https://doi.org/10.2352/EI.2026.38.7.IMAGE-271

Dual Embedding Redefines Zero-Shot Image Classification

Qin Li
School of Computer and Software Engineering, Shenzhen University of Information Technology, Shenzhen, China
Jane You
Department of Computing, The Hong Kong Polytechnique University, Hong Kong, China
Lin Shu
School of Future Technology, South China University of Technology, Guangzhou, China

Abstract

Zero-shot learning (ZSL) aims to classify unseen classes using semantic information from seen classes. However, existing methods often struggle with visual variations within the same attribute, leading to noisy features. We propose CRAE (Class Representation and Attribute Embedding), a novel ZSL method that combines class representation learning and attribute embedding learning for improved robustness and accuracy. CRAE introduces an adaptive softmax activation to normalize attribute feature maps, reducing noise and enhancing discriminability. It also employs attribute-level contrastive learning with hard sample selection and class-level contrastive learning to improve classification performance. Experimental results on CUB, SUN, and AWA2 demonstrate that CRAE outperforms state-of-the-art methods, proving its superiority in zero-shot image classification.

Introduction

Supervised image classification has achieved remarkable accuracy, but it heavily relies on large labeled datasets and can only classify seen classes. Zero-Shot Learning (ZSL) addresses this by transferring knowledge from seen to unseen classes using shared semantic information. ZSL can be divided into **Conventional ZSL (CZSL)** and **Generalized ZSL (GZSL)**, where the latter tests both seen and unseen classes. ZSL methods are typically categorized into **embedding-based** and **generative-based** approaches. Embedding-based methods align visual features with semantic vectors, but they struggle with local feature variability, leading to noise. Generative methods synthesize features for unseen classes but often fail to generate discriminative features and can be unstable.

Recent attention-based methods aim to focus on localized features, but still face challenges with visual representation variability. Hence, developing methods that effectively capture both **class-level and attribute-level features** while minimizing noise is essential.

Zero-Shot Learning Methods

ZSL methods are divided into embedding-based and generative-based approaches. Embedding-based methods align visual features with class semantic vectors using CNNs, but they often include irrelevant background information. Attention-based methods like DAZLE [1] focus on local features to improve classification. Generative methods, such as GANs [2] or VAEs [3], generate features for unseen classes but face instability and challenges in generating high-quality features.

Contrastive Learning in ZSL

Contrastive learning maximizes the similarity between positive pairs and minimizes it between negative pairs. Unsupervised contrastive learning, like SimCLR [4], generates augmented views of images to form positive pairs. Supervised contrastive learning, such as SCL [5], treats images of the same class as positive and those of different classes as negative. Recent methods like EMP [6] and CE-GZSL [7] introduce attribute-level constraints to improve transferability and feature alignment.

We propose CRAE (Class Representation and Attribute Embedding), which combines class representation learning and attribute embedding learning. CRAE uses an adaptive softmax activation to reduce noise and enhance discriminability. Additionally, it leverages attribute-level contrastive learning with hard sample selection and integrates class-level contrastive learning for better feature alignment. Experimental results on CUB, SUN, and AWA2 datasets show that CRAE outperforms state-of-the-art methods in zero-shot image classification.

Method

In this section, we present the notation, problem settings, and the proposed framework for zero-shot learning (ZSL). Our model jointly optimizes class representation learning and attribute embedding learning to acquire highly discriminative visual features.

Notation and Problem Settings

ZSL aims to transfer knowledge from seen classes ($\mathscr{Y}^{\mathscr{S}}$) to unseen classes ($\mathscr{Y}^{U}$), where $\mathscr{Y}^{\mathscr{S}} \cap \mathscr{Y}^{U} = \emptyset$. The training set is $T^s = \{x_i^s, y_i^s, \varphi(y_i^s)\}_{i=1}^{N^s}$, where x_i^s is an image, y_i^s is the class label, and $\varphi(y_i^s)$ is the class semantic vector. The test set is $T^u = \{x_i^u, y_i^u, \varphi(y_i^u)\}_{i=1}^{N^u}$. In conventional ZSL, the goal is to map seen class images to unseen class labels, i.e., $\mathscr{X}^{S} \rightarrow \mathscr{Y}^{U}$. In generalized ZSL, the goal is to map images to both seen and unseen labels, i.e., $\mathscr{X}^{S} \rightarrow \mathscr{Y}^{U} \cup \mathscr{Y}^{\mathscr{S}}$.

