SN Computer Science (2022) 3:142
https://doi.org/10.1007/s42979-022-01024-0

ORIGINAL RESEARCH

MacularNet: Towards Fully Automated Attention-Based Deep CNN for Macular Disease Classification

Sapna S. Mishra[1] · **Bappaditya Mandal**[2] · **Niladri B. Puhan**[1]

Received: 4 August 2021 / Accepted: 5 January 2022

Abstract
In this work, we propose an attention-based deep convolutional neural network (CNN) model as an assistive computer-aided tool to classify common types of macular diseases: age-related macular degeneration, diabetic macular edema, diabetic retinopathy, choroidal neovascularization, macular hole, and central serous retinopathy from normal macular conditions with the help of scans from optical coherence tomography (OCT) imaging. Our proposed architecture unifies refined deep pre-trained models using transfer learning with limited training data and a deformation-aware attention mechanism encoding crucial morphological variations appearing in the deformation of retinal layers, detachments from the subsequent layers, presence of fluid-filled regions, geographic atrophy, scars, cysts, drusen, to achieve superior macular imaging classification performance. The proposed attention module facilitates the base network to automatically focus on the salient features arising due to the macular structural abnormalities while suppressing the irrelevant (or no cues) regions. The superiority of our proposed method lies in the fact that it does not require any pre-processing steps such as retinal flattening, denoising, and selection of a region of interest making it fully automatic and end-to-end trainable. Additionally, it requires a reduced number of network model parameters while achieving higher diagnostic performance. Extensive experimental results, analysis on four datasets along with the ablation studies show that the proposed architecture achieves state-of-the-art performance.

Keywords Age-related macular degeneration · Attention mechanism · Choroidal neovascularization · Diabetic macular edema · Optical coherence tomography

Introduction

Macula is the main sensory region present near the center of retina surrounding the fovea. It is mainly responsible for the central vision. The macular health is affected by a number of diseases, such as age-related macular degeneration (AMD) [23], diabetic macular edema (DME) [28], macular hole (MH) [41], central serous retinopathy (CSR) [25], etc. These diseases, because of their sight-threatening effects and high diagnostic complexity have attracted intensive research efforts in the last few years [32, 39]. The optical coherence tomography (OCT) imaging technique is used to obtain a cross-sectional view of retinal layers and captures the textural and morphological variations [14], making it convenient for macular disease detection. From the onset and as the disease progresses, early diagnosis involves careful visual inspection of retinal layers deformation, detachments from the subsequent layers, presence of fluids, geographic atrophy and drusen. Some sample OCT scans from various classes are shown in Fig. 1.

AMD macula emerges with the accumulation of drusen in the middle of the retinal pigment epithelium (RPE) and the bottom choroid. It causes an irreversible lesion in the retina and damages the macular region causing RPE atrophy, the detachment of layers along with other abnormalities such as choroidal neovascularization (CNV). CNV is detected by the formation of new blood vessels in the choroid layer and this is a typical cause of wet AMD. On the other hand, the leading causes of blindness in adults over the age of 65

✉ Bappaditya Mandal
b.mandal@keele.ac.uk

Sapna S. Mishra
ssm14@iitbbs.ac.in

Niladri B. Puhan
nbpuhan@iitbbs.ac.in

[1] School of Electrical Sciences, Indian Institute of Technology, Bhubaneswar, India

[2] School of Computing and Mathematics, Keele University, Newcastle, UK

Published online: 22 January 2022

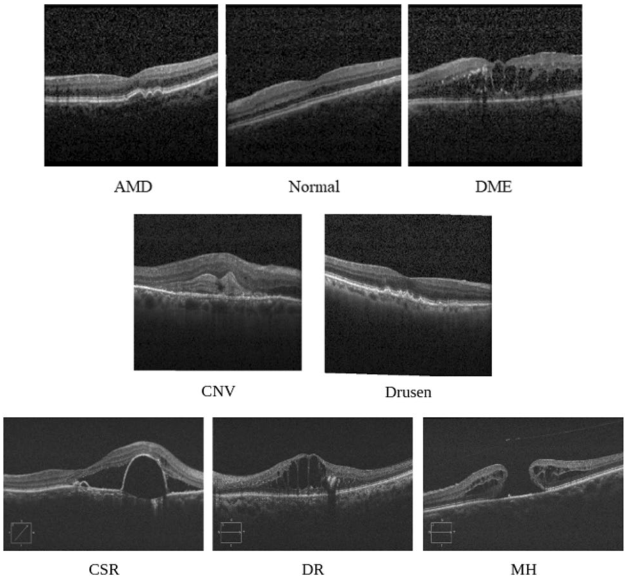

Fig. 1 Sample examples of OCT scans of different macular eye diseases. Top row, shows AMD, normal and DME samples from Duke dataset [39]. Mid row, shows CNV and drusen deformations from UCSD dataset [18]. Bottom row, shows CSR, DR and MH patlogies from OCTID dataset [10]

years is diabetic retinopathy (DR) [37]. Diabetes may lead to breakage, leakage or blockage of retinal blood vessels. DME can occur in any stage of DR causing various deformations to the morphology of retinal layers, such as retinal thickening, microvascular changes, formation of hard exudates and focal retinal detachments. The vascular walls malfunction and liquid starts to accumulate in the region known as fluid-filled regions (FFRs) visible as black blobs in the OCT scans.. The retinal defects caused by MH vary from small hypo-reflective breaks to wide intra-retinal spaces in the OCT images. Other symptoms encompass macular edema, fluid accumulation, and thickened edges with cavities of reduced reflectively [37]. CSR is another common retinal pathology which generally affects the middle age people. It occurs due to the leakage of fluid into the retina through an RPE defect and oftenly is characterized by focal detachments of the retina and the RPE layer [25]. On the OCT scans, CSR is detected by hypo reflective spaces at the sub-retinal and sub-RPE levels along with sub-retinal hyper-reflective deposits in some cases.

The main contribution of this paper is a novel fully automatic system for the classification of macular OCT images. Here, because of the lack of a large number of macular OCT images, we utilize the idea of transfer learning on existing pre-trained deep CNN models We propose a unified framework for integrating fine-tuned pre-trained deep CNN model(s) with a novel deformation-aware attention based mechanism (MacularNet) for the extraction of improved discriminative features resulting in the superior classification of macular OCT images. Our method does not require a separate region of interest (RoI) extraction step nor any other pre-processing steps such as retinal flattening and/or denoising. The developed attention module facilitates the network to focus on the relevant regions automatically, which helps to achieve superior performance with reduced number of model parameters. The main technological contributions are summarized as follows:

- Unified learning mechanism incorporating fine-tuned deep CNN model architecture with the deformation-aware attention mechanism to extract inter-class discriminative features, trainable end-to-end with limited OCT training data which eliminates the need for pre-processing steps and well-suited with multiple deep learned network architectures.
- Extensive experiments and their ablation studies on four datasets. This includes analysis of the proposed architecture without using RoI selection, avoiding denoising and flattening steps, selection of deep CNN model with pre-trained weights, evolution of attention maps over epochs, layers and training procedures, effect of attention over discriminative properties and reduction in the network parameters are performed.

The next subsection discusses some related works. Section "Related Works" describes the proposed attention-based deep CNN macular OCT classification method. In Section "Proposed Method: MacularNet", experimental evaluation and comparative analysis are reported on four OCT datasets. Section "Experimental Results" reports the ablation study on the proposed network architecture along with the attention maps. In section "Ablation Study" we present an extensive analysis of the proposed method with the help of attention maps. Conclusions and future works are discussed in Section "Conclusions and future works".

Related Works

Various works have been reported on macular OCT classification based on traditional machine learning techniques [17, 22, 39, 47], as well as convolutional neural networks [20, 32–34]. Srinivasan *et al.* classified denoised and cropped OCT volumes using histograms of oriented gradients (HoG) descriptor [39]. However, the cropped images can miss the pathology in the peripheral regions and retinal flattening is invalidated for severely distorted RPE layer. In [47], linear configuration patterns (LCPs) based classification is proposed yielding high accuracy, but the motion blurred or shadowed B-scans have not been taken into account. The method in [41] employed sparse coding and dictionary learning, but it again depends on the spatial location-based retinal flattening. These traditional methods are semi-automatic as well as database specific in nature; hence better alternatives are proposed in literature.

In past few years, deep learning methods relying on cascaded neural networks [26, 40] have been proposed for macular OCT classification which extract relevant diagnostic features [29]. In [33] macular OCTs are classified into normal, AMD, and DME using an ensemble model of multi-scale convolutional mixture of expert (MCME) and it required retinal flattening and volume of interest generation. In [34], surrogates are generated using the features of the macular region of B-scans and then CNN is employed to classify these surrogates. An ensemble of four improved ResNet50 architectures has been used in [20] for the classification of macular lesions. In [7], authors have used iterative fusion of CNNs (IFCNN) method for the task of retinal OCT image classification. The method proposed in [4] introduces a multi-scale deep feature fusion (MDFF) based classification approach using CNN. Whereas the work in [5] employs a generative adversarial network (GAN) for the OCT scan classification purpose which additionally addresses the issue of lack of large-scale dataset. On the other hand, works in [6, 8, 24], incorporate attention mechanism for the classification of OCT scans. More recently, in [42], authors have proposed a light weight CNN for macular disease classification but has evaluated it only on one dataset.

