Multi-View Coupled Self-Attention Network for Pulmonary Nodules Classification

Qikui Zhu[1], Yanqing Wang[2⋆], Xiangpeng Chu[4], Xiongwen Yang[3,4], and Wenzhao Zhong[4]

[1] Department of Biomedical Engineering, Case Western Reserve University, OH, USA. QikuiZhu@163.com
[2] Department of Gynecology, Renmin Hospital of Wuhan University, Wuhan, China. yanqingwang543@gmail.com
[3] School of Medicine, South China University of Technology
[4] Guangdong Provincial People's Hospital
Guangdong Academy of Medical Sciences
Guangdong Lung Cancer Institute
Guangdong Provincial Key Laboratory of Translational Medicine in Lung Cancer

Abstract. Evaluation of the malignant degree of pulmonary nodules plays an important role in early detecting lung cancer. Deep learning-based methods have obtained promising results in this domain with their effectiveness in learning feature representation. Both local and global features are crucial for medical image classification tasks, particularly for 3D medical image data, however, the receptive field of the convolution kernel limits the global feature learning. Although self-attention mechanism can successfully model long-range dependencies by directly flattening the input image to a sequence, which has high computational complexity. Additionally, which unable to model the image local context information across spatial and depth dimensions. To address the above challenges, in this paper, we carefully design a Multi-View Coupled Self-Attention Module (MVCS). Specifically, a novel self-attention module is proposed to model spatial and dimensional correlations sequentially for learning global spatial contexts and further improving the identification accuracy. Compared with vanilla self-attention, which has three-fold advances: 1) uses less memory consumption and computational complexity than the existing self-attention methods; 2) except for exploiting the correlations along the spatial and channel dimension, the dimension correlations are also exploited; 3) the proposed self-attention module can be easily integrated with other frameworks. By adding the proposed module into 3D ResNet, we build a classification network for lung nodules' malignancy evaluation. The nodule classification network was validated on a public dataset from LIDC-IDRI. Extensive experimental results demonstrate that our proposed model outperforms state-of-the-art approaches. The source code of this work is available at the https://github.com/ahukui/MVCS.

⋆ Co-corresponding Authors

Keywords: Pulmonary Nodules · self-attention · Multi-View Coupled Self-Attention.

1 Introduction

The accurate and earlier identification of malignant lung nodules from computed tomography (CT) screening images is a critical prerequisite for early detecting and diagnosing lung cancer [2, 8]. Deep learning-based methods [33, 18, 26, 17] have obtained promising results in lung nodules' malignancy identification research with their effectiveness in learning feature representation. For example, Lyu et al. [14] developed a multi-level convolutional neural network (ML-CNN) which consists of three CNNs for extracting multi-scale features in lung nodule CT images to assess the degree of malignancy of pulmonary nodules. Xie et al. [27] proposed a novel Fuse-TSD lung nodule classification algorithm that uses texture, shape and deep model-learned information at the decision level for distinguishing malignant from benign lung nodules. Murugesan et al. [15] created a simple yet effective model for the rapid identification and U-net architecture based segmentation of lung nodules. This approach focuses on the identification and segmentation of lung cancer by detecting picture normalcy and abnormalities. Although these methods have achieved remarkable results, there is still room for improvement in exploiting lung nodules information. Since the receptive field of the convolution kernel used in layer is always small and limited, which also limits the global feature learning during feature extraction and further limits the global information absorbed. However, both local and global features are critical for the malignancy assessment of pulmonary nodules. Assisting the model to obtain global context information from input data can further improve the performance of identification.

Recently, the self-attention mechanism, particularly for Transformer [4], has been recognized as an effective way to exploit the global information and has been successful in natural language processing and 2D image analysis [31], due to the effectiveness of modeling long-range dependencies. For example, Li et al. [13] proposed an induced self-attention based deep multi-instance learning method that uses the self-attention mechanism for learning the global structure information within a bag. Guo et al. [7] proposed a separable self-attention network for video representation learning by investigating the relationship between spatial attention and temporal attention through a sequential self-attention structure. Inspired by PCA, Du et al. [5] proposed an interaction-aware self-attention to further use non-local information in feature maps. By constructing a spatial feature pyramid, the proposed model improves attention accuracy and classification accuracy. Zhang et al. [30] proposed an attention residual learning convolutional neural network model for skin lesion classification in dermoscopy images, which jointly uses the residual learning and novel attention learning mechanisms to improve the discriminative representation ability of DCNNs. However, the existing self-attention module meets two-fold challenges: 1) which builds the dependencies merely by computing the correlations along spatial dimensions and ignores

