach leveraging knowledge graphs is proposed [15]. This approach aims to augment or update pre-constructed graphs, thereby activating context-appropriate knowledge coverage. Despite these advancements, formidable challenges persist in medical report generation, particularly concerning lesion identification accuracy and contextual text generation comprehension. To mitigate these challenges, this study proposes a novel methodology incorporating the Fourier Transform results of cropped lesion-specific images into a Vision Transformer (ViT) for feature learning [6], while employing the transformer's decoder for text generation.

III. Proposed Method

In this paper, we propose a transformer model that combines a lesion detection model and a Fourier transform to generate readouts based on medical image data. In this session, we will describe the proposed model in more detail, and the proposed model is shown in Figure 1. For precise lesion identification in medical image data, a lesion region detection model is used to extract the coordinate values of the lesion. After that, the extracted coordinate values are used to cut out the suspected lesion area in the medical image data, and a Fourier transform is applied to the cut image to convert the image into a new domain. The transformer model is then trained to generate appropriate readings for the medical image data. Session A describes in more detail the feature detection model used for precise lesion identification. Session B describes the image encoder model for lesion feature extraction in more detail, and Session C describes the proposed decoder model for report generation in more detail.

A. Lesion Detection

The medical CT image data and the corresponding readout data are organized as shown in (1) and (2). For medical CT image data, there are multiple cross-sectional images $i_n$in one case of CT image data I, as shown in (1). In the case of reading text data, there is one caption data y_I^c for each case of image data I^n as shown in (2).

$$I = [i_1, i_2, i_3, ..., i_n] \tag{1}$$

$$Y = [y_I^1, y_I^2, y_I^3, \ldots, y_I^c] \tag{2}$$

In this paper, we use the YOLO [16] model to perform lesion region detection. For more precise lesion identification, the CT image data is cropped based on the coordinate values extracted by the lesion region detection model (see (5)). The cropped image is divided into low and high frequencies by Fourier transform using (3). In (3), $f(i, j)$ means image coordinate values, and (k, l) means image coordinates resulting from Fourier transform. N is the image size.

$$DTFT(.) = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} f(i,j) e^{-i2\pi(\frac{k_i}{N} + \frac{l_i}{N})} \tag{3}$$

$$e^i x = \cos x + i \sin x \tag{4}$$

$$(I_x, I_y) = Crop(I, f_d(I, C)) \tag{5}$$

$$I_{cf} = DTFT(I_x, I_y) \tag{6}$$

The final lesion detection model can be represented by equation (6). The lesion detection model is used to extract image training data for the readout generation model.

B. Image Encoder

To extract features from medical CT image data, we use the ViT [6] model as an image encoder. ViT is a model that applies a transformer to image classification. In this paper, we extract the features of the Fourier transform image and use them as input to the decoder for reading text generation without using them for image classification. The cropped Fourier transform image is divided into patches and linearly projected and embedded in D dimensions. In this paper, we fixed the number of dimensions to 64. We then prefix the embedded patch with x_class, like a [class] token. Then we add Positional embedding E_pos to have the position information of the images divided into patches. This is the final output of the image encoder and serves as an image feature. Using the basic transformer encoder structure, a multi-head self-attention layer and a multi-layer perceptron layer are repeated. After that, the nom layer is used before each block, and the residual concatenation is used after every block. In this study 3 layers were stacked. The overall behavior of the encoder is shown in the following equation.

$$z_0 = [x_{class}; x_p^1 E; x_p^2 E; \ldots; x_p^N E] + E_{pos}, \quad E \in \mathbb{R}^{(P^2) \times D}, E_{pos} \in \mathbb{R}^{(N+1) \times D} \tag{7}$$

TABLE I
OVERALL CHARACTERISTICS OF DATA

Characteristics	Values
Case count	202
Average age	62.97 ± 16.56
Male	144
Female	58
Slice thickness	3 (mm)
kVp	120 (keV)
SAH	156
SDH	120
IVH	70
Hemorrhage	85
EDH	44
Acute	99

$$z'_l = MSA(LN(z_{l-1})) + z_{l-1}, \quad l = 1 \ldots L \tag{8}$$

$$z_l = MLP(LN(z'_l)) + z'_l, \quad l = 1 \ldots L \tag{9}$$

In (7), P refers to the size of the patch being split and is fixed to 2 in this study.

