ISSN 2616-5775 Vol. 6, Issue 13: 186-195, DOI: 10.25236/AJCIS.2023.061326

Generation of pathological descriptions with interpretable reasoning via sequential progressive attention network and knowledge based on location relations

Yunan Shi[1,*]

[1]*Jiangsu University, Zhenjiang, China*
[*]*Corresponding author: 1142058227@qq.com*

Abstract: *Deep learning tools have received tremendous attention when applied to the automatic diagnosis of gastric cancer computed tomography (CT) scans. However, the computer is unable to accurately describe the location information and severity based on visual features of the largest cross-section of the tumor alone, making it difficult to find a globally optimal solution. Also the healthcare industry is a high-risk decision-making area with a high demand for interpretable models and risk assessment. In this paper, we propose the Sequential Progressive Attention Network, which has three main contributions: (1) The relative position coding module is designed to align the gastric cavity and obtain the relative position of the tumor to the cavity by using the cavity as a reference object. (2) A dynamically distributed dilated convolution method based on random directional field perturbations is proposed to construct the uncertainty of the model. The method evaluates the impact of different components on the decisions within the local region by locally perturbing the attention region of the dilated convolution. (3) The Long Short Term Memory (LSTM) is applied to analyze the changes in tumor morphology on consecutive CT images. Specifically, a mask-based non-uniform coding module is put forward to reduce the weighting factor of non-tumor regions and reduce the sensitivity of the LSTM to feature changes in non-target regions. (4) The Location relations between the gastric lumen and the tumour is modelled by LSTM to obtain a triadic external knowledge base with relative interpretability, making the model's decisions transparent. Finally, we conduct image-caption experiments on the gastric CT image dataset and apply the BLUE metric to evaluate the effectiveness of the experiment. The experimental results are improved by 4\% compared with the latest models in recent years.*

Keywords: *Sequential Progressive Attention Network, Deep learning, The Long Short Term Memory, computed tomography*

1. Introduction

It is essential that the reading and diagnosis of medical images be conducted by relevant medical personnel who have received professional training. For example, during the process of radiologists diagnosing CT images of gastric cancer and writing pathological descriptions, as the pathological description generally includes specimen type, tumor location, gross type, size and depth of invasion, keywords are given to each part and logical sentences are employed as a means of forming these keywords into structured sentences.

In an era where information technology is prominent, data plays a crucial role. Different hospitals have different standards regarding the writing of case descriptions, which makes it more difficult to share data between hospitals. The lack of a large amount of supporting data can also hinder new research breakthroughs.

The requirements for writing CT image reports are quite strict in some rural-urban fringes where medical levels are relatively poor. Therefore, in order for doctors and researchers to be able to make a correct diagnosis, they must possess strict professional knowledge and have many years of long-term experience[1].

For imaging doctors who skilled in medical and pathology-related knowledge, the writing process can be quite tedious and time-consuming as a result of the repetitive factor of the report[2]. With the

ISSN 2616-5775 Vol. 6, Issue 13: 186-195, DOI: 10.25236/AJCIS.2023.061326

environmental factors and social pressure in the modern environment, the number of patients has increased rapidly[3]. Due to the complexity of the cases, a great deal of time is spent on writing reports, which significantly reduces the work efficiency of relevant staff members[4].

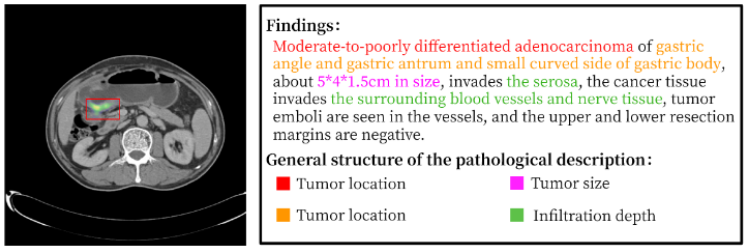

Figure 1: Structure diagram of pathological description.

In terms of the above issues, the idea of using image caption as a means of automatically generating a standardized report on the physical description of gastric cancer cases is proposed. As shown in Figure 1, the report consists of four parts: the pathological type of the tumor, location, size and the depth of tumor invasion. However, there are several key issues that need to be addressed.

1) The location information of the tumor is difficult to judge, because the CT images of gastric cancer from different patients may have significant differences in the morphological distribution of the same tissue part on the images.

2) Analysis of tumor information from a single CT image of gastric cancer to compile a pathological description tends to fall into a local optimal solution, due to the fact that tumor morphology varies slightly on different topographies of CT images of gastric cancer from the same patient.

