Lifelong Zero-Shot Learning

Kun Wei , Cheng Deng* , Xu Yang
School of Electronic Engineering, Xidian University, Xi'an 710071, China
{weikunsk, chdeng.xd, xuyang.xd}@gmail.com

Abstract

Zero-Shot Learning (ZSL) handles the problem that some testing classes never appear in training set. Existing ZSL methods are designed for learning from a fixed training set, which do not have the ability to capture and accumulate the knowledge of multiple training sets, causing them infeasible to many real-world applications. In this paper, we propose a new ZSL setting, named as Lifelong Zero-Shot Learning (LZSL), which aims to accumulate the knowledge during the learning from multiple datasets and recognize unseen classes of all trained datasets. Besides, a novel method is conducted to realize LZSL, which effectively alleviates the Catastrophic Forgetting in the continuous training process. Specifically, considering those datasets containing different semantic embeddings, we utilize Variational Auto-Encoder to obtain unified semantic representations. Then, we leverage selective retraining strategy to preserve the trained weights of previous tasks and avoid negative transfer when fine-tuning the entire model. Finally, knowledge distillation is employed to transfer knowledge from previous training stages to current stage. We also design the LZSL evaluation protocol and the challenging benchmarks. Extensive experiments on these benchmarks indicate that our method tackles LZSL problem effectively, while existing ZSL methods fail.

1 Introduction

In recent years, Zero-Shot Learning (ZSL) [Socher *et al.*, 2013; Xian *et al.*, 2018a; Zhao *et al.*, 2019; Wei *et al.*, 2019; Xu *et al.*, 2019] has gained increasing attention in computer vision [Chang *et al.*, 2020] and machine learning communities [Yang *et al.*, 2019]. Different from traditional classification tasks that require adequate samples of all classes in training phase, ZSL aims to recognize samples of new classes, which have never appeared in the training stage. In the popular ZSL setting, the learning model is only trained on seen classes of a single dataset, and then tested on unseen classes of the same dataset, whose seen and unseen classes are disjoint. However, in many real-world applications, the recognition system is required to have the ability of learning from obtained training data continuously and to improve the system in a lifelong manner.

To meet such a requirement, we propose a more practical ZSL setting, named as Lifelong Zero-Shot Learning (LZSL), which requires the model to accumulate the knowledge of different datasets and recognize the unseen classes of all faced datasets. As illustrated in Figure 1, the model is trained in multiple learning stages, and each stage includes images and semantic embeddings from a new dataset. The semantic embeddings of these datasets are various and complex, e.g., the attribute lists of these datasets are different. After finishing all training stages, the model is evaluated on both seen and unseen testing images of all these datasets.

The mainstream ZSL methods aim to learn a mapping between images and corresponding semantic embeddings. These methods can be divided into three types according to the classification spaces, *i.e.*, visual space, semantic space and common embedding space. Besides, there are some ZSL methods [Felix *et al.*, 2018; Zhu *et al.*, 2018], which train generative models to obtain the features of unseen classes. Then, the visual features of seen classes and the generated visual features of unseen classes are used to train the classifier. These methods convert ZSL tasks to supervised learning tasks. However, these methods cannot effectively deal with LZSL problem, since they lack the mechanism to accumulate knowledge from previously trained tasks without rehearsal.

Aiming to solve aforementioned problems and realize LZSL, we propose a novel method that integrates unified semantic embedding, selective retraining and knowledge distillation strategies seamlessly. Cross and Distribution Aligned VAE (CACD-VAE) [Schonfeld *et al.*, 2019] is selected as the base model, which trains VAEs [Kingma and Welling, 2013] to encode and decode features of visual and semantic embeddings respectively, and uses the learned latent features to train a ZSL classifier. To equip CACD-VAE with the ability of Lifelong Learning, we first use the trained VAEs to obtain unified semantic embeddings in each training stage. With the unified semantic embeddings, the latent space of different tasks is learned and fixed respectively. To ensure the visual features can be projected into the fixed latent space precisely, selective retraining strategy