In the task of medical image analysis, there is an unavailability of huge labelled datasets. Issues of privacy and unacceptability result in collecting a limited amount of medical data for research purposes [43]. In such scenarios, deep CNNs tend to overfit the data. Thus, transfer learning is beneficial which overcomes the obstacle of insufficient training data [43]. It involves fine-tuning a model trained on an uncorrelated dataset with the actual dataset of medical images. Transfer learning has been reported to be a better alternative for several tasks of biomedical image classification [21]. In the specific case of macular OCT classification, fine-tuning of existing deep models have been explored in the works of [16, 17, 35, 45].

Involving pre-processing steps such as denoising, retinal flattening and RoI extraction are the most common problem of existing OCT classification methods and which make these methods database dependent and time-consuming in nature. Although the usage of RoIs yields impressive results, the classification process as a whole becomes less automated and loses its generality for real-world applications. To overcome these challenges, we move towards the concept of attention-based networks. The attention mechanisms were initially introduced in the context of natural language processing [3]. Later, in the field of computer vision, attention mechanisms have been employed to a range of problems for instance, image classification [15], image retrieval [31] and action recognition [1].

In the field of biomedical engineering, attention has been explored for scan segmentation [2], image and text classification [46], and organ localization [48]. For standard medical image classification, the high importance of local information have been exploited in few-research works [8, 30]. In these methods, a hard-attention model is employed or bounding box labels are available or segmented images were utilised to guide the attention. To overcome such manual intervention, soft attention-based mechanism is proposed in [24, 36], where continuous functions are used to assign the attention weights on the input, composing a fully differentiable function and can be trained simultaneously with the complete architecture using the backpropagation technique.

Proposed Method: MacularNet

Motivation

In large-scale clinical evaluation of OCT scans, there is a huge demand for the development of computerized algorithms for detection of the macular eye diseases for better scalability and adaptability. The automated analysis of these images are not only advantageous for better patient diagnosis, but can also provide training to new ophthalmologists. Most of the existing methods in literature use database-specific conventional methods or pre-processing steps as reported in subsection 1.1. It is evident from the literature review that deep CNN-based model is capable of encoding complex spatial information, equivariance to translation and automated feature extraction. CNNs have been proven to be a useful algorithm for many classification purposes, specifically for the task of medical image classification [19]. Hence we have utilized deep CNNs as the base model(s) for encoding crucial discriminative information. Though the deep learning-based models in prior works have reported better performance for macular OCT classification, they require a large number of model parameters, which we have minimised with our proposed approach of deformation-aware attention mechanism.

In our current scenario, there is an unavailability of a large macular OCT dataset or any model pre-trained with a large number of OCT images. To overcome this deficiency, we take the help of transfer learning technique. It serves the purpose of achieving improved classification performance with a limited number of training images and lesser resources. In the proposed network, a novel attention mechanism has been developed to extract local deformation-aware features (as shown in Fig. 1) and classify them into AMD, DME, CNV, macular hole, CSR, DR, and normal macular eye conditions. This helps in eliminating the need for RoI selection, pre-processing steps and make the predictions on relevant features. Our proposed architecture consists of two main modules: a refined, fine-tuned deep learned CNN model(s) and an attention mechanism unified to constitute the MacularNet framework such that the entire configuration is end-to-end trainable. The architecture is shown in Fig. 2 and empirically formulated in (1) for easier interpretability.

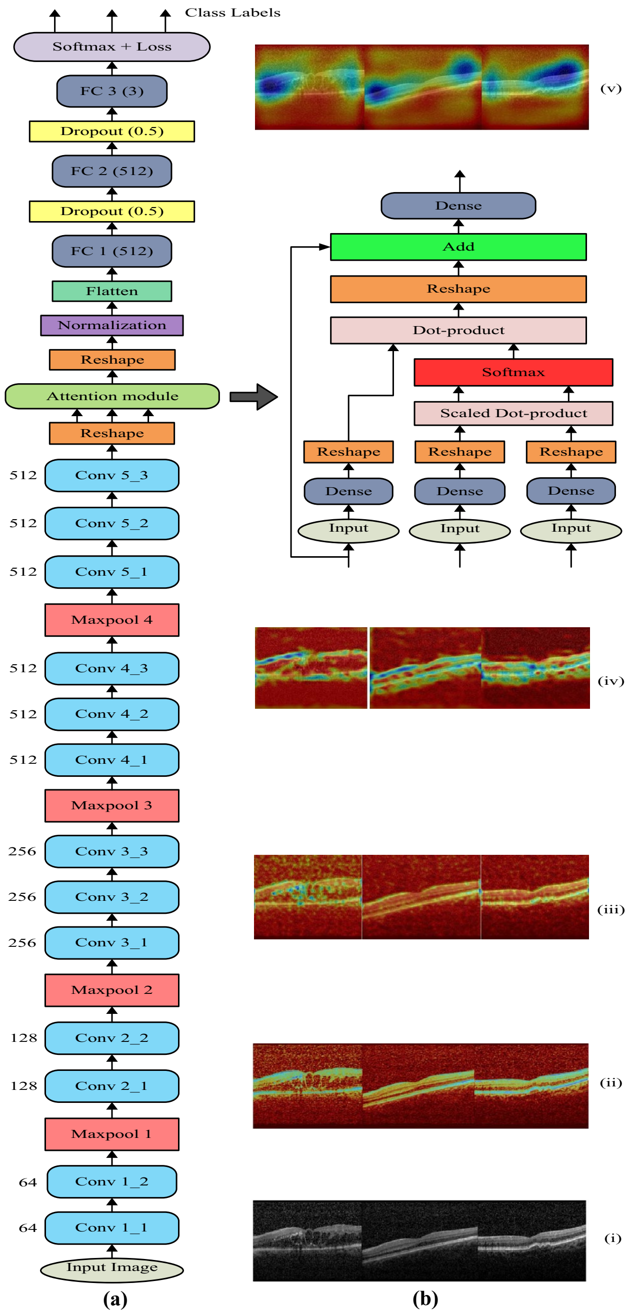

Fig. 2 **a** Architecture of the proposed attention based MacularNet. **b** Some feature maps from the architecture for layers: (i) Input image (ii) Conv 2_1 (iii) Conv 3_2 (iv) Conv 4_3, attention module and (v) FC 3

Preparing the Deep Network

Let, image I(height × width × number of channels) denotes the input data, $C(m, n, s)$ represents the convolutional layer, where m is filter size with stride s and n is number of filters. $M(s)$ denotes max-pooling layer of stride s, $FC(d)$ represents fully connected layer with d number of neurons, F stands for flatten layer. N and $D(r)$ show normalization and dropout layer with dropout rate of r, respectively. The attention module is represented by $Att()$, which is explained in Eq. (9).

$$\begin{aligned}\Phi_M \equiv\ & I(224\times224\times3) \longrightarrow 2\times C(3,64,1) \longrightarrow M(2,2)\\ & \longrightarrow 2\times C(3,128,1)\\ & \longrightarrow M(2,2) \longrightarrow 3\times C(3,256,1) \longrightarrow M(2,2)\\ & \longrightarrow 3\times C(3,512,1)\\ & \longrightarrow M(2,2) \longrightarrow 3\times C(3,512,1)\\ & \longrightarrow R(\{14,14,512\} \Rightarrow \{196,512\})\\ & \longrightarrow Att(\{196,512\};\{196,512\};\{196,512\})\\ & \longrightarrow R(\{196,32\} \Rightarrow \{14,14,32\})\\ & \longrightarrow N \longrightarrow F \longrightarrow FC(512) \longrightarrow D(0.5)\\ & \longrightarrow FC(512) \longrightarrow D(0.5) \longrightarrow FC(3)\end{aligned} \tag{1}$$

In our network architecture, the number of filters per layer increases as we move from first to last convolutional layers as shown in Fig. 2a. These filters have a small receptive field of 3 × 3. Designing the network in this manner adds more non-linearity with each block that makes the decision function more discriminative in nature and the small receptive fields help the model to capture the fine-grain details of input images as shown in Fig. 2b. The final features are extracted from the "Conv 5_3" layer as it was experimentally observed that this layer generates more focused intermediate feature leading to superior classification results. The weights of the dense and attention layers are randomly initialized and are triggered by ReLU function. This randomization enables the model to learn the difference in feature space between the face image dataset and OCT B-scans and improve the generality of the network. The neurons in output layer are activated by softmax function, given by:

$$\sigma(x)_j = \frac{e^{x_j}}{\sum_{i=1}^{q} e^{x_i}},\ for\ j = 1, 2, 3 \ldots, q, \tag{2}$$

where q is the number of neurons and x is the input to the layer. The entire network is trained using backpropagation [13], given by (3), with learning rate η and momentum μ.

