# Cascaded and Dual: Discrimination Oriented Network for Brain Tumor Classification

**Wenxuan Zhang**[†] WEN_XUAN_ZHANG@163.COM
**Dong Zhang**[†] DONGZHANG@NJUST.EDU.CN
**Xinguang Xiang** XGXIANG@NJUST.EDU.CN
*Nanjing University of Science and Technology*

**Editors:** Wee Sun Lee and Taiji Suzuki

## Abstract

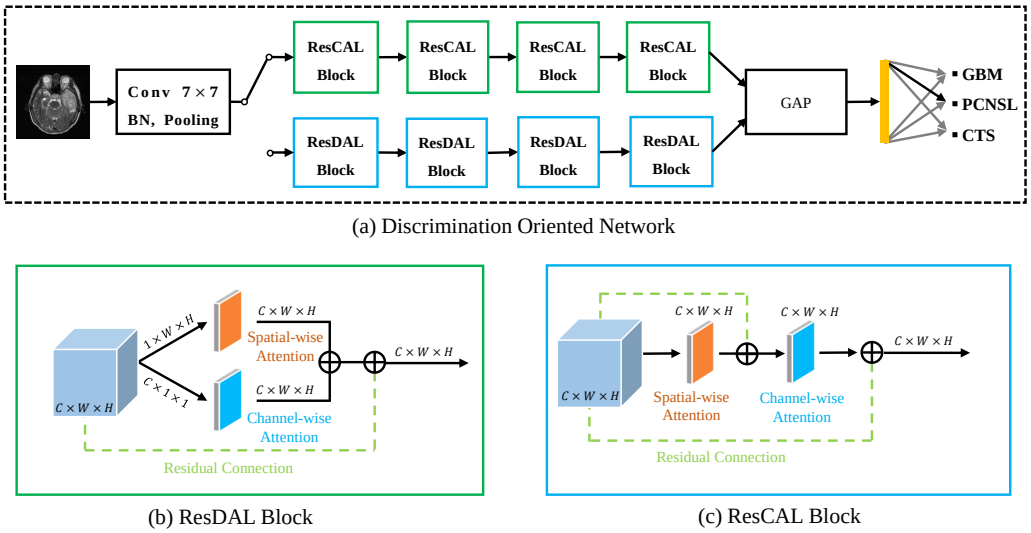
Medical image classification is one of the fundamental research topics in the domain of computer-aided diagnosis. Although existing classification models of the natural image can produce promising results using deep convolutional neural networks in some cases, it is difficult to guarantee that these models can generate promising performance for medical images. To bridge such a gap, we propose a novel medical image classification method for brain tumors in this paper, termed as Discrimination Oriented Network (DONet). Inspired by the attention learning mechanism of the human brain, we first propose two categories of attention learning modules, i.e., the Cascaded Attention Learning (CAL) and the Dual Attention Learning (DAL), which can learn the discrimination information in both the spatial-wise and the channel-wise dimensions in a fine-grained manner. By the CAL and the DAL, the attention information of different dimensions is calculated in a series manner (for cascaded) and a parallel manner (for dual), respectively. To demonstrate the superiority of our proposed modules, we implement the CAL and the DAL on the Deep Residual Network (ResNet) for brain tumor classification. Compared with the ResNet, experimental results show that the DONet has a significant improvement in accuracy. Moreover, compared with state-of-the-art classification methods, the DONet can also achieve better performance.

## 1. Introduction

The brain tumor is a common type of cerebral disease and the advanced one is prone to cancerization, which can usually cause high clinical mortality. Fortunately, a timely and individualized treatment plan can relieve patients' pain and prolong their life expectancy in practice Pandiselvi and Maheswaran (2019). However, it is time-consuming and painful for patients to classify the brain tumor via pathological analysis, such as biopsy or spinal tap Giulioni et al. (2019). To address this problem, the classification methods for brain tumors using the image processing technology with computer vision methods have attracted a lot of attention in recent years Vallée et al. (2018). Particularly, methods based on the Deep Convolutional Neural Networks (DCNNs) have made tremendous progress and outstanding performance.

Although some existing DCNNs methods have proved that the representative learning ability can be further improved with deeper networks within limits, the degradation problem

†. These authors contributed equally.




(a) Discrimination Oriented Network

(b) ResDAL Block

(c) ResCAL Block

Figure 1: Illustration of the Discrimination Oriented Network (DONet) for brain tumor classification. GAP, GBM, PCNSL and CTS means Global Average Pooling, Glioblastoma, Primary Central Nervous System Lymphoma, Common Tissues respectively.

may occur in these deeper networks suffering from the low convergence rate. To address this problem, the "residual connection" is proposed in the Deep Residual Network (ResNet) to skip some unimportant layers spontaneously without additional parameters He et al. (2016). ResNet is an effective network structure, and numerous works have demonstrated its preeminence. In recent years, inspired by the visual perception of the human brain, state-of-the-art methods demonstrated that the performance of DCNNs models can be improved by explicitly modeling the inter-dependencies between the channels or the spatial positions of their convolutional features Wang et al. (2017).

Because of the uncertainty of location, shape, and size of brain tumors, it is a challenging task to classify brain tumors with only the image information. Although many methods have been proposed in recent years Afshar et al. (2018), there are still three intractable problems: (1) Lack of training samples. Usually, there are only few annotated training images in the medical image dataset. Therefore, it is difficult for DCNNs models to achieve satisfactory performance since the scarce training samples cannot make the DCNNs models reach their optimization. (2) The between-class variation (BV) is small, while the within-class variation (WV) is large. The small BV implies that different categories of brain tumors usually have very similar appearances, while the large WV implies that even for the same category of brain tumors, their appearances are not always uniform. (3) Uncertainty of brain tumor status. Because of the extreme complexity of brain structure, the status of a brain tumor is also hypercomplex. There are usually problems of occlusion, interference and invalid information for the classification of brain tumors. To make matters worse, some pivotal information about the lesion may be obscured by normal tissue in the brain, making it difficult to classify tumors.

In this paper, we propose a novel framework for brain tumor classification, termed as the Discrimination Oriented Network (DONet), illustrated in Figure 1. Recalling that the attention learning mechanism has advantages of discriminative learning, we designed two attention learning modules to capture visual features dependencies in the channel and spatial dimensions respectively, i.e., the Cascaded Attention Learning(CAL) block and the Dual Attention Learning (DAL) block. Specifically, the channel attention learning module is used to capture the channel dependencies among different feature maps and assign a weight to each channel with a weighted summation of all channels. At the same time, the spatial attention learning module is used to capture spatial correlations among different coordinates in a specific feature map and assign a weight to each feature grid with the similar weighted summation of all coordinates. The CAL block is composed of the spatial-wise weighted feature maps and the channel-wise weighted feature maps in series, while the DAL block is formed by parallel connection of the spatial-wise weighted feature maps and the channel-wise weighted feature maps. Based on the proposed CAL and DAL block, we can build two different forms for DONet, which can quickly focus on certain spatial locations and channels, whose features have strong discriminative information inside. To demonstrate the superiority of DONet, experiments are carried out on our internal brain tumor dataset, which consists of two categories of common brain tumors and a category of common tissue with magnetic resonance imaging images.

