Semantic-guided Reinforced Region Embedding for Generalized Zero-Shot Learning

Jiannan Ge, Hongtao Xie *, Shaobo Min, Yongdong Zhang
School of Information Science and Technology, University of Science and Technology of China, Hefei, China
{gejn, mbobo}@mail.ustc.edu.cn, {htxie, zhyd73}@ustc.edu.cn

Abstract

Generalized Zero-Shot Learning (GZSL) aims to recognize images from either seen or unseen domain, mainly by learning a joint embedding space to associate image features with the corresponding category descriptions. Recent methods have proved that localizing important object regions can effectively bridge the semantic-visual gap. However, these are all based on one-off visual localizers, lacking of interpretability and flexibility. In this paper, we propose a novel Semantic-guided Reinforced Region Embedding (SR2E) network that can localize important objects in the long-term interests to construct semantic-visual embedding space. SR2E consists of Reinforced Region Module (R2M) and Semantic Alignment Module (SAM). First, without the annotated bounding box as supervision, R2M encodes the semantic category guidance into the reward and punishment criteria to teach the localizer serialized region searching. Besides, R2M explores different action spaces during the serialized searching path to avoid local optimal localization, which thereby generates discriminative visual features with less redundancy. Second, SAM preserves the semantic relationship into visual features via semantic-visual alignment and designs a domain detector to alleviate the domain confusion. Experiments on four public benchmarks demonstrate that the proposed SR2E is an effective GZSL method with reinforced embedding space, which obtains averaged 6.1% improvements.

Introduction

Recently, the deep learning algorithms (Min et al. 2020; Sio et al. 2020; Wang et al. 2020; Xie et al. 2020) have developed rapidly due to the massive increase in labeled data. For example, traditional classification tasks require a large number of images with category labels for training. However, it is infeasible to annotate all kinds of species and learn a classifier for them due to the species diversity in the real world. To handle this issue, Generalized Zero-Shot Learning (GZSL) offers an effective solution, which utilizes category descriptions, such as the representation of color, habit, and some other features, to connect the seen and unseen domain categories (Yang et al. 2016; Cappallo, Mensink, and Snoek 2015). Since the category descriptions for two domains share the same semantic space, the recognition knowledge learned from the seen domain can be directly transferred to the unseen domain.

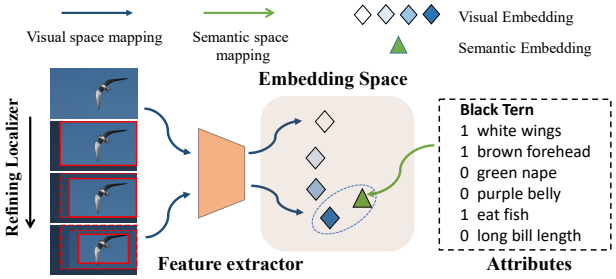

Figure 1: Localizing the bird in an image through a serialized search. As the optimizing of the localizer, the visual description can better match the semantic information in the embedding space, which is indicated by the blue dotted line.

The commonly used semantic descriptor includes human-defined information and learning-based information. The former contains attributes information (Lampert, Nickisch, and Harmeling 2009; Jayaraman and Grauman 2014), and the latter includes label embedding using Word2Vec (Mikolov et al. 2013) or GloVe (Pennington, Socher, and Manning 2014) algorithm. In order to associate the visual image features with corresponding category descriptions, a joint embedding space has been constructed in current methods, where both visual features and category descriptions are projected, and the recognition becomes the nearest neighbor searching problem. Among these methods, some works (Socher et al. 2013; Fu et al. 2015; Morgado and Vasconcelos 2017) span the embedding space directly using category description to preserve the semantic topology, and others (Zhang, Xiang, and Gong 2017; Yu et al. 2018; Annadani and Biswas 2018) select the visual space as the embedding space for better visual discrimination. According to (Fu et al. 2015), learning a latent intermediate space between the visual and semantic space can take both advantages of visual discrimination and semantic preservation, but it is hard to learn. Thus, current methods (Li et al. 2018; Xie et al. 2019) leverage the visual attention mechanism to localize discriminative local regions and select critical visual elements to construct a good intermediate embedding space, thereby improving the semantic-visual association. Nevertheless, these

*Corresponding author.

models are all based on the one-off object localizer with a non-transparent region search process that limits their interpretability and flexibility for further improvement.

In this paper, we propose a novel Semantic-guided Reinforced Region Embedding (SR2E) network that can gradually learn the embedding space according to the reinforcement-learned region embedding mechanism with semantic guidance. As shown in Figure 1, compared with a simple one-off part localizer, our model targets to improve the semantic-visual association mainly by localizing the important regions step by step, e.g., the bird region. SR2E consists of two sub-modules, Reinforced Region Module (R2M) for discriminative feature extraction and Semantic Alignment Module (SAM) for stable knowledge transfer. First, notice that there is no extra annotation for the bounding box as supervision. R2M leverages semantic-driven reward signals designed by category confidence to teach the reinforced region localizer how to localize the semantically related region step by step. It also encodes historical action information to obtain a stable search process that considers long-term benefits. In this sense, R2M can obtain more discriminative features, and thus it is considered well compatible with the corresponding category description. Besides, to avoid the early stopping of region searching, R2M applies different action sets to different localizer states, which can traverse more potential efficient areas for global optimization. Second, SAM performs semantic-visual alignment in the semantic space spanned by visual representations and constructs a domain detector to reduce the bias recognition problem.

