LUNG NODULE SEGMENTATION NETWORK WITH SELF-SUPERVISED LEARNING AND ATTENTION MECHANISMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Pulmonary nodule detection is one of the most important tasks for early lung cancer diagnosis. Especially, end-to-end methods for multi-tasking, including pulmonary nodule detection, false positive detection, and segmentation have been widely used based on supervised learning, leading to significant improvement in performance when detecting pulmonary nodules. However, those methods with confined environments were not able to exploit the representative features comprehensively. Therefore, some self-supervised methods have been proposed to handle the raw dataset. However, they were merely applied to each task, missing rich features of the end-to-end framework. In this paper, we propose a novel adaptation of self-supervised learning to a multi-tasking framework. Additionally, we employed other attention methods, such as Convolutional Block Attention Module(CBAM), and Quartet Attention Mechanism(QAM) to further enhance the performance without significantly increasing the number of parameters to learn.

1 INTRODUCTION

Among various types of cancers, lung cancer stands out as the most lethal disease with the highest mortality rate (Chhikara & Parang, 2023). Therefore, pulmonary nodule detection has gained significance as diagnosing pulmonary nodules in the early stage leads to effective treatment, resulting in a higher chance of patient survival. For detailed detection, computerized tomography (CT) was introduced to identify pulmonary nodules in 3D images.

However, some problems remained unsolved. First of all, detecting lung nodules from various patients puts the burden on radiologists with cumbersome tasks, delaying the overall procedures while deteriorating potential cancers of postponed patients. Secondly, the uncertainty of diagnosing with the naked eye not only decreased diagnostic precision but also slowed down the CT scanning procedure, potentially exposing patients to radiation from CT scans.

Therefore, lots of methods were introduced to take radiologists' hands off and enhance the performance of detection. Especially, to implement multiple tasks such as nodule screening, false positive detection, and segmentation, deep learning-based end-to-end frameworks were introduced and led the performance to a remarkable state. For example, Ronneberger et al. (2015), Setio et al. (2016), and Milletari et al. (2016) proposed evolving convolution networks while performing each of those tasks. However, most of the tasks were based on supervised learning, which essentially requires a labeled dataset. Instead of previously separate tasks, Tang et al. (2019) proposed an end-to-end network to perform detection, false positive detection, and segmentation at once. This provided a glimpse into the potential of utilizing vast amounts of unlabeled datasets for training deep learning models. However, since those methods mainly focused on a single task, they missed the advantages of end-to-end methods, leaving room for improvements.

In this paper, We propose an end-to-end deep learning model utilizing the full potential of the self-supervising method. To improve the currently proposed end-to-end frameworks, we also amend the model by adding various methods such as the Convolutional Block Attention Model(CBAM) or Quartet Attention Model(QAM) on our model. This turned out to improve the performance a lot.