Our main contributions can be summarized as follows: (1) Two novel attention learning blocks are proposed, i.e., CAL and DAL, to capture visual features dependencies in the channel and the spatial dimensions at the same time. (2) We design and implement a novel medical image classification framework with two forms, named as DONet. The learning blocks make the DONet have a strong discriminative learning ability, which is very suitable for the classification task of medical images. (3) To demonstrate the superiority of the DONet, we implement the classification of brain tumors on our internal dataset. Compared with the baseline, DONet can provide a significant improvement in both quantitative and qualitative metrics. Experimental results demonstrate the superiority of DONet.

## 2. Related Work

### 2.1. Medical Image Classification

Most current traditional pathological examinations make patients experience a long and painful diagnostic process for brain tumors Giulioni et al. (2019). To solve this problem, researchers combine medical imaging and image processing technology with traditional pathological diagnosis approaches. In recent years, magnetic resonance imaging images have been widely used for medical image classification Reddy et al. (2019). Recently researchers have proposed many brain tumor classification methods that make use of shallow features such as histogram of oriented gradients, local binary patterns, discrete wavelet transformation Shree and Kumar (2018). They employ classifiers such as support vector machine, random forest to classify brain tumors Shree and Kumar (2018). However, the expression ability of representations generated by classical feature extractors is limited, which makes traditional models unable to make full use of image information. Moreover, the computation cost of classical models increases with the number of data samples, which makes the computation cost beyond acceptance.

### 2.2. Deep Convolutional Neural Networks models

With the development of deep learning, Deep Convolutional Neural Networks (DCNNs) models have shown their potential in many domains, such as image classification, semantic segmentation, and object detection etc Bai et al. (2018). Researchers attempted to train networks with deep structures because previous studies have proven that the increasing depth can enhance the semantic information for the features extracted by networks. Thus, many DCNNs models such as Very Deep Convolutional Networks Simonyan and Zisserman (2014), GoogLeNet Szegedy et al. (2015) or Inceptions Szegedy et al. (2016) have been proposed in recent years. However, classical DCNNs models may suffer from the problems of low convergence rate which become the obstacle of deep networks. The degradation problem can be addressed by Deep Residual Network (ResNet) through the identity skip connections Wu et al. (2018). ResNet provides a stable backbone for many models and makes them achieve developments for many fields.

### 2.3. Self-attention Learning

Under the inspiration of the human visual system, self-attention mechanisms are proposed to learn a distribution of weights through DCNNs models themselves, which helps DCNNs models concentrate on discriminate features Cao et al. (2018). The self-attention mechanisms employed in DCNNs models mostly use feature maps with high semantic information in deep layers to guide the learning process of feature representations with low semantic information in shallow layers. Researchers mainly concentrate on self-attention mechanisms in spatial dimension Jaderberg et al. (2015) and channel dimension Hu et al. (2018). Spatial attention helps focus more attention on regions that play a crucial role in visual tasks, while channel attention is a process for semantic attribute selection.

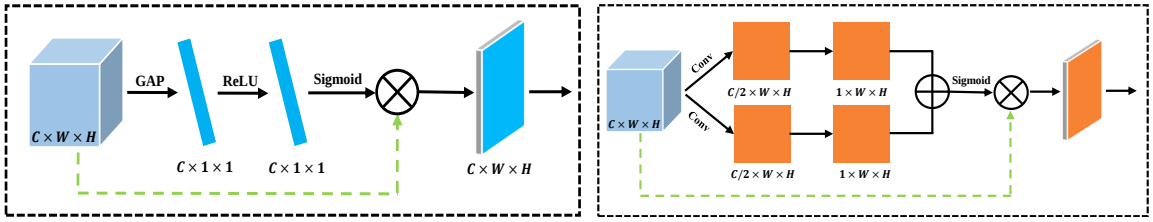
## 3. Our Approach

In this section, based on the channel-wise and the spatial-wise attention learning, we first design two kinds of attention learning blocks for Deep Convolutional Neural Networks (DCNNs), i.e., the Cascaded Attention Learning (CAL) block and the Dual Attention Learning (DAL) block. Then, on the basis of CAL block and DAL block, two different forms of brain tumor classification networks based on attention learning is established. Since CAL and DAL have advantages of discriminative learning, we name the proposed network as Discrimination Oriented Network (DONet).

### 3.1. Cascaded and Dual Attention Learning

To seek out important features with different channels and spatial coordinates at the same time, intuitively, we combine two different attention learning mechanisms. Suppose we have a bunch of deep learning feature maps $\mathbf{U} \in \mathbb{R}^{C\times H\times W}$ as $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_C]$ which are generated by a DCNNs model with an arbitrary input image, where $\mathbf{u}_i \in \mathbb{R}^{H\times W}$, $i = 1, 2, ..., C$ is the channel dimensions. Therefore, the attention weight can be generated by the attention learning function as:

$$\mathbf{W} = Att(\mathbf{U}). \tag{1}$$




(*a*) Channel-wise Attention Learning Module (*b*) Spatial-wise Attention Learning Module

Figure 2: Architectures of the used Channel-wise Attention Learning (CAL) Module and the Spatial-wise Attention Learning (SAL) Module.

Usually, the attention learning function contains two fundamental forms: channel-wise attention and spatial-wise attention. The channel-wise attention learning function is described as:

$$\mathbf{W}_{cha} = \sigma_1(FC_2(\delta(FC_1(GAP(\mathbf{U}), \mathbf{W}_1^c)), \mathbf{W}_2^c)), \tag{2}$$

where $\mathbf{W}_{cha} \in \mathbb{R}^{C\times 1\times 1}$ denotes the weight of channel-wise attention. $GAP$ denotes the Global Average Pooling, $FC$ denotes the Fully Connected layer, $\mathbf{W}_{1,2}^c$ denotes the parameters in channel-wise attention learning module. $\delta$ denotes the $ReLU$ function, $\sigma$ denotes the $Sigmoid$ function. Figure 2 (a) shows an illustration of the channel-wise attention learning.

