Review of Computer Engineering Research
2023 Vol. 10, No. 2, pp. 55-69
ISSN(e): 2410-9142
ISSN(p): 2412-4281
DOI: 10.18488/76.v10i2.3472
© 2023 Conscientia Beam. All Rights Reserved.

A review of few-shot image recognition using semantic information

Liyong Guo[1]
Erzam Marlisah[2+]
Hamidah Ibrahim[3]
Noridayu Manshor[4]

[1,2,3,4]*Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Selangor, Malaysia.*
[1]*Email: guoliyong685@gmail.com*
[2]*Email: erzam@upm.edu.my*
[3]*Email: hamidah.ibrahim@upm.edu.my*
[4]*Email: ayu@upm.edu.my*

(+ Corresponding author)

Article History
Received: 26 June 2023
Revised: 21 August 2023
Accepted: 31 August 2023
Published: 15 September 2023

Keywords
Deep learning
Few-shot learning
FGFS
GAN
Image recognition
Semantic information
Zero-shot learning.

ABSTRACT

In recent years, the utilization of deep learning techniques has been employed in the field of image recognition with the aim of improving performance. However, deep learning demands a substantial amount of labeled data for model training, a process that is both expensive and time-consuming. In order to tackle this particular difficulty, the approach of few-shot learning (FSL) has emerged as a viable alternative. FSL, or Few-Shot Learning, is a computational approach that aims to replicate the cognitive processes observed in humans. By using a small set of examples and experiences, FSL enables the acquisition of new concepts. Research in the field of FSL has investigated many approaches to extracting the highest amount of information from limited data or making use of affordable and easily accessible sources of information. Researchers have been incorporating outside data into FSL techniques more frequently. This paper conducts an in-depth exploration of leveraging semantic information to enhance few-shot learning. By reviewing papers from the last five years in WOS, IEEE, and Science Direct (some papers in arXiv are also used), this study delves into the strategies employed to bridge the gap between visual and semantic information. The review extends to encompass zero-shot learning, which is considered a subcategory of FSL, enriching the analysis. Moreover, this paper identifies the potential of employing semantic information to enhance fine-grained few-shot (FGFS) learning. Techniques such as direct projection and the application of generative adversarial networks (GANs) emerge as promising avenues to accomplish this enhancement.

Contribution/Originality: This work aims to address the existing research gap by conducting a comprehensive assessment of several methodologies that leverage semantic information in order to improve the performance of few-shot learning. In contrast to conventional few-shot learning approaches, semantic-based methodologies priorities the establishment of connections between semantic information and visual representations.

1. INTRODUCTION

The field of computer vision has experienced significant advancements as a result of the use of deep learning techniques, leading to enhanced precision and accuracy in the obtained outcomes [1]. As a part of computer vision, image recognition (IR) can apply recent technology to obtain robust feature representations and increase accuracy [2]. However, deep learning methods require large numbers of labelled examples to train the Convolutional Neural Network (CNN) model. Without sufficient data, it is difficult to train a good classifier and implement accurate IR. Obtaining a large amount of labelled data is not feasible in many cases because doing so is labor-intensive and costly.

© 2023 Conscientia Beam. All Rights Reserved.

For example, it is a challenge to collect enough medical images under time constraints. Therefore, the few-shot learning (FSL) approach was introduced to effectively use a limited number of images to obtain sufficient information and achieve high performance in recognition tasks. There are three methods for FSL: metric learning, meta-learning, and generative (or augmentation)-based methods [3].

This research concentrates its attention on a distinct kind of FSL known as semantic-based FSL. It is a specific augmentation-based technique that uses zero-shot learning (ZSL), which necessitates the classification of new and undiscovered classes. There are no images available for the unseen classes during training; semantic information is the only information that can be used to encode semantic relationships between seen and unseen classes [4]. For ZSL, semantic information is essential. For few-shot learning, this semantic information can be used to improve recognition performance.

The key challenge is to determine how images can be recognized based on a few examples and semantic information. It is difficult to build authentic relationships between images and semantic information using a limited number of samples. In recent years, there have been tremendous developments in research on FSL, especially using semantic information. Our main objective is to summarize researchers' efforts in semantic-based few-shot image recognition, develop a coherent taxonomy to overcome the challenges, understand the characteristics of the new research trends in recent years, and propose some directions for future study.