Proposed Framework

Our framework includes two main components: class representation learning and attribute embedding learning. These components work together to extract discriminative features and align them with corresponding prototypes.

Class Representation Learning: We use a ResNet-101 network to extract local features $f(x)$, then apply global average pooling to obtain global visual features $g(x)$. Class-level contrastive learning enhances the discriminative power of these global features, which are then aligned with their corresponding

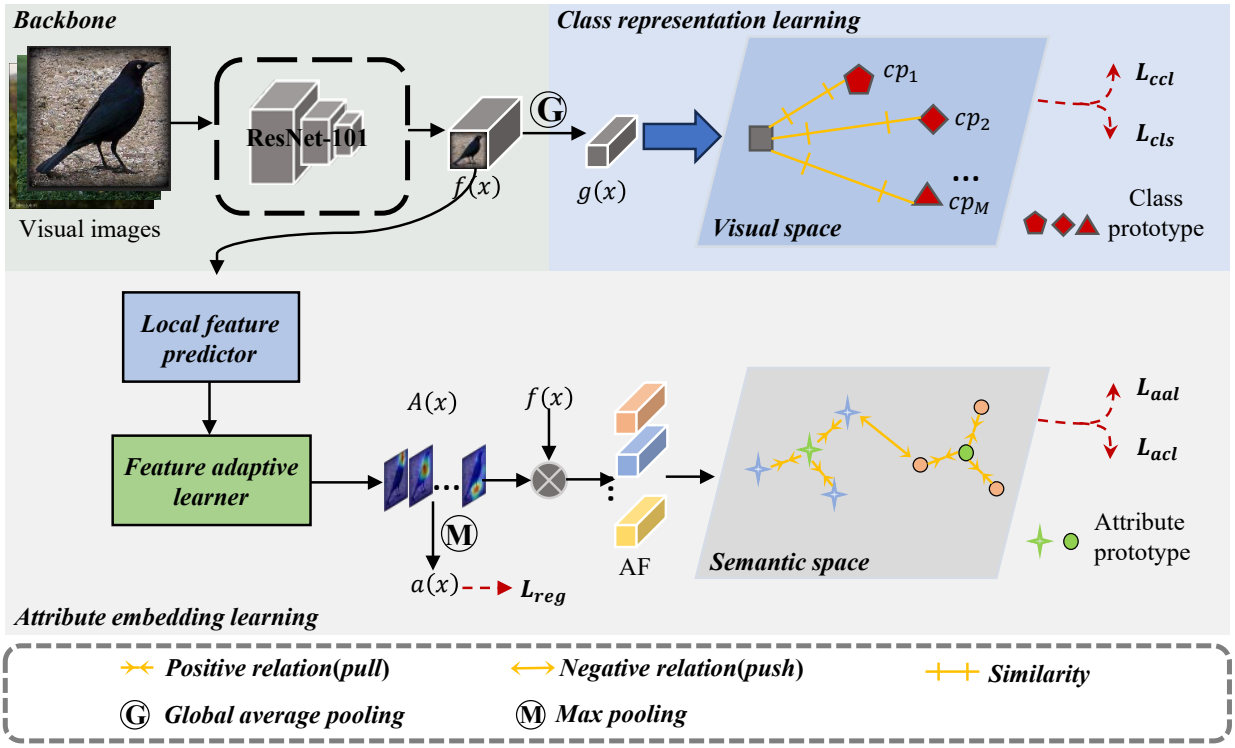


Figure 1: Illustration of our proposed framework.

class prototypes for classification.

Attribute Embedding Learning: Local visual features $f(x)$ are processed through a feature predictor to obtain attribute feature maps $A(x)$. These maps are normalized using an adaptive softmax activation. We then apply attribute-level contrastive learning to improve feature discriminability. Finally, we align the attribute-level features with attribute prototypes to aid knowledge transfer to unseen classes.

Loss Functions

We employ several loss functions to optimize our framework.

Class-level Contrastive Loss: This loss encourages the network to maximize the similarity between features of the same class and minimize it between features of different classes:

$$\mathscr{L}_{ccl} = \frac{1}{B}\sum_{i=1}^{B}\frac{-1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp(\cos(g(i),g(p))/\tau)}{\sum_{a\neq i}^{B}\exp(\cos(g(i),g(a))/\tau)} \quad (1)$$

where $P(i)$ represents the set of positive samples for class i.