$$\mathbf{W}_{spa}^1 = Conv_1^2(Conv_1^1(\mathbf{U}, \mathbf{W}_1^{s,1}), \mathbf{W}_1^{s,2}), \tag{3}$$

$$\mathbf{W}_{spa}^2 = Conv_2^2(Conv_2^1(\mathbf{U}, \mathbf{W}_2^{s,1}), \mathbf{W}_2^{s,2}), \tag{4}$$

$$\mathbf{W}_{spa} = \sigma_2(\mathbf{W}_{spa}^1 + \mathbf{W}_{spa}^2), \tag{5}$$

where $\mathbf{W}_{spa} \in \mathbb{R}^{1\times H\times W}$, $Conv$ denotes the convolutional operation, $\mathbf{W}_{1,2}^s$ denotes the parameters in the spatial-wise attention learning module. For the spatial-wise attention learning function $\mathbf{W}_{spa}$, a spatial-wise $Sigmoid$ function is used to find out the focal spatial coordinates. Figure 2 (b) gives an illustration of the spatial-wise attention learning module used in this work. Based on the attention learning weights generated by Eq.(2) and Eq.(5), we can obtain the weighted feature maps as:

$$\mathbf{U}_{spa} = \mathbf{U} \cdot \mathbf{W}_{spa}, \quad \mathbf{U}_{cha} = \mathbf{U} \cdot \mathbf{W}_{cha}, \tag{6}$$

where $\mathbf{U}_{spa} \in \mathbb{R}^{C\times H\times W}$ denotes the spatial-wise weighted features and $\mathbf{U}_{cha} \in \mathbb{R}^{C\times H\times W}$ denotes the channel-wise weighted features.

Based on Eq.(6), we further propose two forms of enhanced attention learning blocks, as shown in Figure 1 (b) and Figure 1 (c), i.e., the CAL block and the DAL block, which can weight the features in the spatial-wise and the channel-wise simultaneously. Specifically, the CAL block is composed of the spatial-wise attention learning and the channel-wise attention learning in series, while the DAL block is composed of parallel connection of the spatial-wise attention learning and the channel-wise attention learning. The CAL block $\mathbf{U}_{cal}$ is modeled as:

$$\mathbf{U}_{cal} = \beta\ \mathbf{U}_{cha}(\mathbf{U}_{tem}) + \mathbf{U}_{tem}, \tag{7}$$

where

$$\mathbf{U}_{tem} = \alpha\ \mathbf{U}_{spa} + \mathbf{U}. \tag{8}$$

The DAL block $\mathbf{U}_{dal}$ is modeled as:

$$\mathbf{U}_{dal} = \mathbf{W}^T(\alpha\ \mathbf{U}_{spa} + \beta\ \mathbf{U}_{cha}) + \mathbf{U}, \tag{9}$$

where $\mathbf{W}^T$ is the weight parameter to be learned, $\alpha$ and $\beta$ are two hyper-parameters in the range $[0, 1]$ and $\alpha + \beta = 1$. The spatial location information and the channel information of feature maps are taken into account in the CAL block and DAL block at the same time. Therefore, the discriminative ability can be enhanced in these two blocks.

### 3.2. Residual Cascaded and Dual Learning Block

CAL and DAL are two generalized attention learning blocks, which can be used in any deep learning model to enhance its discriminative learning ability. Recall that the Deep Residual Network (ResNet) He et al. (2016) has the advantage of small sample learning, it can be used to solve the problem of scant training samples in medical image classification. Therefore, we try to fuse the CAL and DAL blocks into the residual block for brain tumor classification. As shown in Figure 1 (b) and Figure 1 (c), for the DAL block, the residual concatenation is added into the sum of the two attention modules; for the CAL block, the residual concatenation is added into the result of the spatial-wise attention and the channel-wise attention, respectively. We name the basic residual block with the CAL block and DAL block as ResCAL block and ResDAL block, respectively. In the network implementation, the residual attention learning block is added into the output of the last convolution layer of each residual block from *Stage2* to *Stage5*. The weighted features are merged into the residual features as the output of the current block. Therefore, output with the CAL block $Out_{cal}$ can be expressed as:

$$\mathbf{Out}_{cal} = F(\mathbf{U}) + \mathbf{U}_{ResCal}, \tag{10}$$

where

$$\mathbf{U}_{ResCal} = \beta\ \mathbf{U}_{cha}(\mathbf{U}_{Restem}) + \mathbf{U}_{Restem}, \tag{11}$$

and

$$\mathbf{U}_{Restem} = \alpha\ \mathbf{U}_{spa} + F(\mathbf{U}) + \mathbf{U}, \tag{12}$$

where $F(\mathbf{U})$ is the residual features. Output with the DAL block $Out_{dal}$ is expressed as

$$\mathbf{Out}_{dal} = F(\mathbf{U}) + \mathbf{U}_{dal}. \tag{13}$$

### 3.3. Discrimination Oriented Network

With the ResCAL block and the ResDAL block, we can establish an attention-based image classification DCNNs model with any depth. In this paper, the ResNet18 and the ResNet50 are selected as two baselines for brain tumor classification, which are widely used in many deep learning tasks and their learning abilities have been fully recognized and verified. For

Table 1: Network architectures of ResNet18, ResCALNet18, ResDALNet18, ResNet50, ResCALNet50 and ResDALNet18. " + " means that this is a complete *Stage* rather than a commonplace convolution operation.

| Layer name | Output size | ResNet18 | ResCALNet18 | ResDALNet18 | ResNet50 | ResCALNet50 | ResDALNet50 |
|---|---|---|---|---|---|---|---|
| Conv1+ | $112 \times 112$ | Conv, $7 \times 7$, 64, stride 2 | | | | | |
| Conv2+ | $56 \times 56$ | Max pool, $3 \times 3$, stride 2 | | | | | |
| | | $\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix} \times 2$ | $\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \\ \textbf{CAL} \end{bmatrix} \times 2$ | $\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \\ \textbf{DAL} \end{bmatrix} \times 2$ | $\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$ | $\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \\ \textbf{CAL} \end{bmatrix} \times 3$ | $\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \\ \textbf{DAL} \end{bmatrix} \times 3$ |
| Conv3+ | $28 \times 28$ | $\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix} \times 2$ | $\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \\ \textbf{CAL} \end{bmatrix} \times 2$ | $\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \\ \textbf{DAL} \end{bmatrix} \times 2$ | $\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$ | $\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \\ \textbf{CAL} \end{bmatrix} \times 4$ | $\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \\ \textbf{DAL} \end{bmatrix} \times 4$ |
| Conv4+ | $14 \times 14$ | $\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix} \times 2$ | $\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \\ \textbf{CAL} \end{bmatrix} \times 2$ | $\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \\ \textbf{DAL} \end{bmatrix} \times 2$ | $\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 6$ | $\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \\ \textbf{CAL} \end{bmatrix} \times 6$ | $\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \\ \textbf{DAL} \end{bmatrix} \times 6$ |
| Conv5+ | $7 \times 7$ | $\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix} \times 2$ | $\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \\ \textbf{CAL} \end{bmatrix} \times 2$ | $\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \\ \textbf{DAL} \end{bmatrix} \times 2$ | $\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$ | $\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \\ \textbf{CAL} \end{bmatrix} \times 3$ | $\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \\ \textbf{DAL} \end{bmatrix} \times 3$ |
| GAP | $1 \times 1$ | Global Average Pooling | | | | | |
| Output | 3 | Fully Connected Layer | | | | | |

the sake of distinction, the ResNet that has the ResCAL block is named as ResCALNet, and the ResNet that has the ResDAL block is named as ResDALNet. Since the ResCALNet and the ResDALNet both have a strong discrimination learning ability, we call these two models as DONet.

