Neural Computing and Applications (2019) 31:9295–9305
https://doi.org/10.1007/s00521-019-04281-y

ORIGINAL ARTICLE

Context-aware attention network for image recognition

Jiaxu Leng[1,2] · Ying Liu[1,2,3] · Shang Chen[4]

Received: 4 February 2019 / Accepted: 30 May 2019 / Published online: 18 June 2019

Abstract
Existing recognition methods based on deep learning have achieved impressive performance. However, most of these algorithms do not fully utilize the contexts and discriminative parts, which limit the recognition performance. In this paper, we propose a context-aware attention network that imitates the human visual attention mechanism. The proposed network mainly consists of a context learning module and an attention transfer module. Firstly, we design the context learning module that carries on contextual information transmission along four directions: left, right, top and down to capture valuable contexts. Second, the attention transfer module is proposed to generate attention maps that contain different attention regions, benefiting for extracting discriminative features. Specially, the attention maps are generated through multiple glimpses. In each glimpse, we generate the corresponding attention map and apply it to the next glimpse. This means that our attention is shifting constantly, and the shift is not random but is closely related to the last attention. Finally, we consider all located attention regions to achieve accurate image recognition. Experimental results show that our method achieves state-of-the-art performance with 97.68% accuracy, 82.42% accuracy, 80.32% accuracy and 86.12% accuracy on CIFAR-10, CIFAR-100, Caltech-256 and CUB-200, respectively.

Keywords Convolution neural network · Context learning · Attention transfer

1 Introduction

Image recognition is an important research direction in computer vision. It is the basis of other complex visual problems such as object detection, image segmentation, scene understanding and so on. Although image recognition has been studied for decades, its performance in complex scenes is still unsatisfactory, especially in the face of categories with marginal visual differences.

In recent years, convolutional neural networks (CNNs) have been developed rapidly and have achieved good performance in many fields, such as object tracking [1, 2], image segmentation [3, 4] and object detection [5–7]. In view of the good feature extraction ability of convolution neural network, a substantial number of recognition algorithms based on convolution neural network have been proposed. At present, the popular image recognition methods [7–9] enhance the recognition performance by deepening the depth of convolution neural network with the cost of speed. In addition, it is difficult for these methods to recognize categories with similar characteristics.

The processing of human information acquiring is an important psychological adjustment activity. In general, the actual scene contains not only the target of interest, but also a lot of interference information. Cognitive psychology research shows that the human visual system adopts a serial computing strategy when analyzing complex input scenes, that is, using selective attention mechanism that selects a specific region of the scene according to the local characteristics of the image and moves the region to a high-resolution retina through rapid eye movement scanning so as to make more careful observation and analysis. Visual

✉ Ying Liu
yingliu@ucas.ac.cn

1 School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 101400, China

2 Data Mining and High Performance Computing Lab, Chinese Academy of Sciences, Beijing 101400, China

3 Key Lab of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing 100190, China

4 School of Information and Communication, Guilin University of Electronic Technology, Guilin 541004, China

attention mechanism helps the brain filter out distracting information and focus on the target of interest.

Inspired by the human attention mechanism, the attention mechanism has been widely used in computer vision [10–12]. Attention is used not only to select the location of focus, but also to enhance the feature representations of objects. Despite great success, these methods still do not fully imitate the human visual attention mechanism, which limits the performance. Based on the study of human visual attention mechanism, we propose a context-aware attention network that fully utilizes the valuable contexts and generate different attention regions that contain discriminative features for accurate image recognition. The proposed network mainly consists of two parts: context learning and attention transfer modules. We first design a context learning module to capture the global contexts, which augments the feature representation and benefits for initializing the attention region. Then, an attention transfer module is presented for achieving accurate attention localization. Different from previous attention-based approaches [9, 13, 14], our attention regions are not generated by one glimpse, but by multiple glimpses. In each glimpse, we generate the corresponding attention map and apply it to the next glimpse. This means that our attention is shifting constantly, and the shift is not random but is closely related to the last attention. Finally, we consider utilizing the discriminative features of the located attention regions obtained by multiple glimpses to achieve accurate image classification. In addition to the more discriminative feature representation brought about by our attention module, our model also has attractive features: (1) because different attention regions are captured in different glimpses, the classification performance can be improved by stacking glimpses. (2) It is general and can be applied to various advanced networks.

The main contributions of our work are summarized as follows:

- We propose a context learning module for effectively learning the contextual information, which enhances the performance of image recognition.
- We propose an attention transfer module, which imitates human visual attention mechanism. The proposed attention transfer module can capture multiple attention regions with discriminative features. In addition, a strategy of attention inhibition is adopted in the attention transfer module, which enables us to generate different attention regions in each glimpse.
- We conduct extensive experiments and visual analysis on four datasets: CIFAR-10, CIFAR-100, Caltech-256 and CUB-200. Compared with the current state-of-the-art methods, our method performs better on these four datasets.

The rest of this paper is organized as follows: in Sect. 2, we provide a brief review of related research on image recognition. In Sect. 3, we propose our context-aware attention network. In Sect. 4, we evaluate our proposed context-aware attention network experimentally and compare it with state-of-the-art methods in the literature. Finally, we provide our conclusions in Sect. 5.

2 Related work

Attention in deep learning is derived from the attention mechanism of the human visual system. When the human brain receives external information, such as visual information and auditory information, it does not process and understand all the information, but only pays attention to some of the significant or interesting information, which helps filter out interference information and improve the efficiency of information processing.

