Leveraging attention-based visual clue extraction for image classification

Yunbo Cui1 | Youtian Du1 | Xue Wang1,2 | Hang Wang1 | Chang Su3

1 Faculty of Electronic and Information Engineering, Xi’an Jiaotong University, No. 28, Xianning West Road, Xi’an, China
2 Institute of Advanced Computer Science, Leiden University, Leiden, Netherlands
3 Department of Healthcare Policy and Research at Weill Cornell Medicine, Cornell University, Ithaca, New York, USA

Abstract
Deep learning-based approaches have made considerable progress in image classification tasks, but most of the approaches lack interpretability, especially in revealing the decisive information causing the categorization of images. This paper seeks to answer the question of what clues encode the discriminative visual information between image categories and can help improve the classification performance. To this end, an attention-based clue extraction network (ACENet) is introduced to mine the decisive local visual information for image classification. ACENet constructs a clue-attention mechanism, that is global-local attention, between the image and visual clue proposals extracted from it and then introduces a contrastive loss defined over the achieved discrete attention distribution to increase the discriminability of clue proposals. The loss encourages considerable attention to be devoted to discriminative clue proposals, that is those similar within the same category and dissimilar across categories. The experimental results for the Negative Web Image (NWI) dataset and the public ImageNet2012 dataset demonstrate that ACENet can extract true clues to improve the image classification performance and outperforms the baselines and the state-of-the-art methods.

1 INTRODUCTION

Image classification aims to categorize a set of unlabelled images into several predefined classes based on their visual content. Image classification has become a critical task in multiple related areas, such as object detection/recognition [1], visual concept learning [2] and visual knowledge learning [3]. Image classification has made considerable progress in achieving high performance in recent years. However, there are some important issues existing for the task, such as the interpretability of deep neural network-based approaches.

The progress in image classification can be partitioned into two stages: human-designed feature-based methods and deep learning-based methods. Regarding the former, people design the features in advance in terms of colour, texture and gradient and apply classifiers to these extracted features. Some representative human-designed features, such as the histogram of oriented gradients (HOG) [4], local binary pattern (LBP) [5], the scale invariant feature transform (SIFT) [6] and speeded up robust feature (SURF) [7], have been proposed. In general, these human-designed features have good interpretability but may have limitations in achieving high image classification accuracy, especially for complicated data. Over recent years, convolutional neural networks (CNNs) [8] have made great success in a variety of visual tasks including image classification. CNNs consist of a stack of convolutional layers interleaved with non-linearities and downsampling, and are capable of capturing hierarchical patterns with global receptive fields as powerful image descriptions. Since the introduction of AlexNet [8] in 2012, a variety of new CNN architectures, including visual geometry group network (VGGNet) [9], Inception net [10], residual network (ResNet) [11], dense convolutional network (DenseNet) [12] and NASNet [13], have been proposed. Besides the design of CNN architectures, a number of works have sought to introduce new modules into networks to meet particular requirements [14–16]. The attention mechanism, which adaptively learns the effect of elements of the input on the output, is an important module for deep learning applications, such as image classification and image captioning [17, 18]. Wang et al. [19] proposed a residual attention network for image classification, in which the spatial attention and channel attention on the feature maps are introduced to enhance meaningful contents and suppress
meaningless information while extracting image features. Zhao et al. [20] explored two forms of self-attention and assessed their effectiveness for image recognition.

In general, most previous approaches solve the image classification problem based on the global visual context or important local visual information. Actually, there are three types of visual information causing classification results: (1) the global content of images widely used in a number of approaches; (2) the discriminative local visual clues appearing in a category while not existing in the others, such as some important objects; and (3) the common visual information that appears across different categories and is indiscriminative, such as some backgrounds. Figure 1 shows an example illustrating the effect of the last two types of visual information on image classification. In this figure, we show a few region proposals representing different contents of the images from two categories that is 'sea lion' and 'red wolf'. In the same category, there are region proposals (marked by red bounding boxes) corresponding to the key objects with the same semantics and reflecting the category information while there are also some other region proposals (marked by yellow and blue bounding boxes) with different semantics that are irrelevant to the category across different images. In addition, we find that these region proposals irrelevant to the category may appear in the images from different categories. Hence, a well-formed image classification approach needs to capture and enhance the discriminative features effectively while suppressing the common visual information across different categories.