3) Diagnosing gastric cancer is a high-risk area and the decisions of deep learning black box models lack uncertainty assessment[5]. It is difficult for physicians to trust the decisions made by deep models.

Based on the above motivation, this paper proposes a sequential progressive attention network (SPA-Net). The model is capable of progressively focusing the feature changes of tumors in sequential tomographic CT images. Furthermore, the model provides decision making and risk assessment using interpretable reasoning.

Generally, our main contributions are:

1) A focal position encoding module is proposed that is able to apply the relative position to obtain phase information of the tumor and also to align the gastric cavity.

2) A randomly perturbed dynamically dilated convolution is proposed to allow the model to make diverse decisions based on the characteristics of different distributions.

3) A multi-scale self-attentive non-uniform coding module is designed in this work, which combines multi-scale features while using attention mechanisms to obtain contextual information. Ultimately, the masking mechanism is used to focus on tumor regions and tumor-related contextual information, allowing the LSTM to reduce sensitivity to changes in non-action region features.

4) The concept of knowledge mapping is transferred to images. The gastric cavity, the tumour, and the related phase information are employed to construct a comprehensible external knowledge base, thus making the decision-making process transparent and interpretable.

This paper is organized as follows: section II describes the related work, section III provides an introduction to the proposed method, section IV presents the interpretable evaluation metrics and the experimental results, section V provides a conclusion, and section VI expresses our gratitude to the relevant workers.

2. Related Works

2.1 Position encoding method

Position coding can be divided into absolute position information and relative position information. Absolute position information measures the position information of a single target, while relative

ISSN 2616-5775 Vol. 6, Issue 13: 186-195, DOI: 10.25236/AJCIS.2023.061326

position information measures the positional relationship between different targets. In recent years, researchers have introduced location coding in order to compensate for the inability of attention mechanisms[6] to preserve sequential location information. In 2017, Gehring et al. considered absolute position information as a trainable parameter and built an absolute position vector based on the length of the sentence and the encoding dimension of the word vector[7], while the vector could be updated with the training process. In 2018, to quantify the relative distance between the query vector and the reference vector[8], Shaw proposed to incorporate relative position information into the self-focus mechanism. In recent years, several scholars have introduced Transformer into the direction of computer vision. The emergence of Vision Transformer[9] has led to an increase in the study of positional coding in the field of images. In 2021, Kan Wu et al. found that relative coding methods in images were stronger than absolute coding methods, inserting metrics of orientation and distance into the self-attentive layer, ultimately proving their superiority in tasks related to image classification.

Relative position is a spatial correlation between a subject and an object. Since there is a lack of knowledge beforehand related to self-attention, the query vector in self-attention is subject to change, thus resulting in a global shift in the subject. Hence, it is a key point of exploration in this paper to figure out how to correctly identify the subject and object and implement the process of establishing relative position.

2.2 Multi-scale feature fusion method

Low-level features are of high resolution and contain more detailed information while possessing common semantics[10]. Conversely, high-level features possess strong semantic information, low resolution and a weak ability for perceiving details. In order to preserve image coding as much as possible, Kaiming He (2014) proposed a multi-scale feature fusion method that was based on pyramid pooling[11] to reduce a lot of computation while fusing different scale features. In 2018, Juncheng Li et al. proposed a multi-scale residual network[12], introducing convolution kernels of varying sizes as a means of detecting image features of different scales adaptively. This achieved a significant improvement to the super-resolution of a single image. In 2021, Yao Jianmin et al. proposed a module integrating multi-scale features and attention mechanisms in medical image retrieval research[4]. The module extracts the feature level containing contextual relationships from shallow to deep by means of combining residual network and self-attentive mechanism, which effectively improves the weaknesses of blurred, noisy, and poorly contrasted medical grayscale images.

Although previous methods have largely solved the problem of combining relevant information and feature loss, too much data makes the model unable to focus on the key information of the image. In this paper, proposed model successfully solves this problem.

2.3 Decision Risk Measurement Methods

Explanatory methods can generally be divided into Post-hoc reasoning and Pre-hoc reasoning. Post-hoc explanation represents a unique approach to extract information from a learned model. Although the working principle of the model cannot be accurately elucidated, for a given trained distributed inference model, certain explanation of the model's working mechanism, decision-making behavior, and evidence basis can be made by using explanation methods or constructing explanation models, such as CAM[13], Grad-CAM[14] and Score-CAM[15], which are all applied to understand and reason the behavior of the network through visualization. Pre-hoc explanation models refer to models that are inherently interpretable or integrate interpretable modules into their architecture. For a trained learning model, the decision-making process or decision basis of the model can be understood without additional information, such as knowledge graph, which can explain the model's decision through queries of existing libraries.