The present paper is structured in the following manner: Section 2 of this research presents a comprehensive examination of Few-Shot Learning (FSL) with a specific emphasis on semantic information. This section encompasses the definitions and issues associated with FSL. Moving forward, Section 3 conducts a thorough literature analysis that specifically investigates the methodologies employed in FSL that leverage semantic information. Finally, Section 4 serves as the concluding section of this paper.

2. FEW-SHOT LEARNING AND SEMANTIC INFORMATION

FSL, which stands for Few-Shot Learning, falls under the category of machine learning techniques. Generally, machine learning involves a computer program that learns from prior experiences (E) within a specific set of tasks (T), guided by a performance measure (P). The objective is to enhance performance in task T through learning from experience E, as quantified by measure P [5]. FSL is a distinct class of machine learning problems, sharing common elements of experience E, task T, and performance measure P. However, the differentiating factor between FSL and standard machine learning is experience E, which integrates supervised information tailored to the target task T [6]. FSL relies on limited, supervised information. FSL encompasses sub-categories based on the volume of training data employed. For instance, when N*K samples are utilized for training (N classes, each with K samples), this is termed N-way-K-shot learning. Alternatively, when only a single sample per class is available, it's known as one-shot learning (OSL). In scenarios where at least one sample is unavailable for specific classes, it's referred to as zero-shot learning (ZSL).

The primary limitation lies in traditional deep learning's difficulty in rapidly categorizing with a restricted sample pool [2]. This challenge arises due to the time-intensive process of fine-tuning model parameters for improved performance. Multiple approaches have been proposed to tackle this issue, including metric learning, meta-learning, augmenting sample generation, and utilizing alternative data types [3]. In metric learning, a nonlinear embedding is projected onto a metric space, facilitating the measurement of data similarity or distance. Consequently, image points from the same class cluster closely, while those from different classes remain distant in this metric space, enhancing the manageability of FSL. Meta-learning, or "learning to learn," draws from a diverse array of learning tasks to accelerate the model's capacity to learn effectively, so the algorithm can adapt effectively to new tasks. These tasks are regarded as experiential in nature, and the acquisition of new tasks is expedited. Meta-learning involves the acquisition of knowledge and skills across multiple learning challenges, whereas traditional machine learning

© 2023 Conscientia Beam. All Rights Reserved.

methods often rely on the analysis and modelling of a single task. When faced with a lack of suitable data, the use of external data can be advantageous for FSL acquisition.

Semantic information is a type of external data that is the key to ZSL [7]. It is also useful for other types of FSL [3, 8]. According to Wang, et al. [9], there are four types of semantic information: attributes, lexical items, labels, and text. Attribute and text are the most widely used types. One essential problem emerged with the use of semantic information: how to bridge semantic information and visual images. The most popular method is to learn the embedding or mapping function between visual features and semantic vectors [7]. As a result, there are two types of information (visual and semantic) that can be used to improve learning performance. This paper focuses on how semantic information can be used for FSL.