Inspired by human visual attention mechanism, many algorithms have been proposed to imitate human attention mechanism. Recently, tentative efforts [15–22] have been made toward applying attention into deep neural network. Deep Boltzmann machine (DBM) [23] contains top-down attention by its reconstruction process in the training stage. Attention mechanism has also been widely applied to recurrent neural networks (RNN) and long short-term memory (LSTM) [24] to tackle sequential decision tasks [25–27]. Attention mechanism has various forms of implementation, which can be roughly divided into soft attention and hard attention. One of the most representative hard attentions is recurrent attention model (RAM) [11], which processes the input in time sequence and locates the attention region in the image. The model reduces the interference of unnecessary information and the influence of noise, while reducing the computation cost. Because recognition models based on hard attention need to predict the region of focus, reinforcement learning is usually used in training, which leads to convergence difficulties. The models based on soft attention, that is differentiable, can be trained by back-propagation. Considering the advantage of soft attention that is easy to train, many recognition algorithms [28, 29] based on soft attention are proposed. Two-level attention network (TLAN) [13] applies visual attention to the fine-grained classification problem using DNN. The TLAN integrates three attention models: bottom-up (candidate patch), object-level top-down (certain object-related patch) and part-level top-down to locate discriminative parts. Fully convolutional attention network (FCAN) [14] introduces a reinforcement learning-based fully convolutional attention localization network to adaptively select multiple task-driven visual attention regions. Recurrent attention convolutional neural network (RA-

CNN) [30] proposed by Fu J et al. learns discriminative attention regions and region-based feature representation at multiple scales in a mutually reinforced way. Recently, F. Wang et al. present a residual attention network (RAN) [31] that uses a residual approach similar to ResNet [9] to optimize and learn very deep networks. However, despite these efforts, the improvement on the attention mechanism is still needed.

To better imitate human visual attention mechanism, we propose a context-aware attention network. We first design a context learning module that carries on contextual information transmission along the four directions: left, right, top and down so as to capture valuable contexts. Then, we propose an attention transfer module to capture attention regions that augment the feature representation. Different from previous attention-based approaches, our attention regions are not generated by one glimpse, but by multiple glimpses. In each glimpse, we generate the corresponding attention map and apply it to the next glimpse. This means that our attention is shifting constantly, and the shift is not random but is closely related to the last attention. Finally, we consider utilizing the discriminative features of the located attention regions obtained by multiple glimpses to achieve accurate image classification.

3 Context-aware attention network

In this section, we will introduce our proposed context-aware attention network for accurate image recognition. We consider the attention transfer network with three glimpses as an example in Fig. 1, and more glimpses can be stacked in a similar way. Figure 1 shows the detailed architecture of our proposed context-aware attention network that consists of the context learning module and attention transfer module. First, the input image is fed into our context learning module to capture context-aware feature representation. Second, the output of the context learning module is considered as the input of our attention transfer module to locate the attention regions. There are three glimpses in the attention transfer module, and each glimpse predicts an attention map by fully convolution and sigmoid layers. Finally, our proposed network is optimized to convergence by a softmax classification loss.

To better introduce the proposed context learning module, we resort to a mathematical formulation of the task at hand. Given an input image X, we first extract the context-aware features by feeding the images into our context learning module. The feature representations are denoted as $g(X)$, where $g(\cdot)$ denotes a set of operations of convolution, pooling and activation. Then, we set the output of our context learning module as the input of our attention transfer module that generates an attention map by the operation $h(\cdot)$. Finally, we obtain the feature representation r of the input X for image recognition by the following formulation:

$$r(X) = f(g(X) \odot h(g(X))), \tag{1}$$

where $\odot$ denotes the operation of element-wise product. $f(\cdot)$ represents fully connected layers to map convolutional features to a feature vector that could be matched with the category entries, as well as includes a softmax layer to further transform the feature vector to probabilities.

3.1 Context learning module

Context helps us understand the image! Apart from strong psychological evidence [6, 32, 33], approving the contextual information is important for humans to recognize objects; many empirical studies [4, 34–36] in the field of computer vision have also suggested that recognition algorithms can be improved by proper modeling of context.

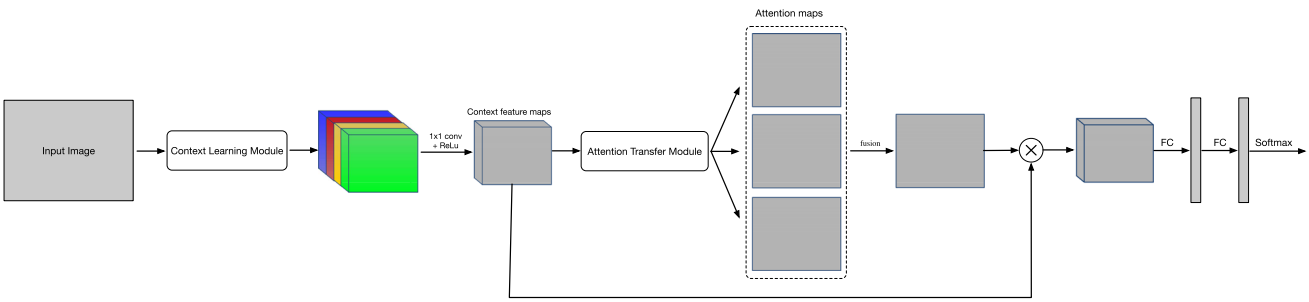

Fig. 1 The architecture of our proposed context-aware attention network. The context learning and attention transfer modules are used to capture the valuable contexts and attention regions in the image, respectively. First, the context learning module carries on context information transmission along the four directions that outputs four kinds of feature maps colored with blue, red, yellow and green. The attention transfer module then generates three attention maps by three glimpses, where each attention map obtains one attention region with discriminative features. Next, the context feature maps and attention map are integrated to generate the final feature maps for image recognition by element-wise product. Finally, the output feature maps are processed by two fully connected (fc) and softmax layers that predict the category (color figure online)

It has been well recognized in the vision community for years that the contextual information enhances object recognition [7, 32, 33].