The experimental results on four public benchmarks show that the proposed SR2E model can outperform state-of-the-art methods significantly. Our main contributions can be summarized into three parts:

- We propose a novel Semantic-guided Reinforced Region Embedding (SR2E) network that can adaptively and gradually localize object regions to construct the semantic-visual embedding space via a reinforcement-based region embedding. To our best knowledge, this is the first work to introduce the reinforcement mechanism into GZSL.
- A Reinforced Region Module (R2M) is designed to teach the localizer how to localize the correct object region step by step under the guidance of semantic category descriptions for more discriminative features.
- A Semantic Alignment Module (SAM) is developed to align semantic and visual information in the visual space and determine the domain distribution for better knowledge transfer.

Related Work

Generalized Zero-Shot Learning

Recent methods of GZSL are generally based on the embedding model which focuses on building a shared embedding space for visual data and corresponding semantic features. For example, ALE (Akata et al. 2015a) trains a bilinear embedding model using a hinge ranking loss, and SJE (Akata et al. 2015b) creates a joint embedding space with the linear combination of multiple compatibility functions. Although these methods display good performance, it is vulnerable to treat attribute space as the semantic space and map the visual features into it for Hubness problem (Shigeto et al. 2015). Some methods (Shigeto et al. 2015; Xian et al. 2018a) use the semantic space spanned by visual representations to solve this problem, and others pay attention to building a shared intermediate space (Yu et al. 2018; Lu 2016). Since there is no limitation of attribute dimensions, these models can retain more visual information and provide better results. Recently, the generative methods (Xian et al. 2018b; Yu et al. 2020) use GANs to produce unseen visual features from semantic attributes, which turns GZSL into a fully supervised task and greatly improve the performance. In this paper, our model is among the embedding-based methods and does not contain any extra generated data.

The attention mechanism has been proven effective in various fields (Tay, Roy, and Yap 2019; Liu et al. 2020). Some current methods (Li et al. 2018; Zhu et al. 2019) also take advantage of this technique to construct useful semantic-visual embedding, therefore obtain excellent results. LDF (Li et al. 2018) combines the original and the cropped region to get stable performance. SGMA (Zhu et al. 2019) focuses on finding the crucial local regions containing the semantic information of the corresponding attributes. Similar to LDF, our model locates the object and extracts global visual features. The difference is that LDF leverages a one-off localizer while our model uses semantic-driven reinforcement learning to build a serialized search, which is proved to improve the performance significantly.

Deep Reinforcement Learning

Reinforcement Learning (RL) (Kaelbling, Littman, and Moore 1996), aiming to find the optimal strategy given a Markov Decision Process (MDP), enables an agent to learn effective behavior through many attempts. Considering the mining sequence process of visual information and long-term benefits, Deep Reinforcement Learning (DRL) has been successfully applied in the field of computer vision. For example, AOL (Caicedo and Lazebnik 2015) proposes the Active Object Localization method to train an intelligent agent to make the bounding box approach the ground truth. DRL-RPN (Duan et al. 2018) uses DRL to mime bitwise relationships and significantly improve the stability of the ambiguous bits. Inspired by applying DRL in object detection, we regard the search process for important regions as an MDP to obtain valid visual information. It is noted that there is no extra annotation for the bounding box in GZSL. Therefore we design a different reward signals by using the prediction of the cropped image. Moreover, a new train strategy of DRL is introduced according to the characteristics of the GZSL task, which is also proved effective.

Semantic-guided Reinforced Region Embedding

Task Definition

In GZSL, the seen domain data is defined by $\mathcal{S} = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$, where $x_i^s \in \mathcal{X}^{\mathcal{S}}$ is the i_{th} image and $y_i^s \in \mathcal{Y}^S$

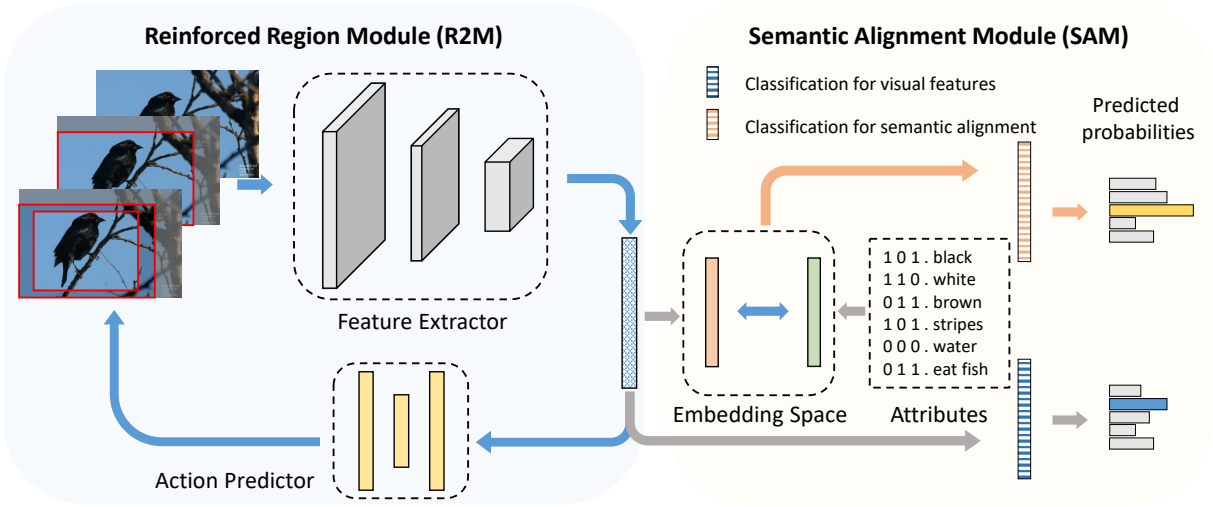

Figure 2: The framework of RL for GZSL (SR2E) model. SR2E achieves serialized search for discriminative regions using RL skills. R2M and SAM modules are included. R2M aims to localize the object region for more discriminative features. SAM focuses on facilitating semantic-visual alignment with a domain detector.