This paper proposes a novel approach called the attention-based clue extraction network (ACENet) to the learning of visual clues in image classification and seeks to answer the question of what clues encode the discriminative visual information between image categories and can help improve classification performance. First, we utilize Faster R-CNN [21] and selective search to obtain the visual clue proposals that consist of objects and backgrounds. Second, we present a clue-attention mechanism, called global-local attention, between images and visual clue proposals to adaptively measure their correlation. To extract the true clues from the proposals and reinforce their discriminability, we introduce a contrastive loss defined over the attention distribution, which encourages those clue proposals that are similar within the same category while being dissimilar across different categories to be given considerable attention. Finally, the image classification is performed based on the combination of the representations of images and the clues. The experimental results conducted on ImageNet 2012 and our NWI dataset demonstrate that our approach is effective in extracting the true visual clues and improving the image classification performance.

In summary, our contributions are three-fold:

1. We present the ACENet approach to improve image classification performance by leveraging local visual clues and seek to answer the question of what clues encode the discriminative visual information between image categories.
2. We propose the global-local attention mechanism, which adaptively measures the correlation between visual clue proposals and images in both training and testing and permits the clue proposals to make varying contributions to image classification.
3. A contrastive loss is introduced to reinforce the discriminability of visual clue proposals and sparsify the attention distribution with consideration of cross-sample information for the extraction of true clues.

The remainder of this paper is organized as follows. Section 2 presents a brief overview of related work. Section 3 describes the details of our proposed method, Section 4 reports the experimental results for the two datasets, and Section 5 concludes the paper.

## 2 RELATED WORK

At present, CNNs have been proven to be the most effective model for solving the image classification problem. Thus, we review the CNN-based methods for image classification in this section.

**CNN architectures.** Since the introduction of AlexNet [8] in 2012, a variety of representative CNN architectures, including VGGNet [9], Inception net [10], ResNet [11], GoogLeNet [10], and DenseNet [12], have been proposed. These variants seek to explore a better model mainly from the following three aspects: (1) different network structures, (2) deeper networks, and (3) high efficiency. For example, Huang et al. [12] designed a new network structure named DenseNet for image classification, which connects each layer to every other layer in a feed-forward fashion and is different from traditional convolutional networks that generally connect each layer and its subsequent layer. Regarding the depth of networks, VGGNet and GoogLeNet demonstrated that a deeper network would extract higher-quality image features. He et al. [11] presented ResNet with a depth of up to 152 layers, a substantially deeper network than those proposed previously, and demonstrated its powerful image classification performance on ImageNet. Convolutional networks are computationally expensive as they
become increasingly deeper and wider. With the growing complexity of networks, the issue of how network architectures can be made more efficient has received considerable attention. In particular, group convolutions [22] have been proven to be a popular method to decrease the number of model parameters and reduce the computational complexity. MobileNet [23] and ShuffleNet [24], which could be viewed as extensions of the grouping operator, constructed efficient network architectures based on the combination of depthwise convolution and pointwise convolution with $1 \times 1$ convolutional filters. Tzelepis et al. [25] introduced a generic model compression method that had minimal impacts on accuracy while reducing the inference time and memory footprint of a network.

**Attention mechanism.** Regarding the progress of deep learning for image classification in terms of different network structures introduced above, the introduction of an attention mechanism module is a significant way to improve classification performance and interpretability. In general, attention can be interpreted as a means of biasing the allocation of available computational resources towards the most informative components of a signal [16, 26–28]. In recent years, spatial attention and channel attention have been widely used for CNNs, especially in visual tasks. Woo et al. [26] proposed a convolutional block attention module (CBAM) for CNNs that combined channel attention and spatial attention for adaptive feature refinement. Hu et al. [27] proposed a network called squeeze-and-excitation network (SENet) and introduced an architecture unit named the ‘squeeze-and-excitation’ block (SE block) to enhance the representational power of deep networks by explicitly modelling the interdependencies between the channels of the convolutional features. The SE block can be regarded as a self-attention function on channels. Attention mechanisms have been widely applied to the image classification and recognition tasks. Jaderberg et al. [16] introduced a new learnable module called the spatial transformer, which allows networks to select the visual regions of an image that are most relevant. Zheng et al. [28] proposed a multi-attention convolutional neural network to learn more discriminative fine-grained features in the fine-grained image classification task. By contrast, our proposed ACENet seeks to find those visual regions from images that encode the most discriminative visual information for image classification with the attention mechanism and enhance the classification performance.

**Image classification applications.** From the view of applications, image classification has been used in various areas, such as retrieval [29], surveillance/monitoring [30, 31] and medical analysis [32]. In these areas, researchers focus on natural images and specific-domain images related to traffic, remote sensing, industry, medical imaging and so forth. A variety of issues have been explored in these image classification applications. For example, as an extension of traditional image classification, fine-grained image classification [29] aims to recognize subcategories under some basic-level categories, where the objects of different subcategories are both semantically and visually similar to each other. Image classification has also been used in visual concept learning to mine a large range of concepts from the vast resources of online data [33]. To accurately implement object detection that is widely employed in surveillance applications, He et al. [34] presented a novel algorithm called the Mask R-CNN that can generate a high-quality segmentation mask and label for an object. Based on the Mask R-CNN, Wu et al. [1] introduced an improved Mask R-CNN to mitigate the performance deterioration when samples are reduced for object detection.

