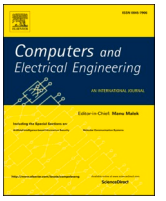
Computers and Electrical Engineering 101 (2022) 108068

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng



Micro-Knowledge Embedding for Zero-shot Classification

Houjun Li [a], Fang Wang [a], Jingxian Liu [a], Jianhua Huang [a], Ting Zhang [b], Shuhong Yang [a,*]

[a] School of Electrical, Electronic and Computer and Computer Science, Guangxi University of Science and Technology, Liuzhou, 545006, Guangxi Zhuang Autonomous Region, China
[b] Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China

ARTICLE INFO

Keywords:
Zero-Shot
Classification Framework
Micro-Knowledge
Self-Attention

ABSTRACT

Zero-shot learning is one of the most challenging machine learning tasks, in which learning stable and transferable knowledge from seen classes plays a pivotal role. To improve the currently unsatisfactory performance of zero-shot object recognition, this paper proposes a novel image representation method, namely, micro-knowledge. In our method, the segmentation of micro-regions and the consequent learning of micro-knowledge are unified by the introduction of a self-attention mechanism. A zero-shot classification framework is carefully designed based on micro-knowledge of images. Under this framework, multiple micro-region descriptions are first obtained by embedding micro-knowledge and then merged to carry out the final classification of unseen objects. Finally, a capsule-unified framework is employed as a graphical programming tool to accomplish the aforementioned tasks. Experiments on public datasets show that the proposed framework can generally achieve competitive results for the classification of unseen objects. Specifically, these results verify that the micro-knowledge learned from one dataset can be directly applied to others without complicated adjustments and demonstrate that using visual features instead of semantic features can result in a decrease in classification error. This research will bring new ideas into the field of zero-shot learning and will serve as an appealing option when addressing the problem of domain shift.

1. Introduction

Programming machines to behave like humans and recognize new objects through association and learning is one of the most challenging tasks in the field of computer vision. Deep neural networks achieve impressive results in the field of computer vision such as image recognition and 3D object retrieval[1,2], but they mostly belong to supervised learning methods, which require a large amount of annotated data to train the models. However, there are no datasets yet that can cover all categories, let alone that many new categories are constantly created with the development of human society. For traditional supervised learning, identifying a new category always requires a large amount of labeled data, which is both expensive and time-consuming for most real-world applications.

Therefore, how to enable machines to recognize new categories never “seen” before is a challenging task of great practical significance. On the one hand, the number of categories in the real world is enormous and growing; on the other hand, such ability of machine learning is inevitably demanded in a wide range of practical applications, such as cross-language translation, unknown image

* Corresponding author.
E-mail address: shuhongyang@gxust.edu.cn (S. Yang).

https://doi.org/10.1016/j.compeleceng.2022.108068
Received 22 February 2022; Received in revised form 1 May 2022; Accepted 4 May 2022
Available online 8 May 2022

synthesis, classification of unseen scenes and so on.

Zero-shot learning (ZSL) was proposed precisely to address the classification problem of unseen objects. Its first proposal dates back to 2008, when Larochelle et al. proposed a zero-data learning method for character classification problems and achieved a classification score of 60% [3]; then, Lampert et al. released the Animals with Attributes Dataset (AWA) and proposed a classic algorithm for attribute-based learning [4], which drew wide attention to ZSL. ZSL maps each image to a predefined semantic feature space through a group of classifiers and then identifies the previously unseen objects according to the descriptions and simple rules of experience [5,6,7]. The basic idea of ZSL can be summarized as follows: it first builds a vectorized semantic feature space, then learns a feature mapping on seen objects by supervised learning, and next it obtains a reasonable mapping from the visual features to the semantic feature space, finally, it extends the trained classifier from the seen to the unseen and gets the job done.

In recent years, the representative works of ZSL can be roughly classified into several categories as follows,

(1) Property-based ZSL methods, such as indirect attribute prediction (IAP) [4], cross-modal transfer (CMT) [8], latent discriminative features [9]. These are pioneer works in the field of ZSL, and their ideas lay the foundation of the following development of ZSL. However, they were limited in that other auxiliary information cannot be well integrated into their models.
(2) Semantic embedding-based ZSL methods, which include the semantic autoencoder [10], semantically consistent regularization [11], and visual center learning [12]. These methods closely combine semantic information with image information, and have certain generalization ability for unseen classes, but it is prone to the problem of semantic gap.
(3) Generalized zero-shot learning (GZSL) methods. Based on the idea of ZSL, GZSL adds the assumption that the samples to be recognized may come from unseen classes, which is therefore faced with an additional issue of model bias. These methods include the dual adversarial semantics-consistent network [13] and the TF-VAEGAN approach [14].
(4) Generative model-based ZSL methods. Examples include the semantic autoencoder method [15], feature generating networks [16], and invertible zero-shot recognition flows [17]. These methods usually use the generative model to embed semantic information into image space. However, such methods also have the problem of domain shift, and usually need large amount of calculation and long training time.

Additionally, ZSL has many other applications in a wide range. To name a few, in target detection, there is zero-shot YOLO, which is put forward based on the fusion of semantic and visual information [18], and zero-shot semantic segmentation [19].

Although research on ZSL has achieved remarkable progress, there is still a large space for further improvement. According to our investigation, factors that bottleneck the classification score in zero-shot conditions are mainly concerned with two issues. First, in the process of semantic embedding, different descriptors are prone to mutual interference between each other when they are considered as a whole. Second, this undesirable situation will worsen as more descriptors are added continually to expand the completeness of semantic space.

Additionally, traditional ZSL methods generally lack stable, detailed, and transferable feature representations of the target object, which limits their efficiency in transferring the knowledge learned from seen objects to unseen ones, and thus may result in large identification bias. In order to overcome these problems, we put forward to exploit the power of micro-knowledge for image representation. Although there are many classification methods based on fine-grained features reported with classification scores of approximately 90% on the bird fine classification database CUB-200-2011 [20], such as [21]. However, if they are applied in zero-shot scenarios their performances will be inevitably unsatisfactory, as the aforementioned performance is achieved by training with labeled data. Micro-knowledge for image representation is presented in detail in Section 2, and it can be easily distinguished from fine-grained features by the description of the latter in related references [21].