2 RELATED METHODS

2.1 NODULENET

NoduleNet (Tang et al., 2019) is an end-to-end framework that takes CT images as inputs and finds the nodules' bounding boxes, possibility to exist, and segmentation masks. This framework integrates otherwise separately operating systems not only to reduce training resources but also to find optimal weights which understand pulmonary CT images thoroughly. In addition, to prevent independent tasks from sharing most of the features and failing to be optimized for each task, the model decouples each task by cropping features according to the bounding box.
Further details are illustrated at 1 and Tang et al. (2019). The U-net like FeatureNet is composed of preBlock and ResBlock3d, both of which include convolution 3d, batch normalization 3d, and Rectified Linear Unit(ReLU) as an activation function. Using the last feature of the network, the Region Proposal Network (RPN) is carried out. Six bounding box regression coefficients and one predicted possibility per nodule (the number of anchor boxes) are yielded, totaling $6k$ outputs, where k is the number of nodules.
Based on these outputs, Regions of Interest (ROI) are computed to conduct false positive reduction. However, instead of cropping the last feature map (feature map 4) on which RPN is implemented, the ROI is found from the middle (down 4 layer) to maintain the separately optimized model. Authors assert that utilization of such distant features prevents possible drawbacks of the multi-task model, the sub-optimization. With 3D ROI pooling applied, the final binary classification is trained on the RCNN head, which is composed of fully connected layers. Note since there are two variables indicating the possibility for nodules, one from RPN and the other from R-CNN head, the average of the two is used as the final possibility for detection in the evaluation stage.
Lastly, based on the $6k$ outputs of the aforementioned RPN, nodule segmentation is implemented. The region of proposals is calculated proportional to the size of features: feature map 4, down 2, and original CT image. These layers are concatenated in the following order with transposed convolution of previous features so that the feature sizes match. The final segmentation result is produced on the same scale as the original input.
NoduleNet successfully achieved State-of-the-Art(SOTA) performance with each targeted task in 2019. The Free Receiver Operating Characteristic curve (Kundel et al., 2008) was 70.82, 78.34, 85.68, 90.01, 94.25, 95.49, and 96.29 for each false positive rate per CT scan (FPs/scan) 0.125, 0.25, 0.5, 1, 2, 4, 8, with mean FROC score 87.27, which is the widely used competition performance metric(CPM) from LUNA16 competition (Setio et al., 2017). Also, the model achieved SOTA 71.85 score on Intersection Over Union(IOU) and 83.10 on the Sørensen-Dice coefficient, for nodules on which four radiologists had consensus. However, NoduleNet has some room for improvement. Although they successfully utilized the shared internal features, the external features that could be obtained from outside of the conventional model are not used, learning the representation of CT images itself.
Therefore we decided to feed this model with some rich information from self-supervised learning before training. Furthermore, we applied extensive ablation studies to achieve better performance.

2.2 MODELS GENESIS

Transfer learning in the medical domain has been restrictively applied since application to 3D images such as CT or Magnetic Resonance Imaging(MRI) was unavailable. To conduct pre-training from conventionally available 2D images, the model had to decompose 3D images and treat them as slices of 2D images, which essentially lost rich 3D information. Therefore, Zhou et al. (2021) proposed Generic Autodidactic Models for 3D Medical Image Analysis(Models Genesis) to solve the above limitation. The Models Genesis first takes some transformations on the given data. Then, it is trained to recover the impaired data to the unharmed data. During the restoration process, various features of the data such as appearance, texture, and context are learned by the model. Essentially, Models Genesis is composed of two parts, transformation and Model part.
In the transformation parts, two distortions and two painting transformations are applied to the input data. The two painting methods are non-linear transformation and pixel shuffling, while the other two painting methods are out-painting, and in-painting. For the model to learn the appearance of the medical images, a non-linear one-to-one function Bézier Curve is applied to the original voxels as the values of the voxels in the medical images indicate the shapes of the organs. Secondly, pixel-shuffling is performed to learn texture. Models Genesis kept the shuffling window smaller than the

receptive field of the original data. The authors state that through the re-ordering process, the model can learn the boundaries and texture. Lastly, out-painting and in-painting are applied. In the out-painting, the outskirts of the image are painted for recovery, whereas internal pixels are painted in the in-painting. Consecutively, a U-Net-like model was trained to recover the ragged input data to fit the original data. L1-norm was used to measure the loss.
It has been shown that the Models Genesis, although it did not outperform ImageNet when both were trained on 2D slices, indicated better performance when trained on a 3D basis. The authors conclude that the increment in the performance is due to two factors. Retaining rich 3D values as well as fully understanding representatives through self-supervised learning. Considering that pre-trained models such as ImageNet are not available on 3D data, Models Genesis has opened the possibility of self-supervised learning at the higher dimension. However, the scope of this model remained limited to a single framework. The lack of application to end-to-end framework lost the potential rich features. Therefore, we proposed alternative methods to apply Models Genesis to an end-to-end framework, which implements the following 3 tasks; nodule screening, false positive detection, and segmentation.