The detailed network structure is shown in Table 1. ResNet18 and ResNet50 have a similar structure except for convolution kernel sizes in some blocks. Both networks start with a $7 \times 7$ convolution kernel, and they are then connected to a maximum pooling layer for down-sampling. After that, four consecutive *Stages*[1] are stacked together, with two blocks (2, 2, 2, 2) for each *Stage* in ResNet18 and three to six blocks (3, 4, 6, 3) for each *Stage* in ResNet50. Moreover, each block of the ResNet18 is composed of two consecutive $3 \times 3$ convolution layers and each block of the ResNet50 is composed of one $1 \times 1$ convolution layer, one $3 \times 3$ convolution layer and one $1 \times 1$ convolution layer. At last, a global average pooling layer and a fully connected layer are used for classification.

## 4. Experiments

In this section, we first briefly introduce the evaluation metrics and the experimental dataset. Then, we present the implementation details in the training stage. After that, experimental results under different settings are compared and analyzed. At last, to make the experimental results more intuitionistic, we visualize features of different attention learning methods.

1. We follow feature pyramid networks Lin et al. (2017) to define that layers producing feature maps with the same spatial size are in the same network *Stage*.

Compared with state-of-the-art image classification methods on brain tumors, the validity of the proposed Discrimination Oriented Network (DONet) is further confirmed.

### 4.1. Evaluation Metrics

We select the quantitative metric and the qualitative metric to evaluate the DONet. By the quantitative metric, we refer to the accuracy, the sensitivity, the specificity, and the Area Under the receiver operating characteristic Curve (AUC). Among these metrics, the accuracy is used to measure the ratio between the number of samples correctly predicted and the total number of samples, which is defined as:

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}, \tag{14}$$

where $TP$, $FN$, $TN$ and $FP$ denote the True Positive, the False Negative, the True Negative and the False Positive respectively. The sensitivity and the specificity are used to determine whether a patient has the brain tumor (actually a binary classification task) which are expressed as:

$$sensitivity = \frac{TP}{TP + FN}, \quad specificity = \frac{TN}{TN + FP}. \tag{15}$$

Besides, AUC is used as a composite metric to reflect the sensitivity and specificity with continuous variables, and each point on the AUC curve can reflect the sensitivity to the same signal stimulus. The definition of AUC is expressed as:

$$AUC = \int_0^1 t_{pr}(f_{pr}) df_{pr} = P(X_1 > X_0), \tag{16}$$

where $t_{pr}$ is the true positive rate, $f_{pr}$ is the false positive rate, $X_0$ and $X_1$ are confidence scores for the negative and the positive instance, respectively.

In addition to the quantitative evaluation metrics mentioned above, by the qualitative metric, we refer to the Classification Activation Mapping (CAM). As proposed in Zhou et al. (2016), we name the feature grid $(i, j)$ in the $C$-th channel from feature maps of the global average pooling layer as $f_c(i, j)$ and we define $\mathbb{M}_t$ as the CAM for class t, where each feature grid $\mathbb{M}_t(i, j)$ can be expressed as:

$$\mathbb{M}_t(i, j) = \sum_c \mathbf{W}_c^t f_c(i, j), \tag{17}$$

where $\mathbf{W}_c^t$ indicates the weight from the $C$-th channel to class $t$. The $\mathbb{M}_t(i, j)$ denotes contributions made by feature grid $(i, j)$ to classifying an image to class $t$. The final CAM can be obtained by up-sampling the current class activation maps with bilinear interpolation operations.

### 4.2. Dataset

Experiments are carried out on our internal brain tumor image dataset, which includes 215 individuals with the resting state Magnetic Resonance Imaging (MRI) images. The data collection began in 2008 and ended in 2018. A total of 10 years of clinical brain tumor

Table 2: Statistical properties of brain tumor dataset.

| | | GBM | PCNSL | CTS |
|---|---|---|---|---|
| Sex | Number | 87 | 63 | 74 |
| | Male | 58 | 34 | 30 |
| | Female | 29 | 29 | 44 |
| Age | 0-30 | 3 | 2 | 5 |
| | 30-60 | 53 | 37 | 47 |
| | $> 60$ | 31 | 24 | 22 |
| | Mean Age | 55.4 | 56.8 | 52.4 |

images were collected, including 80 Glioblastoma (GBM) individuals, 61 Primary Central Nervous System Lymphoma (PCNSL) individuals, and 74 Common Tissue (CTS) individuals. MRI images of each patient are composed of dozens of different modalities, including the T1 Weighted (T1), the T2 Weighted (T2), and the Fluid Attenuated Inversion Recovery (FLAIR), etc. The data of each modality is composed of images from three different angles: coronal view, sagittal view, and horizontal view. All images were investigated on a Magnetom Trio 3T (Siemens AG, Germany) scanner, collected in the Nanjing Brain Hospital, Jiangsu, China. Figure 3 shows some demo images in this dataset under different modalities with the horizontal view, in which the left two columns are the GBM individuals, the middle two columns are the PCNSL individuals and the right two columns are the CTS individuals. Each modality contains a series of images, and each image corresponds to a different slice of space in the brain. For the sake of display convenience, we only select some of the images as demos. All patients volunteered to participate in the study prior to data collection. To protect the patients' privacy, all individual information is hidden in MRI images. The collected parameters for three common modalities are as follows:

- T1: TR=300 ms, TE=2.7 ms, slice thickness=4 mm, 25 slices, flip angle=90, image size=$640 \times 640 \times 25$ px, voxel resolution=$0 : 313 \times 0 : 313 \times 4$ mm$^3$.
- T2: TR=6000 ms, TE=93 ms, slice thickness=5 mm, 20 slices, flip angle=90, image size=$640 \times 640 \times 20$ px, voxel resolution=$0 : 719 \times 0 : 719 \times 2$ mm$^3$.
- FLAIR: TR=8000 ms, TE=97 ms, slice thickness=5 mm, 20 slices, flip angle=150, image size=$640 \times 640 \times 20$ px, voxel resolution=$0 : 449 \times 0 : 449 \times 5$ mm$^3$.