Micro-knowledge for object representation is the knowledge learned to represent specific image regions through sets of descriptors for partial attributes. While lying between global and pixel features, micro-knowledge can be expected to promote both discrimination and stability of feature learning. In fact, it is robust feature, and play a critical role in the similarity measurement [22]. At the same time, there will be more properties (attributes, features) shared among different objects, which will certainly facilitate knowledge transfer for zero-shot classification. Experimental results show that the proposed micro-knowledge can improve the score of zero-shot classification.

Our contributions in this paper are as follows:

1) We propose micro-knowledge as a novel way of representing image-based attributes in an unsupervised manner with the adoption of a self-attention mechanism.
2) We explore a zero-shot recognition framework in which multiple micro-regions descriptions are first obtained by the embedding of micro-knowledge and then merged to carry out the automatic recognition of unseen objects.
3) We make the first attempt to leverage a capsule-unified framework [23] to solve zero-shot recognition problems in a low code manner, which not only improves the development efficiency but also makes the network structure clearer.

2. Methodology

2.1. Micro-Knowledge

Micro-knowledge for images is defined as an image-based mapping from micro-region features to the semantic feature space. Intuitively, when we observe objects with fine enough granularity, they always share similar characteristics, such as the shape of a

claw, the shape of a beak, the color of hair, etc., which can often be used by humans to describe newly discovered objects. Therefore, it is justifiable to believe that micro-knowledge can be employed to recognize unseen objects and is relatively stable.

Suppose the target image X_O contains an object composed of N semantic descriptors, which construct a semantic attribute set W, recorded as $W = \{w_{o,t}\}$, $i = 1, \cdots, N$. The micro-knowledge can be defined as:

$$MK_t : f_t(x_{o,t}) \rightarrow W_{o,t},\ t = 1, \cdots, m \tag{1}$$

where MK_t denotes the t-th micro-knowledge of X_O; $f_t()$ is the function that represents X_O's t-th micro-region feature; and $x_{o,t}$ is the t-th focus region of X_O. Each focus region constitutes an independent feature unit and corresponds to some partial attributes of W, such as words that jointly describe the head or tail of animals. Thus, attribute descriptors can be derived from focus micro-regions, that is, $W_{o,t} = \{w_{o,k} | x_{o,t} \Rightarrow w_{o,k}\} \subset W$. The micro-regions distinguish micro-knowledge from traditional ZSL methods, which need to learn mapping from the overall features of the target object to the set of semantic descriptors. We disassemble the overall features of the target object into a series of focus micro-regions, as

$$f(X_O) \sim \cup_{t=1}^{m} \left(f_t(x_{o,t}) \right) \tag{2}$$

Therefore, the final classification result of the target object can be obtained by combining micro-knowledge from multiple micro-regions in the image. This is mainly inspired by the idea of bottom-up cognition process, i.e. starting with fine-grained details and end up with micro-knowledge learned as universally as possible. In addition, we let $f_{ti} \cap f_{tj} = \varnothing$ to avoid the interference of multiple descriptors with little relevance, which would affect the judgment of unseen objects.

2.2. Zero-shot Classification Framework Based on Micro-Knowledge

Using image micro-knowledge, we designed the zero-shot classification framework shown in Fig. 1, which is divided into two parts: the micro-region features extractor and the micro-knowledge learner.

In the micro-region feature extractor, we incorporate a self-attention mechanism that allows the model to focus on the local details of the target object autonomously. Additionally, the framework combines the segmentation of focused micro-regions with the extraction of regional features, i.e.,

$$f_t(x_{o,t}) = MRF(x_{o,t}),\ where\ x_{o,t} \in \sigma_{Top-m}(AM(featureMap(X_o))) \tag{3}$$

where *featureMap*() denotes the function of a feature map extracting focused micro-region features from the input image; *AM*()

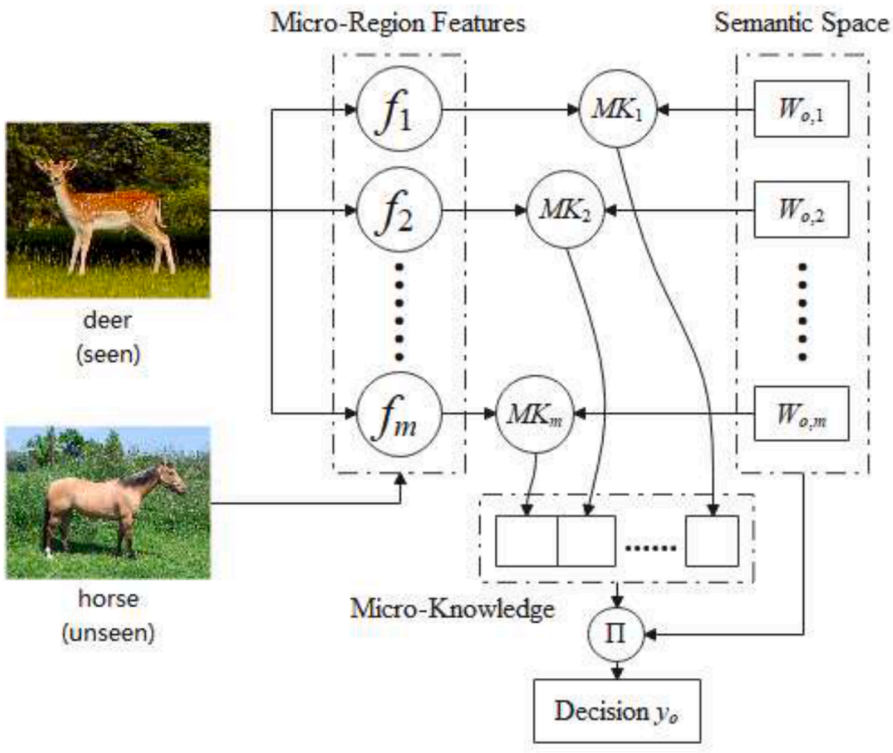


Fig. 1. The zero-shot classification framework based on micro-knowledge.

represents an attention mask obtained by a self-attention mechanism; $\sigma_{Top-m}()$ denotes selecting the top-m self-focus of attention; and $MRF()$ is the micro-region feature extracted from region $x_{o,t}$, which is a function related to $featureMap(X_o)$. Therefore, region segmentation and feature extraction can be completed simultaneously in the same process by using a self-attention mechanism.