## 3 | METHODOLOGY

### 3.1 | Framework

In the task of $C$-class image classification, the data are given as image-label pairs $D = \{(x_i, y_i), i = 1, 2, ..., N\}$, where $x_i$ and $y_i \in \{1, 2, ..., C\}$ denote the sample of images and its label, respectively. We represent $x_i$ with a feature vector $x_i \in \mathbb{R}^l$ and $y_i$ with a one-hot vector $y_i = \mathbbm{1}\{y_i = c\}$, where 1 is an indicator function. Suppose there are $M_p$ clue proposals $\{p^a_i\}_{a=1}^{M_p}$ extracted from $x_i$, where each proposal $p^a_i$ is represented by a feature vector $p^a_i \in \mathbb{R}^l$.

Figure 2 illustrates the framework of the proposed ACENet, which includes four main processes: clue proposal extraction, visual representation, the clue-attention network and classification. ACENet first extracts multiple visual regions from an image and takes them as the clue proposals. The convolutional neural network is then employed to obtain the representation of both the original images and clue proposals. To learn the clues that have important effects on recognizing categories of images, an attention network is introduced to learn the distribution of the attention on the clue proposals and highlight those clue proposals helping to improve the classification with the contrastive loss function. Finally, the classification is performed based on the concatenation of the features derived from the image and clue proposals.

### 3.2 | Clue proposal extraction

In this work, the clue proposals are considered typical local regions in images that have fine and rich visual information, such as important objects and background. We extract the clue proposals based on the cooperation of Faster-RCNN [21] and selective search [35]. First, we implement the Faster R-CNN algorithm to produce $M_r$ object region proposals: $\{p_{r,i}, r_{i,j}\}_{j=1}^{M_r}$, where $p_{r,i}$ describes the $i$th proposal with its coordinates and $r_{i,j}$ denotes the confidence (predicted probability) of this proposal containing an object. In this work, the Faster-RCNN is pre-trained on the MSCOCO dataset and does not perform fine-tuning. Although the datasets used for image classification in this work consist of a proportion of categories that do not appear in MSCOCO, the extracted regions also cover many meaningful objects that can be employed as the clue proposals. In the implementation, we empirically choose 5 proposals with the highest probability from the Faster-RCNN. Second, we utilize a selective search algorithm, which performs the computing in terms of texture, colour, size and overlap, to generate a
Attention-based visual clue extraction

The determination of visual clues for image classification depends on two factors: (1) a specific image and its category, and (2) the discriminability of clue proposals. For the first factor, we present a global-local attention mechanism between an image and the clue proposals extracted from it. Since the category label is unknown for a test image, we do not actually build an attention mechanism between the clue proposals and class labels. Regarding the discriminability, we consider that the true clues for a specific image should be similar to the clues in the images of the same category and dissimilar to the images from the other categories.

First, we introduce the global-local attention mechanism between the original image and clue proposals to generate a discrete attention distribution over clue proposals. Generally, the true clues should be given more attention and lead to a high probability at the corresponding entries of the attention distribution. In general, an attention mechanism can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weight depends on the query, keys, values, and output, respectively; and sim(·, ·) denotes a certain function to measure the correlation of queries and keys. In this work, we have the representation of images and the clue proposals $x_i$ and $\{p_m^i\}_{m=1}^{M_p}$, which correspond to the query and key (here, the key and value are the same) in Equation (4), respectively; and our clue-attention model can be specified as follows:

$$a_i^m = w_d^T \tanh(W_{x}\cdot x_i + W_{p}p_m^i),$$  \hspace{1cm} (2)

$$\alpha_i = \text{softmax}(a_i),$$  \hspace{1cm} (3)

where $a_i^m$, corresponding to the function $\text{sim}(\cdot)$ in Equation (4), encodes the correlation of $x_i$ and $p_m^i$; $\alpha_i$ denotes the discrete attention distribution on all clue proposals for image $x_i$; $W_{x}, W_{p} \in \mathbb{R}^{d \times l}$ are the transformation matrices that map the visual representation into low-dimensional spaces, $w_d \in \mathbb{R}^d$, and $a_i = (a_i^m)^T$. The clue context vector associated with the clue proposals is computed by:

$$\tilde{p}_i = \sum_{m=1}^{M_p} \alpha_i^m p_m^i,$$  \hspace{1cm} (4)

where $\alpha_i^m$ is the $m$th entry of $\alpha_i$, and $\tilde{p}_i$, corresponding to $\text{Attention}(\cdot, \cdot, \cdot)$ in Equation (4), denotes the output of the global-local attention. The final representation of image $x_i$ for image classification is achieved by the concatenation of $x_i$ and $\tilde{p}_i$, that is $\tilde{x}_i = (x_i^T, \tilde{p}_i^T)^T$, where $\tilde{x}_i \in \mathbb{R}^{2l}$.