3. LITERATURE REVIEW

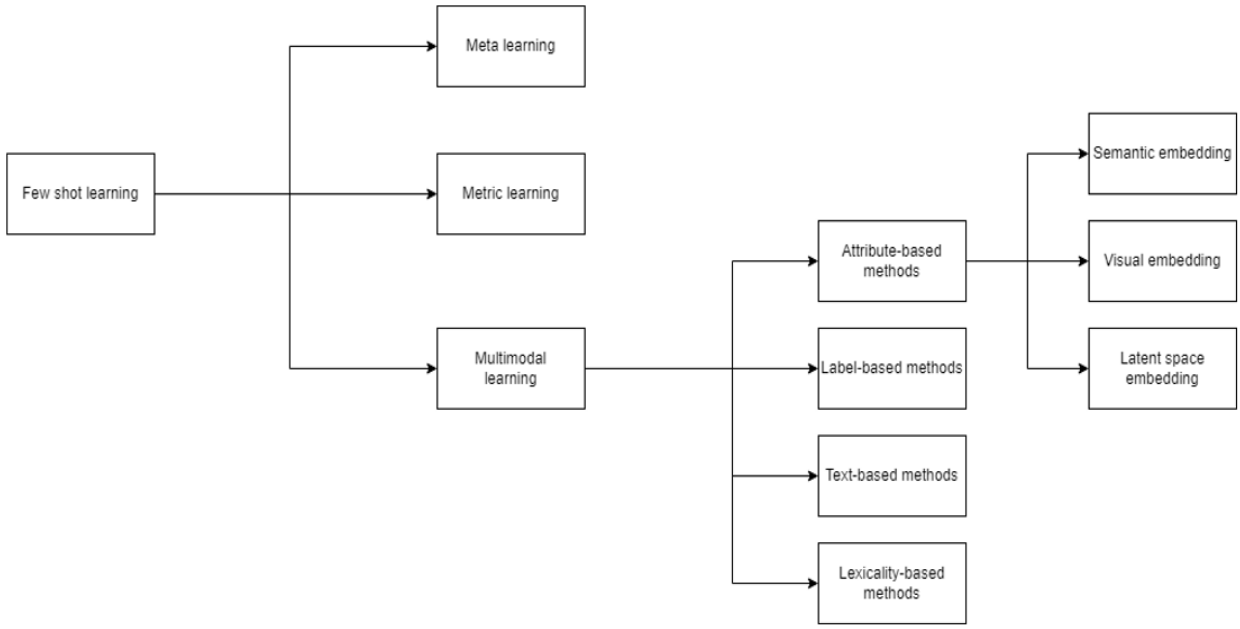


Figure 1. Taxonomy of research literature on few shot learning.

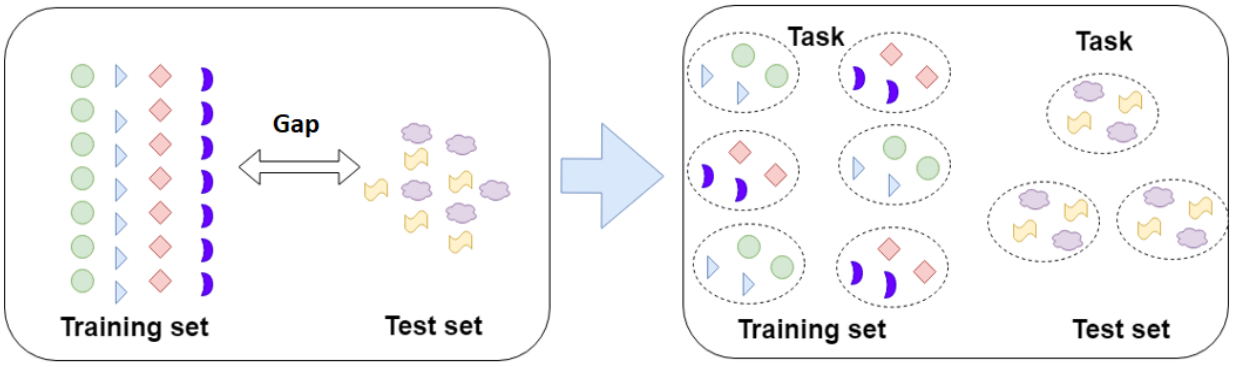
FSL was proposed so machines could mimic the human ability to learn new concepts from a few examples. There was no specific definition of FSL until 2020, when Wang, et al. [6] provided the most widely recognized definition. The most challenging problem for FSL is data sparsity. There are two ways to address this problem. On the one hand, there are ways to use limited data more effectively and extract helpful information. On the other hand, external data can be used to improve FSL performance.

The terms single-modal learning and multimodal learning [10] are used to distinguish these two approaches. In single-modal learning, limited data are used to solve few-shot learning problems with methods such as data augmentation, metric learning, and meta-learning. In contrast, multimodal learning relies on other types of data. For example, image classification using multimodal learning can be based not only on image data but also on text data, attribute data, or some website data.

There are three subtopics in the literature review: meta-learning, metric learning, and multimodal learning. Figure 1 illustrates the taxonomy of few-shot learning, especially for multimodal learning. This includes attribute-based methods, label-based methods, text-based methods, and lexically-based methods.