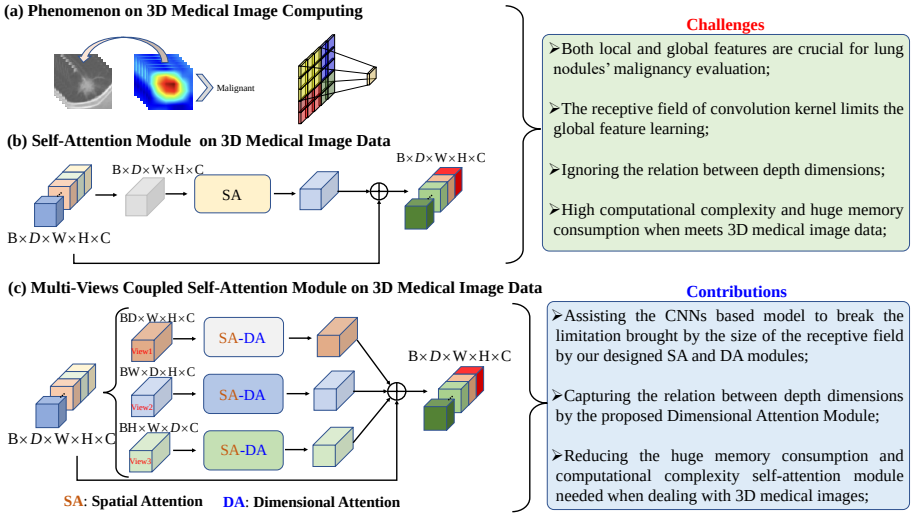

Fig. 1. The distinguishing between the self-attention module and our proposed Multi-View Coupled Self-Attention module.

the relation between depth dimensions; 2) which has high computational complexity and huge memory consumption when meets the date with high spatial size, especially for 3D medical images, as the computational complexity of the self-attention is quadratic with respect to the number of tokens. The above phenomena limit the effectiveness of a vanilla self-attention module [24] in the domain of 3D medical image analysis.

To overcome the above challenges, in this study, we propose a novel self-attention module for assisting the model to model long-range dependencies and extract global information on 3D medical image data. Specifically, we propose a Multi-View Coupled Self-Attention module (MVCS) for completing the above functions with less memory consumption and computational complexity. In our design, different from the vanilla self-attention module, three independent spatial self-attentions are collaborate utilized to investigate the long range dependencies among pixels from three views for modeling the global spatial contextual and dimensional correlation information. Inside MVCS, both local and global spatial contextual information is captured with less memory consumption and computational complexity. Its advance has three-fold: 1) MVCS could model spatial and dimensional correlations sequentially for learning global spatial contexts; 2) MVCS uses less huge memory consumption and computational complexity than the existing self-attention methods when dealing with 3D medical image data; 3) MVCS can be easily integrated with other frameworks. By adding the proposed module into 3D ResNet, we build a nodule classification network for nodules' malignancy evaluation. The nodule classification network was validated on a public dataset LUNA16 from LIDC-IDRI. Extensive experimental results

demonstrate that our proposed model has performance comparable to state-of-the-art approaches.

Summary, our detailed contributions are as follow:

- Our proposed MVCS module solves the problem that the relation between depth dimensions be ignored by a specific designed dimensional attention module.
- Our proposed MVCS module has less memory consumption and computational complexity compared with the vanilla self-attention module when dealing with 3D medical image data.
- Our proposed MVCS module can be easily integrated with other frameworks. By adding the proposed module into 3D ResNet, we build a nodule classification network for nodules' malignancy evaluation. Extensive experimental results demonstrate that our proposed model has performance comparable to state-of-the-art approaches on the public LUNA16 dataset.

2 Related Works

2.1 Pulmonary Nodules Classification

Lung cancer is consistently ranked as the leading cause of tumor-associated deaths all around the world in the past several years due to its aggressive nature and delayed detection at advanced stages. According to the statistics, estimated 10-year postoperative disease-specific survival (DSS) rates were 100% and 100%, and overall survival (OS) rates were 95.3% and 97.8% of patients with resected adenocarcinoma in carcinoma in situ (AIS) and minimally invasive adenocarcinoma (MIA) of the lung [8], respectively. Additionally, the 5-year survival for patients who present with advanced-stage IV non-small cell lung cancer is less than 10%, this percentage increases to at least 71% if the diagnosis is made early [2].

In recent years, deep learning-based methods [34, 32] have obtained promising results on the identification of malignant nodules. For instance, Kumar et al. [11] proposed an autoencoder framework to extract lung nodule deep features for lung nodule classification. Shen et al. [18] proposed a hierarchical Multi-scale Convolutional Neural Networks (MCNN) to capture nodule heterogeneity by extracting discriminative features from alternatingly stacked layers for lung nodule classification. Xie et al. [26] proposed a Multi-View Knowledge-Based Collaborative (MV-KBC) deep model to classify benign-malignant nodules under limited data. Shen et al. [17] proposed a domain-adaptation framework that learns transferable CNN-based features from nodules without pathologically-confirmed annotations to predict the pathologically-proven malignancy of nodules. Jiang et al. [9] presented a novel attentive and ensemble 3D Dual Path Networks for pulmonary nodule classification via contextual attention mechanism and a spatial attention mechanism. Shi et al. [20] proposed a Semi-supervised Deep Transfer Learning (SDTL) framework for benign-malignant pulmonary nodule diagnosis by

utilizing a transfer learning strategy. Additionally, an iterated feature-matching-based semi-supervised method is proposed to take advantage of a large available dataset with no pathological results and a similarity metric function is adopted to iteratively optimize the classification network.