Finally, we classify an unseen object with the proposed framework by using all the micro-knowledge learned from the seen objects. The judgment is defined as follows:

$$y_o = Decision\left(\cup_{t=1}^{m} \overline{W}_{o,t}\right) = \arg\min_{W_k \in W} \| \overline{W}_o - W_k \| \tag{4}$$

where the symbol $\cup$ represents the union of m micro-knowledge and $\overline{W}_{o,t} = \omega^T f_t$ can be learned using a simple model, such as a fully connected network. It is a mapping from features of focused micro-region to attribute descriptors of the object, and $\overline{W}_o = \cup_{t=1}^{m} \overline{W}_{o,t}$.

Note that the zero-shot classification framework in Fig. 1 is modular. The micro-region feature extractor can be easily replaced with other approaches, such as vision transformers. The attribute descriptors can be flexibly expanded. For example, the head of an animal can be described as a whole to learn one micro-knowledge or can be divided into three parts: eyes, nose and mouth to learn three parts of micro-knowledge. Theoretically, when the resolution of an image is high enough, the more detail of the attribute descriptors is, the better stability and transferability of the micro-knowledge can be learned. The number of micro-knowledges can be adjusted according to the granularity of the attribute descriptor. In Section 2.3, we specifically discuss how the self-attention mechanism works in the framework, and in Section 2.4, we discuss how micro-knowledge can be learned.

2.3. Self-Attention Mechanism

Similar to the behavior of human, there is no need to observe all the details of the concerned object to recognize it; what is needed is just to focus on the salient features. The purpose of self-attention is to find the salient feature areas of an image that are often distinctive regions used to recognize objects. Here, we used the nonlocal block [24] shown in Fig. 2 as a self-attention. We represent the nonlocal block using the capsule-unified framework [23], by which the description of deep neural networks can be more easily. In Fig. 2, *N, H, W,* and *C* represent the batch size, height, width, and channels of a capsule, respectively; symbols * and ◁ indicate convolution connection with 1*1 kernel and reshape operation, respectively. The meaning of the standard capsule symbols used in Fig. 2 is shown in Fig. 3.

Furthermore, using a pretrained network and a nonlocal block, the micro-region features of the input image can be extracted with the following steps. First, the pretrained network is used to generate feature maps from the input image; next, a nonlocal block module is added to the feature map to find focused areas; and finally, micro-region features are extracted based on the fusion results of multiple nonlocal block modules. In the experiments, we use ResNet101, which is a model pretrained on ImageNet. The procedure of extracting micro-region features is shown in Fig. 4. Nonlocal block modules 1∼3 are added into the 2nd, 3rd, and 4th residual layers of ResNet101, respectively. Nonlocal block 1 is used as the attention mask, combined with nonlocal block 3 to extract micro-region features. The attention mask can be defined as:

$$\begin{aligned} &AM(NLB_1) = M_k \\ &s.t.\ M_k(i,j) = \frac{1}{k^2} \sum_{D((i,j),k)} \left(\frac{1}{c} \sum_{u} NLB_1(i,j,u) \right) \end{aligned} \tag{5}$$

where NLB_t denotes the nonlocal block t (t=1, 2, 3); $D((i, j), k)$ denotes the k-neighborhood with the center coordinate (i, j); and c is the channel of NLB_1. As shown in Fig. 4, nonlocal block 2 is only used as a transition module for feature map generation and is used for micro-region feature extraction directly. Eq. (3) provides the top-m micro-region features of interest. In fact, due to the characteristics of ResNet101, each central point of the top-m focus regions obtained in the attention mask corresponds to just one of the feature points generated by NLB_3. Therefore,

$$MRF\left(x_{o,t}\right) = NLB_3(i,j,:),\ where\ x_{o,t} \supset D((i,j),k)\ and\ x_{o,t} \in \sigma_{Top-m}(M_k) \tag{6}$$

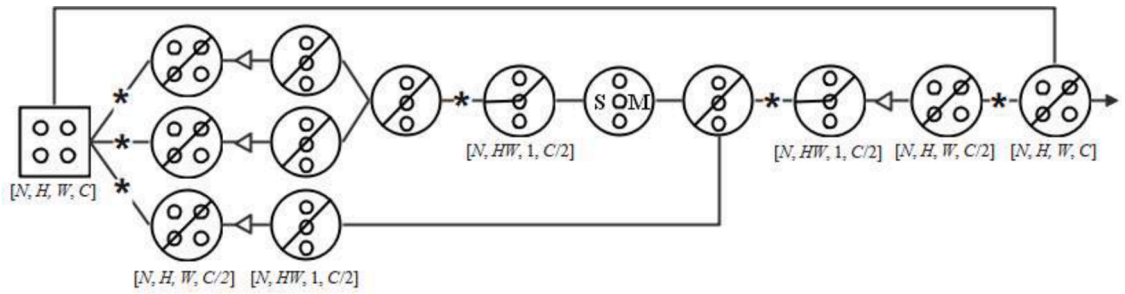


Fig. 2. A capsule graph of the nonlocal block.

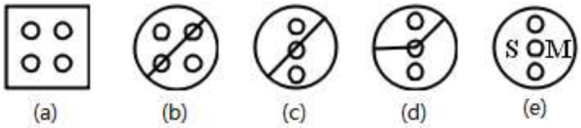


Fig. 3. Standard capsule symbols in Fig. 2: (a) 2D-data capsule, (b) 2D-identical capsule, (c) 1D-identical capsule, (d) 1D-ReLU capsule, (e) 1D-Soft-Max capsule.