© 2023 Conscientia Beam. All Rights Reserved.

3.1. Meta-Learning



a) Traditional deep learning b) Meta learning

Figure 2. The difference between traditional deep learning and meta-learning.

Meta-learning is a learning strategy that enables learners to learn new tasks faster. A wide range of learning tasks are used, and the trained model can quickly adapt to new tasks. Figure 2 illustrates the distinction between traditional deep learning and meta-learning: While traditional deep learning views a dataset as a single task, meta-learning treats it as multiple tasks. The increased number of tasks enables meta-learning to rapidly adapt and learn. Suppose there are two datasets: D_{train} and D_{val} [11]. Meta-learning uses D_{train} to train and obtain prior knowledge and then uses D_{val} to modify the meta-learner. For task $T^{(k)}$, the labelled supporting image is set as $S^{(k)}$, and the query image is set by $Q^{(k)}$. $S^{(k)}$ and $Q^{(k)}$ are from the same set of classes, $C^{(k)}$, which is a subset of C_{train}, C_{val} *or* $C_{train} UC_{val}$. Supporting data is used for learning, and a query set is used for testing for each task.

A simple meta-learning model is model-agnostic meta-learning (MAML) [12]. MAML is used to find the model parameters with good generalization for new tasks. However, random initialization of the model parameter leads to local optima problems during the gradient descent and affects fast learning. A batch has several tasks in MAML training, each with a different gradient descent direction. The model uses the batch's gradient of all task gradients (second-order derivatives) to update the model parameters. The Reptile model [13] is an improved form of MAML. It can ignore second-order derivatives and does not need to update as many parameters as MAML. There are other methods based on MAML, such as MAML++ [14], Imaml [15], and iTAML [16], which were introduced to address some of the drawbacks of MAML: gradient instability, high computational overhead, low generalization performance, and others. Another important method is LSTM (Long Short-Term Memory)-based meta-learning. The process gradient-based update is similar to the cell-state update in LSTM [17]; therefore, the state update formula of LSTM can be used to update the parameters of the meta-learner. Based on the characteristics of LSTM, this meta-learner can capture short-term (for each task) and long-term knowledge (for all tasks), which helps solve the problem of gradient-based optimization failure in FSL. Memory-augmented neural networks (MANN) [18] can quickly learn new information by storing already-seen class information in external memory. This lets them make accurate predictions in FSL.

3.2. Metric Learning

Metric learning involves learning the similarity or distance between labelled and unlabeled samples. The recognition task can be completed by comparing the distance to labelled samples. This method can effectively avoid the problem of over fitting because it does not require additional parameters for new classes [19]. Different distance metrics can be used: Manhattan distance, Euclidean distance, cosine similarity, and others. Siamese neural networks [20] were the first to apply metric learning to one-shot recognition. In this method, there are twin networks that have the same structures and share parameters. Two images are input into the twin networks, and image features or representations are extracted. Then, the L_1 distance, or Manhattan distance, between the twin features is calculated

© 2023 Conscientia Beam. All Rights Reserved.

to rank their similarity. The Siamese neural network is applicable for verification tasks, but it is not suitable for recognition tasks because it only computes the similarities between the twin images. A matching network [21] can be used to address this problem. A label distribution is outputted by calculating the cosine similarities between unlabeled and labelled images. The matching network computes the distance between the support and query samples. However, the number of support samples is limited, so they cannot adequately represent the novel class [19]. Prototypical networks [22] were proposed to address this issue. It is based on the fundamental assumption that all the image representations from the same classes cluster around a prototype representation. The prototype representation can be calculated as the mean of the support example representations for each class [22]. Prototypical networks use Euclidean distance to categorize prototypes and characterize their commonalities. The prototype representation is superior because it is derived from the average of all picture representations for each class.

3.3. Multimodal Learning

In multimodal learning, multiple signals pertaining to the data are used for model building. These modes include pictures, illustrations, audio, speech, writing, print, etc. However, semantic information (attributes, labels, text, and lexicality) is more useful for FSL. According to different types of data, this multimodal learning for FSL can be categorized into four methods: attribute-based, label-based, text-based, and lexicality-based.