Most of the existing interpretable models and methods for interpreting them have been studied in classification tasks and are difficult to transfer to tasks such as segmentation, target detection, etc. Our work is inspired by knowledge graphs and utilizes the gastric cavity, the tumour, and the related phase information to build triadic knowledge graphs with relative interpretability, exploiting the knowledge of the library to query and validate the decisions of the model.

ISSN 2616-5775 Vol. 6, Issue 13: 186-195, DOI: 10.25236/AJCIS.2023.061326

3. Method

3.1 Overview

This work presents a model for generating medical reports based on CT images. As can be seen in Figure 1, the model has three main tasks. First, the phase assessment module acquires the tumor phase information on different topographies and makes the final decision through an expert review mechanism. At the same time, the metric decision module provides the risk assessment for the decision. Secondly, the multi-scale self-retaining non-uniform coding module analyses individual level tumor lesion features, which incorporates multi-scale features of the image while focusing on the background reference information associated with the lesion, and then the LSTM model encodes the changes in the lesion and the background information associated with the lesion at different levels, with the encoded information representing the severity of the lesion. Finally, word generation uses two key pieces of information (phase information and severity) as a guide for generating pathological tumor descriptions.

An overview of the framework is shown in Fig. 2.

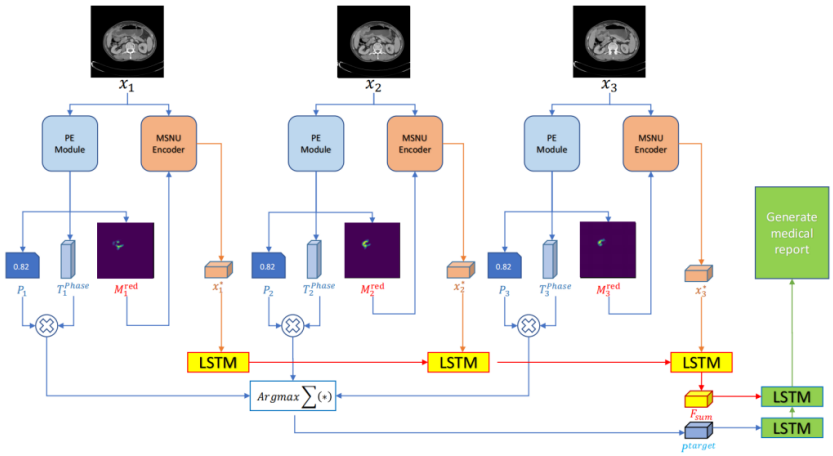

Figure 2: The overall framework of the Aggregated Spatiotemporal Information Network.

3.1.1 Phase Evaluation Module

Accurately delineating the position of a tumor in the stomach is a challenging endeavor due to its diminutive size and indistinct characteristics. Moreover, the medical industry is a high-risk decision area and there is an absence of interpretable ways to evaluate the risk of decisions. In light of these difficulties, we introduce a phase assessment module. As depicted in Figure 3, the phase assessment module has three steps.

3.1.2 Relative position coding

In the initial stage, our experiments complete the alignment of the gastric cavity tissue through relative position coding, correlating each pixel point with the gastric cavity tissue based on the position information.

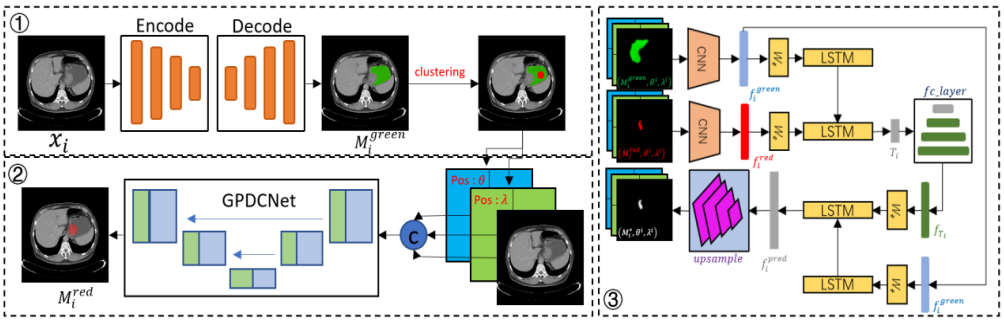

Figure 3: The structure of the Phase Evaluation Module.