$$\begin{aligned}\Delta\mathbf{w_i}^{(T)} &= \mu\Delta\mathbf{w_i}^{(T-1)} - \eta\nabla L_i^{(T-1)}(y,\hat{y}),\\ \mathbf{W_i}^{(T)} &= \mathbf{W_i}^{(T-1)} + \Delta\mathbf{w_i}^{(T)},\end{aligned} \tag{3}$$

where $\mathbf{W_i}$ represents a weight vector and $\nabla L_i(y, \hat{y})$ denotes the gradient of the objective function summed over all the samples of the current minibatch i at every epoch T. The objective function is given by (4). $\Delta \mathbf{w}$ is the change in the weight vector and is initialized with 0. We utilized the categorical cross-entropy (CCE) as the loss function given by:

$$L(y_u, \hat{y}_{u,t}) = -\frac{1}{M} \sum_{c=1}^{M} y_{u,t} log(\hat{y}_{u,t}), \tag{4}$$

where M is the number of classes, $y_{u,t}$ is the target label and $\hat{y}_{u,t}$ is the score of CNN for each output neuron t summed over all the classes, for each sample u. The CNN score can be represented as $\hat{y}_{u,h} = \sigma(\mathbf{S})$, where $\mathbf{S}$ is the input vector to the output layer, activated by the softmax function, $\sigma(x)$, given by (2). The learning rate η is made adaptive such that for every 10 epochs it reduces by a factor of 0.1, if the decrease in loss function is less than 0.001.

Deformation-Aware Attention Mechanism for OCT Classification

Each scan of the OCT volumes is 2-D imaging of retinal layers along with the background. For relevant feature extraction and superior classification, diagnostic systems need to focus on the region of retina that includes major variations of morphological structures, for example, the deformation variations between the disease classes (AMD, DME, CNV, and normal). Because of these macular diseases, deformation occurs in the retinal layers (as shown in Fig. 1). They grow abnormally and detach from the subsequent layers, fluid-filled regions (as black blobs), geographic atrophy, CNV and drusen appear in OCT images. In the literature, researchers [16, 32, 33, 39], have introduced an RoI extraction step prior to feature generation. Although the obtained results are satisfactory, the algorithm as a whole looses its generality and increases the likelihood of ignoring the pathological symptoms which may occur outside the peripheral region. Hence, we hypothesize to develop an attention-based model that would be a better alternative to focus on these regions of deformations, encoding their morphological variations contributing more towards the final classification.

We anticipate that the inclusion of attention modules into a refined fine-tuned CNN improves the classification performance as well as decreases the required number of model parameters in the network architecture. For an attention module, the context vectors, $\mathbf{C_v}$ are generated based on the relevance of the regions of features on the classification. These context vectors are the sum of feature vectors of the input image, weighted by the attention coefficients, given by:

$$\mathbf{C_v} = \sum_{i=1}^{n_v} a_i \mathbf{f}_i \tag{5}$$

where, n_v is the number of feature vectors and $\mathbf{f_i}$ is the ith feature vector.

These attention coefficients, a_i are calculated as the softmax of the predefined alignment score. The score is evaluated by using non-linear function on the weight matrices to be learned in the deformation-aware attention model. Various score functions have been explored in the literature, such as, cosine [11], dot-product [44], tanh [3]. In our approach, we have used the dot-product transformation and the coefficients here are calculated over the feature maps of convolutional layers.

In our attention mechanism, we perform scaled-dot products to compute attention coefficients and the context deformation-aware vectors, resulting in imparting more importance to the relevant parts, thereby, improving its discriminatory abilities. Unlike most of the existing attention based CNN models [15], the proposed attention module focuses only on the spatial features of scans and learns the morphological variations. In this work, soft-attention based attention module is incorporated to the base model in between the "Conv 5_3" and "Normalization" layer along with reshape layers, as shown in Fig. 2.

Details of the proposed attention module are shown in Fig. 3. In this module, the weights are assigned to each region of the feature to generate the context vector, which is updated after each epoch with respect to the relevance of the region in the classification. The attention layer developed here takes three same inputs, which are the features extracted from the previous layer. The dense layer adds additional weights to each input which are then linearly transformed. Each of this reshaped feature is then segregated into h number of channels which in this work is eight. The scaled dot-product between two input features on each of the channels is computed to obtain the alignment scores. This score is passed through the softmax function to yield attention coefficients. The dot-product used here computes the similarity of two inputs, hence providing greater emphasis to higher weighted regions of the inputs and then all the weights are scaled by the scaling factor. The softmax function is used to generate the probabilities of the weights after the dot product to maximize the values of attention coefficients of the relevant features and translating the resultant in the range of 0 to 1.

The vectors obtained after weighing the inputs of the attention module followed by their linear transformation are represented as P^A, P^B and $P^C \in \mathbb{R}^{D \times C \times h}$, where D = 14 × 14 and C = 512/h and 512 is the number of channels in the feature maps obtained from Conv 5_3 layer of the base model. The alignment score generated here is given as:

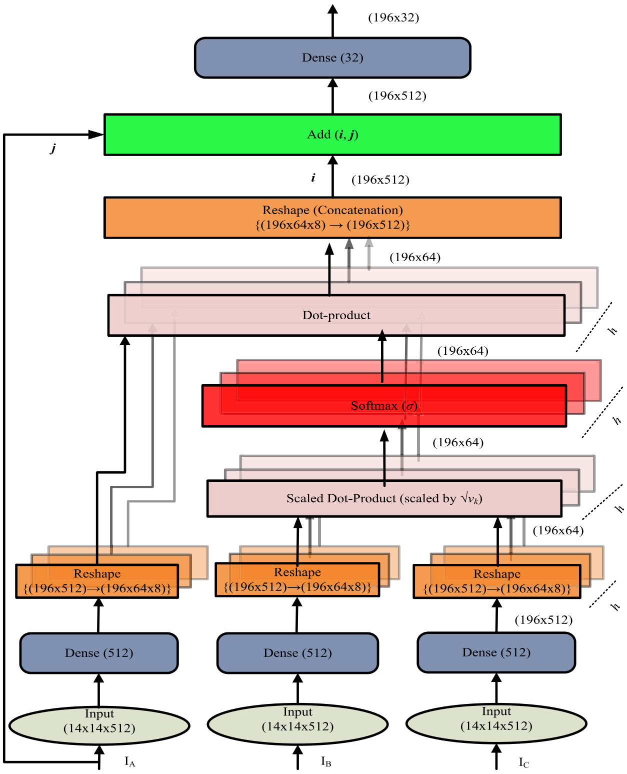

Fig. 3 Architecture of proposed attention module. I_A, I_B and I_C are three inputs to the module. Dense (*d*) represents dense layers with *d* number of neurons. The numbers in the bracket on the right side of arrows denote the output dimensions of respective layers. *h* denotes the number of channels in each layer after transformation. We have considered *h*=8 for our module. The scaling factor, $v_k = 512$. Reshape layer shows the dimensions of tensors before and after reshaping. The last reshape layer performs concatenation of the 8 channels. Add (*i*, *j*) performs additions of two matrices, ***i*** and ***j***

$$
\begin{aligned}
Q_{i,j,k} &= \frac{\sum_{l=0}^{h-1} P^{B}_{i,j,l} \cdot P^{C^T}_{i,l,k}}{\sqrt{v_k}} \quad \in \mathbb{R}^{D \times C \times C}.\\
i &= 0, \ \ldots, \ D-1;\\
j &= 0, \ \ldots, \ C-1;\\
k &= 0, \ \ldots, \ C-1.
\end{aligned}
\tag{6}
$$

where $P^{x^T} \in \mathbb{R}^{D \times h \times C}$ and $x \in \{A, B, C\}$. The obtained score, Q is passed through softmax to yield attention coefficients as shown below:

$$
A = \sigma(Q) \quad \in \mathbb{R}^{D \times C \times C}
\tag{7}
$$

The attention coefficient calculation on eight different channels generates attention coefficients in different feature subspaces, encapsulating finer details of the macular OCT images. The computed coefficients undergo dot product with the third reshaped input to impart the required focus on the significant regions, which is represented as:

$$
\begin{aligned}
V_{i,j,k} &= \sum_{l=0}^{C-1} A_{i,j,l} \cdot P^{A^T}_{i,l,k} \quad \in \mathbb{R}^{D \times C \times h}.\\
i &= 0, \ \ldots, \ D-1;\\
j &= 0, \ \ldots, \ C-1;\\
k &= 0, \ \ldots, \ h-1.
\end{aligned}
\tag{8}
$$

The obtained matrix, $V \in \mathbb{R}^{D \times C \times h}$ is reshaped back to two-dimensional vector. Finally, the output is added to the input feature itself resulting in attended feature generation. The the attended feature is weighted by the dense final dense layer of the attention module to generate a more concise feature with additional trainable weights.

