Semantic Attention guided multi-dimension information complementary network for medical image classification

Yixiang Huang
Aerospace Information Research Institute, Chinese Academy of Sciences

Lifu Zhang (zhanglf@radi.ac.cn)
Aerospace Information Research Institute, Chinese Academy of Sciences

Ruoxi Song
Aerospace Information Research Institute, Chinese Academy of Sciences

Research Article

Keywords: Biomedical image recognition, Deep learning, Self-attention, Multi-dimensional data

Posted Date: November 4th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2205040/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.
Semantic Attention guided multi-dimension information complementary network for medical image classification

Yixiang Huang\textsuperscript{1,2}, Lifu Zhang\textsuperscript{1*} and Ruoxi Song\textsuperscript{1}

\textsuperscript{1*}Aerospace Information Research Institute, Chinese Academy of Sciences, No.20 Datun Road, Chaoyang District, 100101, Beijing, China.

\textsuperscript{2}University of Chinese Academy of Sciences, No.3 Datun Road, Chaoyang District, 100101, Beijing, China.

*Corresponding author(s). E-mail(s): zhanglf@radi.ac.cn;

Abstract

Biomedical image analysis, such as tissue and disease classification by using small-scale data, is a challenging and meaningful problem for clinical diagnosis. In previous works, most classical methods are using 2D convolution to extract feature from a single direction without making information of z-axis into consideration and haven’t considered that information representations are different in every view direction. In this letter, an attention mechanism-based model is proposed to extract and utilize feature more efficiently under the consideration of various information representation from direction to direction. To reach this goal, the work mainly consists of three parts of characteristics. The first characteristic is a plug-and-play structure design with parallel feature extractor in the different axial directions of which is dedicated to solving the problem of losing spatial structure feature by traditional 2D convolution. The second characteristic is intensive skip connection blocks employed to optimize the phenomenon of insufficient transit of information flow which cause the deep layer to fail in learning biomedical attribute. The last step is a designed spatial attention module, which’s applied to capture the long-range dependencies, suppress irrelated information and enhance significant spatial feature globally. The evaluation on various dataset shows that proposed model can achieve a competitive performance with comparison to other existed methods and reach the best 85.18 accuracy on average.
1 Introduction

Medical images classification plays an important role in distinguishing tissue disease. However, conventional method of classification for tissue disease needs well-experienced doctors to read and analyze medical images which is costly for time. With the development of deep learning, methods using neural network have contributed an efficient approach to reach the purpose of time-saving tissue disease analysis, which makes accurate diagnosis, judgement of disease stages and early prediction automatic and practical [1].

In the past decade, deep learning methods for image analysis have expanded application scenarios from the natural domain to biomedical. At present, the convolution neural network (CNN) and attention mechanism-based models are major components. For example AlexNet [2] and ResNet [3], which are representative of CNN network, have been extensively applied in diverse computer vision tasks [4–7]. On the contrary, Transformer as the pioneering work of attention mechanism was first presented in 2017. It has successfully raised and applied self-attention module in natural language processing domain for the first time [8]. Inspired by its design for the sequence modeling and the long range dependencies constructing in kinds of tasks, numerous works have tried their best to adopt attention mechanism to extensive computer vision domain such as semantic segmentation [9–11], image classification [12–14], and object detection [15–17]. More recently in virtue of hierarchical feature maps, Swin Transformer with attention mechanism can be easily used for dense prediction, and a linear computational complexity makes it possible to be regarded as a general-purpose backbone for various vision tasks [18].

With reference to an interdisciplinary domain of medical image analysis, the models transferring to supervision learning in biomedical dataset could be unavailable due to a different convolution process procedure compatibility. The reason is that existing deep learning methods were mostly and initially designed for large-scale dataset in natural image domain such as ImageNet [19] and ADE20K [20]. Although urgently surged CNN-based and attention-based methods [21, 22] have been studied in this field and have achieved outstanding performance in medical segmentation challenge such as The Medical Segmentation Decathlon dataset as well [1], in the context of small-scale 3D medical image analysis, existing methods still encounter several drawbacks. The first drawback is existing models may lose biomedical structure information due to convolutional character. This is because the traditional 2D convolution mainly focuses on plane information and may lack the ability to aggregate z-axis information [22]. While under the circumstance of the dataset is anisotropic, the 3D convolution can cause performance to get worse [23]. Besides, medical images
have the special stereostructure compared to natural images, which contain different spatial information in every axis orientation. Therefore, it still remains unexploited how to gather information from different axis directions effectively. Moreover, models may meet inadequate training stability and low-level convolution fails of gathering information of long-range dependencies.