*Contact Author

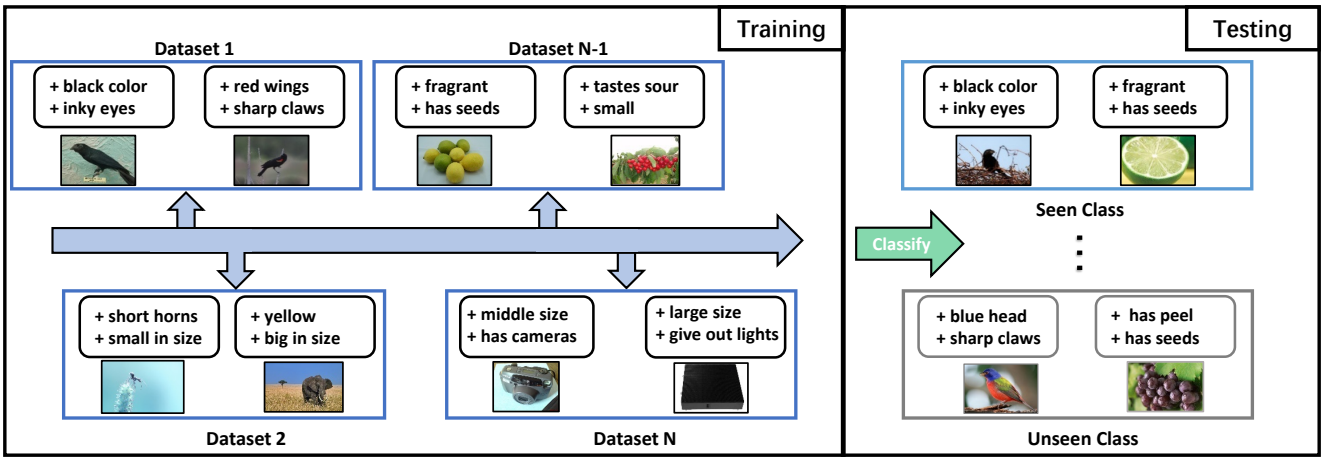


Figure 1: The overview of Lifelong Zero-Shot Learning. When new task arrives, the model learns the new task sequentially, which accumulates the knowledge from all faced tasks. Transferring knowledge from previous tasks to current task promotes the model to classify the unseen classes of different datasets effectively.

is leveraged to promote the similarity among the classification spaces of different tasks, which also avoids negative transfer in the process of capturing the knowledge of new task. Besides, knowledge distillation [Hinton *et al.*, 2015; Chen *et al.*, 2019] is employed to transfer knowledge from previous tasks to current task. Extensive experiments show that our method effectively accumulates knowledge from previous learned tasks and relieves Catastrophic Forgetting, while other state-of-the-art ZSL methods are inoperative. The contributions of our method are summarized as follow:

- To the best of our knowledge, we are the first to propose and tackle Lifelong Zero-Shot Learning problem. The LZSL benchmark and evaluation protocols are also designed in a novel way.
- Aiming to tackle the challenge of isomerism semantic embeddings of different datasets, we employ VAEs to obtain the unified semantic embeddings, which can fix the latent space of corresponding tasks.
- The selective retraining is utilized to promote the similarity among the classification spaces of different datasets, and supervised by knowledge distillation loss, which regularizes the process of transferring the knowledge from previous tasks to current task.
- Extensive experimental results on the proposed benchmark demonstrate the effectiveness of our proposed approach, which significantly outperforms state-of-the-art ZSL methods.

2 Related Work

2.1 Zero-Shot Learning

Zero-Shot Learning [Socher *et al.*, 2013; Zhang *et al.*, 2017; Zhao *et al.*, 2018; Chen *et al.*, 2018] has become a popular research topic, which aims to recognize unseen classes without any labeled training data. In addition, ZSL is a subproblem of transfer learning, whose key point is to transfer knowledge from seen classes to unseen classes. In testing stage, the test samples are captured from visual space, while we only have the semantic embeddings of unseen classes in semantic space. Thus, the mainstream approach of ZSL methods [Chen *et al.*, 2018] is to construct the connection between visual space and semantic space. Typical methods learn functions that maps the visual features and semantic features into a common embedding space, where the embeddings of visual features and semantic features are matched. Recently, generative adversarial networks (GANs) [Goodfellow *et al.*, 2014] had been proposed and successfully introduced to ZSL. The target of generative ZSL methods [Felix *et al.*, 2018; Zhu *et al.*, 2018] is to generate visual features of unseen classes from semantic features, which converts ZSL to traditional supervised classification task. For instance, f-CLSWGAN [Xian *et al.*, 2018b] was proposed by employing conditional Wasserstein GANs, which generated discriminative unseen visual features. Based on f-CLSWGAN, Cycle-WGAN [Felix *et al.*, 2018] leveraged reconstruction regularization that aimed to preserve the discriminative features of classes in transferring process.

