

Research and Empirical Analysis of Zero-Shot Image Classification Methods

Wei Chen

Digital and Information Technology Research Institute, Chinalco Zhengzhou Nonferrous Metals Research Institute Co., Ltd.,
Zhengzhou, Henan, 450000, China

ABSTRACT

With the widespread application of artificial intelligence in the field of image recognition, traditional supervised learning faces the bottleneck of data scarcity when processing tasks involving unseen categories. Zero-Shot Learning (ZSL), as an emerging learning paradigm, has become a research hotspot by introducing semantic information to enable models to recognize unknown categories. This paper focuses on the problem of zero-shot image classification and proposes a deep model that integrates visual feature extraction, semantic representation construction, and cross-modal alignment mechanisms. The model not only combines attribute vectors and text embeddings but also incorporates knowledge graph information to construct a richer semantic space. Meanwhile, contrastive learning and generative mechanisms are introduced to improve the model's generalization ability to unseen categories. Experiments were conducted on several standard datasets (AwA2, CUB, SUN), and the results indicate that the proposed model achieves excellent performance in both conventional ZSL and generalized ZSL (GZSL) tasks. Finally, this paper analyzes the limitations of the model and looks forward to future research directions such as multimodal fusion, graph-based modeling, and large-scale model transfer.

KEYWORDS

Zero-shot learning; Image classification; Semantic alignment; Cross-modal learning; Generative model

1 Introduction

1.1 Research Background

With the rapid development of deep learning technology, image classification has achieved remarkable results in the field of computer vision. However, most traditional image classification methods rely on a large number of labeled samples for supervised training, which faces many challenges in practical applications: on the one hand, the cost of obtaining a large amount of high-quality labeled data is high; on the other hand, a large number of "long-tail categories" or entirely new categories exist in the real world, and these categories often have no samples during the training phase, making it difficult for existing models to generalize.

In the traditional supervised learning framework, model performance is highly dependent on large-scale labeled data. However, in the real world, acquiring sufficient labeled samples for all categories is often expensive, and in some fields (such as medicine, remote sensing, rare species recognition, etc.), it is almost impossible to cover sample information for all categories. Therefore, how to enable models to identify unseen classes has become a critical research issue.

Zero-Shot Learning (ZSL) emerged to address this. Its core goal is to achieve accurate identification of unseen categories by modeling only with the semantic information of categories (such as attribute vectors, word vectors, text descriptions, etc.) when samples of certain categories do not appear in the training set. ZSL not only effectively alleviates the bottleneck of data annotation but also improves the model's generalization ability, thus holding broad application prospects in various fields such as image classification, natural language processing, recommendation systems, and cross-modal retrieval.

In recent years, with the development of deep learning, ZSL methods have continuously evolved from initial attribute transfer methods to more complex model structures based on generative models, adversarial networks, and graph neural networks. Despite this, ZSL still faces several challenges, including the gap between semantic and visual spaces, semantic confusion between categories, and insufficient authenticity and discriminability of generated samples. Therefore, how to effectively utilize semantic information and enhance the model's ability to distinguish unseen categories remains an important direction for current ZSL research.

1.2 Research Problem

This study aims to explore and construct a zero-shot image classification model that integrates visual features and semantic information. It focuses on the following key issues: how to construct a model structure that effectively aligns visual features and semantic embeddings, how to improve the classification accuracy and generalization ability of the model on unseen classes, and how to use real datasets for empirical validation to evaluate the model's performance in ZSL and GZSL scenarios.

1.3 Research Significance

From a theoretical perspective, ZSL challenges the strong dependence of traditional supervised learning on training samples and promotes the development of cross-modal representation learning and transfer learning. From a practical perspective, ZSL provides feasible solutions for fields such as medical image recognition, remote sensing, and biological species identification, especially in scenarios where labeled data is difficult to obtain.

By proposing a ZSL model that integrates image features and semantic embeddings and validating its performance with empirical data, this study is expected to provide theoretical support and practical references for solving few-shot/zero-shot classification problems.

2 Related Methods

2.1 Zero-Shot Learning

Zero-Shot Learning (ZSL) is a learning paradigm designed to solve the problem of identifying unseen categories during the training phase. Its core idea is to achieve knowledge transfer by introducing semantic information between categories. ZSL assumes that the category sets of the training and test sets are disjoint (i.e., seen/unseen class disjoint), connecting visual features and category labels through a shared semantic space (such as attributes, text descriptions, or knowledge graph embeddings).