To address these drawbacks of existing methods, in this paper, a novel multiple direction attention network (MDANet) was proposed for a small-scale medical image dataset classification, namely MedMNIST v2 [24]. On the one hand, MDANet makes a complete consideration of structure information and exhausts axis direction features adequately. A flatten convolution is applied to address the problem of insufficient utilizing of z-axis information and the multiple route structure is designed to acquire direction features. On the other hand, model stability can be improved by more intensive skip connections which enhance information flow and feature reuse. As for insufficient resemblance extracted by low-level convolution, attention mechanism would benefit model by constructing global dependencies to provide pixel-level like internal feature resolution [25].

The contributions of this paper are: (i) Multiple direction spatial aggregation (MDSA) module that can aggregate feature from different orientations is designed to improve perception quality of the input cubic image data. The proposed MDSA enhances spatial feature expression by changing traditional 2D convolution to a flatten convolution. (ii) Spatial Attention (SA) pipeline as the core component of the MDSA is used to extract feature from every single axis orientation and constructed with intensive skip connection blocks and spatial attention mechanism. The former is responsible for hierarchical feature representation and stability of feature transmission. As the complementary, spatial attention mechanism is settled to capture long range dependencies. (iii) Proposed method shows state-of-the-art results in medical images classification performance. Extensive experimental results prove the outstanding achievements of proposed MDANet.

2 Related Works

2.1 Deep Learning in Biomedical Image Analysis

In the domain of biomedical image analysis, deep learning technology has been mainly adopted to medical image segmentation for a relative long time [26, 27]. U-Net was proposed by Olaf Ronneberger et al. to implement end-to-end training schedule from very few images [28]. It consists of a contracting path (left side) and an expansive path (right side) to accomplish the aim of medical image segmentation. HUMUS-Net, which is applied to MRI reconstruction, combines the efficiency and beneficial implicit bias of convolutional networks with the representations of Transformers and their capability to capture long-range pixel dependencies [29]. With the purpose of automatically searching high-performance networks for medical image segmentation. Dong
Yang et al. developed a neural architecture search called DiNTS, which provides a characteristic to support fast gradient-based search within a highly flexible network topology search space [21]. However, methods above still fail to extract abundant features from the concerning that medical image itself behaves different information in every axis observing orientation, where the way of using multiple feature extractor in biomedical classification task remains under-investigated.

2.2 Self-Attention Mechanism

In the domain of natural language processing, attention mechanism is firstly adopted in sequence-to-sequence modeling and able to aggregate entire information from input sequence [8]. After transferring to computer vision domain, attention mechanism-based architecture is also designed to capture long range interactions in natural images analysis [14, 16, 18, 30]. ViT [31] is the prominent figure that makes attention module bring into play. It flattens image patches to a series of sequences and achieves SOTA classification performance on the ILSVRC-2012 ImageNet dataset. Benefit from the transformer encoder-decoder architecture, Nicolas Carion et al. successfully brings attention mechanism into DETR model to explain the relation between detected object and global context to parallel predict output of the final set by a given fixed small set of learned object queries [30]. Ze Liu et al. came up with Swin Transformer V2 using a residual-post-norm method combined with cosine attention to improve training stability. It explored large-scale model in vision task and achieved more efficient performance on ADE20K semantic segmentation task [16]. Therefore, it’s a natural target solution to use attention mechanism to explore global spatial dependencies for boosting performance in biomedical images.

3 MDANet

3.1 Overall Architecture

The overall model architecture and internal module design of proposed MDANet are shown in Figure 1. DenseNet is chosen as the baseline to construct the proposed MDANet, which is appropriate to show multiple direction spatial aggregation (MDSA) module superior performance.

At the stage of data input, cubic medical image $x \in R^{c \times h \times w}$ can be parallelly copied and rotated to three duplicates $x_1, x_2, x_3$ which have different orientation to the next Spatial Attention (SA) Pipeline module. After every duplicate has been prepared, reshape operation is set to make input data change to a flatten shape. Then the data preparing from different direction is formulated as shown below:

$$y_i = flatten(rotate(x)), i = 0, 1, 2,$$ (1)
Fig. 1. The overall model architecture. Multiple direction spatial aggregation (MDSA) module includes 3 parallel spatial attention pipelines (SA) after a flatten input. The output from every SA pipeline is concatenated before the convenient fusion block (CFB).

