Zero-Shot and Few-Shot Learning in Image Recognition Systems

1. Parveen Gorya, Assistant Professor,

Department of Computer Science,

Government College, Narnaul, District

Mahendergarh, Haryana.

2. Manoj Kumar, Assistant Professor,

Department of Computer Science,

Government College, Narnaul, District

Mahendergarh, Haryana.

Abstract

The field of image recognition has witnessed remarkable progress in recent years, largely fueled by deep learning models trained on massive labeled datasets. However, the reliance on extensive annotated data presents significant challenges in real-world scenarios where acquiring such datasets can be expensive, time-consuming, or even infeasible, especially for novel or rare object categories. To address these limitations, researchers have explored alternative learning paradigms: zero-shot learning (ZSL) and few-shot learning (FSL). These techniques aim to enable image recognition systems to generalize to unseen classes or learn from very limited examples, mimicking the remarkable ability of humans to recognize new concepts with minimal exposure. Zero-shot learning tackles the extreme scenario where the image recognition model is expected to classify images belonging to classes that were entirely absent during the training phase. Instead of relying on direct visual examples, ZSL leverages auxiliary information about these unseen classes, often in the form of semantic descriptions, attributes, or word embeddings. The core idea is to establish a relationship between the visual features learned from seen classes and the semantic representations of both seen and unseen classes.

Keywords:

Zero-Shot, Few-Shot, Learning, Image, Recognition

Introduction

Several approaches have been proposed for zero-shot learning (ZSL) in image recognition. Attribute-based methods rely on defining a set of descriptive attributes for each class. The model learns to predict these attributes from the visual input of seen classes and then uses the attribute descriptions of unseen classes to perform classification. Embedding-based methods aim to learn a joint embedding space where visual features and semantic representations are aligned. This allows for direct comparison between the visual embedding of an unseen image and the semantic embedding of its corresponding class. Generative methods take a different approach by generating synthetic visual features for unseen classes based on their semantic descriptions. These generated features can then be used to train a traditional classifier that can recognize both seen and unseen classes. (Walter, 2022)

During training, the model learns to associate visual features with their corresponding semantic descriptions for the seen classes. This creates a shared embedding space where both visual and semantic information can be represented. When presented with an image of an unseen class during inference, the model extracts its visual features and maps them to this shared space.

By comparing the location of the unseen image visual embedding with the semantic embeddings of all classes (including the unseen ones), the model can predict the class based on semantic similarity. For instance, a model trained to recognize dogs and cats using their visual features and textual descriptions ("has fur," "barks," "meows") could potentially recognize a wolf (an unseen class) if it knows the semantic description of a wolf ("has fur," "howls," "looks like a dog").

While ZSL holds immense potential for scenarios with emerging or rare categories, it faces significant challenges. The reliance on accurate and discriminative auxiliary information is crucial, and the performance can be heavily impacted by the quality of these descriptions. Furthermore, the "semantic gap" between high-level semantic descriptions and low-level visual features can be difficult to bridge effectively. Domain shift, where the visual characteristics of seen and unseen classes differ significantly, can also hinder the generalization capabilities of ZSL models. (Zhang, 2020)