The global-local attention mechanism formulates the relationship between the global information (i.e. images) and the local information (i.e. clue proposals) extracted from it. We consider that the global-local attention is different from the following widely used attention mechanisms: (1) self-attention, which describes the importance of elements in images or sentences by measuring the relationship of these elements, and (2) encoder-decoder attention, which generally builds the relationship between the elements in outputs and inputs and has

$$\text{Attention}(Q, K, V) = \text{softmax}(\text{sim}(Q, K)) \cdot V,$$  \hspace{1cm} (1)

where $Q, K, V$ and $\text{Attention}(\cdot, \cdot, \cdot)$ refer to the query, key, value and output, respectively; and $\text{sim}(\cdot, \cdot)$ denotes a certain function to measure the correlation of queries and keys.
Datasets

From two images of different categories, the contrastive loss is defined in the following form:

$$L_{clt} = -\frac{1}{T_p} \sum_{i=1}^{N} \alpha_i \mathbf{W}_{i} \alpha_i + \frac{1}{T_n} \sum_{j=1}^{N} \alpha_j \mathbf{T}_{j} \alpha_j,$$

Subject to: $\alpha_i, \alpha_j \geq 0$, $\|\alpha_i\|_1, \|\alpha_j\|_1 = 1,$

where $i_1$ and $i_2$ refer to two images belonging to the same category; $j_1$ and $j_2$ refer to two images from different categories; $T_p$ and $T_n$ denote the numbers of image pairs from the same category and different categories, respectively; and $\mathbf{W}_{i_1} (\mathbf{W}_{i_2})$ is the similarity matrix of clue proposals from the $i_1 \text{th} (j_1 \text{th})$ image and the $i_2 \text{th} (j_2 \text{th})$ image. We define the $(m_1, m_2)$-entry of matrix $\mathbf{W}_{i_2}$ as follows:

$$\mathbf{W}_{i_2}^{m_1 m_2} = \frac{\mathbf{p}_{i_1}^{m_1} \cdot \mathbf{p}_{i_2}^{m_2}}{\|\mathbf{p}_{i_1}^{m_1}\|_2 \cdot \|\mathbf{p}_{i_2}^{m_2}\|_2},$$

where $\| \cdot \|_2$ denotes the L2-norm. $\mathbf{W}_{i_2}^{m_1 m_2}$ can also be calculated as Equation (6). Clearly, $\mathbf{W}_{i_2}^{m_1 m_2}$ is large for two visually similar clue proposals $\mathbf{p}_{i_1}^{m_1}$ and $\mathbf{p}_{i_2}^{m_2}$ from two images of the same category, and thus the corresponding attention values $\alpha_i^{m_1}$ and $\alpha_i^{m_2}$ tend to be enlarged if we minimize the first term of Equation (5).

In contrast, for the visually similar clue proposals $\mathbf{p}_{i_1}^{m_1}$ and $\mathbf{p}_{i_2}^{m_2}$ from two images of different categories, $\mathbf{W}_{i_2}^{m_1 m_2}$ is also large. In this case, little attention will be given to them when we minimize the second term in Equation (5), and thus those visually different clue proposals are more likely to attract considerable attention. Consequently, minimizing the contrastive loss in Equation (5) encourages the attention being paid to the true clues.

The final loss function is given by combining the cross-entropy loss and the contrastive loss:

$$\mathbf{L} = \mathbf{L}_{ce} + \lambda \mathbf{L}_{clt},$$

where $\lambda$ is a balance parameter. $\mathbf{L}_{ce}$ denotes the cross-entropy loss defined in the following form:

$$\mathbf{L}_{ce} = -\sum_{j=1}^{N} \mathbf{y}_j^T \log \hat{\mathbf{y}}_j,$$

where $\mathbf{y}_j$ and $\hat{\mathbf{y}}_j$ indicate the true label vector and the predicted label vector, respectively. In essence, the contrastive loss can be considered a regularizer of the regularization framework or prior knowledge from the viewpoint of Bayesian learning.