where $y_i$ denotes the parallely processed data waiting to be input into SA pipeline for feature extraction. SA pipeline is responsible for multiple direction feature extraction, which comes to the output feature map $y'_i, i = 0, 1, 2$. After that, these feature map will be aggregated to a fusion feature map by concatenation and the convenient fusion block (CFB) operation for later dense blocks as shown in Figure 1. The process of CFB can be formulated as follows:

$$x_{CBF}^{in} = F_{sigmoid}(F_{BN}(F_{cat}(y'_0, y'_1, y'_2))), i = 0, 1, 2,$$  \(2\)

$$y_{CBF}^{in} = F_{MP}(F_{ReLU}(F_{BN}(F_{7 \times 7\, conv}(x_{CBF}^{in}))), i = 0, 1, 2,$$  \(3\)

where $F_{BN}(\cdot)$ denotes batch normalization operation, $F_{ReLU}(\cdot)$ denotes linear rectification function which is an effective activation function to help decrease computational cost and the stereotyped $F_{MP}(\cdot)$ indicates max pooling operation used for size reduction of signal.

In order to prove the effective and efficient influence of the proposed MDSA module for gathering features from different orientation, it is technically inserted to the part before stacked dense blocks to aggregate shallow feature and hierarchical feature from original data more efficiently. The $F_{DB}(\cdot)$ defines the feature extractor function inspired from dense block. The stacked dense blocks procedure is shown as follow:

$$y'_{out} = F_{DB}^{n-1}(F_{DB}^{n-2}(\cdots(F_{DB}^{1}(F_{DB}^{0}(y_{out}))))).$$  \(4\)

Through all these functional process, hierarchical feature and various direction feature can be abundantly extracted and finally get the classification prediction by using full connected layer as:

$$y_{pred} = F_{FC}(y'_{out})$$  \(5\)

The MDSA module can make proposed model focus on different spatial information which is easily ignored by single route-like linear process units. At the same time, skip connection provides a chance for model to have feature information gathered from MDSA module passed into a deeper dense block feature extractor unit, which’s also able to promote model stability. More details of SA Pipeline are introduced in the following.
3.2 Spatial Attention Pipeline

Spatial attention (SA) pipeline plays an essential part in spatial domain feature learning. The overall structure of SA pipeline is as shown in Figure 2. SA pipeline is constructed with intensive skip connection blocks and spatial attention (SA) module shown in the front part of Figure 2. Here, spatial attention provides model an opportunity to exploit the global spatial correlation in the flatten cubic feature map.

**Fig. 2** Spatial Attention (SA) Pipeline. The dotted line in the left part of SA pipeline indicates intensive skip connection benefitting to keep information flow stable, the construction of bottleneck layer is shown as up-right area which composed with a stack of convolution, bath normalization and ReLU activation function.

### 3.2.1 Intensive Skip Connection Blocks

In intensive skip connection blocks (ISCB), three bottleneck layers was applied before spatial attention module to make model have economical cost for computation while keeping similar time complexity [3]. Compared to residual connection which is raised first to address the problem of network recession, intensive skip connection can provide more fluent feature flow from shallow to deep layer without extra memory cost. Here these skip connections and training for aggregated feature are obliged to avoid the possible failure of feature transmit before the fusion step from all axis orientations.

**Bottleneck Layer.** Assume that the whole bottleneck layer operation is defined as $\beta(\cdot)$, which is the composition of $1 \times 1$ and $3 \times 3$ convolution layer (Conv), batch normalization layer (BN), and followed by ReLU activate function. The relationship of each output linked by skip connection can be formulaically described as:

$$x_l = \beta(F_{cat}(x_0, x_1, \cdots, x_{l-1})), l = 1, 2, 3,$$

where $x_l$ is the output feature produced from $l$-th layer of bottleneck, and $(x_0, x_1, \cdots, x_{l-1})$ indicates feature maps generated from specified bottleneck layer, which is expressed as a consequence-like stack linked by skip connection in Figure 3 as well.
Bottleneck Channel & Expansion. Here expansion refer to a hyper parameter $\epsilon$ used to adjust channel numbers of output feature map together with the bottleneck channel $\tau$. Assume $k_l$ refer to $l$-th output feature map channel from bottleneck layer, therefore channel number of specified layer can be represented as follow:

$$k_l = k_0 + l\epsilon\tau, l = 1, 2, 3,$$  \hspace{1cm} (7)

seeing the formulation, parameter $\epsilon$ in bottleneck layer is used as expansion coefficient number set as 2 by default to decide output channel. And parameter $\tau$ is used for determining the output channel of the first and second cell block in bottleneck layer. Later the influence by both 2 hyper parameters will be discussed in ablation study.