Based on different task settings, ZSL can be divided into: Conventional ZSL (CZSL): The test categories are entirely unseen. Generalized ZSL (GZSL): The test categories include both seen and unseen classes, which is more challenging as the model must balance both.

ZSL typically includes three core elements: Visual Feature Space: High-dimensional visual features extracted from images. Semantic Embedding Space: Semantic representations of categories, such as attribute vectors, word vectors, and knowledge graphs. Mapping or Compatibility Function: Learning the projection or matching relationship from the visual space to the semantic space.

2.2 Zero-Shot Image Classification

Zero-shot image classification methods can be roughly classified into the following categories: Embedding-Based Models: These models learn a mapping function from the visual space to the semantic space, allowing test images to be projected into the semantic embedding for nearest neighbor matching. Examples include DeVISE (2013), ALE, SJE, ESZSL, Relation Net, and Deep Embedding Model. While simple and interpretable, they suffer from the "Projection Bias" problem and have weak generalization for unseen classes. Generative-Based Models: These models use Generative Adversarial Networks (GAN) or Variational Autoencoders (VAE) to generate image features for unseen categories, thereby converting ZSL into a supervised learning problem. Examples include f-CLSWGAN, CVAE-ZSL, GDAN, and Cycle- (U)WGAN. This approach solves the projection bias problem and enhances generalization but requires high stability and training rigor for the generative models. Vision-Language Models (VLMs): Recently, VLMs have made significant progress in multimodal learning. These models are pre-trained on large-scale datasets to efficiently learn joint representations of images and text. CLIP^[2] is a pioneering work in this field, known for its simplicity and efficiency, employing large-scale image-text pair training and contrastive learning. Other VLMs include ALIGN, FILIP, Florence, and REACT.

2.3 Related Research on Model Construction

Image Feature Extraction: Traditional methods are based on CNNs (e.g., ResNet, Inception); newer methods such as CLIP combine image and text joint training for stronger perception of semantic information. Semantic Space Modeling: Static attribute vectors (e.g., the 85-dimensional attributes provided by AwA); text semantics (Word2Vec, GloVe, BERT); knowledge graph embeddings (TransE, DistMult); and multimodal fusion for semantic space construction. Mapping and Matching Mechanisms: Linear projection (DeVISE); non-linear mapping (Deep Neural Networks); contrastive learning/compatibility learning (Triplet Loss, Cosine Similarity); and Attention mechanisms/Transformer structures to enhance alignment capability.

2.4 Current Shortcomings and Challenges

Despite the progress made by existing methods, challenges remain: Insufficient or Incomplete Semantic Information: Some datasets only provide static attributes, which cannot capture complex category semantics. Projection Bias: Since the mapping function is learned on seen categories, it tends to bias toward seen classes for unseen inputs. Poor Generalization: In GZSL scenarios, models often lean heavily toward seen classes, resulting in poor H-mean performance.

Semantic-Visual Gap: The distributions of visual features and semantic spaces are inconsistent, making precise alignment difficult. Inconsistent Evaluation Standards: Different papers use different dataset splits and metrics, leading to a lack of reproducibility.

3 Research Methods

3.1 Research Framework

This study proposes a zero-shot image classification method based on cross-modal alignment, combining visual features, semantic information, and alignment mechanisms. The framework is as follows: Input: Training images and their category labels, semantic information (attributes/text embeddings/knowledge graphs). Visual Feature Extraction: Extracting image features using CNN or Transformer. Semantic Representation Modeling: Constructing the semantic space using text embeddings (Word2Vec, BERT) or knowledge graphs. Cross-modal Alignment: Learning the mapping relationship between visual features and semantic representations. Classification Decision: Performing nearest neighbor search or end-to-end classification in the aligned space.

3.2 Image Feature Extraction Module

To obtain high-quality visual features, this study selects mainstream deep neural networks for image encoding: CNN Structure (e.g., ResNet-101): Using pre-trained models to extract high-level image features. Transformer Structure (e.g., CLIP): Leveraging cross-modal pre-training to enhance semantic consistency. Feature Processing: Global Average Pooling (GAP) and L2 normalization to reduce the impact of uneven data distribution.

Mathematical Expression:

$$V = f_{CNN}(I) \text{ or } V = f_{Transformer}(I)$$

where I is the input image and V is the extracted feature vector.