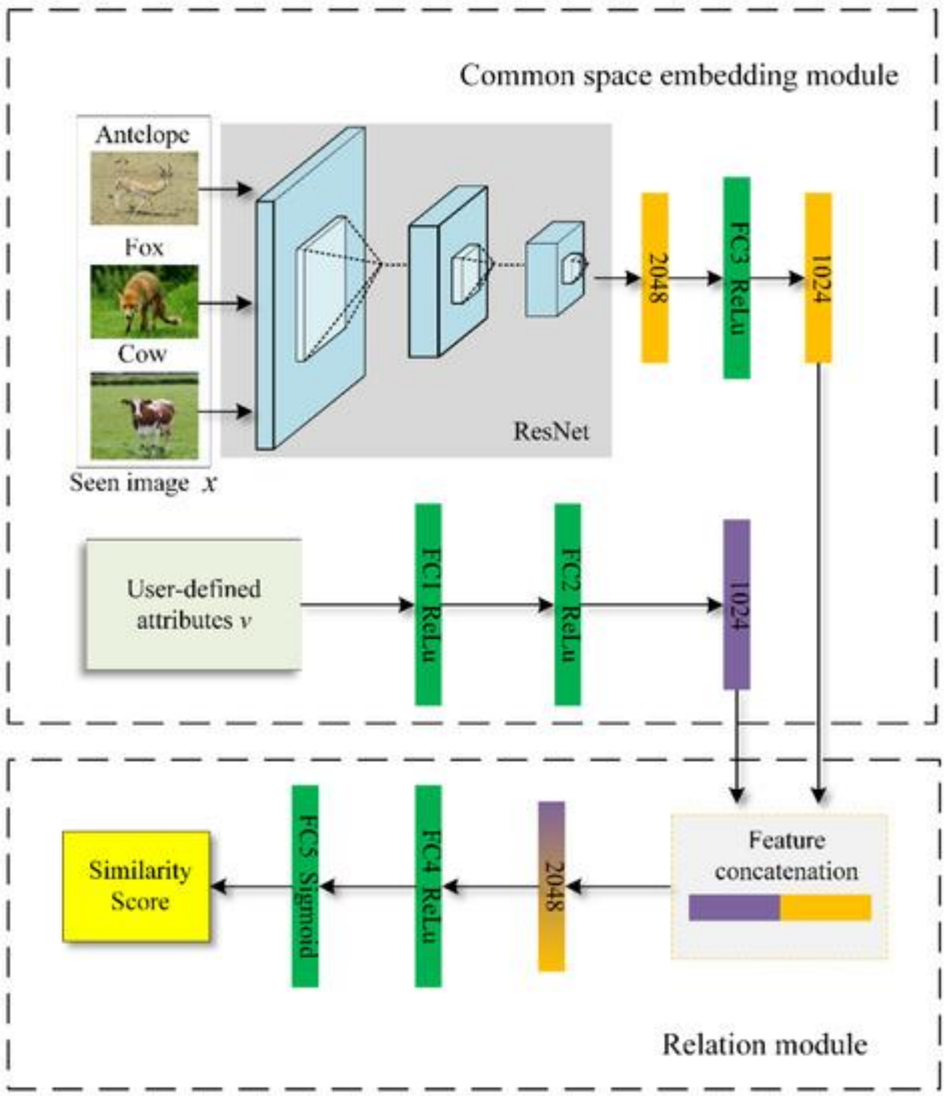

Figure 1. The framework

Few-shot learning addresses a less extreme but still challenging scenario where the image recognition model must learn to classify new classes based on only a handful of labeled examples, typically ranging from one to a few (one-shot learning being a specific case with only one example per class). This setting is more practical in many real-world applications where collecting a large dataset for every possible category is infeasible.

The core idea behind FSL is to leverage prior knowledge learned from a large number of related tasks or classes to quickly adapt to new tasks with limited data. This "learning to learn" paradigm, also known as meta-learning, is central to many FSL approaches.

One prominent family of FSL methods is metric-based learning. These approaches learn a distance metric or similarity function that can effectively compare image features. During training (meta-training), the model is presented with numerous "episodes," each consisting of a support set (a small number of labeled examples from a set of classes) and a query set (unlabeled examples from the same classes).

The model learns to classify the query images based on their similarity to the support set examples using the learned metric. At meta-testing time, the model can classify new classes with only a few support examples using the knowledge gained during meta-training. Examples of metric-based learning include Siamese Networks, Matching Networks, and Prototypical Networks.