4.1 Datasets

We conduct the experiments and test the performance of ACENet over a small dataset we built and the public ImageNet2012 dataset. The small dataset, called negative web image (NWI) dataset, is crawled from social media websites and consists of the data that are undesirable and sometimes harmful to young people. The NWI dataset has a total of 10,500 images and comprises approximately the same number of 6 negative categories, including pornography, luxury watches, luxury cars, luxury bags, cash, and jewellery; and a category of natural images. In the experiment, the dataset is evenly divided into two parts for training and testing. ImageNet2012 comprises 1.28 million training images and 50K validation images from 1000 categories, and has been widely used in the image classification task. We train the networks on the training set and report the performance on the validation set for ImageNet2012. In the training, we follow standard practices and perform data augmentation with random cropping using scale and aspect ratio to a size of $224 \times 224$ pixels (or $299 \times 299$ for the models associated with Inception-ResNet-v2) and perform random horizontal flipping. When evaluating the models we apply centre-cropping so that $224 \times 224$ pixels are cropped from each image, after its shorter edge is first resized to 256 (299 $\times$ 299 from each image whose shorter edge is first resized to 352 for the models associated with Inception-ResNet-v2).

4.2 Implementation details

For each image, we extract 10 visual regions as clue proposals, that is $M_p = 10$. We employ VGG-16, ResNet-50, ResNet-101, Inception-ResNet-v2, DenseNet-201 and Efficient-B7 (without the classification layer) as the backbones for the representation of images and clue proposals. For the four backbones, the representations are 4096-, 2048-, 2048-, 1536-, 1920- and 2560-dimensional feature vectors (i.e. $l = 4096, 2048, 2048, 1536, 1920$ and 2560, respectively). In the clue-attention module, the column dimension of transformation matrices $\mathbf{W}_X$ and $\mathbf{W}_P$ is set to $d = 128$. After the attention module, we add two fully connected layers, FC-1024 and FC-1000 (FC-7), for classification on ImageNet (NWI). In the training process, we use the stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and a minibatch size of 256 for NWI and 1024 for ImageNet. To reduce the computational load, we randomly choose a fixed amount of image pairs in each minibatch when computing the contrastive loss shown in Equation (5). Empirically, we choose $T_p T_n = 4096$ for ImageNet and 1024 for NWI. If the number of image pairs belonging to the same category or different categories in a minibatch is less than the fixed amount,
TABLE 1  Compare with the baselines and state-of-the-art methods on NWI and ImageNet2012 with single-crop testing in terms of the top-1 error and top-5 error (%)

| Dataset         | Method       | Backbone     | #Params | Top-1 error | Top-5 error |
|-----------------|--------------|--------------|---------|-------------|-------------|
| NWI             | VGG-16       | /            | 134M    | 10.39       | 2.91        |
|                 | SENet (2020) | VGG-16       | 136.3M  | 7.85        | 1.90        |
|                 | ACENet       | VGG-16       | 143.4M  | 6.37        | 1.42        |
|                 | ResNet-50    | /            | 24M     | 6.86        | 1.94        |
|                 | SENet (2020) | ResNet-50    | 26M     | 5.47        | 1.45        |
|                 | ACENet       | ResNet-50    | 28.7M   | 4.25        | 0.83        |
|                 | ResNet-101   | /            | 42.7M   | 6.03        | 1.65        |
|                 | SENet (2020) | ResNet-101   | 47.2M   | 4.25        | 0.92        |
|                 | ACENet       | ResNet-101   | 47.4M   | 3.37        | 0.68        |
|                 | Inception-ResNet-v2 | / | 54.5M  | 4.87        | 1.32        |
|                 | SENet (2020) | Inception-ResNet-v2 | 58.3M  | 2.91        | 0.87        |
|                 | ACENet       | Inception-ResNet-v2 | 58M     | 2.16        | 0.55        |
|                 | CSPNet (2020) | DenseNet-201 | 21M     | 5.23        | 1.39        |
|                 | CSPNet (2020) | ResNeXt-50    | 18.5M   | 4.11        | 1.05        |
|                 | ACENet       | DenseNet-201  | 22.5M   | 3.34        | 0.67        |
|                 | BoTNet-S1-128 (2021) | / | 73.1M | 1.94 | 0.50 |
|                 | EfficientNet-B7 (2019) | / | 63.5M | 2.03 | 0.56 |
|                 | ACENet       | EfficientNet-B7 | 69.4M | 1.72 | 0.46 |
| ImageNet2012    | VGG-16       | /            | 138M    | 27.10       | 8.85        |
|                 | SENet (2020) | VGG-16       | 140.3M  | 25.22       | 7.70        |
|                 | ACENet       | VGG-16       | 144.4M  | 25.08       | 7.05        |
|                 | ResNet-50    | /            | 26M     | 24.72       | 7.83        |
|                 | SENet (2020) | ResNet-50    | 28M     | 23.29       | 6.62        |
|                 | ACENet       | ResNet-50    | 29.7M   | 23.02       | 6.47        |
|                 | ResNet-101   | /            | 44.7M   | 23.65       | 7.14        |
|                 | SENet (2020) | ResNet-101   | 49.2M   | 22.38       | 6.07        |
|                 | ACENet       | ResNet-101   | 48.4M   | 22.09       | 6.12        |
|                 | Inception-ResNet-v2 | / | 56M   | 19.96       | 4.92        |
|                 | SENet (2020) | Inception-ResNet-v2 | 59.8M | 19.80 | 4.79 |
|                 | ACENet       | Inception-ResNet-v2 | 59M | 18.25 | 4.03 |
|                 | CSPNet (2020) | DenseNet-201 | 23.1M | 22.3 | 6.4 |
|                 | CSPNet (2020) | ResNeXt-50 | 20.5M | 21.8 | 5.7 |
|                 | ACENet       | DenseNet-201 | 23.5M | 20.62 | 5.35 |
|                 | BoTNet-S1-128 (2021) | / | 75.1M | 15.3 | 3.0 |
|                 | EfficientNet-B7 (2019) | / | 66M | 15.7 | 3.0 |
|                 | ACENet       | EfficientNet-B7 | 70.4M | 14.47 | 2.43 |