The corresponding statistical properties of the investigated individuals are summarised in Table 2. From Table 2, it can be seen that the distribution of this dataset is scientific and there is no imbalance problem, which can be used as a standard experimental dataset. Before using brain tumor images to train the DONet, all original MRI images with Digital Imaging and Communications in Medicine (DICOM) format are transformed into $256 \times 256$-pixel two-dimensional images with jpg format by the RadiAnt DICOM software [2]. The images of each type of brain tumor are randomly divided into 70% training set, 10% validation set and 20% test set according to the upward integer number, i.e. the training set includes 56 GBM individuals, 42 PCNSL individuals and 52 CTS individuals; the validation set includes 8

2. https://www.radiantviewer.com/en/

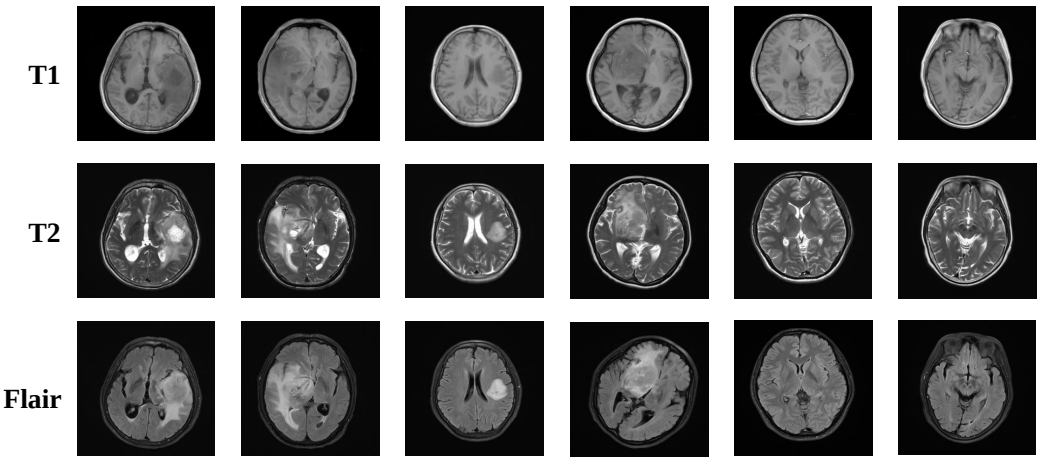



Figure 3: Demo samples of the three modalities with the horizontal view. Images of the same column belong to the same patient.

GBM individuals, 7 PCNSL individuals and 8 CTS individuals; the test set includes 16 GBM individuals, 12 PCNSL individuals, and 14 CTS individuals. In the following part, we will verify the classification performance of DONet on different modalities through comparative experiments.

### 4.3. Implementation Details

Our models are pre-trained on ImageNet Deng et al. (2009). Before training on brain tumor images, we first implement data augmentation as Zhang et al. (2019), which includes random rotation from $-15°$ to $15°$, random horizontal and vertical flip transformation, random scale (the scale factor is selected from [0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0]). After that, all images are resized into $224 \times 224$ with the bilinear interpolation as Sarkar et al. (2019). For the sake of simplicity, all data augmentation procedures are collectively referred to as Multi-Scale (MS) process in this paper. For optimization of ResNet18 and its extensions, we adopt the Adam optimizer and the initial learning rate is 0.00005, rate decay is set to 0.8 in every 5000 iterations. For optimization of ResNet50 and its extensions, we adopt the Nesterov momentum optimizer with momentum=0.9 and the initial learning rate is 0.001, rate decay is also set to 0.8 in every 5000 iterations. The batch size is set to 64 for ResNet18 and its extensions, while the batch size is 32 for ResNet50 and its extensions. The weight decay is 1e-4 and the maximum iteration is set to 160000. For each experiment, we run it for five times and report the mean and the Standard Deviation (STD) to ensure that our results are reliable.

### 4.4. Ablation Study

In the first experiment, we mainly explore the performance of ResCAL block, ResDAL block, MS on ResNet18 and ResNet50, respectively. $\alpha$ and $\beta$ are set to 1. Experiments are carried

Table 3: Classification performance of the ablation study.

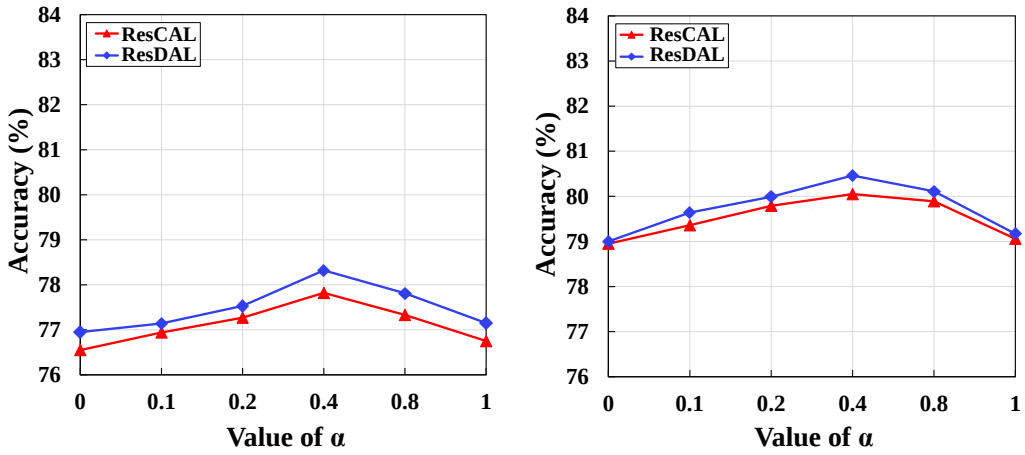
| Methods | ResCAL | ResDAL | MS | ResNet18 | | | | ResNet50 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | ACC | AUC | Sensitivity | Specificity | ACC | AUC | Sensitivity | Specificity |
| Baseline | | | | 75.0±1.92 | 84.2±2.50 | 78.4±3.52 | 79.0±3.91 | 76.9±3.48 | 85.2±1.04 | 79.7±2.95 | 80.3±3.02 |
| Baseline | | | ✓ | 76.2±2.41 | 85.9±3.58 | 80.0±0.92 | 80.2±3.63 | 78.4±1.69 | 87.9±0.57 | 81.0±2.01 | 82.5±2.27 |
| DONet | ✓ | | | 76.6±3.49 | 86.1±3.91 | 80.7±2.90 | 80.5±2.06 | 78.5±2.65 | 87.7±4.20 | 80.9±2.47 | 80.8±3.74 |
| DONet | | ✓ | | 76.8±3.03 | 86.5±3.70 | 81.2±3.10 | 81.0±0.98 | 78.7±1.46 | 87.8±2.51 | 80.4±2.52 | 81.8±3.95 |
| DONet | ✓ | | ✓ | 77.4±2.49 | 88.7±0.39 | 82.3±1.38 | 81.7±4.34 | 79.8±2.47 | 89.8±1.64 | 82.9±2.41 | 83.4±1.24 |
| DONet | | ✓ | ✓ | **77.7±2.94** | **89.1±3.26** | **82.4±4.15** | **82.1±1.53** | **80.1±1.67** | **90.1±2.21** | **83.6±2.67** | **83.7±1.89** |