Although deep learning-based methods have obtained promising results in the study of malignant tumor identification of nodules, the limitations from the CNN itself always affect the performance of the model. First, CNN-based models employ kernel with a fixed and small size, such as 3×3, 5×5, for feature extraction, which hardly extracts global information. Under this configure, only local features can be extracted in each stage of CNNs based model. However, both local and global features are crucial for classification tasks, especially for benign-malignant nodules classification that requires the whole 3D information. Second, effective modeling of long-range dependencies among pixels and making the model pay more attention to the region of interest is essential to capture global and significant contextual information. Therefore, it is still important to improve nodules learning networks to efficiently learn nodules information and further improve the performance of evaluation.

2.2 Self-attention

Recently, the self-attention mechanism including Transformer [4] emerges as an active research area in the computer vision community and has shown its potential to be a viable alternative to CNNs in medical image analysis. For example, Dong et al. [3] proposed a new image polyp segmentation framework, named Polyp-PVT, which utilizes a pyramid vision transformer backbone as the encoder to explicitly extract more powerful and robust features. Zhang et al. [31] combined Transformers and CNNs in a parallel style and proposed a novel paralleling-branch architecture, where both global dependency and low-level spatial details can be efficiently captured in a much shallower manner. Wang et al. [25] proposed an fNIRS classification network based on Transformer, named fNIRS-T, which could explore the spatial-level and channel-level representation of fNIRS signals to improve data utilization and network representation capacity. Shi et al. [22] proposed a unified framework based solely on the attention mechanism for skeleton-based action recognition. The proposed model employed a novel decoupled spatial-temporal attention to emphasize the spatial/temporal variations and motion scales of the skeletal data, resulting in a more comprehensive understanding of human actions and gestures. Wang et al. [23] designed a novel framework Attention-based Suppression and Attention-based Enhancement Net to better distinguish different classes based on attention mechanism and weakly supervised learning for the fine-grained classification of bone marrow cells. Fang et al. [6] proposed a novel attention modulated network based on the baseline U-Net, and explores embedded spatial and channel attention modules for adaptively highlighting interdependent channel maps and focusing on more discriminant regions via investigating relevant feature association. Li et al. [12] proposed a parallel-connected residual channel attention network with less pa-

rameters and a shorter prediction time to enhance the representation ability for remote sensing image SR.

However, the existing self-attention module have huge memory consumption and high computational complexity when meeting data with large spatial sizes, especially for 3D medical images. And the relation between depth dimensions of 3D medical images is also ignored. Thus, effective modeling of long-range dependencies among pixels from both spatial and depth dimensions and making the model pay more attention to the region of interest can overcome the limitations of CNNs in capturing global and significant contextual information, and further improve the performance of the model.

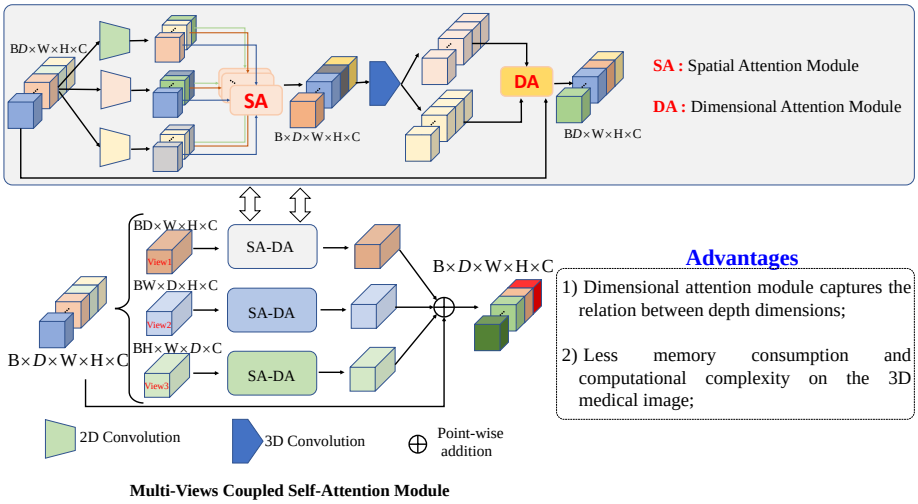

Fig. 2. Diagram of our proposed Multi-view Coupled Self-Attention Module. The proposed method which consists of two separable self-attention: 1) **Spatial Attention**; 2) **Dimensional Attention**.

3 Method

In this section, we first discuss Multi-View Coupled Self-Attention Module in detail, and then give an overview of our proposed pulmonary nodules classification framework.