is the corresponding class label. Similarly, the unseen domain data is defined by $\mathcal{U} = \{(x_i^u, y_i^u)\}_{i=1}^{N_u}$. Notably, the seen classes $\mathcal{Y}^S$ and unseen classes $\mathcal{Y}^\mathcal{U}$ are disjoint, i.e., $\mathcal{Y}^S \cap \mathcal{Y}^\mathcal{U} = \varnothing$. Besides, $\mathcal{A}_s = \{\mathbf{a}_i^s\}_{i=1}^{c_s}$ and $\mathcal{A}_\mathcal{U} = \{\mathbf{a}_i^u\}_{i=1}^{c_u}$ are the semantic vectors which depict the seen and unseen classes. The GZSL task aims to recognize $\mathcal{X}^s$ and $\mathcal{X}^u$ given $\mathcal{S}$ and attributes $\mathcal{A} = \mathcal{A}_\mathcal{S} \cup \mathcal{A}_\mathcal{U}$.

A general paradigm of GZSL is to align the visual features of images with corresponding semantic labels by

$$\mathcal{L} = \sum_{i=1}^{N} d[f(\boldsymbol{x}_i^s), \boldsymbol{a}_i^s], \tag{1}$$

where $f(\cdot)$ is the visual embedding function, and $d(\cdot,\cdot)$ is a metric function, *e.g.,* cosine distance (Annadani and Biswas 2018). Since the semantic labels $\mathcal{A}_s$ and $\mathcal{A}_\mathcal{U}$ share a common semantic space, the semantic-aligned visual representation $f(\cdot)$ can be directly transferred to the unseen domain. Thus, the inference of GZSL can be expressed as

$$y = \arg \min_{y \in \mathcal{Y}_s \cup \mathcal{Y}_u} d[f(\boldsymbol{x}), \boldsymbol{a}(y)], \tag{2}$$

where $\boldsymbol{x} \in \mathcal{X}^s \cup \mathcal{X}^u$ is the input image. Note that $f(\cdot)$ is a core fact that determines the discrimination of embedding space in Eq. (1). Instead of simple fully connections (Li et al. 2018), the proposed Semantic-guided Reinforced Region Embedding (SR2E) is designed to adaptively and gradually select important local regions in $\boldsymbol{x}$ to obtain more discriminative visual features. Figure 2 shows the framework of SR2E consisting of two sub-modules, i.e., Reinforced Region Module (R2M) for $f(\cdot)$ and Semantic Alignment Module (SAM) for semantic-visual alignment with a new domain detector.

The Reinforced Region Module

In order to avoid the non-transparency and inflexibility of the one-off localizer, the Reinforced Region Module (R2M) leverages the semantic guidance to predict the actions of the image cropping process, which allows the locator to approach the object region step by step. Specially, we design semantic-driven reward functions for R2M to teach the localizer how to localize the correct region step by step, thus the network can search for semantically related regions and extract more discriminative features. Moreover, considering that the search space is large and the global image always contains complex background noise, which indicates a local optimum always exit, different action sets are selected in different steps to explore more potential areas containing semantic information.

R2M contains a feature extractor (CNN) and an action predictor. As shown in Figure 2, given the input image I_k, we firstly encode it to the CNN and get the visual features p_k. Then we calculate the current state of the localizer by

$$s_k = f_p(p_k) \oplus f_h(a_k), \tag{3}$$

where $f_p(\cdot)$ is a maximum pooling operation which can obtain the visual state vector from the visual feature, while $f_h(a_k)$ means replacing the oldest 6-dimensional vector with a_k which represents the latest action operation, and then updates the 24-dimensional history state vector which represents the four actions executed in the past. Finally, $f_p(p_k)$ and $f_h(a_k)$ are concatenated as the state s_k. In this sense, the state s_k can remember nearby historical information, which indicates that the action is predicted based on long-term benefits, thereby avoid redundancy and obtain more discriminative features. Then s_k is fed into the action predictor to get the image cropping action. This process is denoted by $a_{k+1} = f_a(s_k)$, where $f_a(\cdot)$ is an action prediction function, which is implemented as a two-layer fully-connected network with 1024 hidden units and six output neurons that represent six actions. Noted that a_{k+1} determines the cropping action performed in I_k and then we get a new input I_{k+1}. To achieve flexible region search, two kinds

of actions are designed including movement action and termination action. Movement action instructs the area selection while termination action means stopping localization. Specifically, the movement action includes zooming to the top left region, top right region, bottom left region, bottom right region and central region. Each action scales the crop window with the scale factor $\alpha \in [0, 1]$, which denotes the ratio of the length of selected regions between two adjacent step. After scaling, the model gets a cropped image and then the whole process is repeated until the model meets a termination action.

Then, R2M assesses the predicted action a_{k+1} according to the reward signal R. Most of the rewards used in the object detection task is guided by the ground-truth bounding box. However, GZSL has only image-level annotations. Thus we use the classification results instead of IoU to design our reward signals. The reward signals consist of two types, reward for movement and reward for termination corresponding to different actions. Also, considering that high confidence should get a high reward, we directly assign the confidence to the reward. Thus the reward for movement is defined as follows:

$$R_a\left(s_k, s_{k+1}\right) = \begin{cases} +P_t(I_{k+1}) & \text{if } P_t(I_{k+1}) > P_t(I_k), \\ -P_t(I_k) & \text{otherwise}, \end{cases} \tag{4}$$

where $P_t(\cdot)$ denotes the prediction score performed by

$$P_t(I_k) = \max_{y \in \mathcal{Y}_s \cup \mathcal{Y}_u} 1 - d[f(I_k), g(\boldsymbol{a}(y))], \tag{5}$$

where $g(\cdot)$ maps attributes to semantic space. We consider the correct classification and the maximum number of iterations as the triggers for termination action. When a termination action is obtained, we compare the predicted result with the corresponding label. If the prediction is correct, the model obtains a positive reward, otherwise a negative reward. However, through experiments we found that it is easy to achieve correct prediction during training that weakens the distinction of reward signals, so we also focus on the change in probability. We preserve the probability of the target label from previous results and compare it with the current prediction result. In the case of the right prediction, the reward function obtains a higher value if the probability result is higher than the previous one. The definition of the reward function for termination action is as follows:

$$R_t\left(s_k, s_{k+1}\right) = \begin{cases} +\eta + 2P_t(I_{k+1}) & \text{if } y^p = y^g, P_t(I_{k+1}) > P_t(I_k), \\ +\eta & \text{if } y^p = y^g, P_t(I_{k+1}) \leq P_t(I_k), \\ -\eta & \text{otherwise}. \end{cases} \tag{6}$$

where η represents the termination reward value, y^p and y^g are the predicted label and the ground-truth label.