However, all the methods mentioned above are only trained on a single dataset, with limited ability to learn various datasets sequentially. To our best knowledge, we are the first to propose and tackle the problem of Lifelong Zero-Shot Learning.

2.2 Lifelong Learning

Lifelong Learning [McCloskey and Cohen, 1989; Rebuffi *et al.*, 2017] is the learning pattern which requires the model to have the ability to learn from a sequence of tasks and to transfer knowledge obtained from earlier tasks to later one. The key challenge for Lifelong Learning is Catastrophic Forgetting, which means the trained model forgets the knowledge of previous task when new task arrives. Many Lifelong learning methods were proposed, which can be divided into three parts, i.e, storing training samples of previous tasks [Rebuffi *et al.*, 2017; Li and Hoiem, 2017], regularizing the pa-

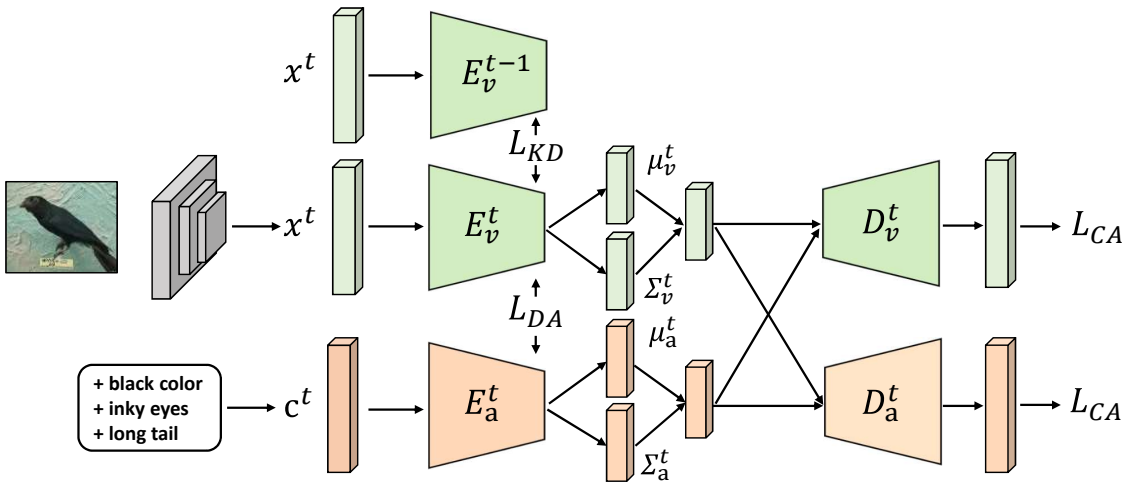


Figure 2: The framework of our proposed method in the t^{th} training stage, which consists of two VAEs and a trained encoder of visual modality in the $(t-1)^{th}$ training stage. Given an image, the feature extractor captures its visual feature x^t, which is mapped into the latent space as μ_v^t and Σ_v^t. Meanwhile, the corresponding semantic embedding c^t is mapped into the latent space as μ_a^t and Σ_a^t. Aiming to achieve latent distribution alignment, the Wasserstein distance between the latent distributions ($\mathcal{L}_{DA}$) is minimized in the training stage. Then, the cross-alignment loss ($\mathcal{L}_{CA}$) is employed to guarantee the latent distributions aligned through cross-modal reconstruction. Besides, we leverage knowledge distillation ($\mathcal{L}_{KD}$) to transfer knowledge obtained from previous tasks to current task.

rameter updates [Liu *et al.*, 2018; Yoon *et al.*, 2017] when new tasks arrives, and memory replay [Shin *et al.*, 2017; Wu *et al.*, 2018] that employs extra generative models to replay training samples of previous tasks.

Different from traditional Lifelong Learning problems, whose training and testing classes are the same in popular Lifelong Learning classification problems, those are disjoint in LZSL.

3 Methodology

To tackle LZSL problems, we propose Lifelong Zero-Zhot Learning, which unifies Lifelong Learning and Zero-Shot Learning seamlessly. The framework of our method is shown in Figure 2. First, we leverage VAEs to obtain the unified semantic embeddings of different datasets. Then, selective retraining strategy is used to approximate the classification space of different datasets and avoid negative transfer. Finally, knowledge distillation is employed to transfer knowledge from previous tasks to current task.