Another important category is optimization-based meta-learning. These methods focus on learning an initialization or optimization strategy that allows a model to quickly adapt to new tasks with a few gradient updates. Model-Agnostic Meta-Learning (MAML) is a representative example, where the model learns an initial set of parameters that can be rapidly fine-tuned for new tasks with minimal data. (Xiang, 2022)

Literature Review

Ren et al. (2022): Model-based meta-learning approaches, on the other hand, utilize architectures with specific inductive biases that facilitate fast learning. For example, memory-augmented neural networks can store and retrieve information from the support set to aid classification of query images.

Huang et al. (2021): Few-Shot Learning **(**FSL) offers a more practical approach than traditional supervised learning in data-scarce scenarios. By leveraging meta-learning techniques, these models can achieve impressive performance with very few examples. However, FSL models can still be sensitive to the quality and representativeness of the few available examples. Furthermore, the performance might degrade if the new tasks or classes are significantly different from those encountered during meta-training.

Deng et al. (2020): In the realm of machine learning, particularly in computer vision and natural language processing, a significant challenge arises when we encounter categories or objects that our models have never been trained on. Traditional supervised learning paradigms necessitate vast amounts of labeled data for each class, rendering them inadequate for novel or rare categories.

Neumann et al. (2021): Zero-Shot Learning (ZSL) emerges as a powerful paradigm, aiming to recognize and classify unseen instances by leveraging prior knowledge about the relationships between seen and unseen classes. Among the various ZSL approaches, attribute-based methods stand out for their intuitive interpretability and their ability to bridge the "semantic gap" between visual features and class labels.

Zero-Shot and Few-Shot Learning in Image Recognition Systems

The fundamental idea behind attribute-based Zero-Shot Learning (ZSL) is to describe both seen and unseen classes using a set of semantic attributes. These attributes are typically human-interpretable properties that characterize the objects or concepts belonging to each class. For example, when classifying animals, attributes could include "has feathers," "has four legs," "is carnivorous," or "lives in water." The key insight is that while a model might not have seen an image of a "penguin" during training, if it has learned to associate visual features with the attributes "has feathers," "has wings," "swims," and "lives in cold climates" from seeing other bird species, it can potentially recognize a penguin based on its attribute profile.

The process in attribute-based ZSL typically involves two main stages. First, a model is trained on seen classes to learn a mapping between visual features extracted from images (e.g., using Convolutional Neural Networks - CNNs) and their corresponding attribute vectors. This mapping essentially learns to predict the presence or absence of specific attributes given a visual input. Second, for unseen classes, only their attribute descriptions are available. To classify a new, unseen instance, its visual features are extracted, and the learned mapping is used to predict its attribute vector. This predicted attribute vector is then compared to the known attribute vectors of the unseen classes, and the instance is assigned to the class whose attribute description is most similar.