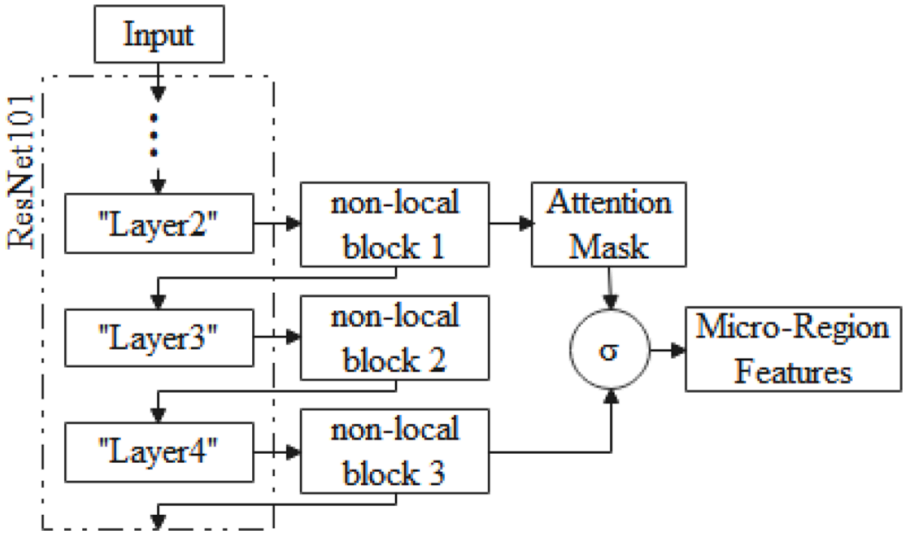


Fig. 4. Calculation of micro-region features.

where (i, j) is the center of both the D neighborhood and $x_{o,t}$. $MRF(x_{o,t})$ can produce a vector with 2048 dimensions.

2.4. Micro-Knowledge Learning

The fully connected network can be used to learn micro-knowledge from micro-region features. To learn different micro-knowledge, we create a series of networks with same structure, which is shown in Fig. 5.

The loss function used to train the fully connected networks is defined as:

$$loss\left(\overline{W}_{o,t},\ W_{o,t}\right) = BCEWithLogitsLoss\left(\overline{W}_{o,t},\ W_{o,t}\right) \tag{7}$$

where *BCEWithLogitsLoss*() denotes the cross-entropy function with an additional sigmoid layer, that is,

$$BCEWithLogitsLoss\left(\overline{W}_{o,t}, W_{o,t}\right) = \frac{1}{M}\sum_{i}[-y_i \log \widehat{y}_i - (1-y_i)\log(1-\widehat{y}_i)] \tag{8}$$

where $\begin{cases} y = Sigmod(W_{o,t}) \\ \widehat{y} = Sigmod(\overline{W}_{o,t}) \end{cases}, i \in \{1, 2, \cdots, M\}$.

The micro-region feature is a sequence extracted from the image feature map with the coordinates (i, j) of each channel. Therefore, the output of $NLB_3$is an eigenvector with 2048 dimensions. When we obtain an M focus from the attention mask, there are M different micro-region features. Therefore, the micro-knowledge includes M fully connected networks trained independently utilizing micro-region features.

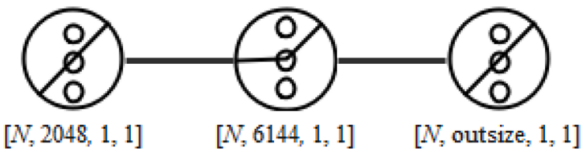


Fig. 5. The fully connected network used for micro-knowledge learning.

3. Experiments

3.1. Experimental Setup

Datasets: To validate the effectiveness of the proposed micro-knowledge representation method for zero-sample classification, we selected two public datasets for experiments. AwA2 [25] contains 30,745 pictures of 50 animals with 85 attributes. CUB200-2011 [20] contains 11788 images of 200 fine-grained birds with 312 attributes.

In experiments, we not only divide seen and unseen classes with the proposed splits (PS) recommended by [25] but also deliberately select a subset of very similar animals or birds as unseen classes to analyze the feasibility of our method. In AWA2, three categories (horse, rat, and bat) were selected as unseen classes, and the remaining objects were employed as seen classes to train the micro-knowledge module. As shown in Fig. 6, the horse is rather different from the other two classes, while the rat and bat are quite similar. In CUB200-2011, ten kinds of birds out of 200 to were selected be unseen classes, while the rest 190 kinds were used as seen classes, as shown in Table 1. The ten unseen classes roughly contain three categories: *woodpecker* (contains two fine-grained birds), *wren* (contains seven fine-grained birds), and *yellowthroat* (contains one fine-grained bird). Note that woodpecker contains six fine-grained birds in the CUB200-2011, so there are four fine-grained birds selected as seen classes for training, which include *American three-toed Woodpecker, pileated woodpecker, red-bellied woodpecker*, and *red-cockaded woodpecker*. Wren and yellow throats do not appear in the seen classes. The reason for this treatment is to make the experimental results more universal and closer to practical application.

Implementation details: In our experiments, the pretrained ResNet101 was downloaded from the official website of PyTorch. To train the three nonlocal blocks shown in Fig. 4, we selected 12 categories from AwA2 as the training set to train the self-attention module. These 12 selected categories contain 7,856 images. They are *killer whale, beaver, dalmatian, skunk, tiger, hippopotamus, gorilla, chimpanzee, giant panda, pig, lion*, and *polar bear*. All the categories can be found in ImageNet and can be recognized with high scores using a pretrained ResNet101. In self-attention module training, the cross-entropy function *CrossEntropyLoss* was used as the loss function. After ten epochs of training, the loss value is reduced to less than 0.001, and an accurate attention mask of the image is obtained with efficient generalization ability. As shown in Fig. 7, the trained self-attention module can be utilized directly in CUB200-2011 and has accurate attention masks. As shown in Fig. 7, after the processing of self-attention, the focus in images basically falls within the target object, and the brightest area is generally concentrated on the head or body of animals, which is in line with the logic of humans. Attention masks of birds work better because the bird images are segmented out of the background beforehand. As seen in the attention masks of birds, their main focus areas are located on the heads, bellies, backs, and tails of the birds, which are easy to extract as micro-regions.