To effectively learn contextual information, we design a context transfer module that carries on context information transmission along the four directions: *left*, *right*, *up* and *down*. The context transfer operation can be expressed as:

$$C_{i,j}^{up} = \max\left(W_{i-1,j}^{up} C_{i-1,j}^{up} + C_{i,j}^{up}, 0\right) \quad (2)$$

The above equation shows the transmission processing in *up* direction, and the other directions conduct the similar operation. In Eq. 1, $C_{i,j}^{up}$ is a cell of the input feature maps and our goal is to update it. $W_{i-1,j}^{up}$ is a transfer parameter and it is restricted between 0 and 1. It is worth noting that the parameter $W_{i-1,j}^{up}$ is learning but not manual setting. By Eq. 2, we can obtain context feature maps $g(X) = concat(C^{left}, C^{right}, C^{up}, C^{down})$ that contains not only the original convolution features, but also the context features transmitted from all the other cells, which is helpful for our image recognition task.

Our proposed context learning module is shown in more detail in Fig. 2. It mainly contains two parts: convolution feature extraction and context transmission. The input image first is convoluted, and output convolution feature maps are set as the input of the context transmission. Then, the convolution feature maps replicated four copies, which are used to learn the contextual information of the four directions (see Eq. 2). For each direction, we obtain the corresponding context feature maps. After that, we concatenate these four directions of context feature maps and finally use $1 \times 1 conv + ReLu$ to fuse these feature maps.

3.2 Attention transfer module

When looking at a picture or a scene, we do not pay our attention evenly across each region. Usually, we first quickly locate some salient regions, and then based on these regions, we spread and divert our attention. In order to imitate this visual mechanism of human beings, we design an attention transfer model, which generates different attention maps by multiple glimpses. Each generated attention map contains its own unique attention region. In addition, these regions from multiple glimpses are not independent of each other, but are mutually restricted and relevant. This means that there exists reasoning relations between attention regions.

Suppose that the context feature maps $g(X)$ are obtained by the proposed context learning module, we feed it into our attention transfer module and generate the expected attention map. It is worth noting that the attention map is not obtained by only one glimpse but by multiple glimpses. Specifically, the process of generating each attention map can be expressed as follow:

$$\begin{aligned} FM_t(X) &= FM_{t-1}(X) * (1 - AM_{t-1}(X)) \\ AM_t(X) &= l(FM_t(X)) \end{aligned} \quad (3)$$

where $FM_t(X)$ presents the input feature maps of the tth glimpse and $AM_t(X)$ is the generated attention map in the tth glimpse. $1 - AM_{t-1}(X)$ is the new attention map obtained by inhibiting $AM_{t-1}(X)$. The pixel values of the attention map are between 0 and 1. Equation 3 shows that the $FM_t(X)$ generated by the tth glimpse are closely related to the previous attention maps generated by the $(t-1)$th glimpse. The previous attention map $AM_{t-1}(X)$ is inhibited by using 1 to subtract each pixel value of the attention map.

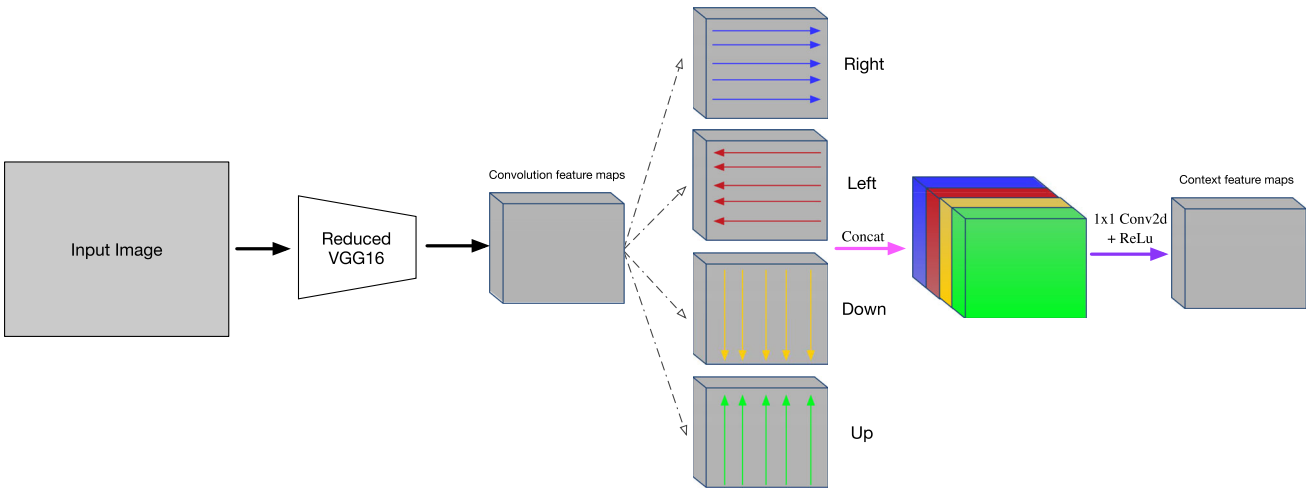

Fig. 2 The details of our context learning module. First, convolution features of the input image are extracted by a set of convolution and max-pooling layers. Then, four kinds of context feature maps are learned by performing context transmission (see the blue, red, yellow and green arrows). We further merge these context feature maps and output the final context feature maps with the same shape as the input convolution feature maps (color figure online)

The feature maps $FM_t(X)$ are then generated by element-wise product of the $FM_{t-1}(X)$ and the new attention maps $(1 - AM_{t-1}(X))$. This operation ensures that our module pays attention to different regions and the attention is constantly shifting, which is benefit for enhancing the recognition performance. $l(\cdot)$ is an operator for learning the attention map and its input is the generated $FM_t(X)$. The $l(\cdot)$ mainly contains a set of convolution, down-sampling and up-sampling operations.