The attention module of our network, shown in Fig. 3, is described in Eq. (9). The $Att(I_A;I_B;I_C)$ layer of (1) is expressed in the equation, where I_A, I_B and I_C are three inputs each of size $k \times l$. The tensor dot product between tensors f and g is denoted by $Bdot(f, g)$, v_k denotes the scaling factor and is derived from number of channels of the output of the final convolutional layer, $\sigma(x)$ indicates the softmax function, shown in (2). The addition operation between two matrices $i, \ j$ is performed by $Add(i, j)$. As mentioned earlier, P^A, P^B and P^C are outputs of reshape layers R_A, R_B and R_C, respectively, which transform one-channel feature into multiple-channel feature subspaces. The variable *out* is the output of the preceding reshape layer which concatenates eight channels to obtain a single feature block. The subscripts in reshape and fully connected layers represent the layer being added to either A^{th} or B^{th} or C^{th} input path, while the absence of subscript denotes that the layer follows the prior computation.

$$
\begin{aligned}
\Phi_{att} \equiv & \big[I_A(196 \times 512), I_B(196 \times 512), I_C(196 \times 512)\big]\\
& \longrightarrow \big[FC_A(512),\ FC_B(512),\\
& FC_C(512)\big] \longrightarrow \big[R_A(\{196 \times 512\}\\
& \Rightarrow \{196 \times 64 \times 8\}), R_B(\{196 \times 512\} \Rightarrow\\
& \{196 \times 64 \times 8\}), R_C(\{196 \times 512\}\\
& \Rightarrow \{196 \times 64 \times 8\})\big] \longrightarrow\\
& Bdot\left(\sigma\left(\frac{Bdot(P^B, P^C)}{\sqrt{v_k}}\right), P^A\right)\\
& \longrightarrow R(\{196 \times 64 \times 8\}\\
& \Rightarrow \{196 \times 512\})\\
& \longrightarrow Add(out, I_A) \longrightarrow FC(32)
\end{aligned}
\tag{9}
$$

Training with MacularNet

The training of MacularNet involves transfer learning of deep network and training from scratch for the attention module. Because of the limited number of training OCT images, we use the concept of transfer learning by initializing and fine-tuning the deep CNN model with pre-trained weights. During the training, all samples are assumed to be independent and identically distributed. In the proposed network, weights of the first 18 layers are pre-trained on the VGG-Face descriptor dataset of 2.6 million face images [27]. The pre-trained model is then unified with the proposed attention model and subsequent dense layers. We have performed several experiments with various deep networks for the classification of OCT images and the analysis of the same have been reported in subsection 4.1. It is evident from our experimental results that the proposed MacularNet architecture is not dependent on any specific deep CNN model.

The deformation-aware attention model is built up and trained from scratch. All the weights in the attention module are randomly initialized and are updated simultaneously with the entire network using the back-propagation algorithm given in (3). It can be observed from Eq. (9) that the proposed attention layer takes all three inputs from the same layer. Hence, the attended feature map as well as the attention coefficients are derived from the same convolutional layer output enabling highly localized deformation-aware attention on global features. The transformations represented in Eq. (9) capture the finer details of the input scans and translates the feature into smaller tensor with relevant details acquired while training after each epoch. Applying soft attention mechanism eliminates the need for any bounding box labels, replaces the non-stochastic hard attention mechanisms' continuous function to assign the attention weights on the entire input image and hence composes an end-to-end trainable CNN. Our proposed deformation-aware attention module increases the focus of the network on the relevant parts of the feature maps and thus eliminates the need for RoI selection, retinal flattening, denoising steps and helps in the reduction of the number of network model parameters. All these efforts result in superiority of our proposed framework over the existing macular OCT classification methods.

Experimental Results

All the OCT images in the dataset are resized to 224 × 224, self-replicated three times and concatenated to generate a tensor of dimension 224 × 224 × 3 to match the input dimensions of the pre-trained network. For training purpose, the augmented dataset is generated by randomly flipping images and translating them by ±40 pixels. This strategy enables the network to tackle the problem of translation. Besides, it degrades the inconsistency due to a different numbers of right and left eyes in the dataset. To counter the effect of inclination incurred in the images, some randomly chosen image samples are rotated.

Experimental Setup

In this work, the experiments are carried out on four datasets: Duke [39], NEH [33], UCSD [18] and OCTID [10]. Two evaluation protocols are used over Duke and NEH datasets. The first protocol, leave patient(s) out (LPO), is followed from [34]. For Duke dataset, the test set is generated by randomly taken out of each case of one patient and the remaining images are segregate into training and validation sets with a 4:1 ratio. For NEH dataset, the test set contains two volumes of each class and the rest of the partition is the same as that of the Duke dataset. For 10 different randomly selected test cases, the experimental process is repeated and the average of these results are reported in this section. On the other hand, the second protocol is fivefold cross-validation (CV), followed from [33] which has been used over the Duke, NEH and OCTID datasets. For the third UCSD dataset, the protocol of [8] has been followed to make a fair comparison with other existing works. Here, the whole UCSD dataset is divided into $\eta = 6$ subsets. The model is trained on one subset and is followed by the testing on the remaining $(\eta - 1)$ subsets. These training experiments are repeated η times such that model is trained by each of the η subset exactly once. Finally, the experimental results obtained over all the folds are averaged and tabulated.

The models are coded in Python 3.5 using Keras package with Tensorflow-GPU v1.8.0 backend. They are trained using 8 GB NVIDIA GeForce GTX 1080 GPU with Cuda v8.0 and cuDNN v6.0 accelerated library on the Linux platform. Following parameters have been used for training and testing of the models: *number of epochs* = 100 for LPO and 50 for each fold of k-fold CV, *batch size* = 64, $decay = 1e - 6$. The categorical cross-entropy loss function given in (4) as considered as a loss function and SGD optimizer is used with $momentum = 0.9$ to update the weights of the network. The adaptive learning rate technique is adopted as training progress. For comparative study, the number of epochs and batch sizes for training have been kept same for the training of all the baseline models and MCME architecture [33]. A similar data augmentation technique was also followed to maintain the consistency with the size of training dataset. For training of the MCME model the custom loss function method used in the original work was replaced with CCE loss here.

The performance of the proposed work is evaluated based on the number of True-Positives (TPs), False-Positives (FPs), False-Negatives (FNs), True-Negatives (TNs) for each ith class of multi-class classification. Primarily, the following three performance metrics have been used for model evaluation in this study. These metrics are defined for ith class which are averaged together to yield final results.

Precision gives the positive prediction value. This value provides information on how efficiently our system avoids false positives. It can be measured as

$$Precision_i = \frac{TPs_i}{TPs_i + FPs_i} \tag{10}$$

Recall, also called sensitivity gives the information about how efficiently the model reduces false negatives. This can be calculated as

$$Recall_i = \frac{TPs_i}{TPs_i + FNs_i} \tag{11}$$

Accuracy is the proximity of measurement results to the true value

$$Accuracy_i = \frac{TPs_i}{TPs_i + FPs_i + TNs_i + FNs_i} \tag{12}$$

Results on Dataset 1

Dataset 1 or the Duke dataset [39] is acquired at Duke University, Harvard University and University of Michigan with approved protocols from the Institutional Review Board and is made freely available to the public for facilitating comparison and future studies by other groups. It contains 45 volumes of OCT acquisitions obtained from 15 subjects of each classes: AMD, DME, and normal, comprising of a total of 3241 B-scan images. Experimental results using the two protocols are reported in Table 1.

Table 1 illustrates the performance obtained for various ways of fine-tuning the baseline architecture of VGG-Face [27]. The output layer and dense layers of the VGG-Face are fine-tuned (in architectures 1 and 2, respectively). In architecture 3, the features for OCT scans are extracted from the "pool5" layer of VGG-Face and then classified using two randomly initialized dense layers of 4096 neurons activated by ReLU function and an output softmax layer. The dense network is trained with the OCT datasets. Furthermore, the proposed MacularNet architecture is evaluated without and with the attention module (architectures 4 and 5, respectively) to illustrate the importance of the mechanism. Table 1 also shows the number of model parameters required in each case. It can be inferred that the addition of the attention module improves the performance and reduces the number of model parameters by 3.5 times approximately. Comparison of the performance of the proposed network with the existing methods [16, 32–34, 47], are reported in Table 2. Our method surpasses the results of the current works without using pre-processing or RoI extraction steps.

To demonstrate the benefits of transfer learning, we conducted an experiment to observe the difference in the results for training the network from scratch and transfer learning. The parameters of the base network are randomly initialized from the truncated normal distribution with zero mean instead of initializing them with the pre-trained weights. The model is then trained for 300 epochs where the validation accuracy converges and obtained results are reported in Table 3. The outcomes of both the training strategies on two versions of our method are tabulated in Table 3. It can be deduced that transfer learning enables the network to perform remarkably well on dataset 1 when compared to the alternate training strategy.