3.2. State-of-the-art Comparison

We compare our method with TF-VAEGAN [14], which outperforms many state-of-the-art approaches. The code of TF-VAEGAN is publicly available, and we modified the feature extractor code based on the methods recommended by [25] using ResNet101 pretrained on ImageNet. We compared different methods using the image recognition accuracy as the classification score. Table 2 shows the results of the Top-1 classification score on two datasets with proposed splits (PS) [25]. For ZSL, the TF-VAEGAN obtains classification scores of 65.7% and 7.7% on AWA2 and CUB200-2011 with segmented images (denoted as CUB200-2011-seg), respectively. On the original CUB200-2011 (denoted as CUB200-2011-org), TF-VAEGAN scores slightly higher, at 8.3%. For TF-VAEGAN, environmental features can improve the classification score. However, this is not conducive to learning the essential characteristics of objects. Our approach uses segmented images and outperforms TF-VAEGAN on CUB200-2011-seg with a classification score of 19.9% but 36.1% on AWA2. Our method is effective for segmented images and can learn the detailed features of objects, that is, micro-knowledge.

In Table 2, the performance of our method is inferior to that of TF-VAEGAN on AWA2 because the 85 attributes given by AwA2 were taken as one partition, that is, setting $m = 1$ in Eq. (1). Many categories in AwA2 contain just a few attributes, and some of those attributes are background words that do not appear in all images. As shown in Fig. 7, there are several brightest areas on the head and body of animals, but we only use mean points to match attribute descriptors. This leads to low classification scores. On CUB200-2011, the attributes given by the CUB200-2011 dataset can be divided into five parts: overall feature, head feature, breast-belly feature, tail feature, and back-wing feature, as shown in Table 3. The third column of Table 3 lists the dimensions of the features of the five parts, which correspond to five different semantic subspaces with non-overlapping attribute descriptors. Therefore, five classes of micro-

Fig. 6. Examples of unseen classes in AwA2.

Table 1
The ten unseen classes selected from CUB200-2011.

No.	Rough classes	Fine-grained classes
1	Woodpecker	Red_headed_Woodpecker
2		Downy_Woodpecker
3	Wren	Bewick_Wren
4		Cactus_Wren
5		Carolina_Wren
6		House_Wren
7		Marsh_Wren
8		Rock_Wren
9		Winter_Wren
10	Yellowthroat	Common_Yellowthroat

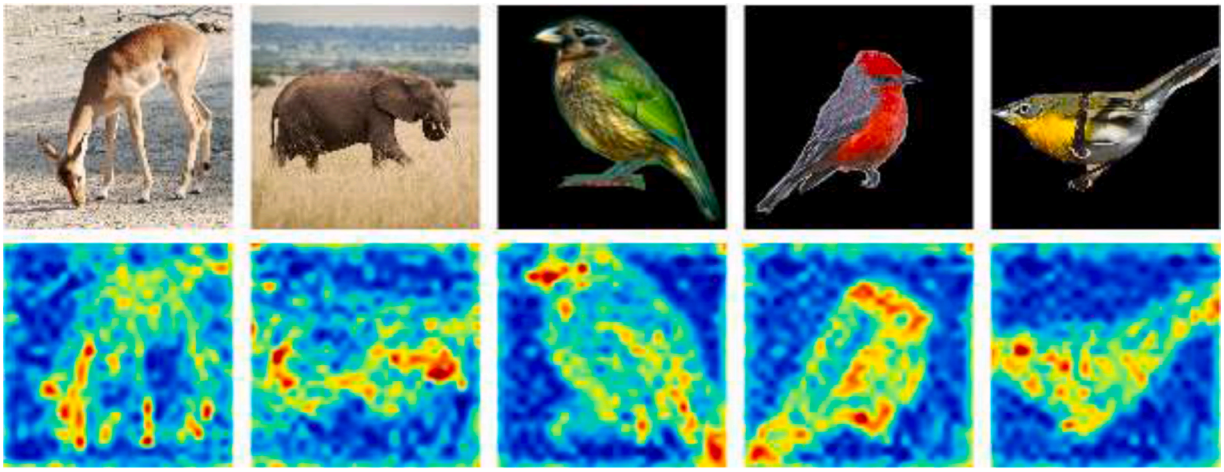


Fig. 7. Attention mask generated by our method.

Table 2
State-of-the-art comparison on two datasets with proposed splits. "CUB200-2011-seg" and "CUB200-2011-org" indicate experiments with segmented and original images, respectively.

Methods	AWA2	CUB200-2011-seg	CUB200-2011-org
TF-VAEGAN	**65.7%**	7.7%	8.3%
Our Approach	36.1%	**19.9%**	——

knowledge can be learned from the images using five fully connected networks. Because the CUB200-2011 dataset provides bird images and segmentation masks, the target objects, i.e., the birds, can be extracted from each image without background, as shown in Fig. 7. Therefore, the classification score is higher than that of TF-VAEGAN. Experiments show the effectiveness of micro-knowledge and we can improve the classification score by increasing the focus of objects.

In addition, TF-VAEGAN achieved lower scores than the reports in [14]. The reason is that we used different versions of the package and different runtime environments.

Experiment in AwA2 with our splits. We first used the pretrained ResNet101 to obtain the feature map of the whole image, then obtained the micro-region feature sequences through the attention mask of the image, and finally embedded them into the semantic

Table 3
Feature classification on CUB200-2011.