3.1.3 Knowledge graph triplet construction

In the third step, our work proposes to train a relatively interpretable triadic knowledge graph by

ISSN 2616-5775 Vol. 6, Issue 13: 186-195, DOI: 10.25236/AJCIS.2023.061326

establishing associations between gastric lumen, tumour and tumour phase information. Ultimately, the model is able to perform a query test on the decision by querying the knowledge graph, as shown in Figure 6.

3.2 Sequential progressive attention network

3.2.1 Multiscale Self-Attention Non-Uniform Encoder

In order to maintain focus on the changes in tumor characteristics across successive tomographies, while reducing the significance of non-tumor regions in influencing model decisions. After this encoder has drawn out the global features through multi-scale approaches and attentional mechanisms, the tumor mask map is implemented to non-uniformly weigh the extracted global features, ultimately concentrating on the tumor region.

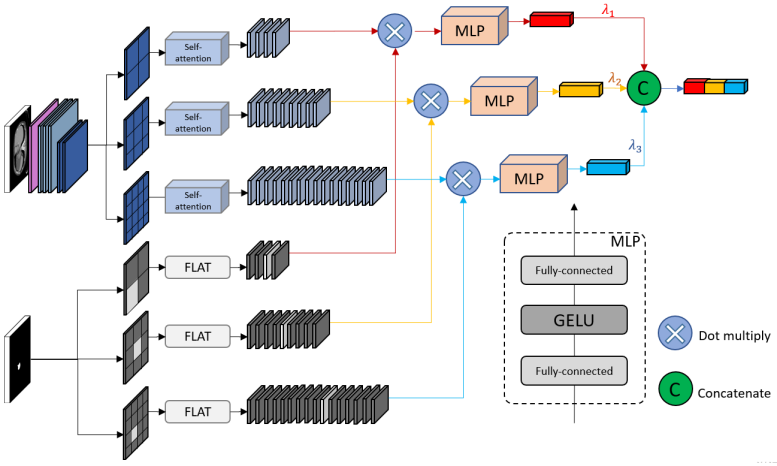

Figure 4: The overall architecture of Multiscale Self-Attention Non-Uniform Encoder.

3.3 Analysis of lesion changes

In order to achieve a globally-optimal solution, we synthetically analyze the morphological variations of tumors on various topographies by transforming a single image recognition task into an video analysis task. Accordingly, we feed the feature encoding F_i^* of the different tomographic images into the LSTM model in the form of sequences, thereby encoding the changes of tumor regions across sequential slices and the additional background information associated with the tumor, and consequently outputting the feature vector F^*.

In this process, the LSTM will be much less sensitive to the feature changes in non-tumor regions due to the reduced weighting factor of non-tumor regions.

3.4 Pathological description generation

The Pathology Description Generator is a multilayer Long Short-Term Memory (LSTM) model which produces a sequence of words describing a pathology condition by employing the phase label vector T^{Target} and the feature vector F^* as the first and second inputs to the model, respectively.

4. Experimental validation]

4.1 Dataset

Our experiments perform extensive assessments on a public medical dataset and a self-built dataset in collaboration with hospitals.

XRay[16] is a collection of chest x-ray images, consisting of 7470 image and report pairs. Each report includes Impressions, Findings, Labels, Comparisons and Indications. On average, each image was associated with 2.2 labels, 5.7 sentences, and each sentence contained 6.5 words.

GC contains 500 CT tomographic images of stomach tumors and diagnostic reports of the tumors

ISSN 2616-5775 Vol. 6, Issue 13: 186-195, DOI: 10.25236/AJCIS.2023.061326

provided by the First People's Hospital of Zhenjiang City, Jiangsu Province. Furthermore, the phase information of the tumors was extracted based on the tumor reports. This dataset contains grayscale images with a resolution of 512 x 512 pixels, which are stored in the DICOM format. In addition, the tumor labeling task for this dataset is carried out by several physicians specialized in imaging. We randomly select 80% of the dataset (i.e. 400 cases) as our training set and the rest as the test set.

4.2 Implementation Details

In the process of building the knowledge graph triad, our experiments construct a triplet external knowledge base composed of gastric cavity, tumor, and phase information applying two short-term LSTM models. The dimensionality of this LSTM hidden layer and the dimensionality of the input features are both set to 768 and 512 respectively. For the coding task of the images, our proposed model carries out feature extraction and flattening utilizing ResNet50, and then the extracted features are unified into a dimension of 512 utilizing a linear fully-connected layer. Moreover, the phase labels have a dimension of 7, which is similarly mapped to a dimension of 512 through a linear fully-connected layer. In the training phase, to produce uncertainty between the two LSTMs while ensuring that the inputs and outputs of both do not affect each other, we trained the two short-time LSTMs independently of each other. In the testing phase, with a view to achieving interpretable risk assessment, we replaced the Mean Squared Error (MSE) with mIou and computed the overlap between the tumour mask map produced by the segmentation task and the lesion mask threshold map obtained via querying the external knowledge base[17-26].