3.1 Multi-View Coupled Self-Attention Module

Given the input data $X \in \mathbb{R}^{B\times D\times H\times W\times C}$ with a spatial resolution of $H \times W$, depth dimension of D (number of slices), C channels and B batch size. The vanilla self-attention maps X into query, key and value embeddings using

three $1 \times 1 \times 1$ convolutions, which are denoted as $X_q \in \mathbb{R}^{B\times D\times H\times W\times C'}$, $X_k \in \mathbb{R}^{B\times D\times H\times W\times C'}$ and $X_v \in \mathbb{R}^{B\times D\times H\times W\times C'}$. The three embeddings are then reshaped to the sizes of $DHW \times C'$, $C' \times DHW$ and $DHW \times C'$, respectively. Afterward, the similarity matrix $M \in \mathbb{R}^{DHW\times DHW}$, which models the long-distance dependency in a global space, is calculated by using $X_q \times X_k$. Finally, the attention map in each location is generated by normalized M through the softmax function.

As the computational complexity of the self-attention is quadratic with respect to the number of tokens. Although the typical self-attention can successfully model long range dependencies by directly flattening the input image to a sequence, which has high computational complexity. Additionally, this simple strategy makes self-attention unable to model the image local context information across spatial and depth dimensions. To address the above challenges, we carefully design a Multi-View Coupled Self-Attention Module (MVCS), which extracts a comprehensive representation of each volume from 2D three views. The main structure of our proposed MVCS is illustrated in Fig. 2, which consists of two separable self-attention:1) **Spatial Attention**; 2) **Dimensional Attention**, to exploit the correlations along the spatial and channel dimension, respectively. The details of the two separable self-attention are described as follows.

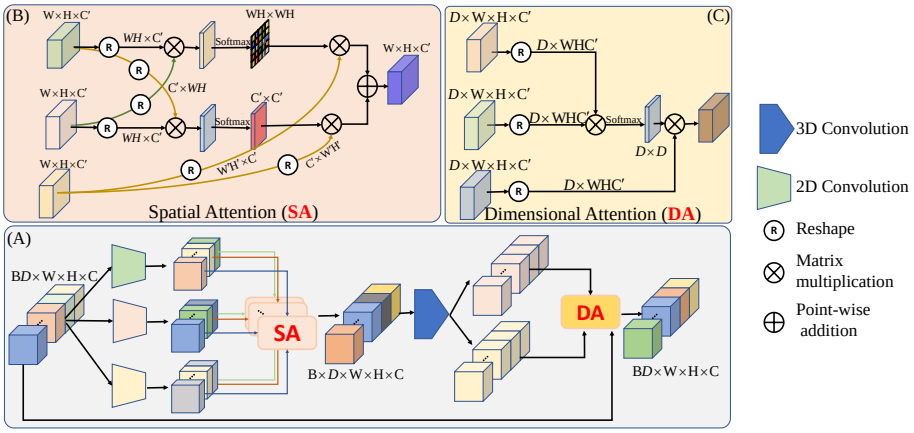

Fig. 3. The details of **Spatial Attention** and **Dimensional Attention**. **Spatial Attention** explores the dependencies along the spatial and channel dimension for computing position-wise attention and channel-wise attention. **Dimensional Attention** is attached after spatial attention, which builds the range correlations along the third dimension for exploiting the dimension correlations.

3.2 Spatial Attention

The input feature X is first converted to three view $X^0 \in \mathbb{R}^{BD\times \mathrm{W}\times \mathrm{H}\times \mathrm{C}}$, $X^1 \in \mathbb{R}^{BH\times \mathrm{W}\times D\times \mathrm{C}}$, $X^2 \in \mathbb{R}^{BW\times H\times D\times \mathrm{C}}$. Each view is mapped into spatial key, query, and value embeddings denoted as X_k^t, X_q^t, and X_v^t using 2D 1×1 convolutions, where t is view index. Then, the embeddings are used to generate spatial attention maps independently. Inside the spatial attention, both position-wise attention and channel-wise attention are computed as shown in Fig. 3(A). Given the embeddings X_q^t and X_k^t, which are first reshaped to the size of $HW \times C'$ and $C' \times HW$. The spatial similarity matrix $M_S^t \in \mathbb{R}^{HW\times HW}$ is generated by $X_q^t \times X_k^t$, which model long range dependencies from spatial view. The channel similarity matrix $M_C^t \in \mathbb{R}^{C'\times C'}$ is generated by $X_k^t \times X_q^t$, which explores the dependencies along the channel dimension. The spatial attention maps for view t are then calculated as:

$$X^t = soft\max(M_S^t) \times X_v^t + soft\max(M_C^t) \times X_v^t \tag{1}$$

3.3 Dimensional Attention

Dimensional attention is attached after spatial attention, which builds the range correlations along the third dimension. The structure of dimensional attention is illustrated in Fig. 3(B). Similar to spatial attention, the input feature X is first mapped into spatial key, query, and value embeddings denoted as $X_k \in \mathbb{R}^{B\times D\times W\times H\times C}$, $X_q \in \mathbb{R}^{B\times D\times W\times H\times C}$, and $X_v \in \mathbb{R}^{B\times D\times W\times H\times C}$ using $3 \times 1 \times 1$ convolution instead. The similarity matrix $M_D^t \in \mathbb{R}^{D\times D}$ along the third dimension is then calculated by reshaped X_q, X_k.