3.3.1. Attribute-Based Methods

An attribute is a type of manually defined data that can describe the features of a class, such as shape and color. The same attribute can be shared among different classes. For example, a bird or horse can have the same attribute: black in color. This characteristic makes it easy to bridge between seen and unseen classes. Because of this, in many recent studies, attributes have been used for FSL, especially for ZSL. Images and attributes are different types of data. Bridging the gap between the two data modes is the most critical issue for this method. Then, a mapping function from the visual space to the attribute space is found to transform visual features into semantic features, a process referred to as semantic embedding. Another approach is visual embedding, which involves a mapping function from the attribute space to the visual space. Latent space embedding refers to the mapping from the visual space or attribute space to a new space. Attribute space is a kind of semantic space that consists of many attributes.

3.3.1.1. Semantic Embedding

The semantic embedding approach projects image features onto the semantic space. It learns an embedding function that maps from the visual space to the semantic space. It uses visual features (x_i^s) and semantics of seen classes (a_i^s) to train the model and minimize the loss function. A Visual feature $x_i^s \in \mathbb{R}^D$ is a D-dimensional vector extracted by some related networks, such as VGG-19 (Visual Geometry Group) or ResNet. $a_i^s \in \mathbb{R}^K$ is a K-dimensional vector of semantic information. It can be obtained by Word2vec or Glove, which inputs the semantic information. A linear regression projection can be described as follows:

$$\min_W \sum_{i=1}^{N_s} ||W^T x_i^s - a_i^s||_2^2 + \lambda ||W||_F^2 \quad (1)$$

Here, λ represents the regularization parameter and W denotes the mapping function from the visual to semantic space. This process of learning entails the discovery of the projection function W that transforms x_i to a_i, aiming to minimize the Euclidean distance between them. The concept of Class Adapting Principal Direction (CAPD) [23] refers to a linear regression model responsible for projecting visual features onto the semantic space. Following this embedding, CAPD utilizes Mahalanobis-derived distances to quantify similarity and facilitate the classification process.

Numerous investigations have been conducted to enhance the efficacy of semantic embedding. One strategy involves refining the discriminative nature of the semantic embedding itself. For instance, a model known as Channel-

© 2023 Conscientia Beam. All Rights Reserved.

wise Mix-Fusion ZSL (CMFZ) [24] exploits the interplay between objects and their environments to emphasize the most distinctive channels. This emphasis on distinctiveness aids in cultivating robust and discriminative semantic embedding tailored for Zero-Shot Learning (ZSL) scenarios. Given that the global feature may lack the requisite discriminative power for achieving optimal performance in ZSL, an attention mechanism is employed to extract local region features. These features are subsequently projected onto the semantic space [25]. Furthermore, a method termed Prototype Adjustment [26] has been proposed to enhance the accuracy and discriminative quality of the semantic representation. This method effectively addresses the challenge of domain shift, which commonly arises in ZSL scenarios.

An alternative strategy involves the utilization of semantic embedding in conjunction with self-reconstruction techniques. Certain researchers have pursued the projection of visual features onto the semantic space, followed by the subsequent reconstruction of these visual features [27-30]. This approach of self-reconstruction serves to bolster the model's generalization capabilities and effectively mitigates the domain shift problem that can arise from such projection. Several studies have detailed methodologies for synthesis. These methodologies focus on the generation of novel samples, leveraging the generated samples in the recognition process. In one such model, a semantic embedding module is employed to extract semantic information from images. Subsequently, a Generative Adversarial Network (GAN) is harnessed to craft an image endowed with specific attributes. This interplay forms a bidirectional mapping that ensures the alignment of the generated images with the intended semantic space [28]. In a distinct study, semantic information is combined with a conditional GAN to produce images that serve to ameliorate challenges related to data scarcity and imbalance [29].

3.3.1.2. Visual Embedding

Visual embedding projects semantic information into the visual space. Visual features have higher dimensions than semantic features, so they are more distinctive. Its ridge regression model can be described as follows:

$$\min_{w} \sum_{i=1}^{N_s} ||x_i^s - W a_i^s||_2^2 + \lambda ||W||_F^2 \qquad (2)$$

Equation 2 illustrates the process of projecting semantic features onto visual features. Here, W represents the projection function from the semantic space to the visual space. x_i Corresponds to the visual features, while a_i corresponds to the semantic features. The symbol λ denotes the regularization parameter. This represents the process of iteratively determining optimal parameters for W in order to minimize the discrepancy between x_i and a_i.