Global Feature Alignment Loss: To align global features $g(x)$ with class prototypes, we use a cosine similarity-based loss:

$$\mathscr{L}_{cls} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\alpha\cos(g(x),c_{p_y}))}{\sum_{i=1}^{M}\exp(\alpha\cos(g(x),c_{p_i}))} \quad (2)$$

Attribute-Level Contrastive Loss: To enhance attribute discriminability, we use a contrastive loss that brings features of the same attribute closer and separates features of different attributes:

$$\mathscr{L}_{acl} = \frac{1}{\hat{K}}\sum_{j=1}^{\hat{K}}\left(-\frac{1}{U}\sum_{u=1}^{U}\log\frac{S(\hat{a}_f^j,a_{f_{ju}}^+)}{\sum_{u=1}^{U}S(\hat{a}_f^j,a_{f_{ju}}^+)}+\sum_{v=1}^{V}S(\hat{a}_f^j,a_{f_{jv}}^-)\right) \quad (3)$$

Zero-Shot Classification

For zero-shot classification, we predict the class label $\hat{y}$ using the closest matching prototype.

Conventional ZSL: For unseen classes, we predict:

$$\hat{y} = \arg\max_{\tilde{y}\in\mathscr{Y}^U}\alpha\cdot\cos(g(x),c_{p_{\tilde{y}}}) \quad (4)$$

Generalized ZSL: We introduce a calibration factor γ for seen classes:

$$\hat{y} = \arg\max_{\tilde{y}\in\mathscr{Y}^U\cup\mathscr{Y}^S}\alpha\cdot\cos(g(x),c_{p_{\tilde{y}}})-\gamma\mathbb{I}[\tilde{y}\in\mathscr{Y}^S] \quad (5)$$

Experiments

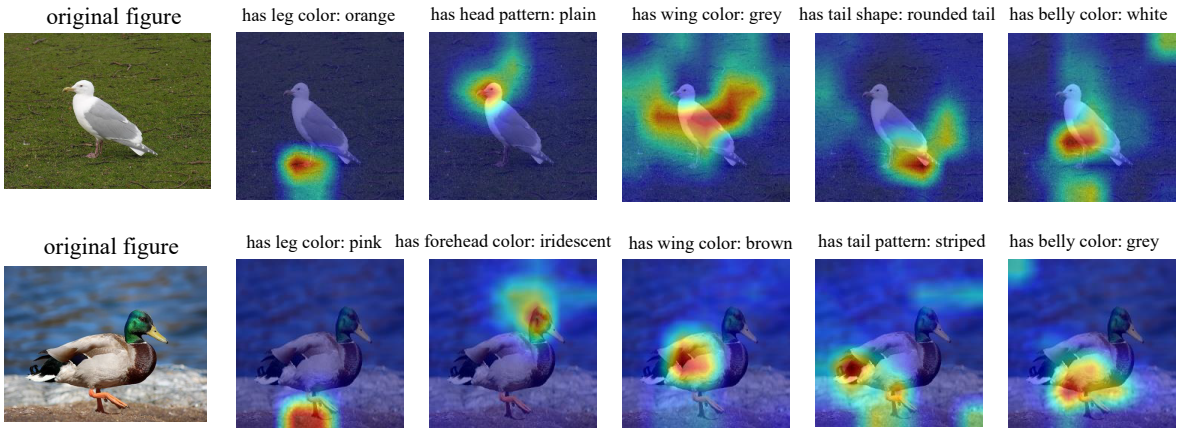
We conduct experiments on three widely used datasets: **CUB-200-2011 (CUB)** [8], **SUN** [9], and **AWA2** [10].

To evaluate our model's performance, we compare CRAE with state-of-the-art algorithms in two categories: generative-based and embedding-based zero-shot learning. Figure 2 visualizes attribute feature maps on the CUB and SUN datasets.