Figure 3 shows the detailed architecture of the proposed attention transfer module. It mainly consists of three parts: attention learning, transfer and fusion. First, the attention map is generated by a fully convolutional neural network that contains contracting path and expanding path. The contracting path is first used to capture local contexts, and then, the corresponding expanding path generates the attention map. The contracting path follows a typical convolution network architecture, i.e., alternate convolution and pooling operations. The feature map is down-sampled, and the number of the feature map is increased layer by layer. Each stage of the expanding path consists of an up-sampling of feature maps and a convolution of the symmetrical position feature maps from contracting path. The expanding path can increase the resolution of the feature maps. For localization, the expanding path combines the up-sampling feature maps with the high-resolution feature maps from the contracting path by the skip connection. The output of the network is the attention map that is a pixel by pixel mask, showing the attention weight of each pixel of the input. Considering that human beings capture useful information by not one glimpse but multiple glimpses that focus on different locations with gaze transfer, our final attention map also is not generated directly. To make each glimpse output different attention regions, an inhibition strategy (see Eq. 3) is used and the inhibited attention map acts on the next glimpse. It can be seen that three attention maps are generated by three glimpses and the whole process of attention transfer is shown in Fig. 3. Finally, we achieve the final attention by merging the three attention maps.

3.3 Implementation details

The proposed method is implemented in Python. We perform the experiments using a Xeon E5-1620 3.70 GHz processor with 32 GB RAM, and all experiments are conducted under the deep learning framework of Tensorflow. In addition, a Titan X is used to accelerate model training in our all experiments.

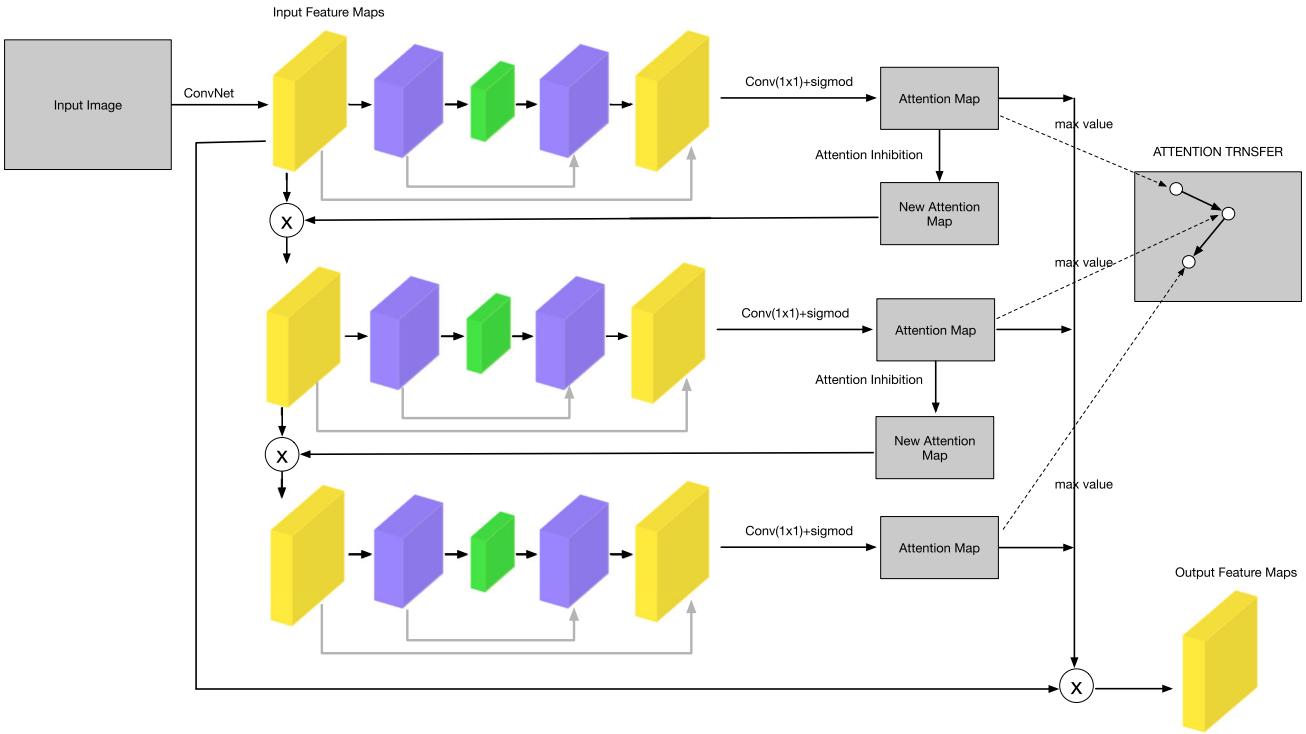

Fig. 3 The details of our attention transfer module. The input of our module is the feature maps after convoluting the image. First, an attention map is generated by a fully convolutional neural network that contains a contracting path and an expanding path. Then, a new attention map obtained by inhibiting the located attention region is transferred to the next glimpse and the new attention map is merged with the convolutional feature maps by element-wise product. Finally, we produce a final attention map by merge three attention maps with different attention regions

To analyze the effectiveness of each module of our proposed context-aware attention network, three versions of the proposed method are provided in our experiments. The first version only contains the proposed context learning module, and the second version only contains the proposed attention transfer module. The final version is the combination of the context learning and attention transfer modules.

The whole training process consists of two stages. For the first stage, we train the first version only containing the context learning module and the second version only containing the attention transfer module, respectively, where the batch size is 32 and the initial learning rate is 10^{-3}. The learning rate is decayed from 10^{-3} to 10^{-5} by 10^{-1}. For each learning rate, we train $20k$, $10k$ and $10k$ iterations separately. Finally, we have two well-trained models with context learning and attention transfer modules, respectively.

For the second stage, we train the final version containing the context learning and attention transfer modules and the training way is similar to the first stage. The trained weights of the first stage first are set as the initial weights, and then, we fine-tune all weights with a batch size of 32. The learning rate is decayed from 10^{-4} to 10^{-6} by 10^{-1}. For each learning rate, we trained $4k$, $2k$ and $2k$ iterations separately.

We also conduct more experiments for the attention transfer modules with different number of glimpses. Experimental results show that the more steps we add, the better results our model achieves. In addition, we find that our attention transfer module presents a nice attention map approaching the optimal solution when the number of glimpses is set to 3. The more details and experimental results will be presented in Sect. 4.