2.3 Convolutional Block Attention Module

Convolutional Block Attention Module (CBAM) (Woo et al., 2018) highlights the important features in terms of channel and spatial perspective consecutively.
In the original paper, from precedent input or hidden layer, which is stacked 2D images $F \in \mathbb{R}^{C\times H\times W}$, the model takes average and max pooling to each channel. The ReLU function is applied respectively. Then, the two pooling layers $F^c_{avg}, F^c_{max} \in \mathbb{R}^{C\times 1\times 1}$ are fed to the same Multi-layer Perceptron (MLP) and sigmoid function, and then added element-wisely.

$$M_c(F) = \sigma(MLP(F^{avg}_c)) \oplus \sigma(MLP(F^{max}_c)) \in \mathbb{R}^{C\times 1\times 1} \tag{1}$$

The σ denotes a sigmoid function and $\oplus$ is an element-wise addition. The attention map $M_c(F)$ is then broadcasted following the spatial dimension. The channel-wise attention model, therefore, basically shows the effect of focusing on the more important channels.
Then it sequentially takes spatial attention to illustrate more important spots of the input data. The average pooling and max pooling are taken again, but this time along the channel. Then $F^s_{avg}, F^s_{max} \in \mathbb{R}^{1\times H\times W}$ are fed into a convolution layer, preserving the size of the input.

$$M_s(F) = \sigma(Conv(Concat(F^s_{avg}, F^s_{max}))) \in \mathbb{R}^{1\times H\times W} \tag{2}$$

$Conv$ denotes convolution network with kernel size 7×7 and $Concat$ means concatenation alongside with channel dimension. The attention map $M_s(F)$ is also broadcasted to the channel and multiplied by the input data. This consequently takes attention to the important features.
Since this model takes element-wise multiplication of input data with channel-wise attention map and spatial-wise attention map respectively, thereby taking attention to more important features, it promises an increase in the performance of CNN blocks. Furthermore, since these attention modules are composed of only two layers of average pooling layer and max pooling layer, they accomplish such an enhancement only requiring a negligible amount of calculation.
However, as the LUNA16 dataset that we deal with is 3D images but not 2D, we properly modified the conventional CBAM to take 3D images as input. The Pooling2D layers have been modified into 3D, and so have the Convolutional layers so that spatial attention is modified to cubic spatial attention.

2.4 Quartet-Attention Module

Inspired by CBAM, the Triplet attention mechanism was developed (Misra et al., 2021). Although admitting that CBAM contributes to a significant increment in performance, the authors asserted that cross-dimensional interaction which CBAM missed contributes to understanding feature representations. They find the reason for the absence of interaction from the sequential process of the channel attention module and spatial attention module within CBAM. Although both modules explain what channels and where the channels are important, as they are computed separately, the model simply fails to understand the inter-dimensional relationship of the features.
Instead, the authors suggest a triplet-attention module to compute the dimensional relationships, that

is, the interaction between (C, W), (C, H), and (H, W). This is done by treating the features with dimension (C, W, H) as a cube. Leaving one path for the identical cube, authors rotate the cube in other paths with regard to H and W dimension with $90°$ anti-clockwise direction. Then for each (un)rotated tensors of dimension (C, W, H), (W, C, H), and (H, W, C), spatial attention module from CBAM is applied. That is, equation 1 is implemented to each tensor. Then formerly rotated 2 out of 3 output tensors $M_s(F)$ are rotated back with $90°$ clockwise direction to match the very same dimension of input tensors so that 3 tensors are in the same shape as the original input tensor (C, W, H). The final output is obtained by taking the element-wise average of the three output features.

The number of parameters CBAM and Triplet Attention Module required to calculate is as follows: $O(6k^2)$ for Triplet Attention module and $O(2C^2/r + 2k^2)$ for CBAM, where $k \ll C$. k is the size of the kernel in the Triplet Attention Module, while C indicates the number of channels of the input features. The Triplet Attention Module, notwithstanding that it has a much smaller number of parameters, has shown a promising increase in terms of feature representation which led to promising performance when attached to various networks such as ResNet, MobileNet, etc..

There were successful trials of employing the Triplet Attention Module to 3D image data (Zhong et al., 2023; Hong et al., 2021), which we call Quartet Attention Module(QAM). Instead of figuring interactions of 3 dimensions (between C, W, H), those trials trained interactions between 4 dimensions (between C, W, H, D). Rotating the 4-dimensional hypercube was achieved by not transposing one and transposing the other three $((C, W), (C, H), (C, D))$ cases. Also, 2D convolution layers and batch normalization were substituted with 3D convolution layers and batch normalization. Since the CT data that we are dealing with is 3D data, we also applied QAM instead of the Triplet Attention Module to fully facilitate cross-dimensional information.