Transition. In order to avoid the unrestricted increasing of channel number and decrease the calculated quantities, before the later spatial attention module, a middle layer called transition layer is applied to squeeze the thickness of feature map. The construction of transition layer is a batch normalization layer for the last stacked bottleneck layer with intensive skip connection, followed by a ReLU activation function settled to guard against overfitting. After that, a convolution operation with $1 \times 1$ kernel size is used to compress thick input feature channel to a relative and acceptable thin feature channel which is defined as out features 3 by default SA pipeline. Transition procedure can be formulated as:

$$x_{\text{transition}} = F_{\text{conv}}^{1 \times 1}(F_{\text{relu}}(F_{\text{BN}}(x_t))),$$ \hspace{1cm} (8)

where $x_t$ indicates input feature map from intensive connected bottleneck stacks, $x_{\text{transition}}$ denotes the compassed feature map by transition layer.

3.2.2 Spatial Attention

Since convolution operation obtains less feature dependencies due to a restricted receptive field, only applied multiple convolutional layers is capable of capturing long range dependencies. By adding spatial self-attention module, the introduced MDANet can seize long range resemblance from a whole spatial map. Here spatial attention module is settled in every SA pipeline to achieve global spatial dependencies from different view. The structure of SA module is shown at Figure 3. It’s quite different to design and implement spatial attention module from various works. However, the basic idea remains unchanged that researchers try to find a method to acquire an attention map for weights distributing which is the core part to extract and enhance feature expression efficiently. The availability of SA module is also evaluated to estimate whether it works for the classification task.

With seeing Figure 3, for the input $x_l \in \mathbb{R}^{c' \times h \times w}$ before proposed spatial attention module, query vector was defined as $v_q$ which is used as a reassignment factor to generate attention map with key vector $v_k$. The generation of $v_q$ is processed by shape of $1 \times 1$ convolution which has channels of feature map.
Fig. 3  Spatial Attention. SA module follows the principle of self-attention to generate $v_k$ and $v_q$ parallelly. The green square indicates attention map corresponding with original input to obtain a new feature map colored in light orange.

Halved then followed by a reshape operation, $v_k$ is created and channel halved by 2D-convolution with $1 \times 1$ kernel size as well followed by adaptive average pooling to reduce resolution to size of $1 \times 1$. Finally, the SoftMax activation function is adopted to obtain the output of $v_k$. Along with created attention map $v_qk$, the global corresponding weights for every unit in vector $v_v$ will be applied so that the last output feature map from spatial attention module is able to be calculated by point-wise production. Mentioned operations are defined formulaically as follows:

$$v_q = F_R(F_{\text{conv}}^{1 \times 1}(x_l)) \in \mathbb{R}^{hw \times c'/2}, \quad (9)$$

$$v_k = F_{\text{softmax}}(F_{\text{AAP}}(F_{\text{conv}}^{1 \times 1}(x_l))) \in \mathbb{R}^{c'/2 \times 1}, \quad (10)$$

$$v_qk = F_R(v_q \odot v_k) \in \mathbb{R}^{h \times w}, \quad (11)$$

where $F_R(\cdot)$ denotes reshape operation and $F_{\text{AAP}}(\cdot)$ refer to adaptive average pooling to polymerize obtained feature map. The matrix multiplication operation is represented as $\odot$. As Figure 5 expresses, $x_l$ is re-calibrated with weight factor. Eventually the spatial attention feature $x_{\text{spatial}}$ is indicated as:

$$x_{\text{spatial}} = v_qk \odot x_l \in \mathbb{R}^{c' \times h \times w} \quad (12)$$

where $\odot$ implies point-wise production for every channel of value vector $x_l$ which means to apply the global dependencies separately and suppress irrelated information while enhancing intrinsic authentic spatial feature.

4 Experiments

4.1 Experimental Setup

Datasets. MedMNIST v2 is introduced into biomedical images classification experiments for the proposed method, which includes 12 datasets for 2D and 6 datasets for 3D. Due to the particularity and focalization of application problem, here 5 datasets for 3D data structure with solid spatial information are chosen to verify the effective performance of the contributed multiple direction
attention network. Every single sub-dataset of 3D MedMNIST v2 have different data samples (from 1300 to 1900) and diverse label classification (binary and multi-class). Tables 1 has shown the detail of the chosen 3D datasets below:

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Modality</th>
<th>Labels</th>
<th>Training/Validation/Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>OrganMNIST3D</td>
<td>Abdominal CT</td>
<td>11</td>
<td>972/161/610</td>
</tr>
<tr>
<td>NoduleMNIST3D</td>
<td>Chest CT</td>
<td>2</td>
<td>1158/165/310</td>
</tr>
<tr>
<td>AdrenalMNIST3D</td>
<td>Chest CT</td>
<td>3</td>
<td>1027/103/240</td>
</tr>
<tr>
<td>VesselMNIST3D</td>
<td>Brain MRA</td>
<td>2</td>
<td>1335/192/382</td>
</tr>
<tr>
<td>SynapseMNIST3D</td>
<td>EM</td>
<td>2</td>
<td>1230/177/352</td>
</tr>
</tbody>
</table>

The OrganMNIST3D acquired from abdominal computed tomography (CT) has the most 11 labels for classification which are body tissues of liver, left kidney, right kidney, right femur and so on; NoduleMNIST3D is the lung nodule dataset collected with LIDC-IDRI equipment, the label is set as benign or malignant; AdrenalMNIST3D consists of shape of masks from 1584 left and right adrenal glands collected by CT, the categories are normal or hyperplasia; VesselMNIST3D is a 3D intracranial aneurysm dataset collected by reconstructing MRA images, the data is labeled as vessel and aneurysm; SynapseMNIST3D is the 3D volume dataset to classify whether a synapse is excitatory or inhibitory which is acquired by a multi-beam scanning electron microscope.

**Implementation Details.** All experiments are performed on a NVIDIA RTX 3060 with 12G video memory. Due to the model structure, it processes $28 \times 28 \times 28$ standard data input to flatten duplicates in 3 spatial pipeline routes simultaneously. In training set procedure, batch size is chosen to set as 32 images, and learning rate $10^{-4}$ with corresponding $10^{-5}$ weight decay has been applied to exhibit a better convergence. Adam optimizer is used here to train the proposed MDANet. Cross entropy loss is chosen as the loss function for the model forward inference, the formulation is shown below:

$$L = \frac{1}{N} \sum_i L_i = -\frac{1}{N} \sum_i \sum_{c=1}^{M} y_{ic} \log (p_{ic})$$ (13)

where $M$ indicates the number of categories, $y_{ic}$ is an indicator function which the value is equal to 1 or 0 depending on whether it’s the true category of sample $i$. The prediction probability belonging to $c$ of sample $i$ is expressed as $p_{ic}$.

In hyper parameter setting part, initial expansion $\epsilon$ and bottleneck channel $\tau$ is placed as 2 and 1 respectively. After every SA pipeline, the output feature is aggregated by concatenation followed by batch normalization to address the change of data distribution [32].
4.2 Evaluation Metrics

For fair and convenient comparison between various labeled classification tasks, accuracy metric and area under curve metric are selected to evaluate and quantize model performance. Both indices are calculated based on confusion matrix and can reveal the overall performance. The larger the value, the better the performance.

**Accuracy (ACC).** In many aspects of classification tasks, accuracy is a relatively simple but effective metric to weigh model presentation. The formulation for accuracy calculation can be indicated as:

\[
\text{acc} = \frac{\sum_{i=0}^{c} TP_i}{\sum_{i=0}^{c} (TP_i + FP_i)}
\]  

(14)

**Area Under Curve (AUC).** AUC comes from ROC curve which uses different threshold value to decide if the prediction is positive so that true positive rate (TPR) and false positive rate (FPR) can be calculated at every specific threshold level. If threshold is set to enough fine and smooth, the group of TPRs and corresponding FPRs can be draw in a coordinate system with a curve which is the so-called ROC curve. Therefore, AUC indicates the area under the ROC curve.

4.3 Methods Comparison

For model comparison with the proposed MDANet, several conventional models were chosen here which have been proved to be effective for large scale image dataset. They are ResNet50 [3], ResNet50 with ACS convolution [33], GoogleNet [34], Inception v3 [35], AlexNet [36]. For fairer comparison, 3 state of the art models which achieved excellent performance on MNIST dataset have been chosen. They are SimpleNet [37], SpinalVGG [38], modelM3 [39]. The quantitative results on 5 sub-datasets of the MedMNISTv2 in terms of ACC and AUC are reported in Table 2. It’s easy to see that proposed MDANet outperforms other methods. Specifically, proposed method shows better performance than the best competitor by 2.99 ACC and 7.86 AUC on average. When compared to conventional SOTA ResNet, GoogleNet, and AlexNet, proposed method shows better 8.55, 3.36, and 12.42 respectively on average. In order to keep original structure of conventional model while making model be able to handle 3-D image data input directly, the axial-coronal-sagittal (ACS) convolution kernel is applied which is under consideration of the suggested operation from original dataset publication, ResNet with ACS convolution is also tested in local computer environment and showed less 7.84 accuracy and 24.1 area under curve than proposed MDANet. When making comparisons to these SOTA models in MNIST dataset which are proven to have ability of handling small-scale data input in actual fact, the proposed MDANet achieved 3.36, 9.7 and 13.54 accuracy performance gain for all cited algorithm respectively. Note that when process model comparison in multi-label dataset
OrganMNIST3D, accuracy shows the most exceeding that the best Inception v3 competitor about 5.74.