3.1 Problem Formulation

During the t^{th} training stage, a dataset $S^t = \{(x^t, y^t, c^t) | x^t \in X^t, y^t \in Y_s^t, c^t \in C^t\}$ is given, consisting of image features x^t extracted by a pre-trained convolution neural network (CNN), class labels y^t of seen classes Y_s^t and semantic embeddings c^t of corresponding classes. Besides, a dataset $U^t = \{(u^t, c_u^t) | u^t \in Y_u^t, c_u^t \in C^t\}$ is available, containing unseen class labels u^t from a set Y_u^t and the semantic embeddings c_u^t of unseen classes. For the most realistic and challenging metric of Generalized Zero-Learning (GZSL), the target is to learn a classifier f_{GZSL}^t: $X^t \rightarrow Y_s^t \cup Y_u^t$. However, our method focuses on learning a generative model through training different datasets sequentially, and then constructs several classifiers corresponding different datasets.

3.2 Background: CADA-VAE

We first introduce a state-of-the-art ZSL methods, Cross and Distribution Aligned VAE (CADA-VAE), which is the basic model of our method. Its goal is to search a common classification space, where the embeddings of semantic features and visual features are aligned. The model contains two VAEs, one for semantic features and the other for visual features, each of which consists of an encoder and a decoder. The objective function of a VAE is the variational lower bound on the marginal likelihood of a given sample, which can be formulated as:

$$\mathcal{L} = \mathbb{E}_{q_\phi(z|x)} \left[\log p_\theta(x|z)\right] - \lambda D_{KL} \left(q_\phi(z|x) \| p_\theta(z)\right), \quad (1)$$

where the first term is the reconstruction loss and the second term is the unpacked Kullback-Leibler divergence to regularize the inference model $q(z|x)$ and $p(z)$. In addition, λ is employed to weight the KL-Divergence. The encoder predicts μ and Σ such that $q_\phi(z|x) = \mathcal{N}(\mu, \Sigma)$, and a latent vector z is obtained by employing the reparametrization trick. The encoders is used to project features into the common space and the decoders are used to reconstruct the original data. The VAE loss of the whole model is the sum of two VAE basic losses:

$$\mathcal{L}_{VAE} = \mathcal{L}_{VAE}^a + \mathcal{L}_{VAE}^v, \quad (2)$$

where $\mathcal{L}_{VAE}^a$ and $\mathcal{L}_{VAE}^v$ represent the VAE losses of semantic modality and visual modality respectively. Besides, aiming to match the embeddings from semantic space and visual space in the common space, the model aligns the latent distributions precisely and needs a cross-reconstruction criterion to ensure. Therefore, the cross-alignment loss (CA) and distribution-alignment loss (DA) are designed and applied.

The cross-alignment loss regulars the reconstructed features from the other modality to be similar to the original modality features. The cross-Alignment loss is:

$$\mathcal{L}_{CA} = |c - D_a(E_v(x))| + |x - D_v(E_a(c))|, \quad (3)$$

where c, D_a and E_a are feature, decoder and encoder of semantic modality, and x, D_v and E_v are feature, decoder and encoder of visual modality.

The distribution-alignment loss is employed to minimize the Wasserstein distance between the latent Gaussian distributions of semantic modality and visual modality, which makes the latent embedding from semantic space and visual space matched. The distance is denoted as:

$$\mathcal{L}_{DA} = \left(\|\mu_a - \mu_v\|_2^2 + \left\| \Sigma_a^{\frac{1}{2}} - \Sigma_v^{\frac{1}{2}} \right\|_{\text{Frobenius}}^2 \right)^{\frac{1}{2}}, \quad (4)$$

where μ_a and Σ_a are predicted by the encoder E_a, while μ_v and Σ_v are predicted by the encoder E_v. The objective function can be denoted as:

$$\mathcal{L}_{CACD-VAE} = \mathcal{L}_{VAE} + \gamma \mathcal{L}_{CA} + \delta \mathcal{L}_{DA}, \quad (5)$$

where γ and δ are the hyper-parameters of the cross alignment and the distribution alignment loss to weight these losses.

3.3 Unified Semantic Embedding

Since the numbers and kinds of attributes are different among datasets, the semantic embeddings of different datasets are various and complex, which is the challenge to be solved first. To solve this problem, we try to find unified semantic embeddings of different datasets. After training the t^{th} task, semantic embeddings c^t can be predicted as μ_a^t and Σ_a^t mapped by E_a^t. The latent vector z is generated by employing the reparametrization trick, the process of which is to generate various latent vectors from point data. The generated latent vectors can be the training data for the final classifier, which contain the discriminative information of the corresponding classes. Based on this, we replace original semantic embedding c^t with μ_a^t and Σ_a^t, from one point data to two point data, which can be viewed as more representative semantic embeddings. After training all tasks, we can employ these new semantic embeddings to replay latent vectors of all datasets, and train robust classifiers.