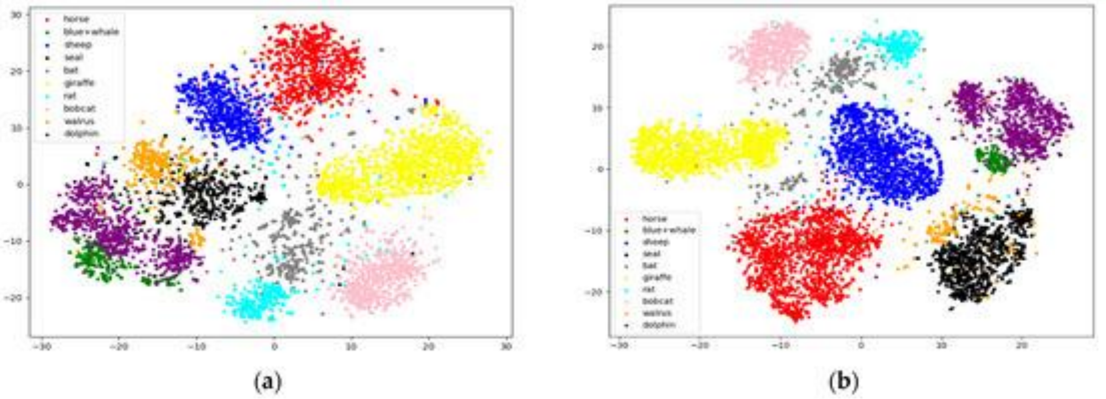

Figure 2. Visualization of the distribution of images

The strength of attribute-based methods lies in their ability to leverage human knowledge and provide a semantic grounding for the classification process. The attributes act as an intermediate layer, disentangling the visual representation from the specific class labels. This allows for knowledge transfer from seen to unseen classes based on shared semantic properties. Furthermore, the use of human-interpretable attributes enhances the explainability of the model's predictions, as we can understand why a particular instance was classified into a specific unseen class based on the presence or absence of certain descriptive features.

However, attribute-based ZSL methods also face several challenges. One significant limitation is the reliance on the quality and completeness of the attribute descriptions. If the chosen attributes are not sufficiently discriminative or if some crucial attributes are missing, the performance on unseen classes can be significantly hampered. Furthermore, defining a comprehensive and universally agreed-upon set of attributes for a wide range of object categories can be a laborious and subjective task.