3.3 Semantic Embedding Module

The following methods are used to construct category semantic information: Attribute-based: Directly using the category attributes provided in the dataset (e.g., 85-dimensional attribute vectors in AwA2). Text-based: Using pre-trained Word2Vec/GloVe/BERT to extract word vectors of category names; calculating the average word vector of category description text as the semantic representation:

$$S = \frac{1}{N} \sum_{i=1}^N BERT(word_i)$$

Graph-based: Using methods like TransE or DistMult to build correlation information between categories; learning category embeddings through Graph Neural Networks (GNN) to improve semantic propagation. Final semantic representation is denoted as:

$$S = f_{semantic}(C)$$

where C is category information and S is the corresponding semantic embedding.

3.4 Cross-modal Alignment Module

This study designs a cross-modal alignment method to learn the mapping relationship: Linear Projection: Learning a linear transformation matrix W to project image features into the semantic space: $\hat{S} = WV$. Using Mean Squared Error (MSE) loss:

$$L_{MSE} = \|\hat{S} - S\|^2$$

Compatibility Function: Using a neural network to learn visual-semantic compatibility:

$$f(V, S) = W_2 \sigma(W_1 V)$$

Optimization using Contrastive Loss:

$$L_{contrastive} = \sum_{(V, S)} \log \frac{\exp(f(V, S))}{\sum_s \exp(f(V, S))}$$

Adversarial Generation: Using GAN/VAE to generate features for unseen categories:

$$G(z, s) = \hat{V}$$

Training with a classifier to perform well on synthetic features. The final alignment goal is $V \approx S$

3.5 Model Training and Optimization

3.5.1 Loss Function

The model is optimized using multiple loss functions: Contrastive Loss: Enhances visual-semantic consistency. Classification Loss (Cross-Entropy): Ensures the discriminability of generated features. Regularization Term (L2, KL divergence): Prevents overfitting. Total Loss Expression:

$$L = \lambda_1 L_{\text{contrastive}} + \lambda_2 L_{\text{class}} + \lambda_3 L_{\text{reg}}$$

3.5.2 Generalized ZSL Optimization

To improve GZSL performance, the study adopts: Semantic Consistency Constraint: Prevents seen class bias. Inter-class Balance Strategy: Adjusting seen/unseen prediction distributions.

4 Analysis and Discussion

4.1 Datasets and Experimental Settings

4.1.1 Dataset Introduction

To evaluate the effectiveness of the proposed model, this paper selects the following three mainstream zero-shot image classification datasets: AwA2 (Animals with Attributes 2). Number of categories: 50 (40 seen / 10 unseen), number of images: 37, 322, attributes: 85-dimensional manually annotated attributes; CUB (Caltech-UCSD Birds-200-2011). Number of categories: 200 (150 seen / 50 unseen), number of images: 11,788, attributes: 312-dimensional fine-grained attributes; SUN, number of categories: 717 (645 seen / 72 unseen), number of images: 14,340, attributes: 102-dimensional scene attributes.

The data split follows the standard PS (proposed split) scheme, and experiments are conducted under the generalized zero-shot learning (GZSL) setting.

4.1.2 Evaluation Metrics

The following indicators are used to evaluate model performance: Conventional ZSL Accuracy: accuracy tested only on unseen classes; GZSL Seen Accuracy: accuracy on seen classes under generalized ZSL; GZSL Unseen Accuracy: accuracy on unseen classes under generalized ZSL; H-Mean (harmonic mean):

$$H = \frac{2 \cdot ACC_{\text{seen}} \cdot ACC_{\text{unseen}}}{ACC_{\text{seen}} + ACC_{\text{unseen}}}$$

Measure the balance of the model between seen and unseen classes.

4.2 Error Analysis and Discussion

Misclassified Sample Analysis: Models confuse fine-grained categories due to high visual similarity and insufficient semantic differentiation. Class Imbalance: Models naturally favor seen classes in GZSL. This is mitigated by semantic consistency loss and feature re-weighting. Generalization Discussion: Fusing generative modeling with graph-augmented semantic spaces effectively improves the generalization of unseen classes.