4 Experiments

In this section, we evaluate the performance of our proposed context-aware attention network on a series of benchmark datasets including CIFAR [37], Caltech-256 [38] and CUB-200 [39]. We first compare the proposed context-aware attention network with state-of-the-art image recognition methods and present the experimental results on these three datasets. Then, we analyze the effectiveness of each module in the context-aware attention network including context learning module and attention transfer module. In addition, we provide the evidence for just generating three attention maps. Finally, we discuss an interesting discovery by visualizing the generated attention maps.

4.1 Datasets

We conduct experiments on three challenging image recognition datasets: CIFAR [37], Caltech-256 [38] and CUB-200 [39]. The statistics of three used datasets are summarized in Table 1.

CIFAR is a dataset for common image recognition. CIFAR datasets are divided into two subsets: CIFAR-10 and CIFAR-100. The CIFAR-10 and CIFAR-100 are labeled subsets of the 80 million tiny images dataset. CIFAR-10 consists of 60,000 three-channel color images with a size of 32×32 and is divided into 10 categories: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck. Each class consists of 6000 images. 50,000 images are used for training, and 10,000 images are used for testing. CIFAR-100 is just like the CIFAR-10, except it has 100 classes containing 600 images each. There are 500 training images and 100 testing images per class. The 100 classes in the CIFAR-100 are grouped into 20 superclasses. Each image comes with a "fine" label (the class to which it belongs) and a "coarse" label (the superclass to which it belongs).

Caltech-256 is an improved version of Caltech-101 [40] that contains 30,607 images and is divided into 256 categories. Compared with Caltech-101, the number of categories has more than doubled, and the minimum number of images in any category has increased from 31 to 80. In addition, a new larger clutter category is introduced, which makes recognition more challenging.

CUB birds contains 11,788 images that are divided into 200 bird species. All images are annotated with bounding boxes, part location and attribute labels. This annotated information will help us verify whether the attention generated by our method is reasonable.

4.2 Comparison with the state-of-the-arts

To evaluate the effectiveness of our proposed context-aware attention network, we compare it with state-of-the-art methods for image recognition. Specifically, the ST-CNN [41], TLAN [13], RAN [42], FCAN [14] and RA-CNN [30] are set as our baselines, and the experimental

Table 1 The statistics of three used datasets

Dataset	Total size	Training size	Testing size	# Categories
CIFAR-10	60,000	50,000	10,000	10
CIFAR-100	60,000	50,000	10,000	100
Caltech-256	30,670	24,485	6122	256
CUB-200	11,788	9340	2358	200

results and analysis will be presented in the following content.

CIFAR-10 and CIFAR-100 We first conduct experiments on the CIFAR datasets and compare it with the five methods mentioned above. The experimental results are shown in Table 2. It can be seen that our proposed context-aware attention network achieves the best performance on CIFAR-10 and CIFAR-100 datasets. When using 40,000 samples to train our model, our proposed method achieves 97.12% and 81.21% accuracy on CIFAR-10 and CIFAR-100, respectively. Compared to these state-of-the-art methods except RA-CNN that are trained with 50,000 samples, our method achieves the best results. This means that our method can learn the discriminative features for image recognition by a small dataset, which benefits from the proposed context learning and attention transfer modules. The proposed context learning module effectively learns the contextual information, which enhances the performance of image recognition. In addition, the proposed attention transfer module captures multiple attention regions to achieve accurate image recognition by imitating human visual attention mechanism. To further validate the effectiveness of our method, we conduct experiments by increasing the training dataset size, and the experimental results are shown in Table 2. It can be seen that the performance is improved by increasing the size of the training dataset. When the size of the training samples is increased to 50,000, our proposed method outperforms all the state-of-the-art methods with 97.68% and 82.42% accuracy on CIFAR-10 and CIFAR-100, respectively. Compared to the RA-CNN, our proposed method achieves a better result with an improvement of 0.56% and 1.21% accuracy on CIFAR-10 and CIFAR-100. Above results suggest that our context learning and attention transfer modules enjoy good performance for image recognition.

Caltech-256 There are 30,607 labeled images in the Caltech-256 dataset. We divide them into two sub-datasets for training and testing. The training and testing datasets contain 24,485 and 6122 images, respectively. We then use the two subsets to evaluate our method and compare it with the baselines.

Table 3 shows the experimental results of different methods on the Caltech-256 dataset. Our proposed method achieves the best performance with 80.32% accuracy. The results of ST-CNN [41], TLAN [13], RAN [42], FCAN [14] and RA-CNN [30] are all listed in Table 3. Among the existing methods, the RA-CNN, which recursively learns discriminative region attention and region-based feature representation at multiple scales in a mutually reinforced way, achieves the best results with 79.24%. However, this method only focuses attention on one discriminative region without considering capturing multiple attention regions to jointly determine image categories. In addition, it does not fully utilize the valuable contextual information, which leads to inaccurate attention localization. Compared with the RA-CNN, our method achieves a 1.08% improvement benefiting from the proposed context learning and attention transfer module.

To illustrate how the proposed attention transfer module works, we show the captured attention regions from multiple glimpses for qualitative analysis. In Fig. 4, we can observe that these localized regions at the second and third glimpses are discriminative to corresponding categories and are easier to be classified than the first glimpse. In addition, we find an interesting phenomenon that the first attention region is always at the top right of image. Our model seems to have developed a habit using training images, which conforms to the human vision habit. The results of attention transfer are consistent with human perception, and it improves the recognition performance.

CUB-200 To further verify the effectiveness of our proposed context-aware attention network, we conduct experiments on more challenging CUB Birds dataset. We divide CUB Birds dataset into two sub-datasets for training and testing. The training and testing datasets contain 9430 and 2358 images, respectively. We then use the two subsets to evaluate our method and compare it with the baselines.