we choose all the image pairs for the computation of the contrastive loss. All the experiments are conducted on a platform with 8 Nvidia Titan V GPUs using PyTorch.

4.3 Experimental results

In the experiments, we adopt four representative CNNs, that is VGG-16, ResNet-50, ResNet-101 and Inception-ResNet-v2, as the baselines and EfficientNet-B7\[37\], CSPNet\[38\], SENet \[27\] and BoTNet\[39\] as the state-of-the-art method for image classification. For SENet and our proposed ACENet, we employ VGG-16, ResNet-50, ResNet-101, Inception-ResNet-v2, DenseNet-201 and EfficientNet-B7 as the backbones. The performance is evaluated with the commonly used metrics ‘top-1 error’ and ‘top-5 error’.

4.3.1 Experimental results on the NWI dataset

Table 1 shows the classification results of our approach and the compared models for the NWI dataset. From this table, we
observe that most of the methods employed here achieve high performance for the small dataset. Compared with the baselines and the state-of-the-art methods, our ACENet achieves the lowest error rates. For example, ACENet obtains a 2.16% top-1 error when using Inception-ResNet-v2 as the backbone, exceeding SENet (2.91%) by 0.75%. ACENet also works well for other backbones: it decreases the top-1 error by 1.48%, 1.22% and 0.88% compared with SENet when using the basic VGG-16, ResNet-50 and ResNet-101 models, respectively. Compared with BoTNet-S1-128, ACENet with the EfficientNet-B7 backbone decreases the top-1 error by 0.22%. Figure 3 illustrates the examples of visual clue extraction for 6 negative categories, where the visual clue proposals with the most attention (i.e. corresponding to the largest $\alpha_m$ for the $i$th image) are marked by red rectangles. For the NWI dataset, we observe that the category of an image is dominantly determined by the important visual objects in it, and thus it is effective to improve the classification performance using visual clue extraction.

### 4.3.2 Experimental results on the ImageNet dataset

Figure 4 shows the change in the top-1 accuracy (i.e. $1 - \text{top-1 error}$) with the balance parameter $\lambda$ in Equation (7) for ACENet on the validation set of ImageNet2012. It is observed that the performance reaches a peak at $\lambda = 10^{-4}$ and then begins to degrade. For all the experiments in this work, we choose $\lambda = 10^{-4}$. In this case, we find that $\lambda L_{cep} / L_{cls} \approx 0.12$ when the training stops, which can be empirically considered reasonable.

We report the classification results of ACENet, the baselines and state-of-the-art models for ImageNet in Table 1. From the table, we observe that our ACENet model largely decreases the error rates compared with the baselines VGG-16, ResNet-50, ResNet-101 and Inception-ResNet-v2 by 2.02%, 1.70%, 1.56% and 1.71%, respectively, in terms of the top-1 error and by 1.80%, 1.36%, 1.02% and 0.89%, respectively, in terms of the top-5 error. This demonstrates that visual clues can improve the classification performance. Compared with SENet, ACENet decreases the top-1 error by 0.14%, 0.27% and 0.29% with the backbones VGG-16, ResNet-50 and ResNet-101, respectively. To our surprise, ACENet achieves a large improvement (1.55%) compared with SENet when using Inception-ResNet-v2 as the basic model. From the table, we also find that ACENet has a minor top-5 error increase (0.05%) compared with SENet when using ResNet-101 as the backbone. From the table, we find that
our approach with the backbone EfficientNet-B7 improves the image classification performance compared with EfficientNet-B7 and exceeds the state-of-the-art methods (SENet is used as state-of-the-art when using the Inception-ResNet-v2 backbone) by 0.83–7.83% in terms of the top-1 error rate. From the table, we notice that ACENet can generally achieve better results with a stronger backbone. Regarding the number of parameters, the increase of parameters in ACENet compared to backbones are distributed in the modules of clue-attention network and two fully connected layers. Take ResNet-50 implemented on ImageNet as an example, ACENet with the ResNet-50 backbone introduces approximately 5.7M and 3.7M additional parameters beyond the parameters in the ResNet-50 backbone (42.7M) and ResNet-50 baseline (44.7M), respectively. As shown in Table 1, we observe that ACENet with the backbone EfficientNet-B7 achieves a decrease of 0.83% in terms of the top-1 error rate compared with BoTNet-S1-128 although it uses less parameters.