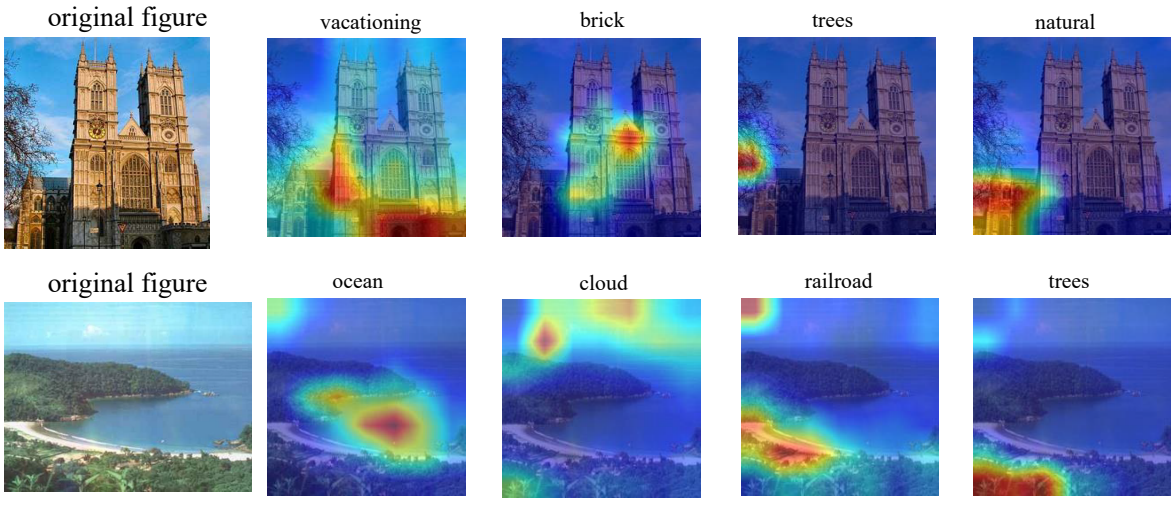
Performance on CZSL: CRAE outperforms state-of-the-art methods on all three datasets, achieving classification accuracies

Table 1: Results (%) of CRAE and other state-of-the-art methods on CUB, SUN, and AWA2. The best results are marked in red. The second-best results are marked in blue.

Type	Methods	CUB				SUN				AWA2			
		CZSL	*U*	*S*	*H*	CZSL	*U*	*S*	*H*	CZSL	*U*	*S*	*H*
Generative	f-CLSWGAN [2]	57.3	43.7	57.7	49.7	60.8	42.6	36.6	39.4	68.2	57.9	61.4	59.6
	f-VAEGAN-D2 [11]	61.0	48.4	60.1	53.6	64.7	45.1	38.0	41.3	71.1	57.6	70.6	63.5
	TF-VAEGAN [12]	64.9	52.8	64.7	58.1	66.0	45.6	40.7	43.0	72.2	59.8	75.1	66.6
	Composer [13]	69.4	56.4	63.8	59.9	62.6	55.1	22.0	31.4	71.5	62.1	77.3	68.8
	HSVA [14]	-	52.7	58.3	55.3	-	48.6	39.0	43.3	-	56.7	79.8	66.3
Embedding	TCN [15]	59.5	52.6	52.0	52.3	61.5	31.2	37.3	34.0	71.2	61.2	65.8	63.4
	AREN [16]	71.8	38.9	78.7	52.1	60.6	19.0	38.8	25.5	67.9	15.6	92.9	26.7
	DAZLE [1]	66.0	56.7	59.6	58.1	59.4	52.3	24.3	33.2	67.9	60.3	75.7	67.1
	RGEN [17]	76.1	60.0	73.5	66.1	63.8	44.0	31.7	36.8	73.6	67.1	76.5	71.5
	CRAE (Ours)	79.4	71.4	78.2	74.6	67.7	49.6	41.2	45.0	75.8	68.6	88.1	77.2



(a) CUB



(b) SUN

Figure 2: Visualization results of attribute feature maps on CUB and SUN datasets.

of 79.4%, 67.7%, and 75.8% for CUB, SUN, and AWA2, respectively. CRAE surpasses the second-best method by 2.2% on AWA2 and 1.6% and 1.7% on CUB and SUN, respectively.

Performance on GZSL: CRAE achieves the highest H metric on all three datasets, with results of 74.6% on CUB, 45.0% on SUN, and 77.2% on AWA2. The model outperforms the second-best method by 1.7% on SUN, demonstrating improved performance on both seen and unseen classes. CRAE effectively balances classification accuracy between seen and unseen classes, mitigating the bias observed in existing methods.