out on three common modalities (T1, T2, FLAIR) at the same time, that is images of three modalities are mixed up without discrimination. ACC is used to measure the classification accuracy of three categories, i.e. the GBM, the PCNSL, and the CTS. AUC, Sensitivity, and Specificity are used to measure the condition of whether there is a brain tumor, i.e. the CTS and others. Experimental results with their STD on the test set are shown in Table 3. The results of the baseline are based on ResNet18 and ResNet50, respectively. From Table 3, we can draw the following conclusions: (1) Compared to ResNet18 and its extensions, ResNet50 and its extensions can achieve a better performance on four evaluation metrics, which implies that ResNet50 and its extensions can achieve a closer fit to the brain tumor images; (2) MS can obviously improve the performance, no matter for the baseline or the DONet. The improvement is 1.3% and 1.4% in the ResCALNet50 and ResDALNet50. It suggests that the input of multi-scale images plays an important role in medical image classification. (3) Compared to ResCALNet, ResDALNet can achieve a better performance both on the individual-scale and multi-scale images. There is a 1.7% improvement in accuracy with the multi-scale images for ResDALNet, which indicates that the ResDAL block is more suitable for the classification of brain tumors.

In the second experiment of the ablation study, we mainly explore the effect of $\alpha$ and $\beta$ on the classification performance with images of three modalities. The weight for the spatial attention branch $\alpha$ is set to [0, 0.1, 0.2, 0.4, 0.8, 1], respectively, and the corresponding weight for the channel attention branch $\beta$ is set to [1, 0.9, 0.8, 0.6, 0.2, 0]. When $\alpha = 0$, it implies that we only consider the channel attention in our model, vice versa. Experimental results of the DONet on the test set are shown in Figure 4. As we can see in Figure 4, the DONet can achieve the best results when $\alpha = 0.4$ and $\beta = 0.6$. Besides, compared with the ResCALNet, the ResDALNet can achieve better performance with the same $\alpha$ and $\beta$. Further, we can conclude that the performance of the classification network is improved by setting from $\alpha = 0$ to $\alpha = 1$. It indicates that the proposed attention learning model and the original ResNet are internally consistent, otherwise the performance of the network will decline to a certain extent.

### 4.5. Results on Different Modalities

MRI images with different modalities can not only reflect the human anatomy but also reflect physiological functions, such as blood flow or cell metabolism. To seek out which modality is more suitable for the classification of brain tumors, we survey classification performance for brain tumors of different modalities in this section. We divide brain tumor images set into 70% training set, 10% verification set and 20% test set according to different modalities, i.e. the common modalities of T1, T2, FLAIR. Except for experimental images, other




(a) DONet18 (b) DONet50

Figure 4: Effects of $\alpha$ on the experimental results with DONet18 and DONet50.

Table 4: Classification performance on different modalities.

| Methods | Basenet | T1 | T2 | FLAIR | ResCAL | | | | ResDAL | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | ACC | AUC | Sensitivity | Specificity | ACC | AUC | Sensitivity | Specificity |
| Baseline | ResNet18 | ✓ | ✓ | ✓ | 87.7±1.38 | 90.1±3.48 | 92.1±3.78 | 84.5±3.20 | 87.7±1.38 | 90.1±3.48 | 92.1±3.78 | 84.5±3.20 |
| DONet18 | ResNet18 | ✓ | | | 76.5±3.23 | 86.5±0.93 | 86.2±1.47 | 76.1±3.09 | 76.2±3.61 | 87.9±1.50 | 86.2±2.69 | 77.8±0.29 |
| DONet18 | ResNet18 | | ✓ | | 77.0±0.57 | 87.1±3.12 | 85.6±3.65 | 84.8±1.38 | 76.8±3.85 | 88.9±1.84 | 84.5±1.39 | 81.7±2.37 |
| DONet18 | ResNet18 | | | ✓ | **88.0±0.28** | **92.1±1.27** | **93.3±2.18** | **86.8±4.24** | **88.3±0.12** | **93.5±0.43** | **93.0±1.06** | 85.2±1.38 |
| DONet18 | ResNet18 | ✓ | ✓ | | 76.2±2.61 | 86.8±1.63 | 86.3±0.91 | 73.1±3.96 | 76.6±0.79 | 87.5±2.45 | 86.2±2.98 | 79.3±4.40 |
| DONet18 | ResNet18 | ✓ | | ✓ | 84.3±1.52 | 88.5±3.73 | 87.4±3.82 | 80.3±1.12 | 84.9±3.90 | 89.2±3.31 | 89.2±1.06 | 85.9±2.10 |
| DONet18 | ResNet18 | | ✓ | ✓ | 86.3±3.75 | 91.8±4.10 | 92.0±0.28 | 81.5±2.42 | 86.7±3.36 | 90.8±4.30 | 91.9±3.41 | **86.4±0.83** |
| Baseline | ResNet50 | ✓ | ✓ | ✓ | 90.2±3.27 | 94.0±1.65 | 92.1±1.74 | 91.1±2.64 | 90.2±3.27 | 94.0±1.65 | 92.1±1.74 | 91.1±2.64 |
| DONet50 | ResNet50 | ✓ | | | 79.2±2.79 | 90.7±1.61 | 86.9±2.08 | 78.9±3.70 | 80.4±3.41 | 90.6±1.86 | 86.0±4.07 | 79.4±2.38 |
| DONet50 | ResNet50 | | ✓ | | 80.8±1.32 | 89.9±3.24 | 83.3±1.35 | 85.0±2.98 | 80.6±2.55 | 88.7±4.11 | 83.9±1.97 | 84.8±2.52 |
| DONet50 | ResNet50 | | | ✓ | **90.4±1.49** | **94.8±1.69** | 91.6±1.39 | **91.7±2.48** | **90.7±2.65** | 95.6±1.43 | **92.3±4.28** | **92.6±3.57** |
| DONet50 | ResNet50 | ✓ | ✓ | | 79.8±3.41 | 89.2±4.34 | 84.0±2.41 | 75.9±1.90 | 79.6±1.66 | 89.9±1.97 | 84.8±1.48 | 76.5±3.36 |
| DONet50 | ResNet50 | ✓ | | ✓ | 86.3±2.58 | 91.6±1.36 | 85.2±4.21 | 87.9±3.67 | 86.4±3.47 | 92.3±2.74 | 86.1±3.27 | 88.5±3.27 |
| DONet50 | ResNet50 | | ✓ | ✓ | 88.7±0.87 | 93.9±0.82 | **93.9±1.93** | 88.5±4.56 | 89.3±1.79 | **96.2±2.67** | 92.2±4.27 | 86.3±3.10 |