In order to facilitate the convergence of R2M and avoid local optimum, different action sets are utilized in different steps. Since the original image always contains a complex background that introduces complex noise, our R2M always meets a local optimum at the beginning of the search process. To alleviate this situation, we remove the termination action from the action sets when we predict the action for the first time. As shown in Figure 3, each line is a possible path. When breaking through the first depth of the search,

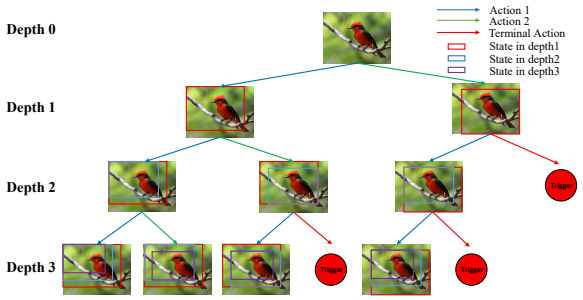

Figure 3: Illustration of the search mechanism. Choose two actions as an example. The termination action is removed in depth 0.

our search process is expanded to enable R2M to explore more regions and find the most representative one. When we only perform different action sets in the first step, we set $d = 1$, where d is the depth for removing the termination action. Moreover, $\epsilon - greedy$ policy (Sutton and Barto 2018) is introduced to expand the search space further.

Semantic Alignment Module

Semantic Alignment Module (SAM) aims to learn a semantic space spanned by visual representations and perform semantic-visual alignment for effective knowledge transfer. We construct two mapping functions, *i.e.,* $f(\cdot)$ and $g(\cdot)$, to map both image $\boldsymbol{x}$ and attribute vector $\boldsymbol{a}$ into visual embedding space. Thus, the inference of semantic prediction is defined as follows:

$$y = \arg \min_{y \in \mathcal{Y}_s \cup \mathcal{Y}_u} d[f(\boldsymbol{x}), g(\boldsymbol{a}(y))], \tag{7}$$

The loss can be expressed by

$$\mathcal{L}_{sem} = \sum_{i=1}^{N} d[f(\boldsymbol{x}_i^s), g(\boldsymbol{a}_i^s)], \tag{8}$$

However, we find that the semantic predictions of unseen images tend to be recognized as the seen categories known as the biased recognition problem (Fu et al. 2014; Min et al. 2019). Therefore, it is necessary to construct a domain detector to determine whether the input sample comes from the seen domain. We discover that the visual classifier output of an unseen sample is smoother than the image from the seen domain through experiments. For this reason, we take the two largest scores of the prediction and calculate their difference $\hat{\mathcal{C}}_{1,2}(\cdot)$, which is used to predict which domain the input image belongs to, to develop a difference-based domain detector. This process is defined as follows:

$$\hat{y} = \begin{cases} \arg\max_{y \in \mathcal{Y}_s} \mathcal{C}(f(\boldsymbol{x})) & \text{if } \hat{\mathcal{C}}_{1,2}(x) > \tau, \\ \arg\min_{y \in \mathcal{Y}_u} d[f(\boldsymbol{x}), g(\boldsymbol{a})] & \text{otherwise}, \end{cases} \tag{9}$$

where $\boldsymbol{x}$ is the input image cropped by R2M, $f(\cdot)$ denotes the feature extractor, $g(\cdot)$ is a multi-layer perceptron, $\mathcal{C}(\cdot)$ is a $|\mathcal{Y}^S|$-way classifier that is trained with cross-entropy loss of seen categories by

$$\mathcal{L}_{vis} = -\frac{1}{N} \sum_{i}^{n} \log \frac{\exp(\mathcal{C}(f(\boldsymbol{x})))}{\sum_c \exp\left(\mathcal{C}(f(\boldsymbol{x}))^c\right)}, \boldsymbol{x} \in \mathcal{X}_{\mathcal{S}} \tag{10}$$

and $\hat{\mathcal{C}}_{1,2}(\cdot)$ is the difference between the maximum and the sub-maximal classification scores in terms of the seen classes $\mathcal{Y}^S$, τ is the threshold for distinguishing the domain. We divide the validation set from the training set to estimate the threshold τ.

Finally, the whole GZSL prediction model is trained by the following loss function:

$$\mathcal{L}_{all} = \mathcal{L}_{sem} + \mathcal{L}_{vis}. \tag{11}$$

Experiments

Experimental Settings

Datasets. We evaluate the proposed SR2E model on four GZSL benchmarks: Caltech-UCSD Birds 200-2011(CUB) (Wah et al. 2011), Animals with Attributes 2(AwA2) (Lampert, Nickisch, and Harmeling 2013), SUN (Xiao et al. 2010) and Yahoo dataset(aPY) (Farhadi et al. 2009). CUB is a fine-grained bird dataset which contains 11,788 bird images from 200 classes, and each class has 312-dimension semantic vector. AwA2 contains 37,322 images of 50 animal categories, each of which has an 85-dimension attribute vector. SUN is a fine-grained dataset including 14,340 images from 717 scene categories. Each class in SUN has a 102-dimension continuous semantic vector. aPY consists of 15,339 images from 32 categories, and 64-dimension is associated with each class. We adopt the splits of seen/unseen classes proposed in (Xian et al. 2018a).