Nevertheless, the utilization of visual embedding gives rise to a hubness problem. This means that a small number of objects become neighbors of most objects in high-dimensional space. A reverse feature projection from semantic to visual space and a cosine distance loss function are used to address the hubness problem in ZSL [31]. Because the prototype is not likely to be affected by some novel or abnormal data, the visual prototype has a better generalization ability than the single visual feature. A prototype is leveraged in multimodal learning to obtain better recognition performance [32]. In this method, the semantic space is projected onto the visual space, and discriminative visual prototypes are calculated. To avoid information loss and alleviate the domain shift problem, attributes are projected onto the visual space and then reconstructed into semantic vectors [33].

3.3.1.3. Latent Space Embedding

Learning an explicit projection between visual features and semantic representations is difficult because they are from two different spaces and have distinct properties. Latent space embedding is used to find a common space in which there are some common properties across different modalities. This method can effectively overcome the domain shift problem of one-way mapping. Figure 3 illustrates how latent space embedding works: visual and semantic features are projected into a new space (latent space), which can be beneficial for the target task.

© 2023 Conscientia Beam. All Rights Reserved.

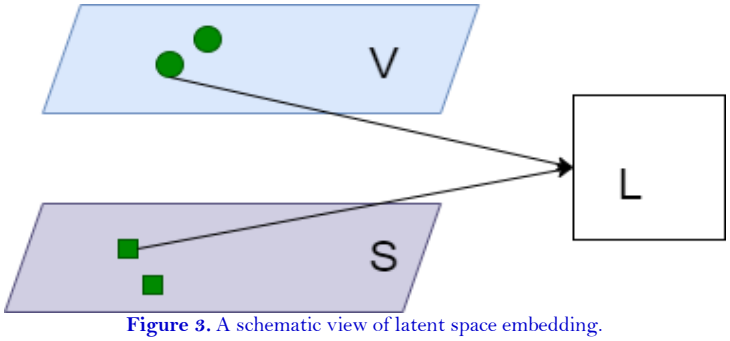


Figure 3. A schematic view of latent space embedding.

Match kernel embedding (MKE) [34] was proposed to bridge the gap between the visual space and the semantic space. Both types of information are projected onto the MKE space to infer similarities between seen and unseen classes. A zero-shot image classification method based on a learnable deep metric (ZIC-LDM) [35] was proposed to learn a common space so that image and semantic features can be mapped onto this space to help with the semantic gap problem. In several papers [36-41], auto-encoders were used to map visual and semantic information onto the latent space. Figure 4 illustrates the process by which an auto-encoder compresses data and reduces its dimensions by filtering out noise from the data.

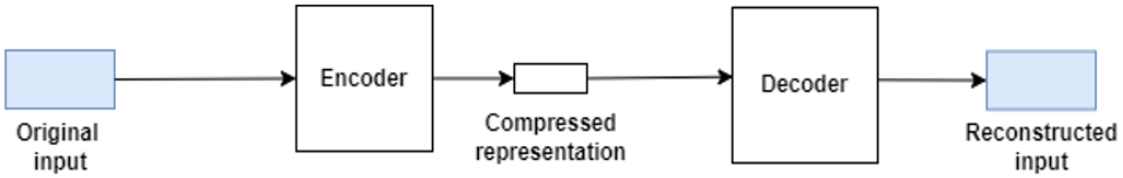


Figure 4. A schematic view of auto encoder.

The difference between direct mapping and using an auto encoder is that the auto encoder has a reconstructed process that can test the effectiveness of the compressed data. The goal of the discriminative dual semantic auto encoder (DDSA) [36] is to build an aligned space with two bidirectional mappings for the visual and semantic spaces. The features in the aligned space are semantic-preserving and discriminative. Figure 5 depicts the processes of DDSA.