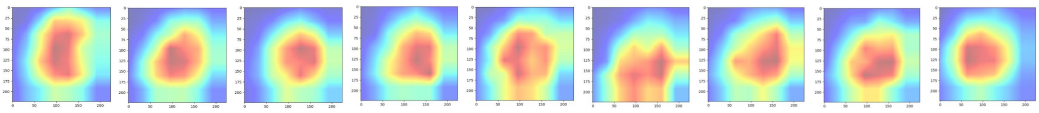
experimental settings in this section are consistent with the previous experiment. Table 4 shows the experimental results on the test set of three common modalities. Two baselines are based on the multi-modality fusion of all modalities. As we can see from Table 4, the FLAIR modality can achieve the best performance among these modalities. The accuracy of the ResDAL block and the ResCAL block based on the ResNet50 is up to 90.7% and 90.4% respectively. T1 and T2 modalities have almost the same classification accuracy, but the classification accuracy decreases when the two are fused together. The accuracy with the ResCAL block of ResNet18 and ResNet50 is 76.2% and 79.8%, and the accuracy with the ResDAL block of ResNet18 and ResNet50 is 76.6% and 79.6%, respectively. Models based on the ResNet50 can achieve a better performance than that of ResNet18.

### 4.6. Comparison to the State-of-the-Art

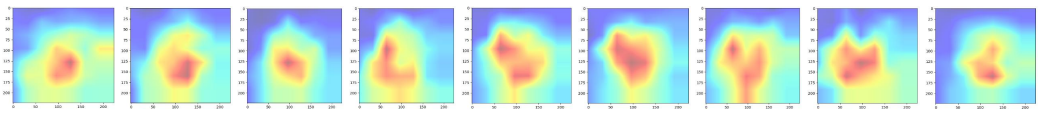
In this subsection, we compare DONet with the existing state-of-the-art image classification methods on Flair modality, including classification models for the natural scene and models

Table 5: Performance comparison among different methods.

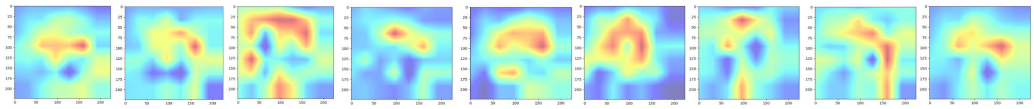
| Methods | Basenet | ACC | AUC | Sensitivity | Specificity |
|---|---|---|---|---|---|
| Baseline He et al. (2016) | ResNet18 | 85.1±0.53 | 89.9±1.84 | 88.9±1.29 | 81.2±0.81 |
| VGGNet19 Simonyan and Zisserman (2014) | VGGNet | 86.5±0.50 | 89.6±2.52 | 88.4±1.62 | 82.2±2.11 |
| GoogLeNet Szegedy et al. (2015) | GoogLeNet | 85.9±2.01 | 89.4±2.85 | 88.8±2.62 | 83.5±1.79 |
| Inception v3 Szegedy et al. (2016) | Inception | 87.3±2.45 | 93.1±1.91 | 88.6±4.31 | 83.0±2.15 |
| ResNeXt18 Xie et al. (2017) | ResNeXt | 87.7±3.65 | 93.1±1.98 | 89.9±2.52 | 85.4±1.79 |
| CANet Hu et al. (2018) | ResNet18 | 87.1±0.30 | 92.4±1.09 | 90.2±1.06 | 85.8±4.24 |
| SANet Jaderberg et al. (2015) | ResNet18 | 86.9±0.33 | 92.0±1.62 | 87.1±2.04 | 83.3±1.03 |
| ResCALNet18 | ResNet18 | 88.0±0.28 | 92.1±1.27 | **93.3±2.18** | **86.8±4.24** |
| ResDALNet18 | ResNet18 | **88.3±0.12** | **93.5±0.43** | 93.0±1.06 | 85.2±1.38 |
| Baseline He et al. (2016) | ResNet50 | 87.8±0.85 | 93.9±1.71 | 89.0±1.54 | 83.5±2.28 |
| FCRN+DRN Yu et al. (2016) | ResNet50 | 88.3±1.36 | 93.7±2.38 | 90.6±3.01 | 84.8±3.65 |
| Color Constancy+ResNet50 Matsunaga et al. (2017) | Ensemble | 88.1±2.07 | 94.9±3.86 | 92.0±1.85 | 84.7±1.08 |
| Modified ResNet Bi et al. (2017) | Ensemble | 88.5±2.76 | 94.7±0.75 | 89.9±2.57 | 84.2±2.55 |
| Multi-Scale+Inception v3 DeVries and Ramachandram (2017) | Ensemble | 88.3±1.31 | 94.3±2.66 | 89.2±2.78 | 84.8±3.54 |
| ARL-CNN Zhang et al. (2019) | ResNet50 | 88.9±2.37 | 95.1±1.23 | 91.7±1.19 | 86.2±1.58 |
| CANet Hu et al. (2018) | ResNet50 | 88.6±1.08 | 93.6±2.03 | 89.1±1.97 | 89.7±1.95 |
| SANet Jaderberg et al. (2015) | ResNet50 | 88.5±1.29 | 94.7±2.17 | 90.5±3.20 | 92.9±3.67 |
| ResCALNet50 | ResNet50 | 90.4±1.49 | 94.8±1.69 | 91.6±1.39 | 91.7±2.48 |
| ResDALNet50 | ResNet50 | **90.7±2.65** | **95.6±1.43** | **92.3±4.28** | **92.6±3.57** |



(*a*) Class activation mappings of ResNet50



(*b*) Class activation mappings of ResCALNet50



(*c*) Class activation mappings of ResDALNet50

Figure 5: Class activation mappings of different models with ResNet50 backbone.

for the medical images. In the case of the natural scene classification models, we refer to the Very Deep Convolutional Networks (VGGNet19) Simonyan and Zisserman (2014), GoogLeNet Szegedy et al. (2015), Inception v3 Szegedy et al. (2016), ResNeXt Xie et al. (2017). In the case of the medical image classification models, we mainly refer to the champion methods used in the International Skin Imaging Collaboration (ISIC) skin lesion classification Codella et al. (2018), such as Yu et al. (2016), Matsunaga et al. (2017), Bi et al. (2017), DeVries and Ramachandram (2017), Zhang et al. (2019). In order to make sure that the comparison models are optimized, we use the consistent hyper-parameters in their papers during training stage, including the learning rate, batch size, image size, max iteration and training epoch. Experimental results on the test set are shown in Table 5. As we can see in Table 5, the DONet can achieve a comparable result. Particularly, DONet based on ResNet50 with the ResDAL block can achieve the best performance among these models. Therefore, these results further verify the superiority of the proposed method.