3 Proposed Methods

To fully exploit the advantages of multi-task learning, attention network, and self-supervised learning simultaneously, we propose a Lung Nodule Segmentation Network with Self-Supervised Learning and Attention Mechanisms. The idea is that through aforementioned self-supervised learning method, Models Genesis, we gain initial weights to be adjusted to the downstream tasks. That is, from the 'CT Image' in figure 1, which is distorted image $\tilde{X}$, we reconstruct feature_map_0, $\tilde{X}'$, and calculate loss with the original CT image X.

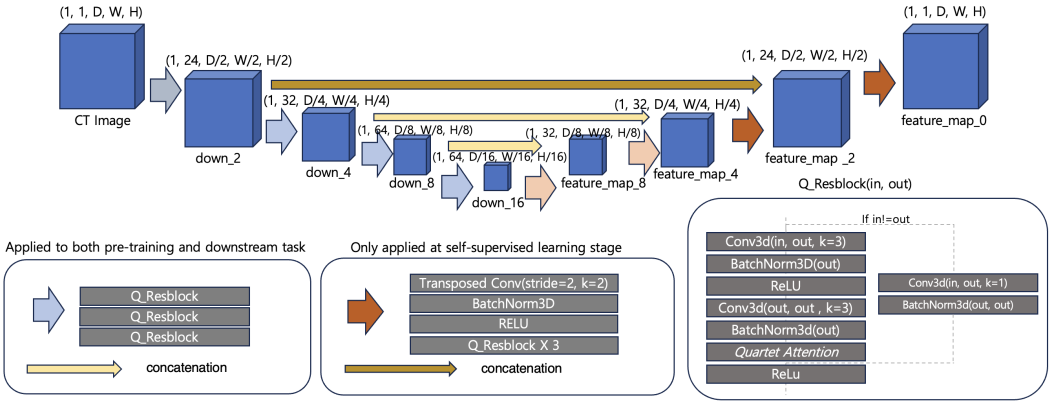

Figure 1: The structure of QAM attached FeatureNet, with extra layers for self-supervised learning.

Note that the procedure from NoduleNet includes cropping ROI to implement false positive reduction and segmentation tasks, which is untrainable in the self-supervised learning stage since corresponding labels for RCNN are not available yet. Therefore, we only take the "FeatureNet" part of our downstream model, which does not require labels for cropping. Then, we transfer those weights to train downstream tasks. The unspecified details of the downstream task can be found at Tang et al. (2019)

After the self-supervised training procedure, the (modified) NoduleNet was implemented based on the pre-trained weight. Although NoduleNet successfully decoupled the features to avoid sub-

Table 1: The sensitivity for each false positive rate per CT scan (FPs/scan) and CPM (The average of each FPs/scan). BASE: NoduleNet, QAM: QAM applied to NoduleNet, and QAM+MG: QAM applied to NoduleNet, trained based on the self-supervised training model, Models Genesis.

Sensitivity	**0.125**	**0.25**	**0.5**	**1**	**2**	**4**	**8**	**avg.**
BASE	0.590	0.662	738	0.784	0.856	0.905	0.941	0.782
QAM	0.652	0.723	0.806	0.861	0.901	0.925	0.957	0.832
QAM+MG	**0.656**	**0.743**	**0.818**	**0.862**	**0.950**	**0.957**	**0.968**	**0.846**

optimal model, it doesn't contain any attention mechanism, thereby missing the information of which channel and which spot to focus on, while this could have increased the performance of the model. Therefore, at each of the ResBlock in the NoduleNet, we added the model with some attention mechanism, i.e., QAM.

4 EXPERIMENTS

4.1 DATASET

To train the self-supervised model and our downstream task, we utilized Lung Nodule Analysis 16 (LUNA16) dataset (Setio et al., 2017), which is part of The Lung Image Database Consortium and Image Database Resource Initiative(LIDC-IDRI) (Armato III et al., 2011). LIDC/IDRI contains 1018 pulmonary CT images from 1010 patients, which are fully annotated by four experienced radiologists. LUNA16 dataset is the subset of the LIDC which contains only 888 cases, removing nodules thicker than 2.5mm. At pre-training stage with Models Genesis, we cropped the input size of $64 \times 64 \times 32$, whereas those CT images were cropped into the size of $128 \times 128 \times 128$ to be trained at the downstream task. During the training and validation stages, we didn't discriminate the dataset between the dataset for self-supervised learning with those for down-stream tasks, as the risk of data labels were not utilized at the self-supervised learning stage.