It is evident from Table 1 that MacularNet produces the best results in all categories of dataset 1. The inclusion of the attention mechanism plays an important role by improving the network performance and reducing the required number of weights of architecture. In addition to this, the comparative results in Table 2 depicts that our method has consistently improved performance on this publicly available dataset even though the pre-processing steps are eliminated.

Results on Dataset 2

In dataset 2, also known as NEH dataset [32], the OCT volumes are acquired from Noor eye hospital, Tehran, collected and made publicly available for the purpose of research on maular OCT classification. This dataset contains OCT volumes of 48 AMD patients, 50 DME patients, and 50 Normal cases, total comprising of 4230 B-scan images. Recently in [8], attention mechanism has been incorporated for macular OCT classification constructed by a series of lesion-attention modules, convolutional and pooling layers; however, this method requires a larger number of model parameters. In the case of dataset 2, our network attains state-of-the-art results using 5-fold CV evaluation, as illustrated in Table 4. To our knowledge, no existing works have been reported on this dataset using the LPO protocol. The evaluation of the proposed MacularNet on both the datasets using both the protocols have been reported separately in Table 5.

The results obtained using both the protocols on this dataset are mentioned in Tables 1 and 5. It is evident that on dataset 2, the proposed method shows superior results when compared to the baseline architectures in all categories. It is evident from Tables 1, 2 and 4 that our MacularNet architecture performs best in all categories when compared to three baseline architectures and other existing works on both the datasets using both the protocols. Incorporation of the attention module has dual advantages of improved classification performance and reduced number of model parameters. It can be inferred from Table 3 that transfer learning helps in improving the performance of the architecture on both datasets 1 and 2, using both the protocols.

Table 1 Experimental results and comparisons with base pre-trained deep learned model with our proposed MacularNet approach on datasets 1 and 2 using two evaluation protocols and the number of parameters involved

Architecture	Total Parameters	Evaluation protocols					
		Leaving patient(s) out (LPO)					
		Accuracy (%)		Precision (%)		Recall (%)	
		Dataset 1	Dataset 2	Dataset 1	Dataset 2	Dataset 1	Dataset 2
1. Fine-tuning output layer of VGG-Face	134,272,835	77.39	73.51	80.63	75.52	74.84	74.71
2. Fine-tuning of dense layers of VGG-Face	134,272,835	83.53	83.93	86.89	84.93	85.21	85.17
3. Feature extraction from VGG-Face and classification	134,272,835	75.07	79.44	69.51	79.35	73.24	79.86
4. Proposed MacularNet without attention module	66,359,619	95.44	91.52	96.77	92.76	95.43	91.98
5. Proposed MacularNet	18,995,107	97.45	94.18	98.27	94.62	97.33	94.69

Architecture	Evaluation protocols					
	5-fold cross-validation					
	Accuracy (%)		Precision (%)		Recall (%)	
	Dataset 1	Dataset 2	Dataset 1	Dataset 2	Dataset 1	Dataset 2
1. Fine-tuning output layer of VGG-Face	87.44 (+/- 1.37)	78.94 (+/- 1.03)	87.91 (+/- 1.64)	79.29 (+/- 0.86)	87.44 (+/- 1.37)	78.94 (+/- 1.03)
2. Fine-tuning of dense layers of VGG-Face	81.39 (+/- 13.71)	67.29 (+/- 7.94)	78.10 (+/- 21.21)	70.42 (+/-3.53)	74.47 (+/- 17.47)	67.29 (+/- 7.94)
3. Feature extraction from VGG-Face and classification	66.01 (+/- 4.88)	67.29 (+/- 7.94)	74.99 (+/- 3.30)	70.42 (+/- 3.53)	58.57 (+/- 5.93)	67.29 (+/- 7.94)
4. Proposed MacularNet without attention module	99.88 (+/- 0.25)	99.34 (+/- 1.05)	99.88 (+/- 0.24)	99.39 (+/- 0.94)	99.91 (+/- 0.19)	99.26 (+/- 1.21)
5. Proposed MacularNet	99.94 (+/- 0.12)	99.79 (+/- 0.37)	99.94 (+/- 0.12)	99.80 (+/- 0.35)	99.95 (+/- 0.09)	99.79 (+/- 0.38)

Table 2 Performance comparison of different methods on dataset 1. Here, + denotes motion blurred and shadowed scans not considered and * denotes RoI extracted. LPO stands for Leave patient(s) out and CV stands for cross-validation

Methods	Accuracy (%)	Precision (%)	Recall (%)
Wang et al. [47]+ (10-fold CV)	98.00	–	98.00
MCME [33]* (5-fold CV)	–	98.33	97.78
MacularNet (5-fold CV)	**99.94 (+/- 0.12)**	**99.94 (+/- 0.12)**	**99.95 (+/- 0.09)**
Surrogate assisted [34] (LPO)	88.45	–	–
MacularNet (LPO)	**97.45**	**98.27**	**97.33**

Table 3 Network performance comparisons on one test case of leave patient's out (LPO) protocol using different training strategies: transfer learning and training from scratch

Performance metric (in %)	MacularNet				MacularNet without attention			
	Transfer learning (epochs = 100)		Scratch (epochs = 300)		Transfer learning (epochs = 100)		Scratch (epochs = 300)	
	Dataset 1	Dataset 2	Dataset 1	Dataset 2	Dataset 1	Dataset 2	Dataset 1	Dataset 2
Accuracy	98.50	97.38	85.77	81.15	95.88	96.34	74.16	86.39
Precision	96.68	97.04	89.61	84.63	95.86	95.53	85.81	87.61
Recall	98.62	97.16	86.94	86.58	96.21	96.81	76.28	90.24

Table 4 Performance comparison of different methods on dataset 2 using 5-fold CV protocol

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score
RoI Extraction Used:				
WCME [32]	–	95.21 (+/- 3.2)	94.6 (+/- 3.4)	0.9458
MCME [33]	–	99.39 (+/- 1.21)	99.36 (+/- 1.33)	0.9934
No pre-processing:				
LACNN [8]	–	99.39 (+/- 1.49)	99.33 (+/- 1.49)	0.993
MacularNet	**99.79 (+/- 0.37)**	**99.80 (+/- 0.35)**	**99.79 (+/- 0.38)**	**0.997**

Table 5 Performance of MacularNet on both the datasets evaluated using two protocols

Parameters	Dataset 1		Dataset 2	
	LPO	5-fold CV	LPO	5-fold CV
Accuracy (%)	97.45	99.94	94.18	99.79
Precision (%)	98.27	99.94	94.62	99.80
Recall (%)	97.33	99.95	94.69	99.79
F1-score	0.973	0.999	0.941	0.997

Table 6 Comparison of performance of MacularNet on dataset 3 (UCSD) [18]

Methods	Accuracy (%)	Precision (%)	Recall (%)
LACNN [8]	90.10 (+/-1.40)	86.20 (+/-2.30)	86.80 (+/-1.30)
Multi-ResNet50 Ensembling [20]	90.40 (+/-1.20)	86.70 (+/-1.80)	87.20 (+/-1.40)
MacularNet	**92.60 (+/- 1.48)**	**90.27 (+/- 2.67)**	**88.52 (+/- 1.47)**

Results on Dataset 3

The UCSD dataset [18], referred to as dataset 3, is composed of 84484 OCT B-scans, comprising of 8866 drusen, 11598 DME, 37455 CNV and 26565 normal B-scans acquired from 4686 patients at the Shiley Eye Institute of the University of California San Diego (UCSD), thereby leading to a four-class classification problem. This is currently the largest and most challenging dataset publicly available in the macular OCT imaging community, obtained under Creative Commons Attribution 4.0 International license. In the case of dataset 3, the evaluation protocol of [8] is followed as mentioned earlier. Experimental results are reported in Table 6. It is to be noted that for the experiments over this dataset, the input size of macular scans were resized to 112 × 112, which reduces the model parameters and requires lesser memory for training and testing which becomes important for larger datasets. The outcomes of the proposed network, reported in Table 6, shows that the proposed method outperforms the existing methods of [8, 20], with respect to all three evaluation metrics. In addition to the performance, MacularNet requires approximately 16.56 millions weights, which is considerably lower than most of the existing deep learned networks. Some sample examples are shown in Table 1.

Results on Dataset 4

OCTID dataset [10], also referred to as dataset 4, consists of 5 classes: four macular pathological conditions and a normal class, obtained under the Creative Commons Attribution 4.0 International (CC BY 4.0) License. It comprises of 102 macular hole, 55 AMD, 107 diabetic retinopathy, 102 CSR and 206 normal retinal images. The 5-fold CV protocol is followed for training and testing of the MacularNet model using dataset 4, the results are noted in Table 7. So far, there is no reported work in the literature which has employed OCTID dataset to validate their classification methods; hence it is difficult to make any comparison with other methods. Nevertheless, the dataset has important macular pathological conditions as classes and our proposed method MacularNet obtains satisfactory results.