In Table 2, we show the results of the ablation study for three configurations: (1) removing the clue-attention network and contrastive loss (i.e. performing classification based on the combination of original images and clue proposals using backbones), (2) keeping only the clue-attention network, and (3) keeping both the clue-attention network and contrastive loss (i.e. the proposed ACENet). The three configurations correspond to the three rows for each backbone in the table. Compared with the baselines in Table 1 that employ only the global image information, we observe that the introduction of clue proposals to the classification (i.e. configuration 1) can enhance the performance. From Table 2, we find that the clue-attention network and contrastive loss can improve the image classification performance. For example, for the backbone ResNet-101, the top-1 error decreases from 23.27% to 22.58% when introducing the clue-attention network only and reaches 22.09% if we further consider the contrastive loss term. The results show that both clue attention and the contrastive loss play important roles in improving the classification performance by strengthening discriminative features and suppressing common visual information across the different categories. In Figure 5, we show the change in the top-1 error rates in the training and testing of our approach and the ResNet-50 and ResNet-101 baselines. From the figure, we can observe that ACENet achieves lower error rates in training and testing processes than the baselines.

Figure 6 illustrates the change in clue-attention distribution $\alpha$ during training, where the height of a bar reflects the degree of attention given to a clue proposal in an epoch. We observe that, as the training epochs increase, the attention distribution changes and converges to the state that the clue proposals that
FIGURE 6 The change in attention distribution during training for an image of the ‘toy terrier’ category, where cp1 to cp10 denote 10 clue proposals extracted from the original image.

FIGURE 7 An example illustration of extracting clues for several categories in the image classification task. The red, green and blue bounding boxes shown in each image represent the top three clue proposals in descending order of attention. From the figure, we observe that these clue proposals, especially the red proposals, possess good discriminability to determine the category to which images should belong. For example, as shown in the second image of the category ‘red wolf’, the top clue proposal (red) accurately corresponds to the region of a red wolf that excludes the background and other objects independent of this category.

look more discriminative are given more attention to e.g., ‘cp10’ in this figure. Clearly, this clue proposal seems to possess good discriminability to classify the image into the ‘toy terrier’ category. In Figure 7, we show an example illustration extracting clues for several categories in the image classification task. The red, green and blue bounding boxes shown in each image represent the top three clue proposals in descending order of attention. From the figure, we observe that these clue proposals, especially the red proposals, possess good discriminability to determine the category to which images should belong. For example, as shown in the second image of the category ‘red wolf’, the top clue proposal (red) accurately corresponds to the region of a red wolf that excludes the background and other objects independent of this category.

5 | CONCLUSION

This paper proposes an attention-based visual clue extraction approach that seeks to determine what clues encode the information of categories and can help improve the image classification performance. In this approach, we first construct a clue-attention mechanism called global-local attention between an image and its clue proposals, which permits the clue proposals to make varying contributions to image classification. Then, we introduce a contrastive loss in the training to encourage considerable attention to be devoted to true visual clues that possess discriminative information. The experimental results show that the proposed approach can effectively extract the visual clues that reflect the category information, and can
improve the classification performance compared with the baselines and the state-of-the-art method.

The visual clue extraction is an important issue in image classification that can make the results more interpretable, and thus can be used in the scenarios with the requirement of high interpretability, such as medical image classification and negative web image recognition. In addition, the current work in the paper is performed on the standard NWI and ImageNet datasets. Actually, there are a considerable amount of data with noisy labels or without labels on the Internet that are useful to enhance the image classification performance. In the future, it will be addressed to incorporate these non-standard data into ACENet under the weakly-supervised or semi-supervised learning paradigm for online intelligence applications.

ACKNOWLEDGEMENTS
This work was supported by the National Key Research and Development Program (2018AAA0101501) of China.