Implementation details. We adopt ResNet101 (He et al. 2016) as the visual feature extractor, which is pretrained on the ImageNet dataset (Russakovsky et al. 2015). The input image is randomly cropped on a 448×448 resized image with random horizontal flipping. In this work, we train the model with two stages: basic model construction and reinforcement feature capture. In the first stage, we get an available model trained on GZSL datasets, and semantic-driven reinforcement learning is introduced in the second stage. After that, we retrain the SR2E model based on the regions extracted by R2M and perform a serialized search again. In this way, our model can capture discriminative features and facilitate the semantic-visual association.

Regarding the parameters, we set learning rate $lr = 1 \times 10^{-4}$ and adopt Adam optimizer with $\beta = (0.5, 0.999)$ and weigh decay 1×10^{-4} in the first stage. In the second stage, the learning rate of the action predictor is fixed and set to 1×10^{-6}, while the termination action reward value η is set to 3. We set $\alpha = 6/7$ for the animal datasets like CUB and AwA2 because their images have specific objects which are useful for classification. SUN and aPY datasets have scene pictures, which means the most areas of the image affect prediction. In this case, we set $\alpha = 9/10$. The maximum number of iterations is set to 6 to avoid invalid search. The ϵ value in $\epsilon - greedy$ is initialized to 1, and decreases by 0.1 at each epoch until the $10th$ epoch.

Evaluation metrics. We only consider generalized settings in our experiments, which require both seen and unseen classes in searched label space. We follow the settings in (Xian et al. 2018a) to adopt Mean Class Accuracy (MCA) as the evaluation indicator. We calculate the MCA of the seen (MCA_s) and unseen (MCA_t) classes separately, and use the harmonic mean (H) to evaluate our model: $H = \frac{2 \times MCA_t \times MCA_s}{MCA_t + MCA_s}$.

Baselines. To verify the effectiveness of different components in SR2E model, three baselines are designed:

- **SE-BS**. This is a general GZSL classification model without RL fine-tuning and domain detector.
- **SE-DD**. Different from SE-BS, we apply the domain detector to alleviate the domain confusion.
- **SR2E**. We add an action predictor based on SE-DD and execute the second stage (reinforcement feature capture).

Comparison with Existing Methods

The results of the GZSL compared with previous methods are reported in Table 1. Compared to the non-generative models, our model has been greatly improved, e.g., SR2E obtains 7.7%, 0.4%, 6.5% and 9.8% improvement in terms of the harmonic mean (H) on CUB, AwA2, SUN and aPY datasets, respectively.

Also, from Table 1, an interesting phenomenon is that some methods like VSE-S and MLSE can achieve a pretty good performance on the seen domain but fail to get a high value of MCA_t. The unbalanced performance on seen and unseen classes makes the metric H of VSE-S much lower than ours on AwA2 dataset (57.2% vs 67.5%). Moreover, the difference between the two domains of SR2E on CUB is only 9%, while MLSE gets 49.3%. The reason is that these models use more visual supervision but lack sufficient knowledge transfer. In contrast, our model imposes visual and semantic supervisions simultaneously, and enhances the effectiveness of these two supervisions through serialized region search and discriminative feature extraction.

Notice that recent region-based methods could achieve excellent performance. Most of them can get high values in seen domain thanks to their effective selection of visual features. However, the regional focus of these models is only performed once in the forward propagation process, and it makes the learning of the localizer insufficient. Different from these methods, our search process is serialized, which is more transparent and precise. We utilize the semantic-driven reward signals to guide the localizer to perform a serialized search so as to capture the import regions more fully, leading to a higher harmonic mean H. Among these methods, LDF obtains the key regions that contain global information, which is most similar to our model. However, it combines the original image with the selected region in a combined method, which is stable but inefficient.

Different from embedding-based methods, generative methods utilize the prior unseen semantic labels to synthetic extra training samples, thereby transforming GZSL into a general supervision problem, which leads to more competitive results. Nevertheless, compared with generative models, our model still has competitiveness. With the help of reinforced region localizer and semantic-visual alignment, SR2E tends to extract better feature representations and perform stable knowledge transfer, resulting in superior performance.

	Methods	CUB			AwA2			SUN			aPY		
		MCA_t	MCA_s	H	MCA_t	MCA_s	H	MCA_t	MCA_s	H	MCA_t	MCA_s	H
GEN	f-CLSWGAN (Xian et al. 2018b)	43.7	57.7	49.7	-	-	-	42.6	36.6	39.4	-	-	-
	SE (Kumar Verma et al. 2018)	41.5	53.3	46.7	58.3	68.1	62.8	40.9	30.5	34.9	-	-	-
	E-PGN (Yu et al. 2020)	52.0	61.1	56.2	52.6	83.5	64.6	-	-	-	-	-	-
NON-GEN	VSE-S(Zhu, Wang, and Saligrama 2019)	33.4	87.5	48.4	41.6	91.3	57.2	-	-	-	24.5	72.0	36.6
	PREN(Ye and Guo 2019)	32.5	55.8	43.1	32.4	88.6	47.4	35.4	27.2	30.8	-	-	-
	MLSE(Ding and Liu 2019)	22.3	71.6	34.0	23.8	83.2	37.0	20.7	36.4	26.4	12.7	74.3	21.7
	LDF*(Li et al. 2018)	26.4	81.6	39.9	9.8	87.4	17.6	-	-	-	-	-	-
	AREN*(Xie et al. 2019)	38.9	78.7	52.1	15.6	92.9	26.7	19.0	38.8	25.5	9.2	76.9	16.4
	SGMA*(Zhu et al. 2019)	36.7	71.3	48.5	37.6	87.1	52.5	-	-	-	-	-	-
	DAZLE*(Huynh and Elhamifar 2020)	56.7	59.6	58.1	60.3	75.7	67.1	52.3	24.3	33.2	-	-	-
	SE-BS (w/o domain detector)	41.0	85.1	55.4	14.8	95.7	25.7	16.9	46.1	24.7	9.54	68.1	16.7
	SE-DD (w domain detector)	58.9	68.9	63.5	55.1	81.8	65.9	42.0	36.2	38.9	35.8	58.5	44.4
	SR2E (w reinforcement learning)	61.6	70.6	**65.8**	58.0	80.7	**67.5**	43.1	36.8	**39.7**	38.4	58.8	**46.4**