Ablation Study

Selection and Integration with Other Deep Networks

The performance of the proposed MacularNet architecture is not dependent on any specific deep learned framework. We conducted many experiments with the popular deep learned models such as VGG-Face [27], VGG16 [38], and ResNet50 [12]. By substituting the VGG-Face network with other deep networks such as, ResNet50 or VGG16, the performance achieved by the network is found to be similar in each case. The experimental results are reported in Table 8. It can be observed from the table that with ResNet50 and VGG16 as the base model, the results obtained in terms of accuracy, precision and recall are s to that of VGG-Face using the 5-fold CV protocol. Hence, we can deduce that the diagnostic accuracy of the MacularNet architecture does not depend on the selection of a deep pre-trained network. We have also reported the number of parameters required for these models for different cases in Table 8. We can notice that the network with VGG-Face and VGG16 have lesser parameters than the one with ResNet50 as the framework whereas the performance of VGG-Face-based network is higher than that of VGG-16. Thus, we chose VGG-Face as our base model for the MacularNet architecture to optimize the classification output and resource requirement.

Table 7 Performance evaluation of MacularNet on dataset 4 (OCTID) [10]

Accuracy (%)	Precision (%)	Recall (%)	AUC
93.12 (+/- 8.59)	92.08 (+/- 10.39)	91.09 (+/- 10.91)	98.51

Evolution of Attention Maps Over Epochs and Layers

The outputs of the intermediate layers of the proposed architecture are shown in Fig. 4, whereas the attention map of Conv 4_3 with respect to training epochs are visualized in Fig. 5. The convergence of the attention maps to the RPE region of macular B-scans can be observed with respect to the training epochs for all three classes. The proposed network prominently focuses on the blue colored regions of the feature maps. The deep architecture of MacularNet has been initialized with pre-trained weights which gives our model a leeway to focus on the uneven layers of the retina within the first few epochs. Figure 5a shows the evolution of the

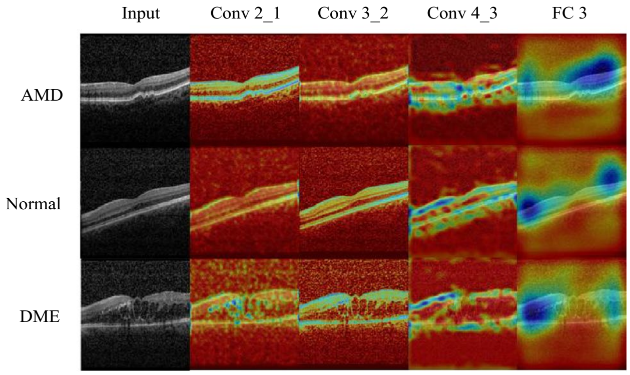

Fig. 4 Examples to show the transformation of feature maps with layers of the network for OCT images from NEH dataset. Names of the corresponding layer are mentioned on top of each column of images and the class of OCT scan is mentioned on left-hand side. Here, blue color denotes the highest attention whereas red denotes the lowest attention

Table 8 Performance of MacularNet with different base pre-trained deep learned networks

Pre-trained Network	Model Parameters (in millions)	Dataset 1			Dataset 2		
		Acuuracy (%)	Precision (%)	Recall (%)	Acuuracy (%)	Precision (%)	Recall (%)
VGG-Face	19	99.94	99.94	99.95	99.79	99.80	99.79
ResNet50	22.6	100	100	100	99.55	99.54	99.54
VGG16	19	99.72	99.79	99.71	98.61	98.74	98.71

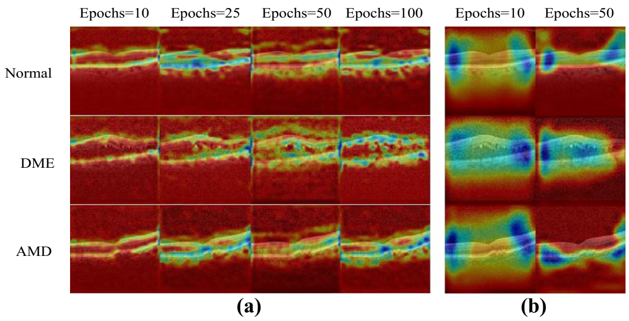

Fig. 5 **a** Shows feature maps across different training epochs (10, 25, 50, 100) for three classes, AMD, DME and normal. The images are extracted from the layer Conv 4_3 of the architecture. (b) Shows feature maps extracted from FC 3 layer after 10 and 50 epochs. It illustrates that the model gradually learns to focus on region in and around the RPE layers. Here, blue color denotes the highest attention being given while red denotes the lowest attention

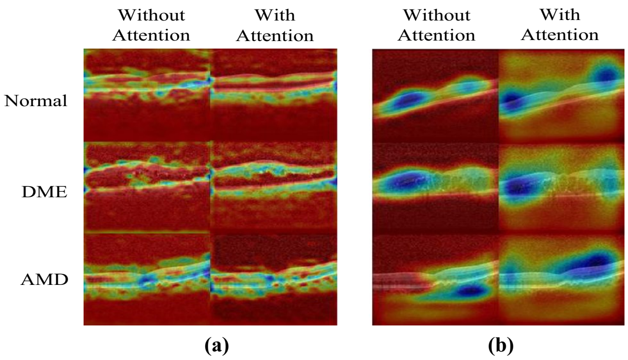

Fig. 6 **a** Examples of three class attention maps extracted from layer Conv 4_3 of MacularNet and the same without the attention module to show the effect of attention mechanism in our architecture. In **b** same analysis is presented for layer FC 3 of the network. Here, blue color denotes the highest attention while red denotes the lowest attention.

feature maps of Conv 4_3 layer (refer to Fig. 2 for entire architecture) after 10, 25, 50 and 100 epochs. Meanwhile, Fig. 5b illustrates the feature maps obtained from the FC 3 layer after 10 and 50 epochs. Figure. 4 depicts the layer-wise progression of the network with the help of feature maps on NEH dataset. The visualization of the intermediate layers shows the localized shift of the attention of the architecture from the first to the final layer. It is evident from Figs. 4 and 5 that the attention coefficients are trained to extract features from the region in and around the RPE layer of retina. It can be analyzed that attention of the the proposed network converges in the relevant regions of the feature maps of input images.

Effect of Attention Mechanism Over Discriminative Properties

In continuation of our investigation, we study the effects of the attention module in the classification task. Fig. 6 demonstrates the impact of the attention mechanism using the feature maps of two different layers. Fig. 6 (a) signifies that the attention coefficients in the Conv 4_3 layer converge towards the RPE layer and its surroundings in the proposed network. Furthermore, Fig. 6 (b) displays the feature maps for FC 3 layer; where again it is observed that, for all three classes, the significant features are given higher priority by the attention-based network. In case of AMD affected macula, the network without attention pays more consideration to non-relevant region below the RPE layer. Whereas, for the attention-based MacularNet, the concentration is in the required region. This shows the appropriate convergence pattern of the developed network. Hence, the inclusion of the attention layer reinforces our proposed method to yield noteworthy outcomes without the requirement of any supplementary measures. It can be observed from Table 1 that, the inclusion of this module improves the performance of the architecture while reducing the number of parameters.

Network Parameters and Computational Complexity

Table 1 shows that the incorporation of attention mechanism reduces the required number of model parameters. MacularNet has around 19 million parameters which are approximately 7 times lesser than baseline VGG network, 3.5 times lower than the MacularNet without attention and 1.24 times lower than the deep CNN employed in [16] for OCT macular image classification. The computational complexity of the proposed network for training is lower than the baseline architectures and the deep CNN of [16]. This is due to the reduced number of model parameters as well as reduced computational complexity. In terms of floating-point operations, MacularNet requires 37 million computations of each sample, whereas MacularNet without attention and VGG-Face need 132 and 289 million operations, respectively. Likewise, the inception model designed for macular OCT classification in [16] requires 47 million similar computations. Hence, even with the added attention module in the proposed model, the complexity is reduced which leads to lesser training and testing time. Moreover, with the exclusion of pre-processing steps, the time complexity is further reduced. For instance, with the denoising step in [34] the total time required for processing of a volume of 97 OCT scans is of 10 minutes approximately. Hence, with the removal of these steps, the processing time of our method is comparatively lower. In the testing phase, each sample of input size $224 \times 224 \times 3$ needs a minimal time of 117 μs for the proposed model whereas the testing time required for the model in [20] is 870 μs (micro seconds) per scan. The reduced computational complexity has an added advantage of lesser computational resource requirement during the training and testing of the model.

Analysis and Significance of the Proposed Approach

Advantages of not Using RoI

Unlike most of the existing works in the literature [16, 32, 33, 39], our approach eliminates the need for RoI extraction. The existing RoI extraction stage results in the network being dependent on the manual localization of the FFRs or irregularities in the RPE and choroid layer below it. It leads to the potential problem of correspondence matching in representation learning. Although the CNNs are translation invariant, the RoI selection step drives the CNN to loose its property of matching ridge points with the features from entire image. Thus, the inclusion of RoI extraction reduces the overall automation and universality of the method. Consequently, we developed a non-RoI extraction-based method which uses the deformation-aware attention mechanism to focus on important discriminative regions in the macular OCT images. The performance of our network even with the inclusion of the RoI selection step is similar to one without RoI selection. Thus, we eliminated this stage to avoid the problem of correspondence matching and automate the entire process without affecting the performance.