### 4.7. Visualization of CAM

As discussed in the above subsections, the DONet can achieve very competitive performance with only the FLAIR modality. Inspired by previous literature, we hypothesize that the performance improvement is due to the ability of self-attention learning modules that can make the model capable of discrimination in the learning process. The ability of discriminative learning enables useful areas to be emphasized in the learning process. That is, self-attention learning modules give crucial areas more weight information, while some areas with weak discriminability are given less weight or even ignored in the learning process. To validate the hypothesis, we visualize the class activation mappings (CAM) with the method proposed in Zhou et al. (2016) obtained by ResNet50 and DONet50 in Figure 5, in which (a) shows the CAM of the ResNet50, while (b) and (c) show the CAM of the ResCALNet50 and the ResDALNet50, respectively. The brighter the region, the stronger the discrimination ability of the region for classification. From Figure 5, we can draw the following conclusions: (1) Both the model based on the self-attention learning mechanism and the common CNN classification model can learn useful areas with the discrimination ability, such as bright areas in Figure 5 (a), Figure 5 (b) and Figure 5 (c). This phenomenon suggests that lesions of brain tumors are mostly found in the middle of the brain, while those near the edges are generally not present. (2) Compared with the CNN model without the attention learning mechanism in Figure 5 (a), the CNN model based on the attention learning mechanism can learn some much smaller discriminating areas as shown in Figure 5 (b) and Figure 5 (c). More importantly, the model with smaller bright areas can achieve better results in the classification of brain tumors, which indicates that the model based on the self-attention learning mechanism can achieve better discrimination learning effect. Some useless areas for the classification of brain tumors are ignored in the learning process, so as to improve the efficiency of the model.

## 5. Conclusion

Inspired by the attention learning mechanism of the human brain, we propose a novel medical image classification method for brain tumors, named Discrimination Oriented Network (DONet) in this paper. We first propose two categories of attention learning mechanisms, i.e., the Cascaded Attention Learning (CAL) and the Dual Attention Learning (DAL), which can learn the discrimination information in both the spatial-wise and the channel-wise dimensions in a fine-grained manner. By the CAL and the DAL, the attention information of different dimensions are calculated in a series manner (for cascaded) and a parallel manner (for dual), respectively. We implement the CAL and the DAL on the Deep Residual Network (ResNet) for brain tumor classification. Compared with the state-of-the-art classification methods, the DONet can achieve satisfactory performance. To make full use of the structural advantages of deep learning networks, we will consider the combination of the self-attention mechanism and some other advanced machine learning methods, such as hierarchical neural architecture search, meta-learning and so on in the future. Besides, since the annotation of medical images is very expensive, we also consider the use of semi-supervised or weakly-supervised methods for the analysis and understanding of medical images.

## References

Parnian Afshar, Arash Mohammadi, and Konstantinos N Plataniotis. Brain tumor type classification via capsule networks. In *ICIP*, pages 3129–3133, 2018.

Hua Bai, Yamei Yang, Yan Liu, Junfa Zhao, and Cheng Zhang. Adaptive detection and correction of fixed pattern noise in scmos cameras. In *ICEEET*, pages 107–111, 2018.

Lei Bi, Jinman Kim, Euijoon Ahn, and Dagan Feng. Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks. *arXiv preprint arXiv:1703.04197*, 2017.

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, and Shengping Liu. Adversarial transfer learning for chinese named entity recognition with self-attention mechanism. In *EMNLP*, pages 182–192, 2018.

Noel CF Codella, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *ISBI*, pages 168–172, 2018.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

Terrance DeVries and Dhanesh Ramachandram. Skin lesion classification using deep multi-scale convolutional neural networks. *arXiv preprint arXiv:1703.01402*, 2017.

Marco Giulioni, Gianluca Marucci, Massimo Cossu, Laura Tassi, Manuela Bramerio, Carmen Barba, Anna Maria Buccoliero, Gianfranco Vornetti, Corrado Zenesini, Alessandro Consales, et al. Cd34 expression in low-grade epilepsy-associated tumors: Relationships with clinicopathologic features. *World Neurosurgery*, 121(1):761–768, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *NeurIPS*, pages 2017–2025, 2015.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017.

Kazuhisa Matsunaga, Akira Hamada, Akane Minagawa, and Hiroshi Koga. Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble. *arXiv preprint arXiv:1703.03108*, 2017.

T Pandiselvi and R Maheswaran. Efficient framework for identifying, locating, detecting and classifying mri brain tumor in mri images. *Journal of Medical Systems*, 43(7):189, 2019.

D Jithendra Reddy, T Arun Prasath, M Pallikonda Rajasekaran, and G Vishnuvarthanan. Brain and pancreatic tumor classification based on glcm—k-nn approaches. In *ICICA*, pages 293–302, 2019.

Subrata Sarkar, Alyson K Fletcher, Sundeep Rangan, and Philip Schniter. Bilinear recovery using adaptive vector-amp. *IEEE Transactions on Signal Processing*, 67(13):3383–3396, 2019.

N Varuna Shree and TNR Kumar. Identification and classification of brain tumor mri images with feature extraction using dwt and probabilistic neural network. *Brain Informatics*, 5 (1):23–30, 2018.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016.

A Vallée, C Guillevin, M Wager, V Delwail, R Guillevin, and J-N Vallée. Added value of spectroscopy to perfusion mri in the differential diagnostic performance of common malignant brain tumors. *American Journal of Neuroradiology*, 39(8):1423–1431, 2018.

Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *CVPR*, pages 3156–3164, 2017.

Songtao Wu, Shenghua Zhong, and Yan Liu. Deep residual learning for image steganalysis. *Multimedia Tools and Applications*, 77(9):10437–10453, 2018.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 1492–1500, 2017.

Lequan Yu, Hao Chen, Qi Dou, Jing Qin, and Pheng-Ann Heng. Automated melanoma recognition in dermoscopy images via very deep residual networks. *IEEE Transactions on Medical Imaging*, 36(4):994–1004, 2016.

Jianpeng Zhang, Yutong Xie, Yong Xia, and Chunhua Shen. Attention residual learning for skin lesion classification. *IEEE Transactions on Medical Imaging*, 1(1):1–12, 2019.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016.