Divided parts	Features	Dimensions
Overall	shape, size, primary_color, upperparts_color, underparts_color	64
Head	head_pattern, eye_color, forehead_color, crown_color, bill_shape, bill_length, bill_color	82
Breast -Belly	breast_color, breast_pattern, throat_color, nape_color, belly_color, belly_pattern	72
Tail	tail_shape, tail_pattern, under_tail_color, upper_tail_color, leg_color	55
Back-Wing	wing_shape, wing_pattern, wing_color, back_color, back_pattern,	39

feature space to obtain micro-knowledge. The target object is recognized through the attribute characteristics of the image. Experimental results show that, in the case of zero-shot classification, our method performs well in three unseen classes, with 1917 out of 2337 images correctly recognized, i.e., a classification score of 82%, higher than many reported results, for example, 80.7% in [25]. This illustrates that the self-attention mechanism performs rather well in zero-shot classification. The classification performance on the three unseen classes is summarized as follows: The total number of images of horses, rats, and bats is 1644, 310, and 383, respectively, and 1661, 292 and 3 images of the corresponding class are correctly recognized, which means the classification score of these three unseen classes is 99.8%, 94.2%, and 0.8%, respectively. From these results, one can find that the classification scores of horse and rat images are rather high, while the score of bat images is undesirable. By examining the images of these three unseen classes carefully, we argue that the main reasons leading to these results may be analyzed as follows:

(1) As shown in Fig. 6, the images of horses generally include horses with larger bodies against rather simple backgrounds, which can offer rich features and thus bring about nearly perfect classification scores. The classification score on rat images is also rather high because these images have rich features. However, the backgrounds are slightly more complex than the horse images, and multiple focal points of similar intensity appear in them on occasion, which may explain why the classification score on rat images is approximately six percent lower than on horse images. However, the objects to be recognized in the third kind of unseen images, i.e., the bats, often appear in the corresponding images with confusing coloration similar to the background. Furthermore, the bats are curled up in most cases, which makes it difficult even for humans to distinguish them from the rats. We argue that all these adverse factors contribute to the low classification score of bat images.
(2) The 85 attributes given by AWA2 include some nonuniversal feature descriptors, such as the living environments and habits of certain animals. Taking the descriptors for the tooth shape as an example, there are not so many images in which an animal's teeth can be observed; this will bring about a small sample problem in the learning of tooth attributes. There are also some words about backgrounds, such as forest, tree, and water, which appear less appropriate as attributes of unseen classes, and also increase the difficulty of classification. In addition, there are some words about "fish", "meet", "new world", "old world", "hunter", and so on. They cannot be observed in the images but need to be obtained by inference. These inappropriate attributes increase the difficulty of recognizing them and are the reasons for the low classification score of bat images.

Experiment in CUB200-2011 with our splits. The results of the experiments showed that there was a 57.2% classification score in unseen classes, higher than the 55.3% reported in [25]. Using the PS, our method performed better than IAP and was close to CMT, as shown in Table 2. In our experiments, the fine-grained classes of wren and yellowthroat do not exist in the training set. Fig. 8 shows the average training curve for micro-knowledge learning. The training curve decreases steadily with the increase of iterations and is less than 0.05 after 700 iterations. In theory, the classification score could be at a higher level. Further analysis of the results found that the main reasons for the low classification score are as follows:

(1) The unseen classes have very similar features. In unseen classes, two fine-grained woodpeckers have a 75.6% similarity. The similarity between seven fine-grained wrens is shown in Table 4, where 3-9 represents the number of fine-grained classes in Table 1. As shown in Table 4, the highest similarity of wrens is 93.5%, the lowest is 74.1%, and the average is 86.5%. This higher similarity makes it difficult to distinguish between them, because of which the classification score of wrens is only 25.8%, whereas the classification score is 70.8% for woodpeckers and 75% for yellowthroats, as shown in Table 5. If only considering

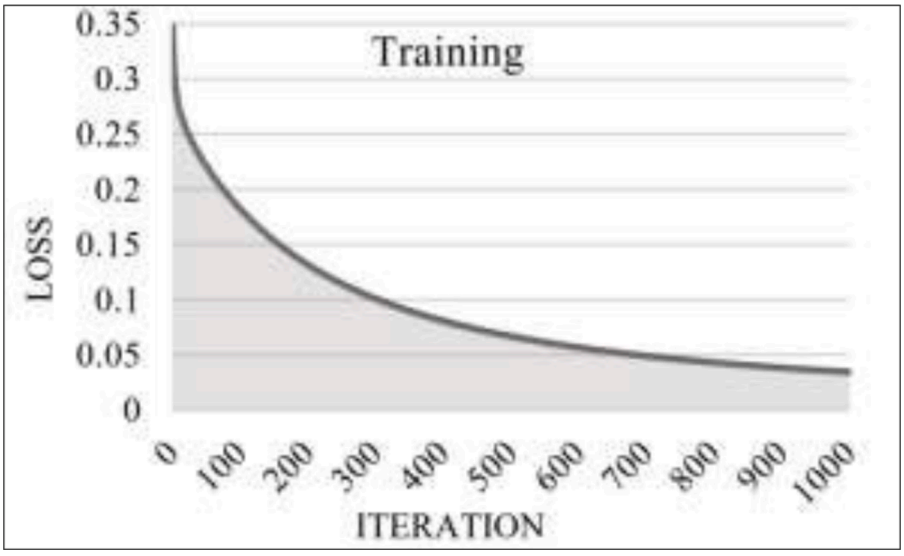


Fig. 8. Curves of the training Training curves.

the rough classification, classifying whether the bird is a *woodpecker, wren*, or y*ellowthroat*, the classification score of unseen classes can reach 91.7%, while both have 100% classification scores in *woodpecker* and *wren*.

(2) Although CUB200-2011 gives 312 attributes, there are multiple descriptors used to describe one feature, such as the attribute "has_wing_color", which has 15 descriptors, all of which are different color words. Therefore, for an individual micro-region, its features are sparse, and its description is relatively rough. This leads to high similarity between birds of the same species and thus result in low classification scores.