Table 9 reports the results for the implementation of the MCME model [33] with and without the RoI extraction stage using categorical cross-entropy as the objective function. It is evident from Table 9 that excluding the RoI cropping step weakens the performance of the MCME method. Our proposed method procures better results without extracting the relevant region manually when compared to the MCME results. Moreover, with RoI selection, our approach performs very similar to that of without using RoI selection, which is 100% accuracy on the dataset 1 and 99.84% accuracy on dataset 2 using 5-fold cross-validation protocol as shown in Table 9. The incorporation of the attention mechanism in the MacularNet removes the requirement RoI selection in our method. It enables the network to concentrate on the deformations encoding their morphological structure variations and put less emphasis on the background without external assistance.

Avoiding the Preprocessing: Denoising and Flattening

In the macular OCT imaging literature, sparse representation based denoising [34] and BM3D denoising [39, 41], have been followed. In our approach, the denoising stage has been avoided to save time and resources without affecting the efficacy of the proposed method. Table 10 shows the classification accuracy of the MacularNet with and without the BM3D denoising step. The experimental results and comparisons using both the protocols show that exclusion of the denoising step does not affect the performance of our proposed framework, which infers that it is robust to the noise introduced while capturing OCT scans.

Retinal flattening is performed to flatten the retinal curvature in [32, 33, 39], which eases the detection of the atrophies in the RPE layer and the region surrounding it. The CNNs are not completely rotation invariant; so the flattening assists the network to counter this issue. Our approach learns to encode the intricate details around the deformed layers of the macula to extract important discriminative information as shown in Fig. 5. The proposed network is trained to self-locate the RPE layer in accordance with its curvature without undergoing any flattening algorithm.

These pre-processing steps are dataset specific and not suitable for the generality of the algorithms for real-world applications. These stages are not scalable in practice which makes it difficult to address affine variations in different datasets. MacularNet circumvents these disadvantages of the

Table 10 Classification accuracies of MacularNet with and without the denoising step on both the datasets (in %)

Datasets	With Denoising		Without Denoising	
	5-fold CV	LPO	5-fold CV	LPO
Dataset 1	99.94	94.89	99.94	97.45
Dataset 2	99.34	93.34	99.79	94.18

Table 9 Performance comparison (%) of the proposed MacularNet with MCME model [33] with and without the RoI extraction step using 5-fold cross validation (CV)

Performance Metrics	Dataset 1				Dataset 2			
	MCME		MacularNet		MCME		MacularNet	
	With RoI	Without RoI	With RoI	Without RoI	With RoI	Without RoI	With RoI	Without RoI
Accuracy	99.04	98.40	100	99.94	90.95	87.31	99.84	99.79
Precision	99.04	98.42	100	99.94	91.08	87.93	99.86	99.80
Recall	99.04	98.40	100	99.95	90.96	87.31	99.83	99.79

existing techniques, encodes the important regions automatically and achieves better classification performance.

Training from Scratch vs. Fine-Tuning

The weights of the base model used in our framework are pre-trained with millions of respective relevant images which allow the model to easily learn the complex, non-linear and uneven features of an image. Refining and fine-tuning help to train the model specifically for the macular OCT scans. To interpret the benefits of transfer learning in our problem, in Figs. 7a & b we visualize the difference in attention maps for the fine-tuning and training from scratch. In Fig. 7a, it can be observed that after 10 epochs of fine-tuning, the network learns to concentrate on the relevant region. During training the model from scratch, it captures the entire macular part of the OCT B-scan. Figure 7b illustrates that the feature maps of the fine-tuned network are more converged to significant discriminative regions than training the network from scratch, even after 100 or 300 epochs. The validation accuracy while training the model from scratch converges in 300 epochs and hence we stopped the training further. It can be inferred from these feature maps that transfer learning helps the proposed architecture to detect the presence of macular pathologies more aptly with lesser time and resource requirement while training. It is also verified from the experimental results shown in Table 3 that transfer learning (fine-tuning) helps in improving the classification accuracy significantly on both the datasets using both the protocols.

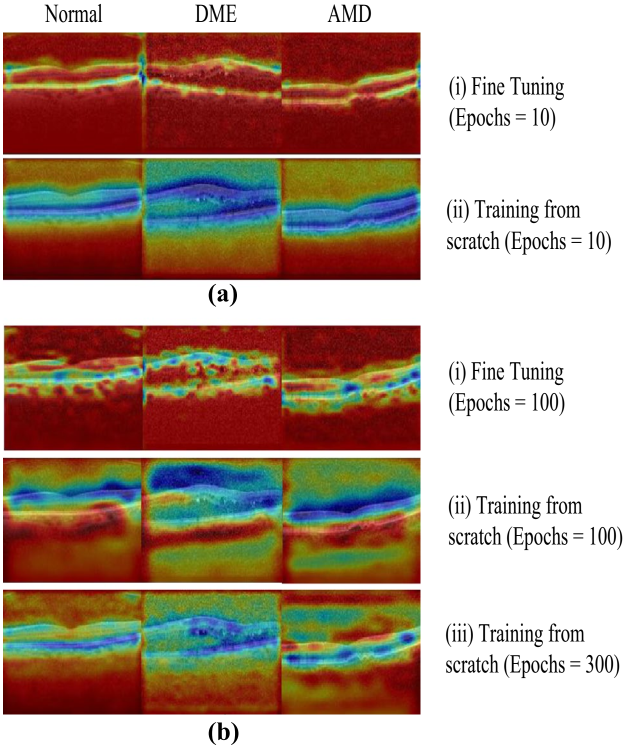
Normal
DME
AMD
(i) Fine Tuning (Epochs = 10)
(ii) Training from scratch (Epochs = 10)
(a)
(i) Fine Tuning (Epochs = 100)
(ii) Training from scratch (Epochs = 100)
(iii) Training from scratch (Epochs = 300)
(b)

Fig. 7 Examples to convey the information of variations in attention maps for the pre-trained base model of MacularNet and training the network from scratch. **a** shows the maps after 10 epochs and **b** depicts the same maps after 100 epochs in (i) and (ii) and after 300 epochs of training from scratch in (iii). It can be interpreted from the images that the mechanism of transfer learning is a better alternative for our problem. Here, blue color denotes the highest attention while red denotes the lowest attention

Conclusions and Future Works

In this paper, we have addressed the problems of classifying macular OCT images into normal, AMD, DME, and CNV. Our proposed approach MacularNet takes advantage of the transfer learning-based deep learned models and attention arising from locally deformation-aware features with improved discrimination ability for classifying macular OCT images. Initialization of a fine-tuned deep CNN model coupled with attention mechanism has helped to extract discriminative features in macular OCT images. The proposed MacularNet has been analyzed carefully with the evolution of attention maps over epochs and layers, over training procedures and effect of attention over discriminative properties.

All these efforts have led to improved classification accuracy, decrease in the training time, usage of lesser computational resources and reduction in the number of parameters of the base CNN model. Unlike most of the existing works, MacularNet does not require any preprocessing steps, such as RoI extraction, denoising or retinal flattening. It is well-suited with multiple deep learning frameworks, end-to-end trainable with a reduced number of network parameters, making it fully automatic, that facilitates efficient learning of relevant local deformation-aware features for classifying OCT images. The proposed attention mechanism does not require any external assistance such as bounding boxes or segmentation maps. Ablation studies with extensive experimental results and analysis on four datasets show that our proposed MacularNet approach achieves state-of-the-art performance.

In future works, the effectiveness of MacularNet with deformation-aware attention-based architecture can be further investigated for detection of several other complicated macular diseases, such as retinal vein occlusion, retinitis pigmentosa and macular telangiectasia [9]. Moreover, specific investigation of our proposed MacularNet method on macular eye diseases occurring at earlier stages (on younger population) might be useful, where the deformations are more subtle and difficult to differentiate and evaluation on larger OCT datasets. We anticipate that the researchers/practitioners can use the deep CNN based attention mechanism proposed within the MacularNet framework for several other tasks of image classification tasks, especially in the field of biomedical engineering, such as analysis of x-ray images for chest/limbs, diagnosis of osteoarthritis through shape or structure analysis of limb joints, and magnetic resonance imaging (MRI) scans of organs and tissues.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest. The authors declare that they have no conflict of interest.