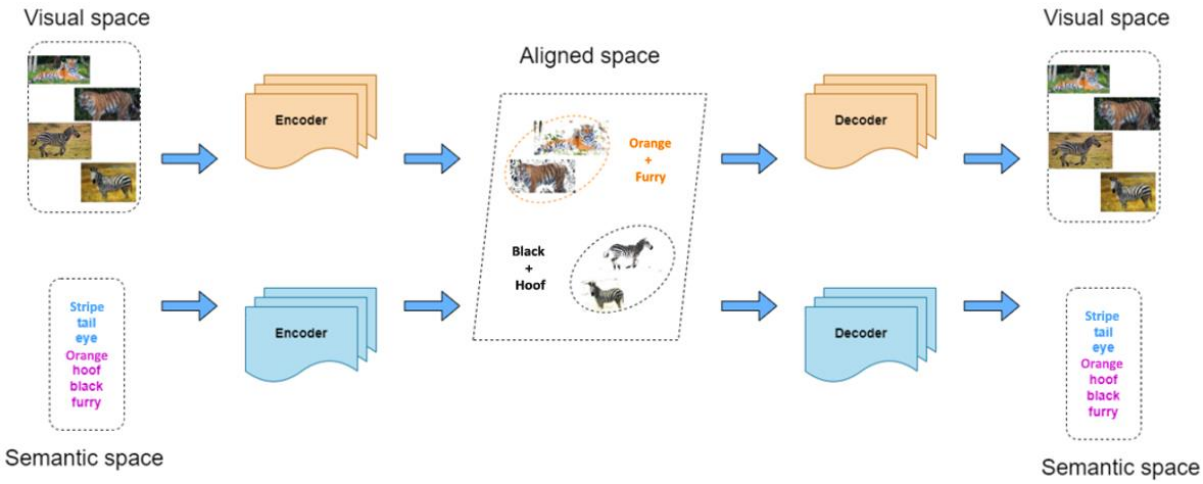


Figure 5. A schematic view of DDSA framework.

The latent or common space learned by the auto-encoder is more discriminative for ZSL [37, 41]. The compressed representations learned from attributes and images are easy to align and mitigate the domain shift problem [38, 41]. Residual attribute extractor generative adversarial networks (ResAttr-GAN) use the encoder-decoder generator and attributes to generate images and achieve better performance for face recognition [40]. We

© 2023 Conscientia Beam. All Rights Reserved.

can conclude that the auto encoder and latent space are always used to address domain shift problems and make the features more discriminative.

Some studies aim to find other latent information to improve the accuracy of ZSL [42]. The traditional attribute is not discriminative enough. As a result, combining the various attributes led to the proposal of a latent attribute [42].

Finally, there are several unique papers that describe how to use attributes in different ways. A self-attention mechanism is used to obtain more discriminative visual features. Then, the discriminative visual features are projected into semantic space and achieve superior performance. Visual features Image features are decomposed into several parts by leveraging attributes [43]. FSL, which uses these different parts, can achieve higher accuracy. Another paper used a salient object detector to obtain a salient map, and the object position helped obtain local features. Global and local image features are extracted and fused to form better visual features. Then, these features are projected into semantic space to predict their labels [44]. The domain of the last paper was structural damage identification [45], and damage attributes were treated as interclass knowledge. This knowledge is transferred from the seen class space to the unseen class space. This paper was also based on meta-learning, so its data were divided into task levels based on the sample levels.

3.3.2. Label-Based Methods

This approach involves finding similarities between class names or labels. The similarity is then used to bridge the relationship between seen and unseen classes. There are two general categories of methods for using labels, depending on how the labels are used. Many of them use graph convolution networks (GCNs) [46-50] to deal with the labels. Some use a tree or hierarchical structure [51, 52] to relate labels.

Graph Convolutional Networks (GCNs) are a computational framework that enables the extraction of relationships between nodes within a graph structure. These extracted relationships have been found to be beneficial in the context of image recognition tasks. The distinction between steamed dumplings and fried dumplings can be challenging, yet it becomes more manageable when we establish a connection between steamed dumplings, fried dumplings, and fried steak.