The experimental results on CUB Birds are presented in Table 4. Different birds are difficult to be recognized, due

Table 2 Comparisons with state-of-the-art methods on CIFAR-10/100

Method	Training size	Testing size	Image size	CIFAR-10 (accuracy%)	CIFAR-100 (accuracy%)
ST-CNN [41]	50,000	10,000	32 × 32	96.58	81.10
TLAN [13]	50,000	10,000	32 × 32	89.62	72.88
RAN [42]	50,000	10,000	32 × 32	96.10	79.55
FCAN [14]	50,000	10,000	32 × 32	95.80	78.32
RA-CNN [30]	50,000	10,000	32 × 32	97.21	81.84
Our approach	40,000	10,000	32 × 32	97.12	81.21
Our approach	50,000	10,000	32 × 32	97.68	82.42

Table 3 Comparisons with state-of-the-art methods on Caltech-256

Method	Training size	Testing size	Image size	Accuracy%
ST-CNN [41]	24,485	6122	224 × 224	78.21
TLAN [13]	24,485	6122	224 × 224	68.82
RAN [42]	24,485	6122	224 × 224	77.72
FCAN [14]	24,485	6122	224 × 224	76.40
RA-CNN [30]	24,485	6122	224 × 224	79.24
Our approach	24,485	6122	224 × 224	80.32

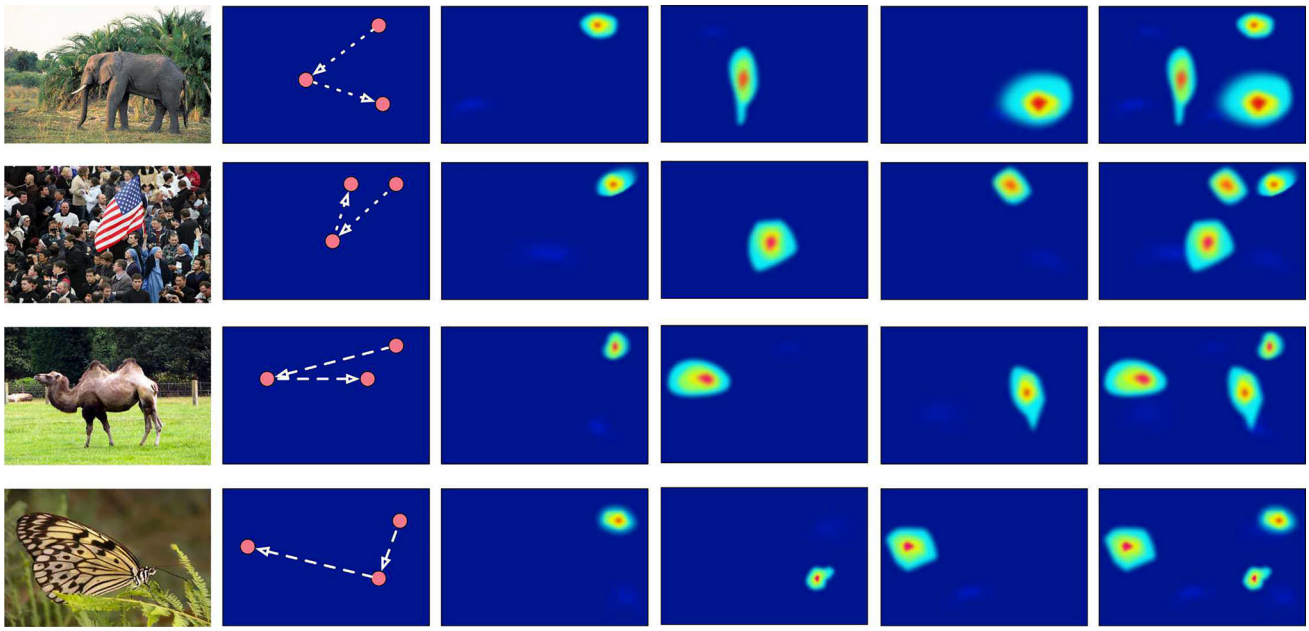

Fig. 4 Examples of the learned attention regions at different glimpses. We can observe clear the attention regions generated by our attention transfer module. The columns in turn display the input image, the attention map from the first glimpse, the attention map from the second glimpse, the attention map from the third glimpse and the final attention map, respectively

Table 4 Comparisons with state-of-the-art methods on CUB-200

Method	Training size	Testing size	Image size	Accuracy%
ST-CNN [41]	9430	2358	224 × 224	84.10
TLAN [13]	9430	2358	224 × 224	77.90
RAN [42]	9430	2358	224 × 224	83.52
FCAN [14]	9430	2358	224 × 224	82.04
RA-CNN [30]	9430	2358	224 × 224	85.31
Our approach	9430	2358	224 × 224	86.12

to the subtle differences. Our proposed context-aware attention network achieves the best performance with the recognition accuracy of 86.12% by leveraging the power of region localization and combination, which captures multiple attention regions with discriminative features. Experimental results show that the proposed attention transfer network is capable of localizing the representative attended regions. Compared with the state-of-the-art methods, our approach surpasses ST-CNN [41], TLAN [13], RAN [42], FCAN [14] and RA-CNN [30], which benefits from the use of our context learning and attention transfer modules.

4.3 Ablation studies

To verify the effectiveness of different components of our proposed context-aware attention network, we explore the variants of our architecture and justify our design decisions with experiments performed on Caltech-256 and CUB-200.

Table 5 Ablation study: effectiveness of each module of our approach on Caltech-256 and CUB-200

Method	Caltech-256 (accuracy%)	CUB-200 (accuracy%)
VGG19 [43]	70.62	77.80
VGG19 + ATM	78.51	85.47
GoogLeNet [44]	72.42	79.31
GoogLeNet + ATM	78.68	85.74
ResNet101 [9]	75.14	81.32
ResNet + ATM	79.10	85.91
CLM + ATM	80.32	86.12

Contributions of each module The proposed methods consist of two modules: Context learning module (CLM) and attention transfer module (ATM), which are utilized to learn contextual information and capture attention regions with discriminative features, respectively. We run models with different settings and record their evaluations in Table 5. The VGG19 [43] is set as our baseline and it provides 70.62% and 77.80% accuracy on Caltech-256 and CUB-200, respectively. The ATM in our proposed approach is design to capture multiple attention regions with discriminative features. When adding the ATM to the VGG19 recognition framework, it provides 78.51% and 85.47% on Caltech-256 and CUB-200. Compared with the baseline VGG19, there are 7.89% and 7.67% improvements in accuracy. Furthermore, replacing the VGG19 with our CLM that is design to learn valuable contexts, we find that there is a significant performance increase. This further demonstrates the effectiveness of our context learning module. Finally, combining the CLM and ATM, it provides the best performance with 80.32% and 86.12% accuracy on Caltech-256 and CUB-200, respectively.