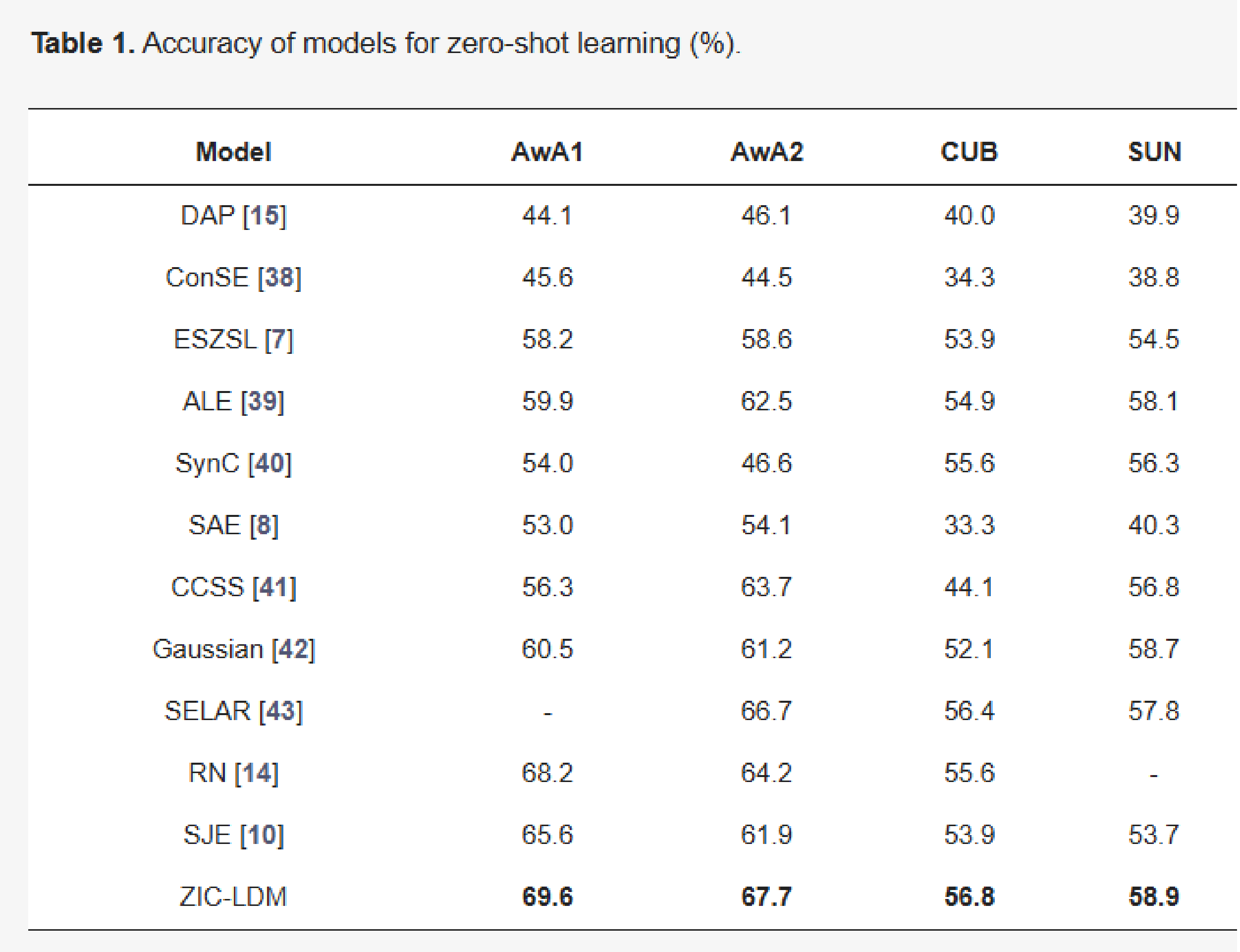

Table 1. Accuracy of models for zero-shot learning (%).

Model	AwA1	AwA2	CUB	SUN
DAP [15]	44.1	46.1	40.0	39.9
ConSE [38]	45.6	44.5	34.3	38.8
ESZSL [7]	58.2	58.6	53.9	54.5
ALE [39]	59.9	62.5	54.9	58.1
SynC [40]	54.0	46.6	55.6	56.3
SAE [8]	53.0	54.1	33.3	40.3
CCSS [41]	56.3	63.7	44.1	56.8
Gaussian [42]	60.5	61.2	52.1	58.7
SELAR [43]	-	66.7	56.4	57.8
RN [14]	68.2	64.2	55.6	-
SJE [10]	65.6	61.9	53.9	53.7
ZIC-LDM	**69.6**	**67.7**	**56.8**	**58.9**

Another challenge lies in the "semantic gap" itself. While attributes aim to bridge this gap, the relationship between visual features and semantic attributes might not always be straightforward or linear. Visual features can be complex and high-dimensional, and mapping them accurately to a lower-dimensional attribute space can be challenging. Additionally, the visual appearance of objects within the same attribute category can vary significantly, making it difficult to learn robust attribute predictors.

Recent research has explored various techniques to address these limitations. One direction involves learning more effective mappings between visual and attribute spaces, often using sophisticated deep learning architectures and loss functions. Another approach focuses on automatically discovering or refining attributes from data, reducing the reliance on manual annotation. Generative models have also been employed to synthesize visual features for unseen classes based on their attribute descriptions, which can then be used to train traditional classifiers. Furthermore, exploring richer forms of semantic knowledge beyond simple attribute vectors, such as word embeddings for knowledge graphs, is an active area of research.

In conclusion, zero-shot attribute-based methods offer a compelling approach to tackling the challenge of recognizing unseen objects. By leveraging human-interpretable semantic descriptions, they enable knowledge transfer from seen to unseen classes, providing a degree of interpretability that is often lacking in other ZSL techniques. While challenges related to attribute quality, the semantic gap, and the complexity of visual-semantic relationships remain, ongoing research continues to refine and enhance these methods, paving the way for more robust and generalizable intelligent systems capable of understanding and categorizing the ever-expanding world around us. As the need to handle novel and rare categories grows across various applications, attribute-based zero-shot learning will undoubtedly continue to be a vital area of investigation in the pursuit of more flexible and adaptable machine learning models.

In the realm of machine learning, the quest for models that can learn effectively from limited data has gained significant momentum. This is largely driven by real-world scenarios where acquiring large, labeled datasets is often expensive, time-consuming, or even impossible. Few-Shot Learning (FSL) emerges as a powerful paradigm to address this challenge, aiming to train models that can generalize to new tasks or classes with only a handful of examples. Among the diverse approaches within FSL, metric-based learning stands out as an intuitive and effective strategy.