ORCID
Youtian Du https://orcid.org/0000-0002-1714-3433

REFERENCES
1. Wu, M., et al.: Object detection based on RGC mask R-CNN. IET Image Process. 14(8), 1502–1508 (2020)
2. Du, Y., et al.: Fundamental visual concept learning from correlated images and text. IEEE Trans. Image Process. 28(7), 3598–3612 (2019)
3. Chen, X., Shrivastava, A., Gupta, A.: Enriching visual knowledge bases via object discovery and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, Ohio, USA, 2027–2034 (2014)
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA. 886–893 (2005)
5. Heikkilä, M., Pietikäinen, M., Schmid, C.: Description of interest regions with local binary patterns. Pattern Recognit. 42(3), 425–436 (2009)
6. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60(2), 91–110 (2004)
7. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded up robust features. In: Proceedings of the European Conference on Computer Vision. Graz, Austria. 404–417 (2006)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. Lake Tahoe, Nevada, USA. 1097–1105 (2012)
9. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:14091556 (2014)
10. Szegedy, C., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 1–9 (2015)
11. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, Nevada, USA. 770–778 (2016)
12. Huang, G., et al.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 4700–4708 (2017)
13. Zoph, B., et al.: Learning transferable architectures for scalable image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. 8697–8710 (2018)
14. Mnih, V., et al.: Recurrent models of visual attention. In: Advances in Neural Information Processing Systems. Montreal, Quebec, Canada. 2204–2212 (2014)
15. Ge, W., Lin, X., Yu, Y.: Weakly supervised complementary parts models for fine-grained image classification from the bottom up. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA. 3034–3043 (2019)
16. Jaderberg, M., et al.: Spatial transformer networks. In: Advances in Neural Information Processing Systems. Montreal, Quebec, Canada. 2017–2025 (2015)
17. Xu, K., et al.: Show, attend and tell: Neural image caption generation with visual attention. In: Proceedings of the International Conference on Machine Learning. Lille, France. 2048–2057 (2015)
18. Chen, L., et al.: 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. Honolulu, HI, USA. 5659–5667 (2017)
19. Wang, F., et al.: Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 3156–3164 (2017)
20. Zhao, H., Jia, J., Koltun, V.: Exploring self-attention for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA. 10076–10085 (2020)
21. Ren, S., He, K., et al.: Faster R-CNN: Towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39(6), 1137–1149 (2017)
22. Zhang, T., et al.: Interleaved group convolutions. In: Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy. 4373–4382 (2017)
23. Howard, A.G., et al.: MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv:170404861 (2017)
24. Zhang, et al.: ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. 6848–6856 (2018)
25. Tzeteas, et al.: Deep neural network compression for image classification and object detection. In: Proceedings of the IEEE International Conference on Machine Learning and Applications. Boca Raton, Florida, USA. IEEE, pp. 1621–1628 (2019)
26. Woo, S., et al.: CBAM: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision. Munich, Germany. 3–19 (2018)
27. Hu, J., et al.: Squeeze-and-excitation networks. IEEE Trans. Pattern Anal. Mach. Intell. 42(8), 2021–2025 (2020)
28. Zheng, H., et al.: Learning multi-attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy. 5209–5217 (2017)
29. Mafla, A., et al.: Fine-grained image classification and retrieval by combining visual and locally pooled textual features. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Snowmass Village, Colorado, USA. 2950–2959 (2020)
30. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. 6479–6488 (2018)
31. Zhang, S., Li, H., Kong, W.: Object counting method based on dual attention network. IET Image Process. 14(8), 1621C–1627 (2020)
32. Lipiens, G., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88 (2017)
33. Divvala, S.K., Farhadi, A., Guestrin, C.: Learning everything about anything: Webly-supervised visual concept learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, Ohio, USA. 3270–3277 (2014)
34. He, K., et al.: Mask R-CNN. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Venice, Italy. 2961–2969 (2017)
35. Uijlings, J.R., et al.: Selective search for object recognition. Int. J. Comput. Vis. 104(2), 154–171 (2013)
36. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems. Long Beach, CA, USA. 5998–6008 (2017)
37. Tan, M., Le, Q: EfficientNet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the International Conference on Machine Learning. Long Beach, CA, USA. 6105–6114 (2019)
38. Wang, C.Y., et al.: CSPNet: A new backbone that can enhance learning capability of CNN. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Seattle, WA, USA. 390–391 (2020)
39. Srinivas, A., et al.: Bottleneck transformers for visual recognition. arXiv:210111605 (2021)

How to cite this article: Cui Y, Du Y, Wang X, Wang H, Su C. Leveraging attention-based visual clue extraction for image classification. IET Image Process.. 2021;1–11. https://doi.org/10.1049/ipr2.12280