C. Medical Report Generation Decoder

We propose a transformer model-based decoder for medical report generation. The transcript caption data is subjected to word embedding for data generation and each position information is added to the embedded transcript data. Then, similar to the conventional encoder-decoder, it performs masked self-attention to avoid referring to the input value at a future point in time. The extracted values are then used as the key and value of the transformer model and passed on to the next multi-head attendance. The overall decoder equation is as follows.

$$y_e = W_e Y \tag{10}$$

$$C_1 = y_e + PE(y_e) \tag{11}$$

$$C'_{l-1} = MultiHeadAtt([C_{l-1}, C_{l-1}, C_{l-1}]) \tag{12}$$

$$C_l = FCN(MultiHeadAtt([z_l, C_{l-1}, C_{l-1}]) \tag{13}$$

In (10), W_e means the embedding weight value and y_e means the embedded caption value. In (11), PE performs the process of embedding the position information of the caption by position embedding.

IV. EXPERIMENT

A. Experimental setup

Dataset: Images of patients who visited KU Guro Hospital from January 1, 2020 to December 31, 2022 and were diagnosed with cerebral hemorrhage through head and neck CT scans. A total of 200 data were collected, and the characteristics of the data are shown in Table I.

TABLE II
BLEU SCORE COMPARISON

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
CNN+LSTM	0.17	0.303	0.0063	0.0015
ViT+Decoder (w/o FT)	0.17	0.032	**0.0069**	0.0018
ViT+Decoder	0.17	**0.034**	**0.0069**	**0.002**

The data were collected along with the radiologist's diagnosis of the disease. The readings and image data were reviewed by a radiology data expert and used in the study after preprocessing personal information. The research design and data were approved by the Institutional Review Board (IRB) of Guro Hospital, Korea University. (Approval number: 2020GR0452)

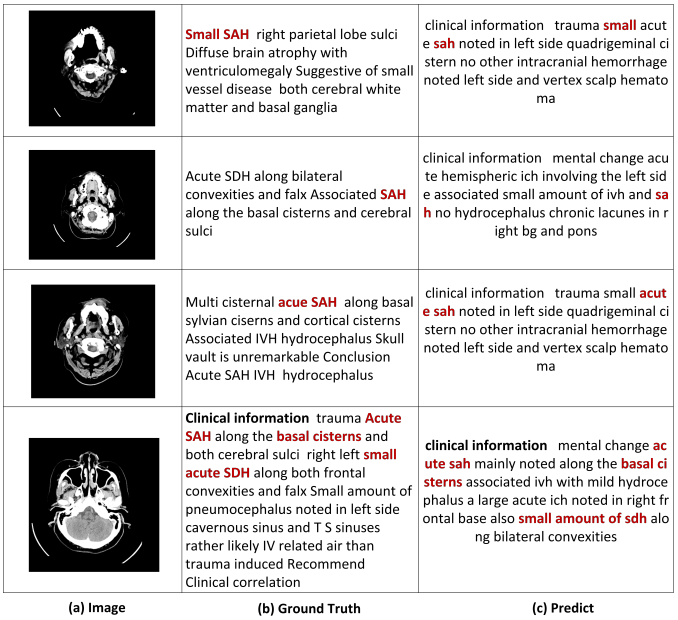

Fig. 2. Visualize the predictions of our proposed model

B. Result

Figure 2 visualizes the results of our proposed model. Figure 2 (a) shows a sample of head and neck CT image data. (b) in Figure 2 shows a written reading of the head and neck CT image data by a radiologist. In this study, we used this reading as the correct answer. Figure 2 (c) shows the predicted reading from our proposed model. The red text in Figure 2 shows when the correct reading and the generated reading are the same. This shows that the sentences are generated based on the correct disease.