At its core, metric-based learning in FSL focuses on learning a similarity metric or a distance function between data points. The fundamental idea is that if a model can accurately measure how similar or dissimilar two examples are, it can then classify a new, unseen example by comparing it to the few labeled examples it has for each class. The class of the labeled example that is most similar to the new example is then predicted as the label for the new example. This approach elegantly bypasses the need to learn complex decision boundaries directly from scarce data, instead leveraging the power of similarity comparisons in a learned embedding space.

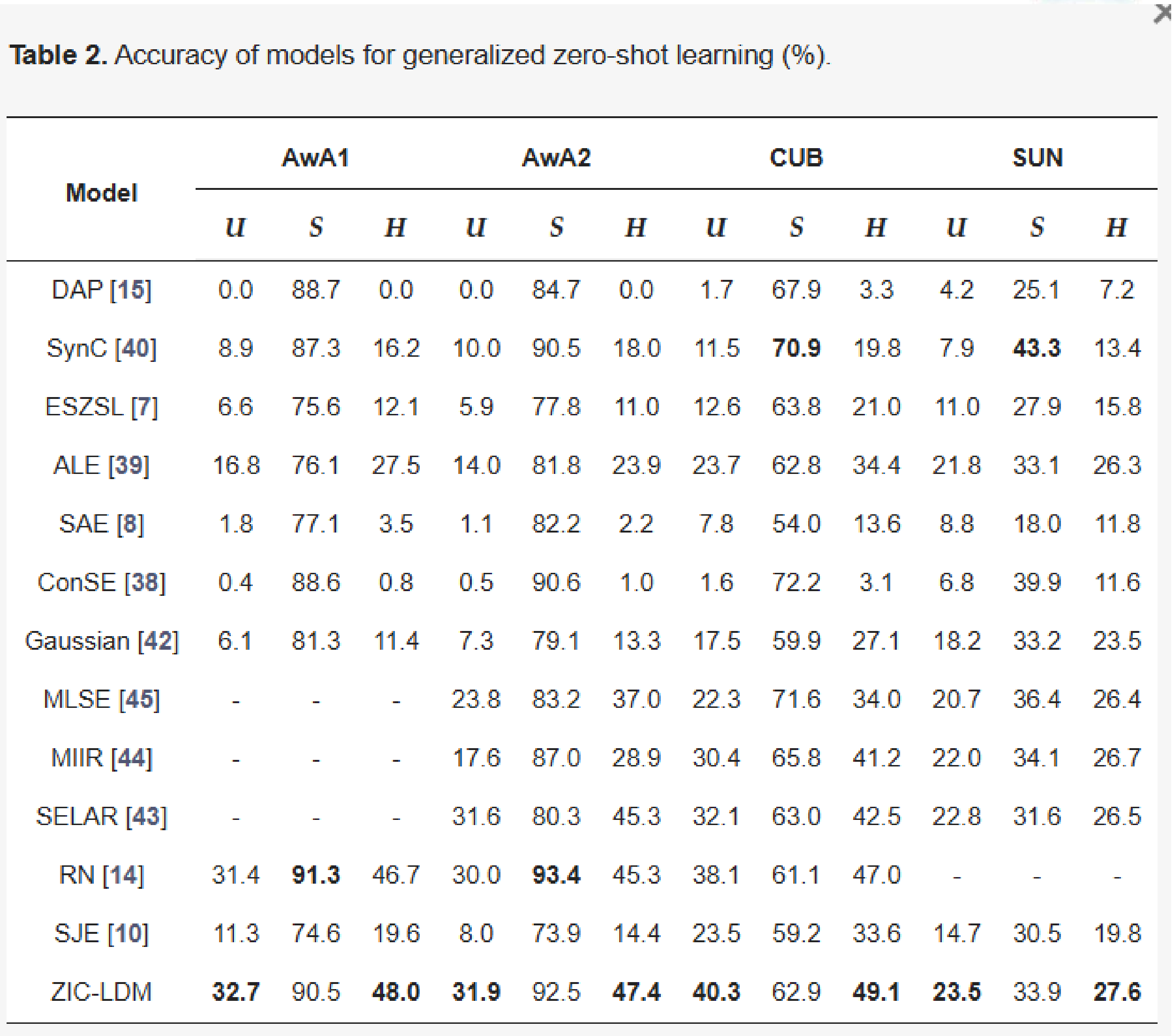

Table 2. Accuracy of models for generalized zero-shot learning (%).

Model	AwA1			AwA2			CUB			SUN		
	u	*S*	*H*	*u*	*S*	*H*	*u*	*S*	*H*	*u*	*S*	*H*
DAP [15]	0.0	88.7	0.0	0.0	84.7	0.0	1.7	67.9	3.3	4.2	25.1	7.2
SynC [40]	8.9	87.3	16.2	10.0	90.5	18.0	11.5	**70.9**	19.8	7.9	**43.3**	13.4
ESZSL [7]	6.6	75.6	12.1	5.9	77.8	11.0	12.6	63.8	21.0	11.0	27.9	15.8
ALE [39]	16.8	76.1	27.5	14.0	81.8	23.9	23.7	62.8	34.4	21.8	33.1	26.3
SAE [8]	1.8	77.1	3.5	1.1	82.2	2.2	7.8	54.0	13.6	8.8	18.0	11.8
ConSE [38]	0.4	88.6	0.8	0.5	90.6	1.0	1.6	72.2	3.1	6.8	39.9	11.6
Gaussian [42]	6.1	81.3	11.4	7.3	79.1	13.3	17.5	59.9	27.1	18.2	33.2	23.5
MLSE [45]	-	-	-	23.8	83.2	37.0	22.3	71.6	34.0	20.7	36.4	26.4
MIIR [44]	-	-	-	17.6	87.0	28.9	30.4	65.8	41.2	22.0	34.1	26.7
SELAR [43]	-	-	-	31.6	80.3	45.3	32.1	63.0	42.5	22.8	31.6	26.5
RN [14]	31.4	**91.3**	46.7	30.0	**93.4**	45.3	38.1	61.1	47.0	-	-	-
SJE [10]	11.3	74.6	19.6	8.0	73.9	14.4	23.5	59.2	33.6	14.7	30.5	19.8
ZIC-LDM	**32.7**	90.5	**48.0**	**31.9**	92.5	**47.4**	**40.3**	62.9	**49.1**	**23.5**	33.9	**27.6**