4.2 EXPERIMENTAL SETTINGS

There have been 3 other paths, NoduleNet (baseline), NoduleNet with QAM (Zhong et al., 2023; Hong et al., 2021), and NoduleNet with QAM and Models Genesis. Especially, QAM was added in the Residual Block of the FeatureNet, so as for each residual blocks to focus on what features are more important than those which is not.
While same procedures of Zhou et al. (2021) was implemented, note "FeatureNet" of modified NoduleNet ends with weight smaller than its original size. Therefore, when training the self-supervised pre-training model, we added some more transposed convolutional layers as well as output transition block, where transposed convolutional layers are as same as those at decoder so that consistent weights can be obtained. We utilized L1-norm distance which is the same loss function from Models Genesis. (Zhou et al., 2021).
For the downstream task, the loss function is as follows:

$$L_{total} = L_{rpn} + L_{rcnn} + L_{mask} \tag{3}$$

where L_{rpn} takes a probability and 6 coefficient of z, y, x coordinates, depth, height and width for bounding box regression as the inputs, which is the same loss function from Ren et al. (2016). Equivalently, L_{rcnn} also employed same function as L_{rpn}. Lastly, L_{mask} indicates a soft dice loss between predicted semantic segmentation and labeled data. The validation is implemented on the same denoted loss, while evaluation CPM of LUNA16 competition (Setio et al., 2017) based on FROC is used for test metric.

The final results are as follows 1. Attaching quartet attention model to conventional NoduleNet increased the performance significantly, and when it is given self-supervised pre-trained weight as initial weight, the performance was the highest.

5 CONCLUSION

In this research, we explored the possibility of self-supervised pre-trained model, Models Genesis, on the pulmonary lung detection to find out that self-supervised model can be applied to integrated detection framework such as NoduleNet. This is significant as vast number of unlabeled pulmonary dataset can be utilized for enhancing performance and robustness to nodule detection. Furthermore, applying quartet attention methods on basic model, we could further increase the possibility of the model.

REFERENCES

Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011.

Bhupender S Chhikara and Keykavous Parang. Global cancer statistics 2022: the trends projection analysis. *Chemical Biology Letters*, 10(1):451–451, 2023.

Luminzi Hong, Risheng Wang, Tao Lei, Xiaogang Du, and Yong Wan. Qau-net: Quartet attention u-net for liver and liver-tumor segmentation. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2021.

HL Kundel, K Berbaum, DD Dorfman, D Gur, CE Metz, and RG Swensson. Receiver operating characteristic analysis in medical imaging. *ICRU Report*, 79(8):1, 2008.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, 2016. doi: 10.1109/3DV.2016.79.

Diganta Misra, Trikay Nalamada, Ajay Uppili Arasanipalai, and Qibin Hou. Rotate to attend: Convolutional triplet attention module. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3139–3148, 2021.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241. Springer, 2015.

Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Geert Litjens, Paul Gerke, Colin Jacobs, Sarah J Van Riel, Mathilde Marie Winkler Wille, Matiullah Naqibullah, Clara I Sánchez, and Bram Van Ginneken. Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks. *IEEE transactions on medical imaging*, 35(5):1160–1169, 2016.

Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, Moira SN Berens, Cas Van Den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. *Medical image analysis*, 42: 1–13, 2017.

Hao Tang, Chupeng Zhang, and Xiaohui Xie. Nodulenet: Decoupled false positive reductionfor pulmonary nodule detection and segmentationdecoupled false positive reductionfor pulmonary nodule detection and segmentation. *CoRR*, abs/1907.11320, 2019. URL `http://arxiv.org/abs/1907.11320`.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Liming Zhong, Zeli Chen, Hai Shu, Yikai Zheng, Yiwen Zhang, Yuankui Wu, Qianjin Feng, Yin Li, and Wei Yang. Qacl: Quartet attention aware closed-loop learning for abdominal mr-to-ct synthesis via simultaneous registration. *Medical Image Analysis*, 83:102692, 2023.

Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. Models genesis. *Medical image analysis*, 67:101840, 2021.