Table 2 compares the results between our proposed model and other models. We use BLEU as the evaluation metric. All experiments were conducted on the same dataset, and for the proposed model, the ViT+Decoder model, the results were compared using video data with and without Fourier transform. Table 2 shows that the accuracy of the ViT+Decoder model is higher than that of the CNN+LSTM model. It can also be seen that the highest score is achieved when the proposed model and Fourier transform image data are utilized. In the case of reading texts, they are written by humans, and the composition of the texts may be different for each author, and the composition of the texts is complex. For this reason, the BLEU score is not as high as that of a typical image captioning model.

V. CONCLUSION

In this study, we demonstrated that our proposed model outperforms existing composite embedding frameworks. We also substantiated that the application of Fourier-transformed image data yields higher levels of accuracy in comparison to conventional image data. Although the present study manifests promising outcomes, numerous avenues for enhancement remain. As part of our future work, we posit that there are multifaceted strategies to elevate the precision of caption generations, including the acquisition of supplemental data and the exploration of alternative datasets. Further research is warranted to augment both the volume and quality of data, as procuring more comprehensive and diverse medical datasets will bolster the model's robustness and generalizability. It is anticipated that the employment of data augmentation techniques can optimize the utilization of existing data pools and refine the model's training regimen. To glean a comprehensive spectrum of image features, ensuing research will explore an approach that generates high-quality captions by assimilating various types of image data into the Vision Transformer (ViT); this will include entire CT slides encapsulating cerebral hemorrhage lesions, cropped lesion-specific images demarcated with bounding boxes, and images translated into various domains such as Fourier, Radon, and Hough Transforms. While BLEU scores serve as a prevalent metric for evaluating captioning models, the inclusion of context, semantics, and domain-specific accuracy is imperative in the medical field. As such, we advocate for the formulation of a holistic evaluation metric encompassing these variables. Future extensions of this research could address a broader array of neurological conditions, and additional model development targeting ailments like stroke, brain tumors, and Alzheimer's disease could diversify the range of applicability of the proposed methodology.

REFERENCES

[1] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[2] M. Li, R. Liu, F. Wang, X. Chang, and X. Liang, "Auxiliary signal-guided knowledge encoder-decoder for medical report generation," *World Wide Web*, vol. 26, no. 1, pp. 253–270, 2023.

[3] Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Hybrid retrieval-generation reinforced agent for medical image report generation," *Advances in neural information processing systems*, vol. 31, 2018.

[4] A. K. Bharodiya, "Feature extraction methods for ct-scan images using image processing," *Computed-Tomography (CT) Scan*, p. 63, 2022.

[5] H. S. Bhadauria and M. Dewal, "Efficient denoising technique for ct images to enhance brain hemorrhage segmentation," *Journal of digital imaging*, vol. 25, pp. 782–791, 2012.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[8] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, 2014.

[9] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," *arXiv preprint arXiv:1511.06732*, 2015.

[10] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4651–4659.

[11] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[12] Z. M. Ziegler, L. Melas-Kyriazi, S. Gehrmann, and A. M. Rush, "Encoder-agnostic adaptation for conditional language generation," *arXiv preprint arXiv:1908.06938*, 2019.

[13] X. Wu, S. Yang, Z. Qiu, S. Ge, Y. Yan, X. Wu, Y. Zheng, S. K. Zhou, and L. Xiao, "Deltanet: Conditional medical report generation for covid-19 diagnosis," *arXiv preprint arXiv:2211.13229*, 2022.

[14] L. Xu, B. Liu, A. H. Khan, L. Fan, and X.-M. Wu, "Multi-modal pre-training for medical vision-language understanding and generation: An empirical study with a new benchmark," *arXiv preprint arXiv:2306.06494*, 2023.

[15] M. Li, B. Lin, Z. Chen, H. Lin, X. Liang, and X. Chang, "Dynamic graph enhanced contrastive learning for chest x-ray report generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3334–3343.

[16] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.