4.3 Implementation Details

In the process of building the knowledge graph triad, our experiments construct a triplet external knowledge base composed of gastric cavity, tumor, and phase information applying two short-term LSTM models. The dimensionality of this LSTM hidden layer and the dimensionality of the input features are both set to 768 and 512 respectively. For the coding task of the images, our proposed model carries out feature extraction and flattening utilizing ResNet50, and then the extracted features are unified into a dimension of 512 utilizing a linear fully-connected layer. Moreover, the phase labels have a dimension of 7, which is similarly mapped to a dimension of 512 through a linear fully-connected layer. In the training phase, to produce uncertainty between the two LSTMs while ensuring that the inputs and outputs of both do not affect each other, we trained the two short-time LSTMs independently of each other. In the testing phase, with a view to achieving interpretable risk assessment, we replaced the Mean Squared Error (MSE) with mIou and computed the overlap between the tumour mask map produced by the segmentation task and the lesion mask threshold map obtained via querying the external knowledge base.

In the attention multiscale non-uniform coding module, experiments use 1 and 0.3 to set the factors of the weight matrix, respectively. The three parameters lambda_1,lambda_2, and lambda_3 depicted in Figure 4 are initialised with random numbers as the weighting factors for the different scale feature maps.

In the text generator module, the dimensionality of the hidden state and word embedding of the LSTM is set to 512. The output pathology report utilises the phase tagging feature of the lesion and the encoding after incorporating changes in the lesion features.

Our proposed model applies the Adam optimizer[27] for parameter learning. The learning rates for CNN (ResNet-50[28]) and hierarchical LSTM are 1e-4 and 1e-5, respectively.

4.4 Evaluation metrics

The model proposed in this paper involves three tasks: image segmentation, label classification, and report generation. Next, our expriements apply some evaluation indicators utilized for different tasks.

In the comparison experiment of paragraph generation, our work employs the Blue metric, the CIDEr metric, and the Rouge-L metric, which are important evaluation metrics in NLP, in order to make a comprehensive comparison with other models.

In evaluating the ablation experiments of the proposed positional encoding in the image segmentation stage and the prediction label stage, our expriements evaluate the accuracy of image segmentation using DICE and Recall, which are commonly applied in medical segmentation tasks, and

ISSN 2616-5775 Vol. 6, Issue 13: 186-195, DOI: 10.25236/AJCIS.2023.061326

the accuracy of prediction using Precision[29-31].

4.5 Comparative experiments

Our experiments make use of the text generation evaluation metric BULE to assess the performance of sentence generation (Table 1). Our proposed method is distinct from one-to-one models that generate case descriptions from a single image, in that it is a many-to-one model that produces case descriptions reliant upon sequential changes in image. The results of our experiments demonstrate that the performance of the one-to-one model is significantly inferior to that of the many-to-one model.