$$X^t = soft\max(M_D^t) \times X_v^t \tag{2}$$

Summary, the generated output feature can be described as:

$$\mathrm{X}=\sum_{t=0}^{2}(soft\max(M_S^t) + soft\max(M_C^t)+soft\max(M_D^t)) \times X_v^t \tag{3}$$

Summary of the advantages: 1)our proposed MVCS module solves the problem that the relation between depth dimensions be ignored by the dimensional attention module. 2)our proposed MVCS module breaks the limitation brought by the size of the receptive field and assists the model to extract both local and global information in each stage. 3)solving the challenges that huge memory consumption and computational complexity self-attention module needed when dealing with 3D medical images. 4) Last but not least, the proposed MVCS module is portable for other tasks models.

3.4 Pulmonary Nodules Classification

Following previous works [28, 21], we also employ a 3D CNN as the backbone in this work. We choose 3D ResNet framework as our baseline. Meanwhile, we

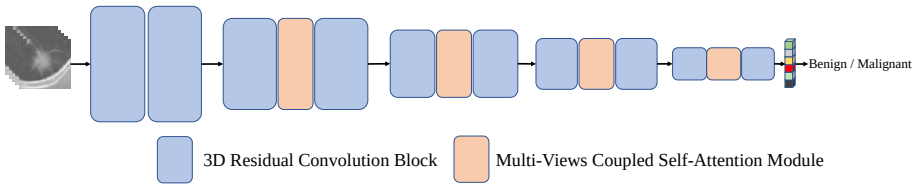

Fig. 4. Overview of multi-view coupled self-attention network for pulmonary nodules classification. The 3D ResNet framework as baseline.

insert Multi-View Coupled Self-Attention (MVCS) module into different layers to create our proposed Multi-View Coupled Self-Attention Network as shown in Fig. 4. In this architecture, MVCS module assists the model to capture both global and local contextual information in different stages to further improve the performance of the model. Remarkably, MVCS module can also be easily added in other 3D based architecture.

4 Experiment

4.1 Datasets and Implementation Details

To comprehensively evaluate the classification performance of our model, the LUNA16 dataset is employed in our experiments. Especially, LUNA16 dataset is a subset of LIDC-IDRI database from the Cancer Imaging Archive. Inside LUNA16 dataset, the CTs with slice thickness greater than 3mm, the annotated nodules of size smaller than 3mm, slice spacing inconsistent or missing slices from LIDC-IDRI dataset are removed, and explicitly gives the patient-level 10-fold cross validation split of the dataset. Finally, there are totally 1004 nodules left, in which 450 nodules are positive and 554 nodules are negative.

All the lung nodules were cropped from raw CT images for training and testing. Then, 3D nodule patch with size $32 \times 32 \times 32$ pixels was cropped from CT images around the centers of the lung nodules. Afterward, each 3D nodule patch is normalized by the z-score standardization method. The mean and std values are set as -400 and 750, respectively. During training, randomly adding Gaussian noise, horizontal flip, vertical flip, z-axis flip the data are utilized for data augmentation. We implement our framework by using Pytorch and two GTX 2080Ti GPUs. Adam optimizer with a minibatch size of 48 was applied for optimization. The learning rate and weight decay were set to 1e-4 and 0.01, respectively. Additionally, linear warmup with cosine annealing was also used for learning rate adjusting.

Following the settings in [9, 35], we also evaluate our method on folds 1–5 and the average performance of 5 folds as final results. The metrics of the AUC, Specificity, Sensitivity, Accuracy, Precision, and F1-score were calculated for comprehensively evaluating the classification performance of model. Sensitivity (Recall) denotes the percentage of correctly predicted malignant nodules and

is crucial for CAD; Accuracy evaluates the percentage of correctly predicted malignant/benign nodules; Precision is the percentage of correctly predicted benign nodules; F1-score evaluates the trade-off between Sensitivity and Precision. Defined as follows:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{4}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{5}$$

$$Specificity = \frac{TN}{FP + TN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$F1 - score = \frac{2TP}{2TP + FP + FN} \tag{8}$$

where True-Positive (TP) is the number of correctly predicted malignant nodules; False-Positive (FP) denotes the number of predicted malignant nodules that are actually benign; True-Negative (TN) represents the number of correctly predicted benign nodules; False-Negative (FN) is the number of predicted benign nodules that are actually malignant.

Table 1. Quantitative evaluation results of proposed models and other state-of-the-art methods on LUNA16 dataset.