The process typically involves training a neural network to embed the input data (e.g., images, text) into a lower-dimensional feature space. The key objective during training is to learn an embedding such that examples from the same class are mapped to nearby points in this space, while examples from different classes are pushed far apart. This is often achieved through episodic training, where the model is presented with multiple "few-shot" tasks during training. Each task consists of a support set (the few labeled examples) and a query set (unlabeled examples to be classified). The model learns to classify the query examples based on their similarity to the support set examples.

Siamese networks consist of two identical neural networks that process two input examples separately and then compare their resulting embeddings using a distance metric (e.g., Euclidean distance, cosine similarity). The network is trained to minimize the distance between embeddings of same-class pairs and maximize the distance between embeddings of different-class pairs.

Triplet Networks take three inputs: an anchor example, a positive example (from the same class as the anchor), and a negative example (from a different class). The network learns embeddings such that the distance between the anchor and the positive example is smaller than the distance between the anchor and the negative example by a certain margin.

Matching networks directly learns a similarity function between a query example and each example in the support set. It uses attention mechanisms to weigh the contributions of different support examples when classifying the query. The final classification is a weighted sum of the labels of the support set, where the weights are determined by the learned similarity scores. Prototypical networks compute a prototype for each class in the support set by taking the mean of the embeddings of the examples belonging to that class. A query example is then classified by finding the prototype that is closest to its embedding in the learned feature space.