3.3. Effectiveness of Micro-Knowledge

To further demonstrate the effectiveness of micro-knowledge for zero-shot classification, we conducted experiments using only the "Overall" feature in Table 3. On CUB200-2011, the classification score was only 15.5%. Compared with the results of Table 2, using micro-knowledge showed an improvement in the classification score. As shown in Eq. (1), our network can learn a large amount of common micro-knowledge for the image dataset containing animals. As shown in Eq. (2), each image is represented by a specific amount of micro-knowledge. For the experiments on the CUB200-2011, we took only five kinds of micro-knowledge, but needed to classify 50 categories. Too little micro-knowledge is also the main reason for the low Top-1 classification score. We then calculate the Top-5 and Top-10 classification scores, both of which are greatly improved, at 41.6% and 63.9%, respectively, as shown in Fig. 9.

Considering the defects of the semantic features in CUB200–2011, such as multiple words used to describe the same attribute and coarse descriptors, we randomly selected an image from each class and replaced the semantic feature vector with the visual features as the feature representation of that class. That is, we substituted the dataset's 312-dimensional semantic characteristics with 2048-dimensional micro-knowledge derived from the model in Fig. 4. We do not require any training in CUB200-2011, but we actually achieve a Top-1 classification score of 21.3% on CUB200-2011, which is higher than the case of using 312-dimensional semantic features. This law also exists for the Top-5 and Top-10 classification scores, which reach 48% and 72.3%, respectively, as shown in Fig. 9. On the CUB200-2011, we also conducted studies in the seen and unseen category divisions. To divide the images, we employed a voting mechanism based on Top-K classification scores, and the results are described in Table 6. Obviously, when using the Top-5 results for division, the classification score reached 65.9%, and when using the Top-20 results, the classification score increased to 71.5%. This also demonstrates that micro-knowledge can distinguish between seen and unseen categories and has high transferability. This allows us to achieve good results without retraining when applied to a new dataset.

4. Conclusion

In the presented work, we propose micro-knowledge as a novel way to represent image-based attributes in an unsupervised manner with the adoption of a self-attention mechanism. We carefully design a zero-shot classification framework based on micro-knowledge, and employ the capsule-unified framework as a graphical programming tool to draw a deep neural network in an intuitive and natural way. To the best of our knowledge, this is the first application of the capsule-unified framework. Extensive experiments are carried out on AWA2 and CUB200-2011, and the results demonstrate that micro-knowledge can improve zero-shot classification scores. On AWA2, we achieved a Top-1 classification score of 36.1% while achieving a Top-1 classification score of 82% among the three unseen categories we selected. On CUB200-2011, using five micro-knowledge can increase the Top-1 classification score from 15.5% to 19.9% and achieve a Top-5 classification score of 53.1%. For these results, the feature extracting model was just trained on AWA2 and then directly applied to CUB200-2011. Note that no birds are included in AWA2, but we also achieve competitive results for bird classification. This verifies that the prosed micro-knowledge is well transferable and cross-domain adaptable, i.e., the micro-knowledge learned in one dataset can be directly applied to the other new datasets without complicated adjustment.

However, due to the small amount of micro-knowledge available in the experiments and a lack of semantic features that constitute micro-knowledge, the Top-1 classification score on the CUB200-2011 is unsatisfactory, which leaves space for future improvement. In fact, the classification framework we designed for the zero-sample problem is modular, in that for different problems it can be easily replaced with different attention models such as transformers, and various micro-knowledge learning approaches can be adopted to achieve better classification results. This is also one of the directions of our future work.

Author Statement

Houjun Li: Writing- Original draft preparation, modeling, Methodology.
Fang Wang: programming and debugging.
Jingxian Liu: Visualization, Validation.
Jianhua Huang: Data Curation, debugging.
Ting Zhang: Language editor
Shuhong Yang: Supervision, reviewing and editing.

Declaration of Competing Interest

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

Table 4
Characteristic similarity matrix of Wren.

Wren	3	4	5	6	7	8	9
3	0	74.1%	**92.8%**	**90.1%**	**92.6%**	84.9%	86.4%
4	74.1%	0	74.8%	77.6%	83.0%	84.1%	80.8%
5	**92.8%**	74.8%	0	**92.6%**	**91.5%**	82.0%	89.6%
6	**90.1%**	77.6%	**92.6%**	0	**93.5%**	89.3%	**93.5%**
7	**92.6%**	83.0%	**91.5%**	**93.5%**	0	**92.3%**	**90.2%**
8	84.9%	84.1%	82.0%	89.3%	**92.3%**	0	81.2%
9	86.4%	80.8%	89.6%	**93.5%**	**90.2%**	81.2%	0

Table 5
Experiment results in CUB200-2011.

Unseenclasses	Fine-grainedclassification	Roughclassification
Woodpecker	70.8%	100%
Wren	25.8%	100%
Yellowthroat	75.0%	75.0%
AVERAGE	57.2%	91.7%