References

1. Aghaei A, Nazari A, Moghaddam ME. Sparse deep lstms with convolutional attention for human action recognition. SN Comput Sci. 2021;2(3):1–14.
2. Alzahrani Y, Boufama B. Biomedical image segmentation: A survey. SN Computer Science. 2021;2(4):1–22.
3. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate 2014.
4. Das V, Dandapat S, Bora P. Multi-scale deep feature fusion for automated classification of macular pathologies from oct images. Biomed Signal Process Control. 2019;54:101605. https://doi.org/10.1016/j.bspc.2019.101605.
5. Das V, Dandapat S, Bora PK. A data-efficient approach for automated classification of oct images using generative adversarial network. IEEE Sens Lett. 2020;4(1):1–4.
6. Das V, Prabhakararao E, Dandapat S, Bora PK. B-scan attentive cnn for the classification of retinal optical coherence tomography volumes. IEEE Signal Process Lett. 2020;27:1025–9. https://doi.org/10.1109/LSP.2020.3000933.
7. Fang L, Jin Y, Huang L, Guo S, Zhao G, Chen X. Iterative fusion convolutional neural networks for classification of optical coherence tomography images. J Vis Commun Image Rep. 2019;59:327–33.
8. Fang L, Wang C, Li S, Rabbani H, Chen X, Liu Z. Attention to lesion: Lesion-aware convolutional neural network for retinal optical coherence tomography image classification. Image IEEE Trans Med. 2019.
9. Fauw J, Ledsam J, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, Askham H, Glorot X, O'Donoghue B, Visentin D, Driessche G, Lakshminarayanan B, Meyer C, Mackinder F, Bouton S, Ayoub K, Chopra R, King D, Karthikesalingam A, Ronneberger O. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med 24, 2018. https://doi.org/10.1038/s41591-018-0107-6
10. Gholami P, Roy P, Parthasarathy MK, Lakshminarayanan V. Octid: Optical coherence tomography image database. Comput Electr Eng. 2020;81:106532.
11. Graves A, Wayne G, Danihelka I. Neural turing machines. arXiv preprint arXiv:1410.5401; 2014.
12. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Confe. comput. vision and pattern recognit., pp. 770–778; 2016.
13. Hinton GE. Training products of experts by minimizing contrastive divergence. Neural Comput. 2002;14(8):1771–800.
14. Huang D, Swanson EA, Lin CP, Schuman JS, Stinson WG, Chang W, Hee MR, Flotte T, Gregory K, Puliafito CA, et al. Optical coherence tomography. Science. 1991;254(5035):1178–81.
15. Jetley S, Lord NA, Lee N, Torr PH. Learn to pay attention. arXiv preprint arXiv:1804.02391; 2018.
16. Ji Q, He W, Huang J, Sun Y. Efficient deep learning-based automated pathology identification in retinal optical coherence tomography images. Algorithms. 2018;11(6):88.
17. Karri SPK, Chakraborty D, Chatterjee J. Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration. Biomed Opt Exp. 2017;8(2):579–92.
18. Kermany D, Goldbaum M, Cai W, Valentim C, Liang HY, Baxter S, McKeown A, Yang G, Wu X, Yan F, Dong J, Prasadha M, Pei J, Ting M, Zhu J, Li C, Hewett S, Dong J, Ziyar I, Zhang K. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell. 2018;172:1122-1131.e9. https://doi.org/10.1016/j.cell.2018.02.010.
19. Kolluru C, Prabhu D, Gharaibeh Y, Bezerra H, Guagliumi G, Wilson D. Deep neural networks for a-line-based plaque classification in coronary intravascular optical coherence tomography images. J Med Imaging. 2018;5(4):044504.
20. Li F, Chen H, Liu Z, dian Zhang X, shan Jiang M, zheng Wu Z, qian Zhou K. Deep learning-based automated detection of retinal diseases using optical coherence tomography images. Biomed Opt Express 2019; 10(12): 6204–6226. https://doi.org/10.1364/BOE.10.006204. http://www.osapublishing.org/boe/abstract.cfm?URI=boe-10-12-6204
21. Lima DM, Rodrigues-Jr JF, Brandoli B, Goeuriot L, Amer-Yahia S. Dermadl: advanced convolutional neural networks for computer-aided skin-lesion classification. SN Comput Sci. 2021;2(4):1–13.
22. Liu YY, Chen M, Ishikawa H, Wollstein G, Schuman JS, Rehg JM. Automated macular pathology diagnosis in retinal OCT images using multi-scale spatial pyramid and local binary patterns in texture and shape encoding. Med Image Anal. 2011;15(5):748–59.
23. Mehta S. Age-related macular degeneration. Primary Care: Clin Off Pract. 2015;42(3):377–91.
24. Mishra SS, Mandal B, Puhan NB. Multi-level dual-attention based CNN for macular optical coherence tomography classification. IEEE Signal Process Lett. 2019;26(12):1793–7.
25. Nicholson B, Noble J, Forooghian F, Meyerle C. Central serous chorioretinopathy: update on pathophysiology and treatment. Surv Ophthalmol. 2013;58(2):103–26.
26. Panda R, Puhan NB, Mandal B, Panda G. Glauconet: patch-based residual deep learning network for optic disc and cup segmentation towards glaucoma assessment. SN Comput Sci. 2021;2(2):99. https://doi.org/10.1007/s42979-021-00491-1.
27. Parkhi OM, Vedaldi A, Zisserman A, et al. Deep face recognition In: BMVC. 2015;1:6.
28. Pershing S, Enns EA, Matesic B, Owens DK, Goldhaber-Fiebert JD. Cost-effectiveness of treatment of diabetic macular edema. Ann Internal Med. 2014;160(1):18.

29. Qiu J, Sun Y. Self-supervised iterative refinement learning for macular oct volumetric data classification. Comput Biol Med. 2019;111:103327. http://www.sciencedirect.com/science/article/pii/S0010482519301969
30. Rangrej SB, Sivaswamy J. Assistive lesion-emphasis system: an assistive system for fundus image readers. J Med Imaging. 2017;4(2):024503.
31. Rao SS, Ikram S, Ramesh P. Deep learning-based image retrieval system with clustering on attention-based representations. SN Comput Sci. 2021;2(3):1–16.
32. Rasti R, Mehridehnavi A, Rabbani H, Hajizadeh F. Wavelet-based convolutional mixture of experts model: An application to automatic diagnosis of abnormal macula in retinal optical coherence tomography images. In: 2017 10th Iranian Conf. Machine Vision and Image Process. (MVIP), 2017; pp. 192–196. IEEE.
33. Rasti R, Rabbani H, Mehridehnavi A, Hajizadeh F. Macular OCT classification using a multi-scale convolutional neural network ensemble. IEEE Trans Med Imaging. 2017;37(4):1024–34.
34. Rong Y, Xiang D, Zhu W, Yu K, Shi F, Fan Z, Chen X. Surrogate-assisted retinal OCT image classification based on convolutional neural networks. IEEE J Biomed Health Inf. 2018;23(1):253–63.
35. Roy K, Chaudhuri SS, Roy P, Chatterjee S, Banerjee S. Transfer learning coupled convolution neural networks in detecting retinal diseases using oct images. In: Intelligent Computing: Image Processing Based Applications, pp. 153–173. Springer 2020.
36. Schlemper J, Oktay O, Schaap M, Heinrich M, Kainz B, Glocker B, Rueckert D. Attention gated networks: learning to leverage salient regions in medical images. Med Image Anal. 2019;53:197–207.
37. Schuman JS, Puliafito CA, Fujimoto JG, Duker JS. Optical coherence tomography of ocular diseases. Slack New Jersey: 2004.
38. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014.
39. Srinivasan PP, Kim LA, Mettu PS, Cousins SW, Comer GM, Izatt JA, Farsiu S. Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images. Biomed Opt Exp. 2014;5(10):3568–77.
40. Sultana NN, Mandal B, Puhan NB. Deep regularized discriminative network. SN Comput Sci. 2021;2(4):235. https://doi.org/10.1007/s42979-021-00647-z.
41. Sun Y, Li S, Sun Z. Fully automated macular pathology detection in retina optical coherence tomography images using sparse coding and dictionary learning. J Biomed Opt. 2017;22(1):016012.
42. Sunija A, Kar S, Gayathri S, Gopi VP, Palanisamy P. Octnet: a lightweight cnn for retinal disease classification from optical coherence tomography images. Comput Methods Programs Biomed. 2021;200:105877.
43. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, Liang J. Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans Med Imaging. 2016;35(5):1299–312.
44. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Adv. Neural Inf. Process. Sys., 2017; pp. 5998–6008.
45. Wang D, Wang L. On oct image classification via deep learning. IEEE Photon J. 2019;11(5):1–14.
46. Wang X, Peng Y, Lu L, Lu Z, Summers RM. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In: Proc. IEEE conf. comput. vis. pattern recognit., 2018; pp. 9049–9058
47. Wang Y, Zhang Y, Yao Z, Zhao R, Zhou F. Machine learning based detection of age-related macular degeneration (amd) and diabetic macular edema (dme) from optical coherence tomography (OCT) images. Biomed Opt Exp. 2016;7(12):4928–40.
48. Wen G, Rodriguez-Niño B, Pecen FY, Vining DJ, Garg N, Markey MK. Comparative study of computational visual attention models on two-dimensional medical images. J Med Imaging. 2017;4(2):025503.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.