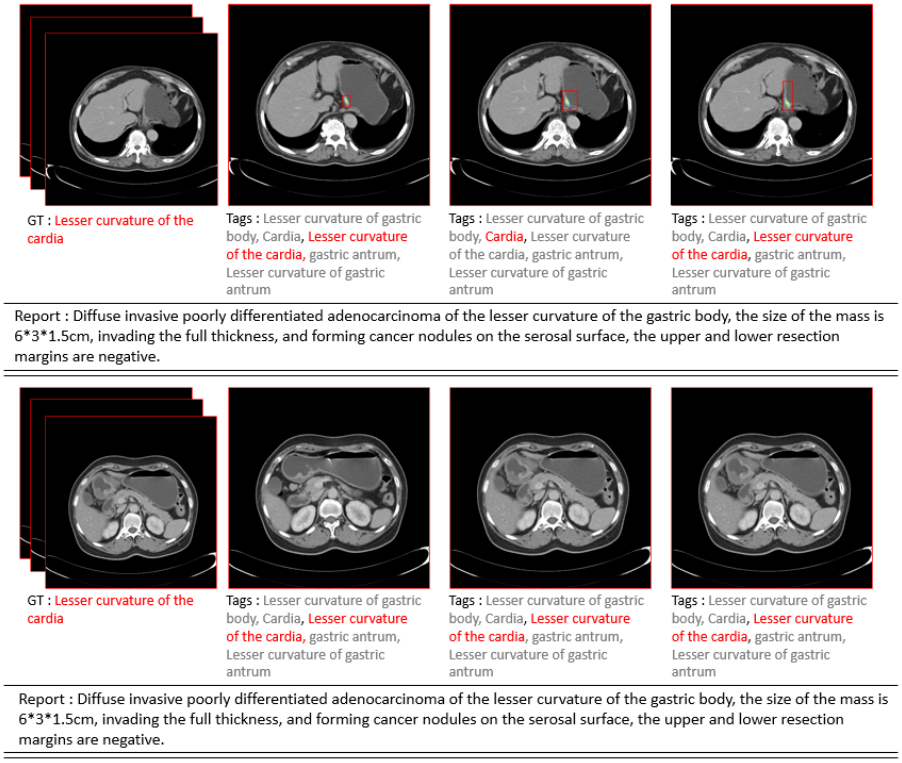

Figure 5: Example of Phase Labels Predicted from Sequence CT Images and Generated Tumor Case Description Report.

It is not difficult to understand that the structure of the stomach and the morphology of the lesion is altered in various tomographic scans. Models based on a single image can only focus on the static distribution of lesion morphology, leading to decisions that tend to fall into local optimal solutions.

Table 1: Main results for paragraph generation on the Gastric cancer lesion dataset and IU-XRay dataset.

Dataset	Method	year	Bule-1	Blue-2	Blue-3	Blue-4	Blue-5	CIDEr	ROUGE-L
Gastric cancer lesion dataset	CNN-RNN[16]	2015	0.447	0.381	0.333	0.344	0.362	1.580	0.578
	LRCN[17]	2015	0.495	0.402	0.369	0.308	0.274	1.589	0.577
	Soft-ATT[19]	2015	0.608	0.495	0.487	0.424	0.351	1.568	0.576
	ATT-RK[20]	2016	0.612	0.543	0.451	0.412	0.376	1.564	0.544
	Transformer [21]	2020	0.542	0.354	0.313	0.301	0.251	1.589	0.588
	MT[21]	2020	0.570	0.421	0.368	0.288	0.333	1.688	0.641
	PPKED[22]	2021	0.583	0.451	0.414	0.421	0.318	1.599	0.541
	AlignTransformer[23]	2021	0.614	0.412	0.453	0.405	0.358	2.140	0.612
	ASIN	Ours	0.654	0.547	0.464	0.402	0.357	2.250	0.623
IU-Xray	HRNN[24]	2017	0.439	0.281	0.190	0.133	0.102	0.261	0.342
	CoAtt[25]	2017	0.455	0.288	0.205	0.154	0.134	0.277	0.369
	HRGR-Agent[26]	2018	0.438	0.298	0.208	0.151	0.121	0.343	0.322
	CMAS-RL[27]	2020	0.464	0.301	0.210	0.154	0.103	0.275	0.362
	Transformer[21]	2020	0.396	0.254	0.179	0.135	0.094	0.204	0.342
	MT[21]	2020	0.470	0.304	0.219	0.165	0.115	0.229	0.371
	PPKED[22]	2021	0.483	0.315	0.224	0.168	0.123	0.351	0.376
	ASIN	Ours	0.498	0.336	0.219	0.161	0.139	0.362	0.374

Our experiments utilize an expert review mechanism that combines N decision scores from consecutive tomographic CT images to determine the final output. Scoring of decisions is made by

ISSN 2616-5775 Vol. 6, Issue 13: 186-195, DOI: 10.25236/AJCIS.2023.061326

querying the knowledge base of location relations as shown in Figure 6, the red mask is from the segmentation model while the yellow mask indicates the overall range of lesion occurrences at the corresponding location, which implies that the decision is more risky when the red part surpasses the threshold. The full model achieves optimal results on all evaluation metrics, thus validating the efficacy of the proposed model.