3.4 Selective Retraining

For the new task, a natural way would be fine-tuning the entire model. However, fine-tuning the entire model would change the affected weights of previous tasks, leading to Catastrophic Forgetting of neural network. Thus, we employ selective retraining strategy to fine-tune the whole model. When the unified semantic embeddings are obtained, the classification spaces for different datasets are fixed, which are also the latent spaces for previous tasks. Therefore, the model that is the projection from the visual space to the classification space, is the encoder of visual modality E_v^t. We denote W^t as the parameter of E_v^t and W_l^t is denoted as the model parameter at layer l, the number of whose layer is L. When a new task arrives, we first froze the parameters W_L^{t-1} and fine-tune the model to obtain the connections between the output unit o_t and the hidden unit at layer $L-1$. Then, we can select all units and weights that are affected in the training process, and remain the part that are not connected to output unit o_t unchanged. The selective operation can be viewed as giving the model an initialization, ensuring that the direction of optimization is to protect the classification spaces of previous tasks. Finally, we only fine-tune the selected weights, which is denoted as W_S^t. Algorithm 1 describes the selective retraining process.

Algorithm 1 The Process of Selective Retraining

Input: Dataset S^t, Previous parameter W^{t-1}
Output: Selected parameter W_s^t

```
Froze parameter W_L^{t-1}, S^t = {o_t}
Fine-tune the network
for l = L, ..., 1 do
    Add neural i to S^t if there exists some neural j ∈ S such that W_{l,ij}^{t-1} ≠ 0
end for
Fine-tune the selected parameter W_S^t
```

3.5 Knowledge Distillation

Through selective retraining, the selective neurons change and other neurons are frozen, but the optimization direction of the whole model, which motivates the model to preserve the knowledge of previous tasks, is not ensured. Aiming to transfer the knowledge from previous tasks to current task, we adopt knowledge distillation strategy. When the t^{th} task arrives, we hope the outputs of E_v^t is similar to the outputs of E_v^{t-1} with the same input x^t, which would ensure the classification spaces of the t^{th} task and the $(t-1)^{th}$ task are approximate. After training all datasets sequentially, the final E_v have the ability to predict the similar μ_v^t and Σ_v^t as the E_v^t when inputting the same image feature x^t. The distillation loss is denoted as :

$$\mathcal{L}_{KD} = \left\| \mu_v^t - \widehat{\mu_v^t} \right\|_1 + \left\| \Sigma_v^t - \widehat{\Sigma_v^t} \right\|_1, \quad (6)$$

where μ_v^t and Σ_v^t are predicted by E_v^t, while $\widehat{\mu_v^t}$ and $\widehat{\Sigma_v^t}$ are predicted by E_v^{t-1}.

When $t > 1$, the objective function is denoted as:

$$\mathcal{L} = \mathcal{L}_{CACD-VAE} + \beta \mathcal{L}_{KD}, \quad (7)$$

where β is the hyper-parameter to weight the knowledge distillation loss and set to 1.

3.6 Training and Inference

In training, we train the datasets sequentially and save the unified semantic embeddings of all classes. After the training stage of VAEs, we employ the saved semantic embeddings to replay the latent vectors of all classes. The process of generating latent vectors is repeated n_s times for every seen class and n_u for every unseen class. n_s and n_u are set to 200 and 400, respectively. These latent vectors contain the discriminative information of these classes. We use the latent vectors of different datasets to train softmax classifiers respectively.

Dataset	Semantics Dim	Image	Seen Classes	Unseen Classes
APY	64	15339	20	12
AWA1	85	30475	40	10
CUB	312	11788	150	50
SUN	102	14340	645	72

Table 1: Datasets used in our experiments, and their statistics.

In testing stage, the test visual features of seen classes and unseen classes are projected as the latent vectors by the encoder of visual modality E_v. Then the test features are fed to the trained classifier to get the results on different datasets.

4 Experiment

In this section, involved datasets, evaluation metrics and the implementation details are introduced in detail. Then, we will present several state-of-the-art competitors as well as the experimental results of our method. Finally, the ablation studies will prove the effectiveness of our proposed approach.