Table 1: Results on GZSL. The harmonic mean (H) is the comprehensive evaluation of MCA_s and MCA_t. The best results are marked in bold font. GEN indicates generative methods, which utilize GANs to synthetic unseen data for training. * indicates the models based on one-off object localizer.

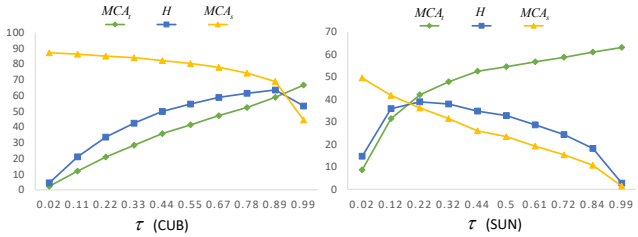

Figure 4: The performance of the domain detection.

Methods	CUB			aPY		
	MCA_t	MCA_s	H	MCA_t	MCA_s	H
SR2E-Scale A	58.6	72.3	64.7	35.8	58.5	44.4
SR2E-Scale B	61.6	70.6	65.8	37.8	57.7	45.7
SR2E-Scale C	58.2	73.3	64.9	38.4	58.8	46.4

Table 2: Results on different scale factors α. Scale A, Scale B and Scale C indicate $\frac{4}{5}$, $\frac{6}{7}$ and $\frac{9}{10}$, respectively.

Ablation Studies

Effects of domain detection. A general GZSL model always meets the biased recognition problem, and thus limiting the performance of the classifier. As shown in Table 1, SE-BS can only achieve modest performance on all four datasets. However, combined with the difference-based domain detector, SE-DD achieves more than 20% improvement on average in terms of a harmonic mean H compared with SE-BS. The reason is that SE-DD can effectively distinguish the categories of different domains, thereby obtain stable transfer ability and more effective semantic-visual alignment. An interesting phenomenon is that our SE-DD raises H from 25.7% to 65.9% over SE-BS on AwA2 datasets, which is mainly contributed by the dramatically 40.3% improvement of MCA_t. This is because there are sufficient training samples for AwA2, leading to strong supervision in the seen domain but weak knowledge transfer in the unseen domain, while domain detector alleviates this phenomenon.

Effects of varying τ. As the τ is the factor that determines the performance of our difference-based domain detector, we calculate the results of different values and evaluate their influence. As shown in Figure 4, we can observe that high MCA_t is always accompanied by low MCA_s. The reason is that higher τ prefers to divide the sample into the unseen domain, causing MCA_s to decrease. We consider the trade-off between both two metrics and find that H reaches the highest value when τ is set to 0.89 in CUB datasets. However, we observe that the intersection of the three lines appeared early in SUN datasets. This phenomenon is caused by insufficient training samples. SUN has only 16 samples per category on average, resulting in a smoother output even when the input image is from the seen domain. To achieve better performance, the τ is set to 0.89, 0.97, 0.81 and 0.22 for CUB, AwA2, aPY and SUN, respectively.

Effects of reinforcement feature capture. Different from SAM that improves the semantic-visual alignment, R2M improves the performance by extracting more discriminative features via reinforced region embedding. In the second stage (reinforcement feature capture), we get the SR2E model. This model adds another coarse-to-fine processing based on SE-DD model by regarding the search process for important regions as the Markov Decision Process (MDP). As shown in Table 1, the SR2E gets a higher harmonic mean (H) than SE-DD for all four datasets (65.8% in CUB, 67.5% in AwA2, 39.7% in SUN, 46.4% in aPY). Also, we can observe that the reinforcement feature capture promotes both MCA_s and MCA_t in CUB, SUN and aPY datasets, which proves the validity of efficient features. It is noted that R2M produces less performance improvement than SAM. The reason is that the domain detector can adjust MCA_s and MCA_t to obtain a higher H. This improvement is often relatively large due to the mitigation of the biased recognition problem.

Effects of scale factor α. For the SR2E model, the search scale factor α can be set to different values. Large α means a small range of variation for the cropped region between two steps, and this careful search can avoid losing potential