	Accuracy[%]	Sensitivity[%]	Specificity[%]	Precision[%]	AUC[%]	F1-score[%]
HSCNN [16]	84.20	70.50	–	–	85.60	–
3D CNN [29]	87.40	89.40	85.20	–	94.70	–
Multi-crop CNN [19]	87.14	77.00	93.00	–	93.00	–
Local-Global [1]	88.46	88.66	–	87.38	95.62	88.37
Deep+visual features [27]	88.73	84.40	90.88	82.09	94.02	83.23
Dual Path Networks [9]	90.24	92.04	–	–	–	90.45
DeepLung [35]	90.44	81.42	–	–	–	–
NASLung [10]	90.77	85.37	95.04	–	–	89.29
Our	91.25	89.10	93.39	91.59	91.25	90.19

4.2 Experimental Results

Compared with other state-of-the-art methods: We compare our proposed model with state-of-the-art methods, including Multi-crop CNN [19], Vanilla 3D CNN [29], DeepLung [35], Ensemble 3D Dual Path Networks [9], and NASLung [10], where all models use the same dataset with the same number of samples.

The results of our proposed method and compared methods are shown in Table 1. As it can be seen from the Table 1, our proposed model achieves the highest Accuracy and Precision compared with other state-of-the-art methods, which represents our model possesses more powerful nodules representation learning capability and can classify benign and malignant features accurately. And MVCS could further boost the diagnosis accuracy and justify the diagnosis process. The results confirmed two aspects of feature learning for lung nodule classification. First, building the dependencies correlations along both spatial and depth dimensions could assist the model to extract the feature representation. Second, absorbing the global spatial contextual information could further improve the performance of the model.

Table 2. Quantitative evaluation results of 3D ResNet with various configures on LUNA16 dataset (the 5th fold).

	Accuracy[%]	Sensitivity[%]	Specificity[%]	Precision[%]	AUC[%]	F1-score[%]
Baseline	82.61	83.78	81.82	75.61	82.80	79.49
SANet	85.87	78.39	90.91	85.29	84.63	81.69
*VSANet	86.96	83.78	89.09	83.78	86.44	83.78
DANet	88.04	91.89	85.45	80.95	88.67	86.08
MVCSNet	92.39	94.59	90.91	87.50	92.75	90.91

Effectiveness analysis using MVCS: To demonstrate the effectiveness of the proposed MVCS, we analyze the influence of each part on the classification results by adding the split MVCS module inside the base network. Three other baseline methods and one vanilla self-attention based method are included as follows (Notably, in this section, we use the 5th fold as the testing dataset and all the other folds as the training dataset.):
(1) 3D ResNet (baseline): The classification is achieved by directly using 3D ResNet without attention mechanism. The classification results as a baseline.
(2) 3D ResNet + Spatial Attention (SANet): Different from (1), here the classification is acquired by inserting Spatial Attention into 3D ResNet.
(3) 3D ResNet + Dimensional Attention (DANet): Dimensional Attention is inserted into 3D ResNet for evaluating the effectiveness of Dimensional Attention.
(4) 3D ResNet + MVCS (MVCSNet) : Our proposed model.
(5) 3D ResNet + Vanilla Self-Attention (VSANet) : To compare the effectiveness of the self-attention module, we also insert the vanilla self-attention module which directly flattens the feature maps to model long-range interactions and spatial relationships into ResNet3D.

The Table 2 lists the classification performance of the methods described above. From the Table 2, we can notice that the innovations in our framework bring significant enhancements. The baseline (3D ResNet) has a poor performance. The major reason is that the typical convolution block construed by

stacked convolution layers cannot fully learn nodules' representation. The spatial attention module assists the baseline model to obtain a 3.26%, 9.09%, 9.68%, 1.83% and 2.20% improvement on Accuracy, Specificity, Precision, AUC and F1-Score, respectively. When utilizing dimensional attention, the Accuracy, Sensitivity, Specificity, Precision, AUC and F1-Score achieve 5.43%, 8.11%, 3.63%, 5.34%, 5.87% and 6.59%, respectively. Those improvements proved the effectiveness of the two attention modules in modeling long-range dependencies among pixels and capturing global and significant contextual information. What's more, when using these two modules in combination, the Accuracy, Sensitivity, Specificity, Precision, AUC and F1-Score are increased by 9.78%, 10.81%, 9.09%, 11.89%, 9.95% and 11.42% due to the multi-view information adaptive fusion. Above extensive experiments with promising results reveal the power of the MVCS module and its significance in improving the performance of the model.

Meanwhile, in this section, we also compare the effectiveness of the vanilla self-attention module with our proposed MVCS. From Tabel 2, we can notice that both the vanilla self-attention module and MVCS module could improve the performance of ResNet. Specifically, the vanilla self-attention module assists the baseline model to obtain a 4.35%, 7.27%, 8.17%, 3.64% and 4.29% improvement in Accuracy, Specificity, Precision, AUC and F1-Score, respectively. Compared with the vanilla self-attention module, MVCS obtains a higher improvement, 5.43%, 10.81%, 1.82%, 3.72%, 6.31% and 7.13% in Accuracy, Sensitivity, Specificity, Precision, AUC and F1-Score, respectively. And from the Tabel 2, we can notice that SANet obtains a similar performance to SANet, which proves that vanilla self-attention ignores the correlations along depth dimensions, and absorbs the global spatial contextual information from the depth dimension could improve the performance of the model. Additionally, those results confirmed three aspects of feature learning for lung nodule classification: 1) extracting the correlations along spatial and depth dimensions is advanced in nodule features learning. 2) the improvement over the vanilla self-attention module could be explained that exploiting the relation between depth dimensions is also significant. 3) less memory consumption and computational complexity are advanced in improving the performance of the model.