Effectiveness of the ATM using different number of glimpses as shown in Fig. 3, there are three glimpses in our attention transfer module and each glimpse localizes an attention region. We also conduct experiments with different number of glimpses. The results are presented in Table 6, where Attention-N means our attention transfer module with N glimpses. The experimental results show that Attention-3 provides the best accuracy. When the number of glimpse is set to 1 or 2, the attention region localized is not sufficient for accurate image recognition. Although Attention-1 and Attention-2 can roughly locate the object in the image, they cannot capture the discriminative parts of the object, which limits the recognition performance. In addition, we also design the attention transfer module with more than three glimpses. Attention-4 and Attention-5 obtain similar results with 80.10% and 79.32% accuracy on Caltech-256. Compared to Attention-3, there are a 0.22% decrease and a 1.09% decrease in accuracy. Theoretically, the greater the number of glimpses, the higher the accuracy. However, this is not the case in our model. At least on the CIFAR, Caltech-256 and CUB-200 datasets, the best choice of attention transfer module is the Attention-3. As the number of glimpses increases, we will force more attention regions to be located, which results in a lot of useless regions being integrated into our final attention map. These useless regions will interfere with the extracted discriminative features and decrease the recognition accuracy.

Effectiveness of combining the context feature maps in different ways a key step is to fuse the four kinds of context feature maps learned in *left*, *right*, *up* and *down* directions. As shown in Fig. 2, we first concatenate the four feature maps and then use $1 \times 1 Conv2d + ReLu$ to generate our final context feature maps. In what order to concatenate the feature maps is discussed. In addition, we also try some other fusion strategies such as $element - wise\ sum$ and $element - wise\ product$. Table 7 shows the results of different versions on the Caltech-256 and CUB-200 datasets. It can be seen that the order of concatenating the four kinds of context feature maps has a tiny effect on the recognition results. The CLM with $element - wise\ sum$ achieves a 78.42% accuracy and a 84.20 on the Caltech-256 and CUB-200, respectively. The CLM with $element - wise\ product$ achieves a 74.64% accuracy and a 83.62 on the Caltech-256 and CUB-200, respectively. Among different versions of the CLM, the CLM using the $1 \times 1 Conv2d + ReLu$ achieves the best results. This means that the fusion way learned by the network is better than the fusion ways of our manual design (Table 7).

Table 6 Ablation study: effectiveness of the ATM using different number of glimpses on Caltech-256 and CUB-200

Method	Caltech-256 (accuracy%)	CUB-200 (accuracy%)
Attention-1	75.32	82.22
Attention-2	77.42	84.18
Attention-3	80.32	86.12
Attention-4	80.10	85.20
Attention-5	79.23	84.32

Table 7 Ablation study: effectiveness of the CLM using different fusion ways on Caltech-256 and CUB-200

Method	Order	Fusion way	Caltech-256 (accuracy%)	CUB-200 (accuracy%)
CLM	$right \rightarrow left \rightarrow down \rightarrow up$	$1 \times 1 Conv2d + ReLu$	80.32	86.12
CLM	$up \rightarrow down \rightarrow left \rightarrow right$	$1 \times 1 Conv2d + ReLu$	80.24	86.13
CLM	$right \rightarrow up \rightarrow down \rightarrow left$	$1 \times 1 Conv2d + ReLu$	80.34	85.98
CLM	$right \rightarrow left \rightarrow down \rightarrow up$	$element - wise\ sum$	78.42	84.20
CLM	$right \rightarrow left \rightarrow down \rightarrow up$	$element - wise\ product$	74.64	83.62

Table 8 Ablation study: effectiveness of the attention map learned by different activation functions on Caltech-256 and CUB-200

Method	Activation function	Caltech-256 (accuracy%)	CUB-200 (accuracy%)
ATN	*softmax*	80.32	86.12
ATN	*sigmoid*	78.12	84.32

Effectiveness of generating the attention map using various activation functions as shown in Fig. 2, there is a step using the activation function *sigmoid* to generate the attention map. The *sigmoid* makes each pixel value of the feature map is in range of 0–1. We also conduct more experiments on the CUB-200 dataset to discuss the choice of activation functions for generating the attention map. The experimental results are presented in Table 8. From Table 8, we can see that different choices make different results. The ATM with the activation function *softmax* achieves a 80.32% accuracy and a 86.12% accuracy on the Caltech-256 and CUB-200 datasets, respectively. Compared with the choice of *sigmoid*, there are a 2.2% decrease and a 1.9% decrease in accuracy. To explain the results, we visualize the attention map of all glimpses and see that the attention map generated by the *softmax* cannot capture the discriminative regions. All pixel values of the attention map learned by the activation function *softmax* are so small that the learned attention map is useless to extract discriminative features from the convolutional feature maps.

5 Conclusion

In this paper, we present a context-aware attention network for accurate image recognition, which fully utilizes the contextual information and discriminative parts to achieve state-of- the-art performance. The main contributions of the proposed method are as follows. First, we propose a context learning module for effectively learning the contextual information, which enhances the performance of image recognition. Second, we propose an attention transfer module, which imitates human visual attention mechanism and can capture multiple attention regions to achieve accurate image recognition. In addition, a strategy of attention inhibition is adopted in the attention transfer module, which enables us to generate different attention regions in each glimpse. Finally, we conduct extensive experiments on CIFAR-10, CIFAR-100, Caltech-256 and CUB-200. The results show that our method achieves state-of-the-art performance.