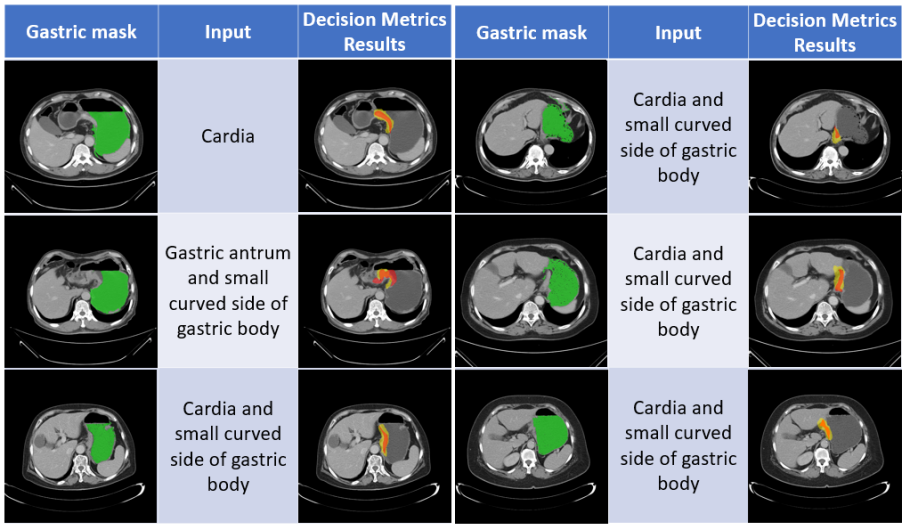

Figure 6: Schematic diagram of the decision metric, the green part represents the mask of the gastric cavity, the yellow part represents the result of evaluating the decision, and the red part represents the tumor segmentation result.

5. Ablation Results

Table 2 shows the accuracy of the respective label predictions and the matching degree of the generated paragraphs evaluated by different encoding methods, including our relative position embedding, our model without position embedding, our proposed model with an absolute position, and the attention mechanism that comes with Patch PosEmbedding. In this experiment, our positional encoding method achieves the highest accuracy in predicting labels, with some improvement in paragraph-sentence matching.

It is evident that different labels and morphological variations in lesion characteristics can have an effect on the resulting pathological description. Every case description is based on the lesion's phase information, depth, and size. The experimental results indicate that the model lacking location encoding cannot accurately identify the location information of lesions around the stomach. On the other hand, by integrating our position encoding module, the model can accurately analyze the positional relationship between the lesion and the stomach and capture phase information of the lesion.

Table 2: Illustration of paragraph generated by Ours-PosEmbedding, Ours-no-PosEmbedding

Methodl	Acc	Bule-1	Blue-2	Blue-3	Blue-4	Blue-5	CIDEr	ROUGE-L
Ours-PosEmbedding	0.742	0.694	0.655	0.564	0.542	0.601	2.850	0.623
No-PosEmbedding	0.451	0.523	0.489	0.477	0.462	0.485	1.980	0.512
Absolute location	0.508	0.568	0.572	0.544	0.505	0.490	2.065	0.545
Patch with PosEmbedding	0.680	0.642	0.623	0.552	0.548	0.536	2.250	0.590

6. Conclusion

In this paper, we present a sequential progressive attention network for the generation of CT diagnostic reports of gastric cancer. Firstly, our experiments propose relative position information based on polar coordinates to help the model learn the orientation and distance relationships between the gastric cavity and the tumor in the images. Simulating sensory uncertainty by introducing a dynamically expanding convolutional kernel allows the model to perceive different features to make personalized decisions. Thirdly, we introduced a multi-scale self-aware non-uniform encoder in the analysis of tumor changes on serial tomography. The encoder focuses on the tumor region and

ISSN 2616-5775 Vol. 6, Issue 13: 186-195, DOI: 10.25236/AJCIS.2023.061326

contextual information related to the tumor and reduces the sensitivity of the LSTM model to changes in non-tumor regions.

Fourthly, a three-tuple external knowledge base is generated by associating the gastric cavity, tumor and the location of the tumor, and the model's decisions are interpreted and evaluated via the query of the knowledge base. Finally, with the gastric cancer medical CT dataset, the effectiveness of the proposed method can be demonstrated through quantitative and qualitative studies.