4.1 Benchmark and Evaluation Metrics

We evaluate our method on four dataset: Attribute Pascal and Yahoo dataset (aPY) [Farhadi *et al.*, 2009], Animals with Attributes 1 (AWA1) [Xian *et al.*, 2018a], Caltech-UCSD-Birds 200-2011 dataset (CUB) [Wah *et al.*, 2011], and SUN Attribute dataset (SUN) [Patterson and Hays, 2012]. Statistics of the datasets are presented in Table 1. For all datasets, we extract 2048 dimensional visual features using the pre-trained 101-layered ResNet. The sequence of training dataset is aPY, AWA1, CUB and SUN, which is alphabetical order.

Following the Generalized Zero-Shot Learning setting, we employ the same evaluation metrics for LZSL:

- u: average per-class classification accuracy on test images from the unseen classes with the prediction label set, which is used to measure the capacity of recognizing unseen classes.
- s: average per-class classification accuracy on test images from the seen classes with the prediction label set, which is used to measure the capacity of recognizing incremental seen classes.
- H: the harmonic mean of u and s, which is formulated as

$$H = \frac{2 \times u \times s}{u + s}. \tag{8}$$

H balances the performance between u and s metrics, which is the most important metrics for our task. All results of the three metric are measured after the training of all datasets.

4.2 Implementation Details

All encoders and decoders are multilayer perceptrons with one hidden layer. We use 1560 hidden units for the image feature encoder and 1660 for the decoder. The attribute encoder and decoder have 1450 and 660 hidden units, respectively. δ is increased from epoch 6 to epoch 22 by a rate of 0.54 per epoch, while γ is increased from epoch 21 to 75 by 0.044 per epoch. The weight λ of the KL-divergence is increased by a rate of 0.0026 per epoch until epoch 90. Besides, we use the L1 distance as reconstruction error, which obtains better results than L2.

For every dataset, the number of epochs is set to 100, and the batch size is set to 50. The learning rate of VAEs is set to 0.00015, which is set to 0.001 for classifiers. In addition, our method is implemented with PyTorch and optimized by ADAM optimizer.

4.3 Comparison to Exisiting Baselines

Baseline Models. Since there is no previous work for Lifelong Zero-Shot Learning, we compare the baselines, which combine CACD-VAE with traditional lifelong methods. (a) Sequential Fine-tuning (SFT): The model is fine-tuned when a new task arrives sequentially, the parameters of which is initialized from the model trained/fine-tuned on the previous task. (b) L2 regularization (L2): at each task t , W^t is initialized as W^{t-1} and continuously trained with L2-regularization between W^t and W^{t-1}. (C) L1 regularization (L1): at each task t , W^t is initialized as W^{t-1} and continuously trained with L1-regularization between W^t and W^{t-1}.

Results and Analysis. Table 2 summarizes the results of all the comparing methods and our method under three evaluation metrics on the four benchmark datasets. For ZSL methods on GZSL metrics, the H is the most important metric to evaluate the performance of ZSL methods, which balances the performance of u and s metrics.

The "Base" in Table 2 denotes the model is trained sequentially without any lifelong strategy and the "Original" denotes the models, which train the datasets respectively. Obviously, we can find the results of base obtain the worst performance of previous datasets, which do not have the ability to accumulate the knowledge of previous datasets when a new task arrives. Besides, the model with sequential fine-tuning strategy also obtain the worse results compared with those without such a strategy, which indicates the existence of Catastrophic Forgetting in ZSL.

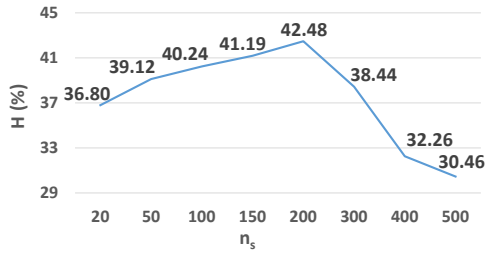
Compared with other baselines, our method obtains the best performances of three evaluation metrics in previous three datasets. On aPY, our model achieves 29.11% in u, 43.29% in s and 34.81% in H, with improvements of 2.69% in u, 13.50% in s and 6.80% in H. On AWA1, our model achieves 51.17% in u, 63.66% in s and 56.73% in H, with improvements of 1.53% in u, 4.59% in s and 3.14% in H. On CUB, our model achieves 38.82% in u, 45.81% in s and 42.03% in H, with improvements of 3.29% in u, 11.07% in s and 7.68% in H. Although our method do not obtain best results in SUN datasets, the drop of results is little compared with the improvement in other datasets, whose reason is that our method balances the ability of accumulating knowledge from previous tasks and capturing knowledge of current task better. We also calculate the average H results of these methods on four datasets. The average H results are 10.2%, 36.73%, 38.03%, 36.73% and 42.48% for base, SFT, L1, L2 and our method, with improvements of 4.45% in average H results. In conclusion, our method obtains a balanced performance of previous tasks and current task, which notably outperforms the baselines.