Since the ResNet classification model and others can not handle with 3D data input directly, for every model input who need conventional natural image size input, convolution and up-sampling operations are settled to change data shape available. When comes to traditional methods with deep and complicated model architecture. It can be intuitively categorized to 3 types, AlexNet with thought of deeper layers, ResNet with skip connections, Both GoogleNet and Inception v3 facilitated with multi-scale feature aggregation. When making a comparison with AlexNet and ResNet, it can be observed that the performance of AlexNet is not ideal in most situation. This result could result from losing information flow and skip connection becomes necessary. This could explain why MDANet with more intensive skip connections outperforms its sister model. When having MDANet compared to GoogleNet and Inception v3, it can be told that both are constructed for extract multi-scale feature but lacking sufficient information flow with various size of convolution kernel before a concatenation for feature combination. Corresponding to the feature fusion with different scale, the proposed MDANet concatenates feature representation from different axis orientation, model extracts hierarchical features and concatenates them for the next bottleneck layer which may be the reason leading to more efficient feature representation for deeper feature extraction layer. As for the last 3 models, SimpleNet, SpinalVGG and modelM3, they are designed and tested for the MNIST recognition task which can be naturally assumed to benefit small-scale data input from well-developed model structure. Among them, it could be noticed that modelM3 perform worst due to its plain convolution and batch normalization stacking probably which may be burdened from insufficient information transition. For the models of SpinalVGG and SimpleNet, both are applied gradual input or expansion to gain partial multi-scale feature gradually, it’s reasonable to believe learning from different scale can benefit model efficiently. The proposed model can accomplish it by skip connections with concatenation.

The parameters contained by every model are compared as well shown in Table 2. Compared to the 4 conventional SOTA methods, MDANet retain competitive economy for 6.98M parameters. ResNet, Inception v3 and especially AlexNet shows worse performance while having much more parameters which means these model structures and abundant network layers aren’t suitable for feature extraction in MedMNIST due to the gap from natural image to biomedical image. This results also indicates extra and unnecessary design may cause inefficient computational and memory cost. Although GoogleNet has less parameters, MDANet achieves more 3.36 ACC and 8.07 AUC results which make sacrifice acceptable. With the same reason, compared to those 3 methods designed for small-scale image tasks, MDANet achieves considerable progress with a slightly more model consuming which indicates it’s necessary to maintain a certain level complexity.
Table 2  The ACC (left entry in each cell) and AUC (right entry in each cell) results of model comparison on 5 datasets