Conclusion

This work addresses limitations in zero-shot image classification, particularly visual variations and noise from irrelevant features. We propose CRAE, a framework that combines class representation and attribute embedding learning.

CRAE optimizes class-level and attribute-level features for better classification. Attribute-level contrastive learning with hard sample selection reduces discrepancies, while adaptive softmax activation mitigates noise. Class-level contrastive learning enhances global feature discriminability, improving classification accuracy.

Experiments on CUB, SUN, and AWA2 datasets show CRAE outperforms state-of-the-art methods, demonstrating its effectiveness for real-world zero-shot classification.

References

[1] Dat Huynh and Ehsan Elhamifar. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4483–4493, 2020.

[2] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5542–5551, 2018.

[3] Jingjing Li, Mengmeng Jing, Ke Lu, Zhengming Ding, Lei Zhu, and Zi Huang. Leveraging the invariant side of generative zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7402–7411, 2019.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607, 2020.

[5] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

[6] Yang Zhang and Songhe Feng. Enhancing domain-invariant parts for generalized zero-shot learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6283–6291, 2023.

[7] Zongyan Han, Zhenyong Fu, Shuo Chen, and Jian Yang. Contrastive embedding for generalized zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2371–2381, 2021.

[8] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010.

[9] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2751–2758, 2012.

[10] 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 transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018.

[11] Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. f-vaegan-d2: A feature generating framework for any-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 10275–10284, 2019.

[12] Sanath Narayan, Akshita Gupta, Fahad Shahbaz Khan, Cees GM Snoek, and Ling Shao. Latent embedding feedback and discriminative features for zero-shot classification. In *European Conference on Computer Vision*, pages 479–495, 2020.

[13] Dat Huynh and Ehsan Elhamifar. Compositional zero-shot learning via fine-grained dense feature composition. *Advances in Neural Information Processing Systems*, 33:19849–19860, 2020.

[14] Shiming Chen, GuoSen Xie, Yang Liu, Qinmu Peng, Baigui Sun, Hao Li, Xinge You, and Ling Shao. Hsva: Hierarchical semantic-visual adaptation for zero-shot learning. *Advances in Neural Information Processing Systems*, 34:16622–16634, 2021.

[15] Huajie Jiang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Transferable contrastive network for generalized zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9765–9774, 2019.

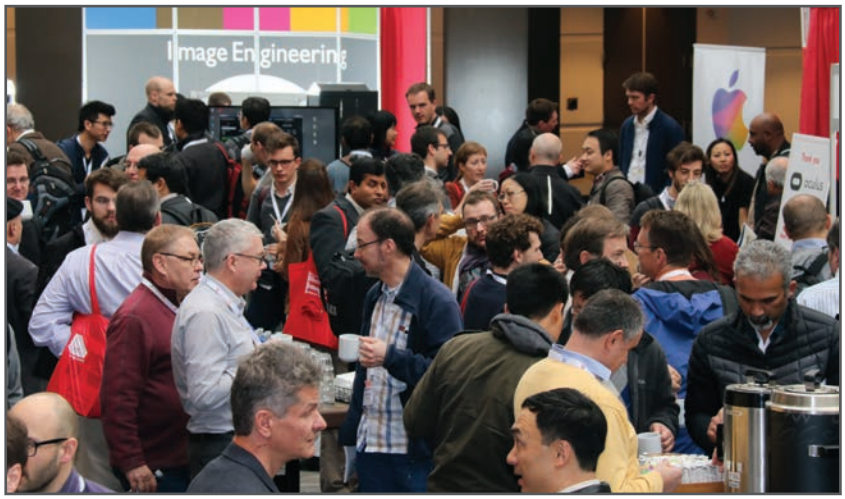
[16] Guo-Sen Xie, Li Liu, Xiaobo Jin, Fan Zhu, Zheng Zhang, Jie Qin, Yazhou Yao, and Ling Shao. Attentive region embedding network for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9384–9393, 2019.

[17] Guo-Sen Xie, Li Liu, Fan Zhu, Fang Zhao, Zheng Zhang, Yazhou Yao, Jie Qin, and Ling Shao. Region graph embedding network for zero-shot learning. In *European conference on computer vision*, pages 562–580, 2020.

JOIN US AT THE NEXT EI!

electronic IMAGING

Imaging across applications . . . Where industry and academia meet!



• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

ADVANCING IMAGING SCIENCE 75 1947 IS&T 2022 imaging.org