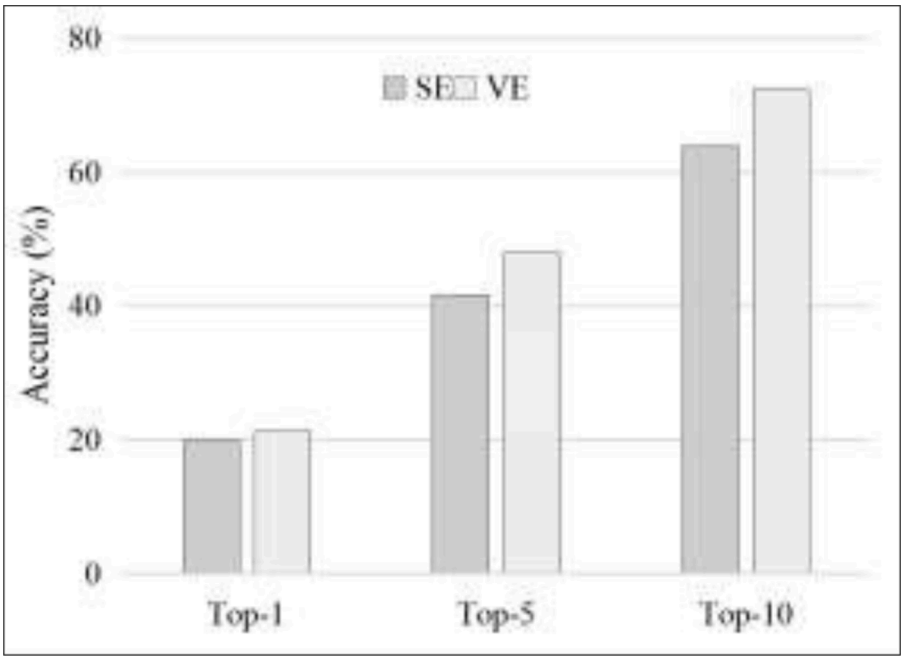


Fig. 9. Experimental results using PS for CUB200-2011. SE = semantic features, VE = visual features.

Table 6
Division of seen and unseen categories using *Top-K* results.

Top-K	classification scores
5	65.9%
10	64.9%
15	71.0%
20	71.5%

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62166002, by the Natural Science Foundation of Guangxi under Grants 2019GXNSFAA245033, 2019GXNSFAA245049, and 2018GXNSFAA050020, and by 2019

Guangxi Education Department Program 2019KY0372.

References

[1] C. Szegedy, S. Ioffe and V. Vanhoucke. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. AAAI Conference on Artificial Intelligence, 2017.
[2] Gao Z, Xue H, Wan S. Multiple Discrimination and Pairwise CNN for view-based 3D object retrieval. Neural Networks 2020:290–302.
[3] H. Larochelle, D. Erhan and Y. Bengio. Zero-data learning of new tasks. AAAI Conference on Artificial Intelligence, 2008: 646-651.
[4] Lampert CH, Nickisch H, Harmeling S. Attribute-Based Classification for Zero-Shot Visual Object Categorization. IEEE Transactions on Pattern Analysis & Machine Intelligence 2014;36(3):453–65.
[5] Rohrbach M, Stark M, Schiele B. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2011. p. 1641–8.
[6] Akata Z, Reed S, Walter D, Lee H, Schiele B. Evaluation of output embeddings for fine-grained image classification. In: IEEE Conference on Computer Vision and Pattern Recognition; 2015. p. 2927–36.
[7] Romera-Paredes B, Torr PHS. An Embarrassingly Simple Approach to Zero-Shot Learning. Visual Attributes. Cham: Springer; 2017. p. 11–30.
[8] Socher R, Ganjoo M, Manning CD, Ng AY. Zero-Shot Learning Through Cross-Modal Transfer. In: 26th International Conference on Neural Information Processing Systems; 2013. p. 935–43.
[9] Li Y, Zhang J, Zhang J, Huang K. Discriminative Learning of Latent Features for Zero-Shot Recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018. p. 7463–71.
[10] Kodirov E, Tao X, Gong S. Semantic Autoencoder for Zero-Shot Learning. In: IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 4447–56.
[11] Morgado P, Nuno V. Semantically Consistent Regularization for Zero-Shot Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 2037–46.
[12] Wan Z, Chen D, Li Y, Yan X, Zhang J, Yu Y, Liao J. Transductive Zero-Shot Learning with Visual Structure Constraint. Advances in Neural Information Processing Systems 2019;32:9972–82.
[13] Ni J, Zhang S, Xie H. Dual Adversarial Semantics-Consistent Network for Generalized Zero-Shot Learning. In: 33rd International Conference on Neural Information Processing Systems; 2019. p. 6146–57.
[14] S. Narayan, A. Gupta, F.S. Khan, C. G. M. Snoek and L. Shao. Latent Embedding Feedback and Discriminative Features for Zero-Shot Classification. 2020 European Conference on Computer Vision. 2020: 479-495.
[15] Kodirov E, Xiang T, Gong S. Semantic autoencoder for zero-shot learning. In: IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 3174–83.
[16] Xian Y, Lorenz T, Schiele B, Akata Z. Feature generating networks for zero-shot learning. In: IEEE Conference on Computer Vision and Pattern Recognition; 2018. p. 5542–51.
[17] Shen Y, Qin J, Huang L. Invertible zero-shot recognition flows. In: 2020 European Conference on Computer Vision; 2020. p. 614–31.
[18] Gupta D, Anantharaman A, Mamgain N, Sowmya KS, vineeth NB, Jawahar CV. A Multi-Space Approach to Zero-Shot Object Detection. In: 2020 IEEE Winter Conference on Applications of Computer Vision; 2020. p. 1198–206.
[19] Gu Z, Zhou S, Li N, Zhao Z, Zhang L. Context-aware Feature Generation For Zero-shot Semantic Segmentation. In: 28th ACM International Conference on Multimedia; 2020.
[20] C. Wah, S. Branson, P. Welinder, P. Perona and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Computation & Neural Systems Technical Report, CNS-TR-2011-001.
[21] Li X, Wu J, Sun Z, Ma Z, Cao J, Xue J. BSNet: Bi-Similarity Network for Few-shot Fine-grained Image Classification. IEEE Transactions on Image Processing 2021;30:1318–31.
[22] Gao Z, Li Y, Wan S. Exploring Deep Learning for View-Based 3D Model Retrieval. ACM Transactions on Multimedia Computing, Communications, and Applications. 2020;16(1):1–21.
[23] Li Y, Shan C, Li H, Ou J. A capsule-unified framework of deep neural networks for graphical programming. Soft Computing 2021;25(5786):1–23.
[24] Wang X, Girshick RB, Gupta AK, He K. Non-local Neural Networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018. p. 7794–803.
[25] Xian Y, Lampert CH, Schiele B, Akata Z. Zero-Shot Learning—A Comprehensive Evaluation of the Good, the Bad and the Ugly. IEEE Transactions on Pattern Analysis and Machine Intelligence 2019;41:2251–65.