<table>
<thead>
<tr>
<th>Model</th>
<th>Organ3D</th>
<th>Nodule3D</th>
<th>Adrenal3D</th>
<th>Vessel3D</th>
<th>Synapse3D</th>
<th>Avg.</th>
<th>Para.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet</td>
<td>65.24, 81.25</td>
<td>79.35, 57.52</td>
<td>76.84, 34.53</td>
<td>88.74, 56.44</td>
<td>73.01, 46.39</td>
<td>76.63, 55.22</td>
<td>23.52M</td>
</tr>
<tr>
<td>ResNet-ACS</td>
<td>73.44, 95.16</td>
<td>80.00, 81.17</td>
<td>76.84, 48.86</td>
<td>88.74, 35.75</td>
<td>70.17, 51.53</td>
<td>77.83, 62.49</td>
<td>25.61M</td>
</tr>
<tr>
<td>GoogleNet</td>
<td>83.27, 97.38</td>
<td>85.16, 83.13</td>
<td>79.19, 73.07</td>
<td>88.21, 74.53</td>
<td>73.29, 61.73</td>
<td>81.82, 77.96</td>
<td>5.61M</td>
</tr>
<tr>
<td>Inception v3</td>
<td>83.93, 96.95</td>
<td>86.45, 83.83</td>
<td>78.85, 80.74</td>
<td>88.74, 69.44</td>
<td>73.01, 59.93</td>
<td>82.19, 78.17</td>
<td>21.81M</td>
</tr>
<tr>
<td>AlexNet</td>
<td>45.90, 72.22</td>
<td>79.35, 50.00</td>
<td>76.84, 37.90</td>
<td>88.74, 30.79</td>
<td>73.01, 50.00</td>
<td>72.76, 48.18</td>
<td>57.05M</td>
</tr>
<tr>
<td>SimpleNet</td>
<td>82.95, 98.45</td>
<td>86.77, 90.14</td>
<td>78.52, 82.45</td>
<td>88.74, 63.14</td>
<td>72.15, 65.34</td>
<td>81.82, 79.90</td>
<td>5.50M</td>
</tr>
<tr>
<td>SpinalVGG</td>
<td>67.21, 93.99</td>
<td>81.93, 78.84</td>
<td>76.51, 69.57</td>
<td>85.07, 76.52</td>
<td>72.44, 50.86</td>
<td>76.63, 73.95</td>
<td>3.63M</td>
</tr>
<tr>
<td>modelM3</td>
<td>37.54, 81.81</td>
<td>84.83, 84.97</td>
<td>75.83, 63.04</td>
<td>89.00, 69.54</td>
<td>71.02, 53.76</td>
<td>71.64, 70.62</td>
<td>1.14M</td>
</tr>
<tr>
<td><strong>MDANet</strong>*</td>
<td><strong>89.67, 98.93</strong></td>
<td><strong>86.77, 85.99</strong></td>
<td><strong>81.54, 83.95</strong></td>
<td><strong>92.93, 90.11</strong></td>
<td><strong>75.00, 71.18</strong></td>
<td><strong>85.18, 86.03</strong></td>
<td><strong>6.98M</strong></td>
</tr>
</tbody>
</table>
4.4 Ablation Study

To verify the effectiveness of every design in MDANet and impact made by hyper parameters, several experiments of ablation study have been designed with careful consideration, which are (1) study of module effectivity, (2) the influence by the spatial attention module and the intensive skip connection, and (3) impact of hyper parameters. In the part of the validation of module effectivity, the experiment is designed to make comparison with MDANet and corresponding model baseline. For the part of analysis of influence by spatial attention module, experiments are launched to verify if module works in every SA pipeline by removing it or not. As for the hyper parameters, a set of parameter values are chosen by manual to evaluate the impact of them. More details are shown as follow.

**Module Effectivity.** To verify the significance of proposed MDSA module, ablation study has been designed and tested in the same 5 chosen sub-datasets as well. Compared to DenseNet baseline, proposed module successfully improves 3.75 accuracy on average. When it comes to Organ-MNIST3D dataset, the model performance is improved considerably from baseline 85.08 to 89.67 ACC while AUC indicator keeps slightly rising. A similar achievement is found in NoduleMNIST3D as well, which model illustrate 6.77 ACC gain by module with corresponding 5.23 AUC gain. For the rest of datasets, the betterment is relative mild and achieves 2.01, 0.26 and 5.12 ACC progress respectively. However, AUC shows more considerable results which are 6.19, 7.62 and 5.52 increasing compared to baseline. The detailed results of promotion are illustrated in Figure 5.

**Fig. 4** Comparison results of baseline, without SA and MDANet in 5 datasets. The blue line indicates ACC improvement from model to model and red line is referring to AUC improvement.

The MDSA module is the core component for proposed model. It is mainly regard as a construction with 3 characters. The first character is multi route to gain information from different axis orientation. It’s not the first thought by MDANet to take special data structure and abundant information of cubic
space into consideration. For example, ACS convolution is initially proposed to make existing models be able to deal with stereoscopic data input and model compatibility while avoiding changing model structure and may be comparatively full of reference significance. However, the implementation of ACS convolution is using heterogeneous 3D convolution kernel with one dimension be set as 1 and sagittal views. Different from this, proposed module is facilitated with flatten stereoscopic inputs which are from different axis presentation to address the insufficient feature aggregation in z-axis by ACS convolution. The second character is intensive skip connected bottleneck layers. Appropriate skip connections can increase feature reuse and information flexibility which have been proved empirically. As for bottleneck layer, the design of it follows rule of keeping feature map’s resolution unchanged and channel increasing by expansion coefficient. The last component is spatial attention module for long range dependencies expression and the irrelevant background information inhabitation. These ablation study results illustrate that proposed model with MDSA can learn feature from different axis directions effectively. For module analysis, it is settled before an original feature extraction backbone which collects information from shallow representation. The considerable improvement of model efficiency is influenced by the whole three joint characters.