Method	aPY			AWA1			CUB			SUN		
	u	s	H	u	s	H	u	s	H	u	s	H
Base	6.69	0.59	1.09	5.14	0.92	1.56	0.87	0.67	0.76	**43.40**	**33.95**	**38.10**
SFT	24.24	23.21	23.71	47.27	55.18	50.92	35.46	34.74	35.10	38.47	36.10	37.20
L1	26.42	29.79	28.01	49.64	58.23	53.59	35.11	32.31	33.65	40.14	34.11	36.88
L2	24.08	23.61	23.84	46.71	59.07	52.17	35.53	33.24	34.35	42.08	32.33	36.56
Ours	**29.11**	**43.29**	**34.81**	**51.17**	**63.66**	**56.73**	**38.82**	**45.81**	**42.03**	42.43	31.78	36.34
Original	30.36	59.36	40.18	57.30	72.80	64.10	53.50	51.60	52.40	35.70	47.20	42.60

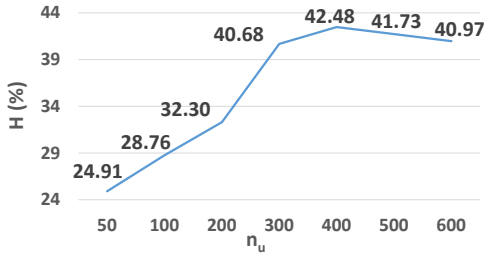
Table 2: Classification accuracy (%) of Lifelong Zero-Shot Learning with the three evaluation metrics on the four datasets.

Method	aPY			AWA1			CUB			SUN		
	u	s	H	u	s	H	u	s	H	u	s	H
Base	24.24	23.21	23.71	47.27	55.18	50.92	35.46	34.74	35.10	38.47	**36.10**	**37.20**
KD	26.47	35.09	30.17	**56.95**	52.67	54.73	37.65	42.92	40.11	41.53	32.48	36.45
SR	25.63	40.62	31.43	53.70	56.94	55.27	**40.94**	40.64	40.79	40.42	31.98	35.70
Ours	**29.11**	**43.29**	**34.81**	51.17	**63.66**	**56.73**	38.82	**45.81**	**42.03**	**42.43**	31.78	36.34

Table 3: Ablation study: classification accuracy (%) with different modules, "KD" and "SR" respectively indicate knowledge distillation and selective retraining.



(a) n_s



(b) n_u

Figure 3: The average H results with different n_s and n_u hyper-parameters.

4.4 Ablation Study

We conduct two groups of ablation experiments to study the effectiveness of our method.

The results of our basic model added different modules are presented in Table 3. The base model is CACD-VAE with sequential fine-tuning training strategy. Based on the base model, we add knowledge distillation and selective retraining modules, which are represented as "KD" and "SR" respectively. As shown in Table 3, both knowledge distillation and selective retraining can improve the performance on the previous three datasets. The improvement of adding "KD" indicates knowledge distillation can transfer the knowledge of previous task to the current task, which remits the unfavourable influence of Catastrophic Forgetting to some extent. Besides, the improvement of adding "SR" indicates selective retraining can preserve the affected weights of previous tasks and avoid negative transfer, since neurons that are not selected will not get affected by the retraining process. When adding all modules, our method performs best.

We perform an experiment to discuss the influence of the numbers n_s and n_u for replaying, whose average H results are shown in Figure 3. The best performance is achieved when n_s and n_u are set as 200 and 400. Obviously, we can notice the phenomenon that the average H increases with the increasing of n_s and n_u before achieving the peak performance of the average H.

5 Conclusion

Tto our best knowledge, this paper strikes the first effort to introduce and tackle Lifelong Zero-Shot Learning. Firstly, we employ VAEs to obtain the unified semantic embeddings, which bridges the gaps among the semantic embeddings of different datasets. Then, the selective retraining strategy are leveraged to preserve the projection to a great extent, which is constructed in previous training stage. Finally, we distillate the knowledge from previous tasks and transfer to current training stage. Experiments show that our method outperforms previous methods by a large margin on four benchmark datasets.