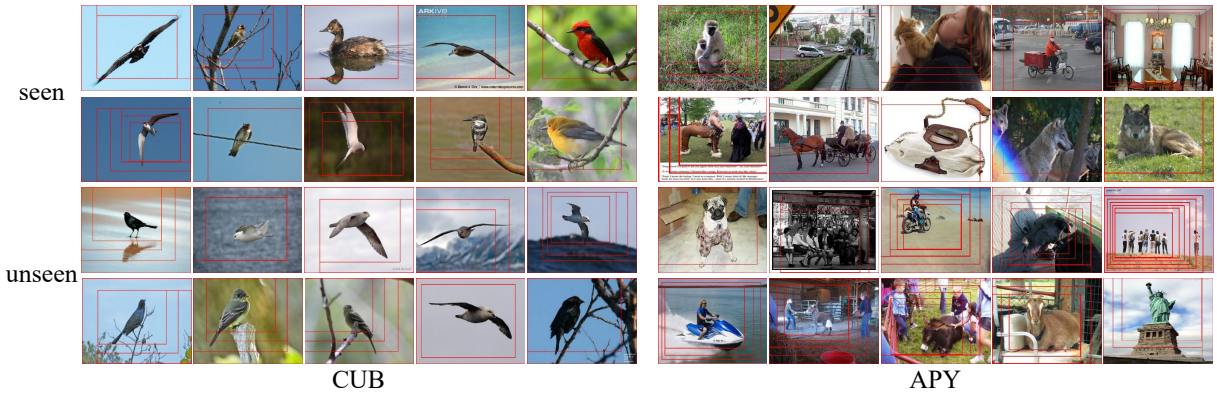

Figure 5: The visual samples of SR2E on CUB and aPY datasets. The first two rows are the seen samples, and the other rows are the unseen samples. Each image has multiple red bounding boxes, and these bounding boxes from large to small show the serialized search process for the key regions. The scale factors on CUB and aPY are set to $\frac{6}{7}$ and $\frac{9}{10}$, respectively.

	CUB			aPY		
Methods	MCA_t	MCA_s	H	MCA_t	MCA_s	H
SR2E-ID0	58.8	72.0	64.8	35.7	60.6	45.0
SR2E-ID1	61.6	70.6	65.8	38.4	58.8	46.4
SR2E-ID2	58.9	72.5	65.0	37.7	58.3	45.8

Table 3: Results of depth d. ID0, ID1 and ID2 indicate $d = 0$,$d = 1$ and $d = 2$, respectively. $d = 0$ means SR2E does not perform different action sets.

areas. However, there will be more time to pay and thus a termination action will be triggered without fully expanding. Based on this, it is worth choosing a proper α to search for suitable areas. According to the observation of the GZSL datasets, we choose three different scales: $\{\frac{4}{5}, \frac{6}{7}, \frac{9}{10}\}$. The results in CUB and aPY datasets are shown in Table 2. It can be observed that the harmonic mean (H) of the SR2E using Scale A achieves 64.7% in CUB, while gets no improvement in aPY (using the performance of SE-DD). The reason is that the object of an image from CUB is usually in a suitable area, while the image from aPY has more noise and scale change. From Table 2, we also observe that SR2E-Scale B gets the best performance in CUB (65.8%) and SR2E-Scale C obtains the highest H in aPY (46.4%). The results show a suitable α can achieve a balance between the search speed and search space.

Effects of depth d. We study the appropriate depth d to better guide the localizer to find more discriminative regions. Through Table 3, it can be seen that the technique of selectively ignoring termination action can effectively improve the performance of the SR2E model. For example, on the aPY dataset, the SR2E-ID1 model raises a harmonic mean (H) from 45.0% to 46.4% compared to the SR2E-ID0 model, which is mainly attributed to the contribution of 2.7% improvement on the unseen domain. Further, SR2E-ID1 obtains the best performance in terms of metric H on both CUB and aPY dataset, which shows that the depth set to 1 is enough to expand the search space and find a more accurate object region.

Visualizations of the serialized search process. To obtain the process of the serialized search for discriminative regions, we save the coordinates of the cropped region at each step and draw the bounding box on the corresponding image. The results are shown in Figure 5. On the CUB datasets, our model can catch the object in the first few steps because most unrelated areas are relatively single. We can see that most positioning results are finer, which also confirms that the redundant background of the image has an impact on classification confidence. Additionally, we know that the bounding box annotations can be provided by CUB datasets, while our model can find the location of the object automatically without these annotations, which illustrates that our model has an outstanding performance on extracting key semantic information. On aPY datasets, we can see that there are some small objects, e.g., people and the donkey, which indicates that aPY is more complex. Nonetheless, our model can localize the key region through more steps, showing the robustness and accuracy of our model.

Conclusion

In this paper, we proposed a novel Semantic-guided Reinforced Region Embedding (SR2E) network for challenging generalized zero-shot image recognition. Different from existing methods focusing on a one-off localizer to directly select partial regions, we develop a serialized search guided by category description which is interpretable and stable to construct semantic-visual embedding space. By regarding the search process for important regions as a Markov Decision Process, we propose the R2M module that can adaptively and gradually localize the semantically related object region. Semantic-driven reward signals and different action sets are designed to guide the R2M extract more discriminative features, which is proved to greatly improve the performance of our model. To further boost the semantic-visual alignment, SAM with a domain detector is provided, which reduces the bias recognition problem and improve the knowledge transfer. And the experiment results show our model outperforms existing state-of-the-art approaches.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (2018YFB0804203), the National Nature Science Foundation of China (61525206, 62022076, U1936210), the Youth Innovation Promotion Association Chinese Academy of Sciences (2017209). We also acknowledge the support of GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC.

References

Akata, Z.; Perronnin, F.; Harchaoui, Z.; and Schmid, C. 2015a. Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence* 38(7): 1425–1438.

Akata, Z.; Reed, S.; Walter, D.; Lee, H.; and Schiele, B. 2015b. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2927–2936.

Annadani, Y.; and Biswas, S. 2018. Preserving semantic relations for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7603–7612.

Caicedo, J. C.; and Lazebnik, S. 2015. Active object localization with deep reinforcement learning. In *Proceedings of the IEEE international conference on computer vision*, 2488–2496.

Cappallo, S.; Mensink, T.; and Snoek, C. G. 2015. Image2emoji: Zero-shot emoji prediction for visual media. In *Proceedings of the 23rd ACM international conference on Multimedia*, 1311–1314.

Ding, Z.; and Liu, H. 2019. Marginalized Latent Semantic Encoder for Zero-Shot Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6191–6199.

Duan, Y.; Wang, Z.; Lu, J.; Lin, X.; and Zhou, J. 2018. GraphBit: Bitwise interaction mining via deep reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8270–8279.