In the future, it could be of interest to add a segmentation task into our recognition task and utilize the labeled segmentation information to help us accurately locate the discriminative regions. Furthermore, it is worth studying how to transform the attention with the help of human being visual attention mechanism.

Acknowledgements This project was partially supported by Grants from Natural Science Foundation of China 71671178, 91546201. It was also supported by University of Chinese Academy of Sciences Project Y954016XX2, and by Guangdong Provincial Science and Technology Project 2016B010127004.

Compliance with ethical standards

Conflict of interest We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

References

1. Bertinetto L, Valmadre J, Henriques JF, Vedaldi A, Torr PH (2016) Fully-convolutional siamese networks for object tracking. In: ECCV. Springer, pp 850–865
2. Nam H, Han B (2016) Learning multi-domain convolutional neural net-506 works for visual tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4293–4302
3. Chen LC, Papandreou G, Kokkinos I et al (2018) DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell 40(4):834–848
4. Girshick RB, Donahue J, Darrell T et al (2016) Region-based convolutional networks for accurate object detection and segmentation. IEEE Trans Pattern Anal Mach Intell 38(1):142–158

5. Redmon J, Farhadi A (2016) YOLO9000: better, faster, stronger. arXiv preprint, 1612
6. Santoro A, Raposo D, Barrett DG et al (2017) A simple neural network module for relational reasoning. In: Advances in neural information processing systems, pp 4974–4983
7. Leng J, Liu Y (2018) An enhanced SSD with feature fusion and visual reasoning for object detection. Neural Comput Appl. https://doi.org/10.1007/s00521-018-3486-1
8. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. ICLR
9. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: CVPR
10. Itti L, Koch C (2001) Computational modelling of visual attention. Nat Rev Neurosci 2(3):194
11. Mnih V, Heess N, Graves A et al (2014) Recurrent models of visual attention. In: NIPS
12. Zhao B, Wu X, Feng J, Peng Q, Yan S (2016) Diversified visual attention networks for fine-grained object classification. arXiv preprint arXiv:1606.08572
13. Xiao T, Xu Y, Yang K, Zhang J, Peng Y, Zhang Z (2015) The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In: CVPR, pp 842–850
14. Liu X, Xia T, Wang J, Lin Y (2016) Fully convolutional attention localization networks: efficient attention localization for fine-grained recognition. CoRR arXiv:1603.06765
15. Ji Y, Zhang H, Wu QMJ (2018) Salient object detection via multi-scale attention CNN. Neurocomputing 322:130–140
16. Zhang H, Ji Y, Huang W et al (2018) Sitcom-star-based clothing retrieval for video advertising: a deep learning framework. Neural Comput Appl. https://doi.org/10.1007/s00521-018-3579-x
17. Xu K, Ba J, Kiros R et al (2015) Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning, pp 2048–2057
18. Chen L, Zhang H, Xiao J et al (2017) SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5659–5667
19. Seo PH, Lin Z, Cohen S et al (2016) Progressive attention networks for visual attribute prediction. arXiv preprint arXiv:1606.02393
20. Das D, George Lee CS (2018) Sample-to-sample correspondence for unsupervised domain adaptation. Eng Appl Artif Intell 73:80–91
21. Das D, George Lee CS (2018) Unsupervised domain adaptation using regularized hyper-graph matching. In: 2018 25th IEEE international conference on image processing (ICIP). IEEE
22. Courty N et al (2017) Optimal transport for domain adaptation. IEEE Trans Pattern Anal Mach Intell 39(9):1853–1865
23. Larochelle H, Hinton GE (2010) Learning to combine foveal glimpses with a third-order Boltzmann machine. In: Advances in neural information processing systems, pp 1243–1251
24. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
25. Kim JH, Lee SW, Kwak D et al (2016) Multimodal residual learning for visual QA. In: Advances in neural information processing systems, pp 361–369
26. Noh H, Hong S, Han B (2015) Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE international conference on computer vision, pp 1520–1528
27. Srivastava RK, Greff K, Schmidhuber J (2015) Training very deep networks. In: Advances in neural information processing systems, pp 2377–2385
28. Jaderberg M, Simonyan K, Zisserman A (2015) Spatial transformer networks. In: Advances in neural information processing systems, pp 2017–2025
29. Xiao T, Xu Y, Yang K et al (2015) The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 842–850
30. Fu J, Zheng H, Mei T (2017) Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. CVPR 2:3
31. Wang F et al (2017) Residual attention network for image classification. In: CVPR
32. Divvala SK, Hoiem D, Hays JH, Efros AA, Hebert M (2009) An empirical study of context in object detection. In: CVPR
33. Galleguillos C, Rabinovich A, Belongie S (2008) Object categorization using co-occurrence, location and appearance. In: CVPR
34. Uijlings JR, De Sande KE, Gevers T et al (2013) Selective search for object recognition. Int J Comput Vis 104(2):154–171
35. He K, Zhang X, Ren S et al (2014) Spatial pyramid pooling in deep convolutional networks for visual recognition. In: European conference on computer vision, pp 346–361
36. Girshick RB (2015) Fast R-CNN. In: International conference on computer vision, pp 1440–1448
37. Krizhevsky A, Hinton G (2009) Learning multiple layers of features from tiny images. Technical report, University of Toronto
38. Griffin G, Holub A, Perona P (2007) Caltech-256 object category dataset
39. Welinder P, Branson S, Mita T, Wah C, Schroff F, Be-longie S, Perona P (2010) Caltech-UCSD Birds 200. Technical report CNS-TR-2010-001, California Institute of Technology
40. Fei-Fei L, Fergus R, Perona P (2004) Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In: IEEE CVPR 2004, workshop on generative-model based vision
41. Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K (2015) Spatial transformer networks. In: NIPS, pp 2017–2025
42. Wang F, Jiang M, Qian C et al (2017) Residual attention network for image classification. arXiv preprint arXiv:1704.06904
43. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: ICLR, pp 1409–1556
44. Szegedy C et al (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).
Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.
These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.
We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.
While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.
These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.
To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.
Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.
If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com