Relation Networks learn a relation module that takes the embeddings of a query example and a support example as input and predicts a relevance score indicating their similarity. This allows for learning more complex, non-linear relationships between the embeddings.

The strengths of metric-based learning in few-shot scenarios are manifold. Firstly, by focusing on learning a generalizable similarity metric, these methods can effectively handle new classes not seen during training. The learned embedding space captures semantic similarities that can be applied across different categories. Secondly, the episodic training paradigm closely mimics the few-shot learning scenario encountered during evaluation, leading to better generalization. Finally, these approaches are often conceptually simpler and easier to implement compared to other FSL techniques like meta-learning algorithms that explicitly learn an optimization process.

However, metric-based learning also has its limitations. The performance heavily relies on the quality of the learned embedding space and the chosen distance metric. Designing an effective embedding architecture and loss function that captures the essential features for distinguishing between a wide range of classes can be challenging. Furthermore, these methods might struggle when the novel classes have significantly different characteristics or lie in a different data distribution compared to the classes seen during training.

Despite these challenges, metric-based learning remains a cornerstone of few-shot learning research and has demonstrated remarkable success in various applications, including image classification, object recognition, and natural language processing with limited data. Ongoing research continues to explore novel embedding architectures, more sophisticated similarity metrics, and strategies to improve the robustness and generalizability of these approaches. As the demand for machine learning models that can learn efficiently from scarce data grows, metric-based learning will undoubtedly continue to play a crucial role in pushing the boundaries of artificial intelligence.

Conclusion

Zero-shot and few-shot learning represent significant steps towards building more adaptable and data-efficient image recognition systems. By moving beyond the traditional paradigm of training on large, labeled datasets, these techniques open up possibilities for recognizing novel objects and learning from limited supervision, making image recognition applicable to a wider range of real-world scenarios. While both ZSL and FSL face their own set of challenges, ongoing research continues to refine these approaches, explore new methodologies, and integrate them with other learning paradigms. As the demand for intelligent systems capable of understanding the visual world in diverse and dynamic environments grows, zero-shot and few-shot learning will undoubtedly play an increasingly crucial role in the future of image recognition.

References

1. Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2022; pp. 770–778.
2. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 22–25 July 2021; pp. 4700–4708.

3. Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2020; pp. 4690–4699.

4. Liu, C.; Zoph, B.; Neumann, M.; Shlens, J.; Hua, W.; Li, L.; Fei, L.; Yuille, A.; Huang, J.; Murphy, K. Progressive neural architecture search. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2021; pp. 19–34.

5. Frome, A.; Corrado, G.S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.; Mikolov, T. Devise: A deep visual-semantic embedding model. In Proceedings of the Conference and Workshop on Neural Information Processing Systems (NIPS), Harrahs and Harveys, Lake Tahoe, NV, USA, 7 December 2022; pp. 2121–2129.

6. Zhang, Z.; Saligrama, V. Zero-shot learning via semantic similarity embedding. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 14–17 December 2020; pp. 4166–4174.

7. Romera-Paredes, B. An embarrassingly simple approach to zero-shot learning. In Proceedings of the 32nd International Conference on Machine Learning (ICML), Lille, France, 6–11 July 2021; pp. 2152–2161.

8. Kodirov, E.; Xiang, T.; Gong, S. Semantic autoencoder for zero-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 22–25 July 2022; pp. 3174–3183.

9. Zhang, L.; Xiang, T.; Gong, S. Learning a deep embedding model for zero-shot learning. Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Honolulu, HI, USA, 22–25 July 2020; pp. 2021–2030.

10. Akata, Z.; Reed, S.; Walter, D.; Lee, H.; Schiele, B. Evaluation of output embeddings for fine-grained image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2022; pp. 2927–2936