Memory consumption and computational complexity: In this section, we compare our proposed MVCS with the vanilla self-attention module in the memory consumption and computational complexity. Given an input data $X \in \mathbb{R}^{D\times H\times W\times C}$ with size $D \times H \times W \times C$, the per-layer complexity of the vanilla self-attention module is $O(D^2H^2W^2C)$ and the attention matric size is $DHW \times DHW$. The per-layer complexity of three views inside the MVCS is $O(D^2H^2C)$, $O(D^2W^2C)$, $O(H^2W^2C)$ and the attention matric size is $DH \times DH$, $DW \times DW$, and $HW \times HW$, respectively. We can notice that our proposed model significantly reduces memory consumption and computational complexity.

5 Conclusion

In this paper, we propose a Multi-View Coupled Self-Attention module to assists the CNNs based models to break the limitation brought by the size of the receptive field inside the convolutional layer. In specific, two types of self-attention mechanisms are designed to investigate the relationship between spatial attention and dimensional attention, and a view-complementary manner is proposed to model both local and global spatial contextual information. The proposed model solves two challenges in self-attention: 1) builds the dependencies merely by computing the correlations along spatial dimensions and ignoring the relation between depth dimensions; 2) huge memory consumption and high computational complexity in 3D medical image data. Additionally, our proposed module can be easily integrated with other frameworks. By adding the proposed module into 3D ReseNet, we build a nodule classification network for nodules' malignancy evaluation. The nodule classification network was validated on a public dataset LUNA16 from LIDC-IDRI. Extensive experimental results demonstrate that our proposed model has performance comparable to state-of-the-art approaches.

Acknowledgements This work was supported by the National Natural Science Foundation of China (81872510); Guangdong Provincial People's Hospital Young Talent Project (GDPPHYTP201902); High-level Hospital Construction Project (DFJH201801); GDPH Scientific Research Funds for Leading Medical Talents and Distinguished Young Scholars in Guangdong Province (No.KJ012019449); Guangdong Basic and Applied Basic Research Foundation (No.2019B1515130002).

References

1. Al-Shabi, M., Lan, B.L., Chan, W.Y., Ng, K.H., Tan, M.: Lung nodule classification using deep local–global networks. International journal of computer assisted radiology and surgery **14**(10), 1815–1819 (2019)
2. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A.: Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **68**(6), 394–424 (2018)
3. Dong, B., Wang, W., Fan, D.P., Li, J., Fu, H., Shao, L.: Polyp-pvt: Polyp segmentation with pyramid vision transformers. arXiv preprint arXiv:2108.06932 (2021)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. Du, Y., Yuan, C., Li, B., Zhao, L., Li, Y., Hu, W.: Interaction-aware spatio-temporal pyramid attention networks for action classification. In: Proceedings of the European conference on computer vision (ECCV). pp. 373–389 (2018)
6. Fang, W., Han, X.h.: Spatial and channel attention modulated network for medical image segmentation. In: Proceedings of the Asian Conference on Computer Vision (2020)