References

[1] B. Jing, P. Xie, E. Xing, ``On the automatic generation of medical imaging reports," arXiv preprint arXiv:1711.08195 (2017).
[2] B. Yan, M. Pei, M. Zhao, C. Shan and Z. Tian, "Prior Guided Transformer for Accurate Radiology Reports Generation," in IEEE Journal of Biomedical and Health Informatics, 2022, doi: 10.1109/ JBHI.2022.3197162.
[3] L. Zhang et al., "Multi-Focus Network to Decode Imaging Phenotype for Overall Survival Prediction of Gastric Cancer Patients," in IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 10, pp. 3933-3942, Oct. 2021, doi: 10.1109/JBHI.2021.3087634.
[4] L. Zhou, J. Yao and Q. Yan and Z. Lin, ``Medical image retrieval with multi-scale features and attention mechanism," \emph{Chinese Journal of Liquid Crystal \& Displays], vol. 36, 2021.
[5] A. Kendall, Y. Gal, ``What uncertainties do we need in bayesian deep learning for computer vision?" \emph{Advances in neural information processing systems], 30 (2017).
[6] A. Vaswani, N. Shazeer, et al, ``Attention is all you need," Advances in neural information processing systems, 30 (2017).
[7] J. Gehring, M. Auli, et al, ``Convolutional sequence to sequence learning," International conference on machine learning, PMLR, 2017.
[8] P. Shaw, J. Uszkoreit, A. Vaswani, ``Self-attention with relative position representations,"arXiv preprint arXiv:1803.02155 (2018).
[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al, ``An image is worth 16x16 words: Transformers for image recognition at scale,"arXiv preprint arXiv: 2010. 11929 (2020).
[10] K. Wu, H. Peng, M. Chen, J. Fu, ``Rethinking and improving relative position encoding for vision transformer," Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.
[11] K. He, X. Zhang, et al, ``Spatial pyramid pooling in deep convolutional networks for visual recognition," \emph{IEEE transactions on pattern analysis and machine intelligence], 37.9 (2015): 1904-1916.
[12] J. Li, F. Fang, K. Mei, et al, ``Multi-scale residual network for image super-resolution," \emph{Proceedings of the European conference on computer vision (ECCV)], 2018.
[13] B Zhou, A Khosla, et al, ``Learning deep features for discriminative localization," \emph{In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2921–2929], 2016.
[14] M Cogswell, A Das, ``Grad-cam: Visual explanations from deep networks viagradient-based localization," \emph{In Proceedings of the IEEE international conference on computer vision, pages 618–626], 2017.
[15] H Wang, Z Wang, ``Score-cam: Score-weighted visual explanations for convolutional neural networks," \emph{In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pages 24–25,], 2020.
[16] D. Demner-Fushman, MD. Kohli, et al, ``Preparing a collection of radiology examinations for distribution and retrieval," \emph{J Am Med Inform Assoc], 2016 Mar;23(2):304-10. doi: 10.1093/jamia/ocv080. Epub 2015 Jul 1. PMID: 26133894; PMCID: PMC5009925.
[17] J. Donahue, L. Anne Hendricks, et al, ``Long-term recurrent convolutional networks for visual recognition and description," \emph{Proceedings of the IEEE conference on computer vision and pattern recognition], 2015.
[18] A. Karpathy, L. Fei-Fei, ``Deep visual-semantic alignments for generating image descriptions," \emph{Proceedings of the IEEE conference on computer vision and pattern recognition], 2015, pp.3128-3137.
[19] K. Xu, J. Ba, et al, ``Show, attend and tell: Neural image caption generation with visual attention," \emph{International conference on machine learning. PMLR], 2015, pp.2048-2057.
[20] Q. You, H. Jin, et al,``Image captioning with semantic attention," \emph{Proceedings of the IEEE conference on computer vision and pattern recognition], 2016, pp.4651-4659.
[21] Z. Chen, Y. Song, et al, ``Generating radiology reports via memory-driven transformer,"arXiv

ISSN 2616-5775 Vol. 6, Issue 13: 186-195, DOI: 10.25236/AJCIS.2023.061326

preprint arXiv:2010.16056 (2020).
[22] F. Liu, X. Wu, S. Ge, et al, ``Exploring and distilling posterior and prior knowledge for radiology report generation," \emph{Proceedings of the IEEE conference on computer vision and pattern recognition], 2021, pp.13753-13762.
[23] D. You, F. Liu, S.Ge, X. Xie, et al, ``Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation," \emph{International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham], 2021, pp.72--82.
[24] J. Krause, J. Johnson, R. Krishna and L. Fei-Fei, "A Hierarchical Approach for Generating Descriptive Image Paragraphs," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3337-3345.
[25] B. Jing, P. Xie and E.Xing, ``On the automatic generation of medical imaging reports," arXiv preprint arXiv:1711.08195 (2017).
[26] Y. Li, et al, ``Hybrid retrieval-generation reinforced agent for medical image report generation," Advances in neural information processing systems 31 (2018).
[27] B. Jing, Z. Wang, E. Xing, ``Show, describe and conclude: On exploiting the structure information of chest x-ray reports," arXiv preprint arXiv:2004.12274, 2020.
[28] DP. Kingma, J. Ba, ``A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.
[29] K. He, X. Zhang, S. Ren, J. Sun, ``Deep residual learning for image recognition," Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 770-778.
[30] J. Long, E. Shelhamer, T. Darrell, ``Fully convolutional networks for semantic segmentation," Proceedings of the IEEE conference on computer vision and pattern recognition, 2015: 3431-3440.
[31] O. Ronneberger, P. Fischer, T. Brox, ``U-net: Convolutional networks for biomedical image segmentation," International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.