**Spatial Attention Impact.** The effect of SA module is discussed by removing it or not. Since the SA module is parallelly applied into every pipeline, the operation of removing is set for them at the same time. Then the construction of SA pipeline would leave stacked bottleneck layer merely. As Fig. 4 also shows, models whether are optimized by SA module have been compared. When making comparison with model without SA and MDANet, on average, added SA module can bring 1.42 accuracy improvement for all 5 datasets. It can be acknowledged that SA module performs best on OrganMNIST3D which improved 2.3 accuracy with corresponding 0.38 AUC gain followed by achievements on NoduleMNIST3D illustrating 1.61 ACC and 0.48 AUC. By all these ablation study research, the effectiveness of SA module has been proved for certain.

**Intensive Skip Connection Impact.** After the removing operation, the model will be left with intensive skip connection blocks to extract feature in a hierarchical way. Therefore, it can be used to analyze the effects of intensive skip connection by making model without SA module compare with baseline. In most situations, the intensive skip connection would benefit model effectiveness, such as model obtains 2.29 accuracy improvement with corresponding 0.55 area under curve advancement. However, a phenomenon is observed that it fails to improve prediction performance while AUC keeps the trend of rising by 6.55 in VesselMNIST3D dataset.

**Hyper Parameter Impact.** Here the influence of both hyper parameters, which are expansion $\epsilon$ and bottleneck channel $\tau$, will be discussed. $\tau$ controls input channel number of bottleneck block and $\epsilon$ is responsible for expansion coefficient which determine the output channel number. The performance of
MDANet is analyzed when $\epsilon = 1, 2, 3, 4$ in Table 3. The bigger expansion brings worse biomedical image classification performance and it’s obviously to find that when $\epsilon = 2$ proposed model show its best performance which indicates 2 times of expansion is sufficient to transit feature, less or more would burden information flow or have feature redundant.

<table>
<thead>
<tr>
<th>Metric</th>
<th>$\epsilon = 1$</th>
<th>$\epsilon = 2$</th>
<th>$\epsilon = 3$</th>
<th>$\epsilon = 4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACC</td>
<td>86.72</td>
<td>89.67</td>
<td>84.26</td>
<td>80.98</td>
</tr>
<tr>
<td>AUC</td>
<td>98.5</td>
<td>98.93</td>
<td>97.72</td>
<td>97.66</td>
</tr>
</tbody>
</table>

As Table 4 shows, influence by $\tau = 1, 2, 4, 8$ has been tested. With the increasing of bottleneck channel from $\tau = 8$, the proposed MDANet demonstrates better performance. However, width of model may blindly lead to computational complexity. Fortunately, when $\tau = 1$ model shows the best performance which can effectively economize computational cost. This result demonstrates that for a flatten feature input it is suitable for information storage by setting channel of bottleneck as 1.

<table>
<thead>
<tr>
<th>Metric</th>
<th>$\tau = 1$</th>
<th>$\tau = 2$</th>
<th>$\tau = 4$</th>
<th>$\tau = 8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACC</td>
<td>89.67</td>
<td>83.93</td>
<td>84.59</td>
<td>86.22</td>
</tr>
<tr>
<td>AUC</td>
<td>98.93</td>
<td>97.43</td>
<td>97.81</td>
<td>98.52</td>
</tr>
</tbody>
</table>

5 Conclusion

In this work, a novel model called MDANet is proposed for biomedical image classification task with small-scale data structure. The proposed model can significantly utilize and aggregate information from stereoscope data shape. With extensive experiments on 5 different datasets and comparison with sufficient classification model, the usefulness of proposed approach has been confirmed.

This model considers information from axis direction that brings representation enhancing, intensive skip connections which stabilize information flow and attention mechanism benefitting from long range and global resemblance, which may lead to better results than single route extractor. In fact, its simplicity and plug-and-play design makes it even acceptable and applicable on the top of other models.

In the future, the aim of study would focus on segmentation of 3-Dimensional dataset and try to analyze if a plug-and-play module is enough
Semantic Attention Multi-dimension Information Network

to gather information for inputting to existent model directly. Moreover, proposed approach is learning from scratch, pre-training or proxy tasks from more extensive biomedical data is still uninvestigated.

Declarations

Ethical Approval. Not applicable.

Competing interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors’ contributions. Yixiang Huang wrote the main manuscript text and did experiments. Lifu Zhang and Rouxi Song made specifically critical review, commentary and revision.

Funding. This work is supported by the National Natural Science Foundation of China (grant no. 41830108).

Availability of data and materials. The dataset that support the findings of this study are available in the MedMNIST repository, https://medmnist.com.

References


Semantic Attention Multi-dimension Information Network