7. Guo, X., Guo, X., Lu, Y.: Ssan: Separable self-attention network for video representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12618–12627 (2021)
8. Hussein, S., Cao, K., Song, Q., Bagci, U.: Risk stratification of lung nodules using 3d cnn-based multi-task learning. In: International conference on information processing in medical imaging. pp. 249–260. Springer (2017)
9. Jiang, H., Gao, F., Xu, X., Huang, F., Zhu, S.: Attentive and ensemble 3d dual path networks for pulmonary nodules classification. Neurocomputing **398**, 422–430 (2020)
10. Jiang, H., Shen, F., Gao, F., Han, W.: Learning efficient, explainable and discriminative representations for pulmonary nodules classification. Pattern Recognition **113**, 107825 (2021)
11. Kumar, D., Wong, A., Clausi, D.A.: Lung nodule classification using deep features in ct images. In: 2015 12th Conference on Computer and Robot Vision. pp. 133–138. IEEE (2015)
12. Li, Y., Iwamoto, Y., Lin, L., Chen, Y.W.: Parallel-connected residual channel attention network for remote sensing image super-resolution. In: Proceedings of the Asian Conference on Computer Vision (2020)
13. Li, Z., Yuan, L., Xu, H., Cheng, R., Wen, X.: Deep multi-instance learning with induced self-attention for medical image classification. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 446–450. IEEE (2020)
14. Lyu, J., Ling, S.H.: Using multi-level convolutional neural network for classification of lung nodules on ct images. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 686–689. IEEE (2018)
15. Murugesan, M., Kaliannan, K., Balraj, S., Singaram, K., Kaliannan, T., Albert, J.R.: A hybrid deep learning model for effective segmentation and classification of lung nodules from ct images. Journal of Intelligent & Fuzzy Systems (Preprint), 1–13 (2022)
16. Shen, S., Han, S.X., Aberle, D.R., Bui, A.A., Hsu, W.: An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. Expert systems with applications **128**, 84–95 (2019)
17. Shen, W., Zhou, M., Yang, F., Dong, D., Yang, C., Zang, Y., Tian, J.: Learning from experts: Developing transferable deep features for patient-level lung cancer prediction. In: International conference on medical image computing and computer-assisted intervention. pp. 124–131. Springer (2016)
18. Shen, W., Zhou, M., Yang, F., Yang, C., Tian, J.: Multi-scale convolutional neural networks for lung nodule classification. In: International conference on information processing in medical imaging. pp. 588–599. Springer (2015)
19. Shen, W., Zhou, M., Yang, F., Yu, D., Dong, D., Yang, C., Zang, Y., Tian, J.: Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. Pattern Recognition **61**, 663–673 (2017)
20. Shi, F., Chen, B., Cao, Q., Wei, Y., Zhou, Q., Zhang, R., Zhou, Y., Yang, W., Wang, X., Fan, R., Yang, F., Chen, Y., Li, W., Gao, Y., Shen, D.: Semi-supervised deep transfer learning for benign-malignant diagnosis of pulmonary nodules in chest ct images. IEEE Transactions on Medical Imaging (2021). https://doi.org/10.1109/TMI.2021.3123572
21. Shi, F., Chen, B., Cao, Q., Wei, Y., Zhou, Q., Zhang, R., Zhou, Y., Yang, W., Wang, X., Fan, R., et al.: Semi-supervised deep transfer learning for benign-malignant diagnosis of pulmonary nodules in chest ct images. IEEE Transactions on Medical Imaging (2021)

22. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Decoupled spatial-temporal attention network for skeleton-based action recognition. Proceedings of the Asian Conference on Computer Vision (2020)
23. Wang, W., Guo, P., Li, L., Tan, Y., Shi, H., Wei, Y., Xu, X.: Attention-based fine-grained classification of bone marrow cells. In: Proceedings of the Asian Conference on Computer Vision (2020)
24. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794–7803 (2018)
25. Wang, Z., Zhang, J., Zhang, X., Chen, P., Wang, B.: Transformer model for functional near-infrared spectroscopy classification. IEEE Journal of Biomedical and Health Informatics **26**(6), 2559–2569 (2022). https://doi.org/10.1109/JBHI.2022.3140531
26. Xie, Y., Xia, Y., Zhang, J., Song, Y., Feng, D., Fulham, M., Cai, W.: Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest ct. IEEE transactions on medical imaging **38**(4), 991–1004 (2018)
27. Xie, Y., Zhang, J., Xia, Y., Fulham, M., Zhang, Y.: Fusing texture, shape and deep model-learned information at decision level for automated classification of lung nodules on chest ct. Information Fusion **42**, 102–110 (2018)
28. Xu, X., Wang, C., Guo, J., Gan, Y., Wang, J., Bai, H., Zhang, L., Li, W., Yi, Z.: Mscs-deepln: Evaluating lung nodule malignancy using multi-scale cost-sensitive neural networks. Medical Image Analysis **65**, 101772 (2020)
29. Yan, X., Pang, J., Qi, H., Zhu, Y., Bai, C., Geng, X., Liu, M., Terzopoulos, D., Ding, X.: Classification of lung nodule malignancy risk on computed tomography images using convolutional neural network: A comparison between 2d and 3d strategies. In: Asian Conference on Computer Vision. pp. 91–101. Springer (2016)
30. Zhang, J., Xie, Y., Xia, Y., Shen, C.: Attention residual learning for skin lesion classification. IEEE transactions on medical imaging **38**(9), 2092–2103 (2019)
31. Zhang, Y., Liu, H., Hu, Q.: Transfuse: Fusing transformers and cnns for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 14–24. Springer (2021)
32. Zhu, Q., Du, B., Yan, P.: Boundary-weighted domain adaptive neural network for prostate mr image segmentation. IEEE transactions on medical imaging **39**(3), 753–763 (2019)
33. Zhu, Q., Du, B., Yan, P.: Self-supervised training of graph convolutional networks. arXiv preprint arXiv:2006.02380 (2020)
34. Zhu, Q., Wang, Y., Du, B., Yan, P.: Oasis: One-pass aligned atlas set for medical image segmentation. Neurocomputing **470**, 130–138 (2022)
35. Zhu, W., Liu, C., Fan, W., Xie, X.: Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification. In: 2018 IEEE winter conference on applications of computer vision (WACV). pp. 673–681. IEEE (2018)