5 Conclusion and Outlook

5.1 Research Conclusion

This study focuses on the construction and empirical analysis of image classification methods based on zero-shot learning, aiming to solve the problem of image recognition in the absence of target category samples, and proposes a zero-shot image classification model that integrates visual features, semantic representations, and cross-modal alignment. Through theoretical modeling and empirical analysis, the main research conclusions are as follows:

In terms of model construction: a zero-shot image classification framework is proposed, which integrates deep visual feature extraction, semantic embedding modeling, and cross-modal alignment mechanisms. The semantic space not only includes attribute information, but also introduces text embeddings and knowledge graph structures to achieve richer semantic representations. Contrastive learning and generation mechanisms are adopted to enhance the model's generalization ability on unseen categories. In terms of experimental verification: evaluations are conducted on three mainstr

In summary, the method proposed in this paper has significant advantages in improving the model's zero-shot recognition performance, enhancing semantic expression ability and cross-modal alignment ability, and possesses good scalability and application potential.

5.2 Limitation Analysis

Although this study has achieved certain results, it still has the following limitations. The model performance relies heavily on the quality of semantic information such as attributes, text and knowledge graphs, and deviations in semantic embedding will affect recognition accuracy. The introduction of a generation mechanism has increased the training cost of the model, making it difficult to deploy on platforms with limited computing resources. In addition, fine-grained category recognition still faces difficulties: on fine-grained datasets such as CUB, small semantic differences between different categories lead to model confusion, which requires further optimization of semantic modeling capabilities.

5.3 Future Directions

Although the current research has achieved certain results, there are still some deficiencies. The model performance is highly dependent on the quality of semantic information such as attributes, text and knowledge graphs, and deviations in semantic embedding will affect recognition accuracy. The generation module has high training complexity. The introduction of the generation mechanism increases the training cost of the model, making it difficult to deploy on platforms with limited computing resources. Fine-grained category recognition is still challenging. In fine-grained datasets like CUB, semantic differences between different categories are small, leading to model confusion, which requires further optimization of semantic modeling capabilities.

Based on the current research results and existing problems, future research can be further carried out in the following aspects: multi-modal semantic fusion, which constructs a more robust semantic space by combining multi-modal information such as image captions, knowledge graph contexts, and large language models; robust alignment mechanism design, which explores more robust alignment strategies such as graph matching and multi-hop reasoning to further reduce the problem of projection deviation; large model transfer and fine-tuning, which explores stronger zero-shot perception capabilities by leveraging large-scale cross-modal models such as CLIP, BLIP, and DINOv2; real-world application expansion, which applies the model to real-world scenarios such as medical image recognition, wildlife monitoring, and satellite image analysis to verify its practical value; and low-resource self-supervision mechanism, which introduces a self-supervised learning framework to reduce dependence on labeled data and external knowledge, and improve the generalization and scalability of the model.

References

- [1] ANWAAR M U, PAN Z, KLEINSTEUBER M. On leveraging variational graph embeddings for open world compositional zero-shot learning[J]. Proceedings of the 30th ACM International Conference on Multimedia, 2022, 2022(1): 4645-4654.
- [2] MISRA I, GUPTA A, HEBERT M. From red wine to red tomato: Composition with context[J]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, 2017(1): 1792-1801.
- [3] XU G Y, KORDJAMSHIDI P, CHAI J. Prompting large pre-trained vision-language models for compositional concept learning[J]. arXiv preprint arXiv:2211.05077, 2022.
- [4] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[J]. International Conference on Machine Learning, 2021, 2021(1): 8748-8763.
- [5] ATZMON Y, KREUK F, SHALIT U, et al. A causal view of compositional zero-shot recognition[J]. Advances in Neural Information Processing Systems, 2020, 33(1): 1462-1473.
- [6] BAO W T, CHEN L C, HUANG H, et al. Prompting language-informed distribution for compositional zero-shot learning[J]. Proceedings of the European Conference on Computer Vision, 2024, 2024(1).
- [7] CHEN C F R, FAN Q F, PANDA R. Crossvit: Cross-attention multi-scale vision transformer for image classification[J]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, 2021(1): 357-366.
- [8] CHEN S F, GE C J, TONG Z, et al. Adaptformer: Adapting vision transformers for scalable visual recognition[J]. Advances in Neural Information Processing Systems, 2022, 35(1): 16664-16678.
- [9] FU J L, ZHENG H L, MEI T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition[J]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, 2017(1): 4438-4446.
- [10] ISOLA P, LIM J J, ADELSON E H. Discovering states and transformations in image collections[J]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, 2015(1): 1383-1391.
- [11] HU X M, WANG Z L. A dynamic learning method towards realistic compositional zero-shot learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(3): 2265-2273.