Farhadi, A.; Endres, I.; Hoiem, D.; and Forsyth, D. 2009. Describing objects by their attributes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 1778–1785. IEEE.

Fu, Y.; Hospedales, T. M.; Xiang, T.; Fu, Z.; and Gong, S. 2014. Transductive multi-view embedding for zero-shot recognition and annotation. In *European Conference on Computer Vision*, 584–599. Springer.

Fu, Y.; Yang, Y.; Hospedales, T.; Xiang, T.; and Gong, S. 2015. Transductive multi-label zero-shot learning. *arXiv preprint arXiv:1503.07790* .

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huynh, D.; and Elhamifar, E. 2020. Fine-Grained Generalized Zero-Shot Learning via Dense Attribute-Based Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4483–4493.

Jayaraman, D.; and Grauman, K. 2014. Zero Shot Recognition with Unreliable Attributes. *CoRR* abs/1409.4327.

Kaelbling, L. P.; Littman, M. L.; and Moore, A. W. 1996. Reinforcement Learning: A Survey. *J. Artif. Int. Res.* 4(1): 237–285. ISSN 1076-9757.

Kumar Verma, V.; Arora, G.; Mishra, A.; and Rai, P. 2018. Generalized zero-shot learning via synthesized examples. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4281–4289.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 951–958.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2013. Attribute-based classification for zero-shot visual object categorization. *IEEE transactions on pattern analysis and machine intelligence* 36(3): 453–465.

Li, Y.; Zhang, J.; Zhang, J.; and Huang, K. 2018. Discriminative learning of latent features for zero-shot recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7463–7471.

Liu, C.; Xie, H.; Zha, Z.-J.; Ma, L.; Yu, L.; and Zhang, Y. 2020. Filtration and Distillation: Enhancing Region Attention for Fine-Grained Visual Categorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11555–11562.

Lu, Y. 2016. Unsupervised Learning on Neural Network Outputs: With Application in Zero-Shot Learning. In *IJCAI*, 3432–3438.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; and Dean, J. 2013. Distributed Representations of Words and Phrases and their Compositionality. *CoRR* abs/1310.4546.

Min, S.; Yao, H.; Xie, H.; Zha, Z.-J.; and Zhang, Y. 2019. Domain-Specific Embedding Network for Zero-Shot Recognition. In *Proceedings of the 27th ACM International Conference on Multimedia*, 2070–2078.

Min, S.; Yao, H.; Xie, H.; Zha, Z.-J.; and Zhang, Y. 2020. Multi-Objective Matrix Normalization for Fine-Grained Visual Recognition. *IEEE Transactions on Image Processing* 29: 4996–5009.

Morgado, P.; and Vasconcelos, N. 2017. Semantically consistent regularization for zero-shot recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6060–6069.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.;

et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115(3): 211–252.

Shigeto, Y.; Suzuki, I.; Hara, K.; Shimbo, M.; and Matsumoto, Y. 2015. Ridge regression, hubness, and zero-shot learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 135–151. Springer.

Sio, C. H.; Ma, Y.-J.; Shuai, H.-H.; Chen, J.-C.; and Cheng, W.-H. 2020. S2SiamFC: Self-supervised Fully Convolutional Siamese Network for Visual Tracking. In *Proceedings of the 28th ACM International Conference on Multimedia*, 1948–1957.

Socher, R.; Ganjoo, M.; Manning, C. D.; and Ng, A. 2013. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, 935–943.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

Tay, C.-P.; Roy, S.; and Yap, K.-H. 2019. Aanet: Attribute attention network for person re-identifications. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7134–7143.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset .

Wang, Y.; Xie, H.; Zha, Z.-J.; Xing, M.; Fu, Z.; and Zhang, Y. 2020. ContourNet: Taking a Further Step toward Accurate Arbitrary-shaped Scene Text Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11753–11762.

Xian, Y.; Lampert, C. H.; Schiele, B.; and Akata, Z. 2018a. Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence* 41(9): 2251–2265.

Xian, Y.; Lorenz, T.; Schiele, B.; and Akata, Z. 2018b. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5542–5551.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3485–3492. IEEE.

Xie, G.-S.; Liu, L.; Jin, X.; Zhu, F.; Zhang, Z.; Qin, J.; Yao, Y.; and Shao, L. 2019. Attentive region embedding network for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9384–9393.

Xie, H.-X.; Lo, L.; Shuai, H.-H.; and Cheng, W.-H. 2020. AU-assisted Graph Attention Convolutional Network for Micro-Expression Recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2871–2880.

Yang, Y.; Luo, Y.; Chen, W.; Shen, F.; Shao, J.; and Shen, H. T. 2016. Zero-shot hashing via transferring supervised knowledge. In *Proceedings of the 24th ACM international conference on Multimedia*, 1286–1295.

Ye, M.; and Guo, Y. 2019. Progressive ensemble networks for zero-shot recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 11728–11736.

Yu, Y.; Ji, Z.; Han, J.; and Zhang, Z. 2020. Episode-Based Prototype Generating Network for Zero-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14035–14044.

Yu, Y.; Ji, Z.; Li, X.; Guo, J.; Zhang, Z.; Ling, H.; and Wu, F. 2018. Transductive zero-shot learning with a self-training dictionary approach. *IEEE transactions on cybernetics* 48(10): 2908–2919.

Zhang, L.; Xiang, T.; and Gong, S. 2017. Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021–2030.

Zhu, P.; Wang, H.; and Saligrama, V. 2019. Generalized zero-shot recognition based on visually semantic embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2995–3003.

Zhu, Y.; Xie, J.; Tang, Z.; Peng, X.; and Elgammal, A. 2019. Semantic-guided multi-attention localization for zero-shot learning. In *Advances in Neural Information Processing Systems*, 14943–14953.