Acknowledgments

Our work was supported in part by the Key R&D Program-The Key Industry Innovation Chain of Shaanxi under Grant 2019ZDLGY03-02-01, and in part by the National Key R&D Program of China under Grant 2017YFE0104100 and 2016YFE0200400.

References

[Chang *et al.*, 2020] Dongliang Chang, Yifeng Ding, Jiyang Xie, Ayan Kumar Bhunia, Xiaoxu Li, Zhanyu Ma, Ming Wu, Jun Guo, and Yi-Zhe Song. The devil is in the channels: Mutual-channel loss for fine-grained image classification. *IEEE Transactions on Image Processing*, 29:4683–4695, 2020.

[Chen *et al.*, 2018] Long Chen, Hanwang Zhang, Jun Xiao, Wei Liu, and Shih-Fu Chang. Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In *CVPR*, pages 1043–1052, 2018.

[Chen *et al.*, 2019] Hanting Chen, Yunhe Wang, Chang Xu, Zhaohui Yang, Chuanjian Liu, Boxin Shi, Chunjing Xu, Chao Xu, and Qi Tian. Data-free learning of student networks. In *ICCV*, pages 3514–3522, 2019.

[Farhadi *et al.*, 2009] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *CVPR*, pages 1778–1785. IEEE, 2009.

[Felix *et al.*, 2018] Rafael Felix, Vijay BG Kumar, Ian Reid, and Gustavo Carneiro. Multi-modal cycle-consistent generalized zero-shot learning. In *ECCV*, pages 21–37, 2018.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Nips*, pages 2672–2680, 2014.

[Hinton *et al.*, 2015] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv*, 2015.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv*, 2013.

[Li and Hoiem, 2017] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Trans. Pattern Anal. Mach. Intell*, 40(12):2935–2947, 2017.

[Liu *et al.*, 2018] Xialei Liu, Marc Masana, Luis Herranz, Joost Van de Weijer, Antonio M Lopez, and Andrew D Bagdanov. Rotate your networks: Better weight consolidation and less catastrophic forgetting. In *ICPR*, pages 2262–2268. IEEE, 2018.

[McCloskey and Cohen, 1989] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

[Patterson and Hays, 2012] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, pages 2751–2758. IEEE, 2012.

[Rebuffi *et al.*, 2017] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, pages 2001–2010, 2017.

[Schonfeld *et al.*, 2019] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *CVPR*, pages 8247–8255, 2019.

[Shin *et al.*, 2017] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *Nips*, pages 2990–2999, 2017.

[Socher *et al.*, 2013] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *Nips*, pages 935–943, 2013.

[Wah *et al.*, 2011] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[Wei *et al.*, 2019] Kun Wei, Muli Yang, Hao Wang, Cheng Deng, and Xianglong Liu. Adversarial fine-grained composition learning for unseen attribute-object recognition. In *ICCV*, pages 3741–3749, 2019.

[Wu *et al.*, 2018] Chenshen Wu, Luis Herranz, Xialei Liu, Joost van de Weijer, Bogdan Raducanu, et al. Memory replay gans: Learning to generate new categories without forgetting. In *Nips*, pages 5962–5972, 2018.

[Xian *et al.*, 2018a] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *IEEE Trans. Pattern Anal. Mach. Intell*, 2018.

[Xian *et al.*, 2018b] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *CVPR*, pages 5542–5551, 2018.

[Xu *et al.*, 2019] Xinyi Xu, Huanhuan Cao, Yanhua Yang, Erkun Yang, and Cheng Deng. Zero-shot metric learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 3996–4002. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[Yang *et al.*, 2019] Xu Yang, Cheng Deng, Feng Zheng, Junchi Yan, and Wei Liu. Deep spectral clustering using dual autoencoder network. In *CVPR*, June 2019.

[Yoon *et al.*, 2017] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. *arXiv*, 2017.

[Zhang *et al.*, 2017] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *CVPR*, pages 2021–2030, 2017.

[Zhao *et al.*, 2018] Bo Zhao, Xinwei Sun, Yanwei Fu, Yuan Yao, and Yizhou Wang. Msplit lbi: Realizing feature selection and dense estimation simultaneously in few-shot and zero-shot learning. *ICML*, 2018.

[Zhao *et al.*, 2019] Bo Zhao, Xinwei Sun, Xiaopeng Hong, Yuan Yao, and Yizhou Wang. Zero-shot learning via recurrent knowledge transfer. In *WACV*, pages 1308–1317. IEEE, 2019.

[Zhu *et al.*, 2018] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *CVPR*, pages 1004–1013, 2018.