Food images do not have a distinctive appearance or prominent layout structure. To obtain a more discriminative and robust classifier, a GCN was introduced to capture interclass relations [49]. A two-head model was proposed to learn a CNN-based classifier and a GCN-based classifier, and then the two results are fused into one result to improve accuracy [48]. GCN was used to map semantic embedding into interdependent classifiers so that global label correlation could be taken into account and the performance of multi-label ZSL could be improved [47]. A labelled graph was inputted into a GCN-based network to learn the embedding vector, which was inputted into a metric learning network to learn similarities in low-dimensional space between each node [46]. This approach can be used to address the limited labelled data problem for document and image classification. Semantic label embedding and knowledge graphs were exploited to augment the visual features [50]. They used a GCN as the semantic-visual mapping network.

A tree or hierarchical structure may be the simplest way to connect different types of labels. A Meta-Concept method [52] was designed involving a concept graph, and Figure 6 illustrates the structure of the concept graph, which takes the form of a tree structure. This model trains not only the image classifier but also the concept classifier. It can adapt quickly and exhibit high performance by inferring the abstract concept and dealing with a few labelled samples.

© 2023 Conscientia Beam. All Rights Reserved.

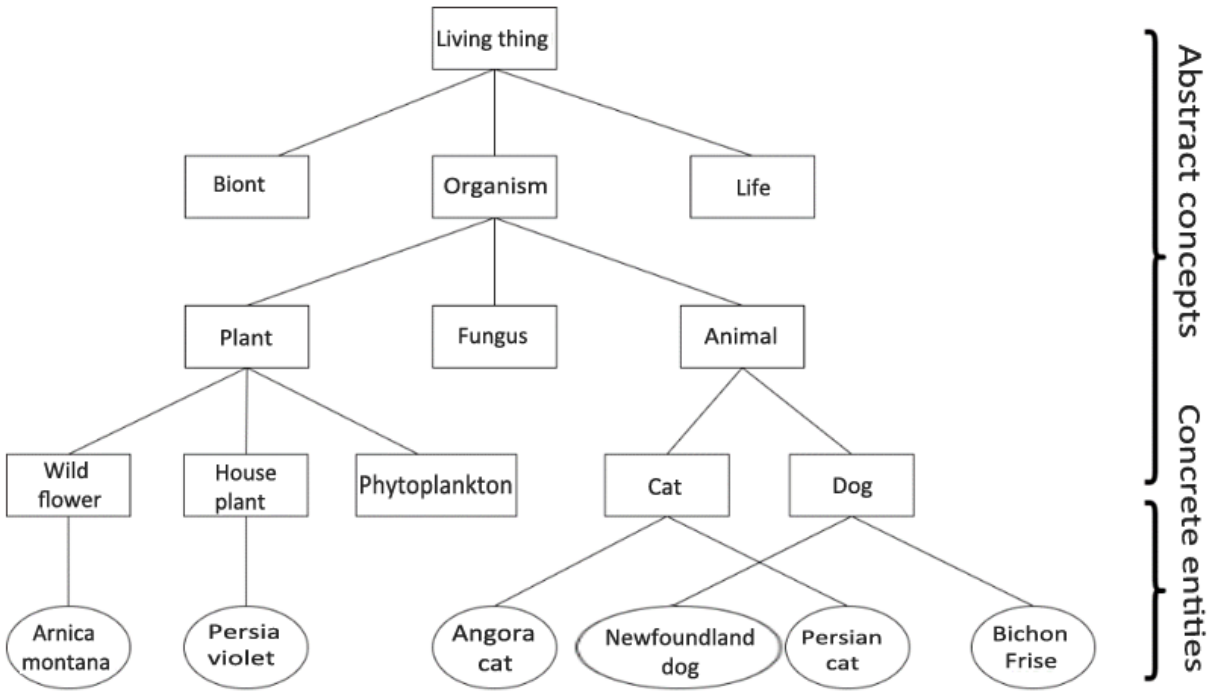


Figure 6. A schematic view of meta concept method with concept graph.

Chinese character recognition requires studying the relations between the characters and their radicals and obtaining a tree layout for primitives [51]. The hierarchical decomposition of the Chinese character is depicted in Figure 7, illustrating the character's breakdown. The character is comprised of five primitives, showcased in the nodes. With the help of these radicals and structures, the Chinese character can be recognized without a labelled sample from the training data.

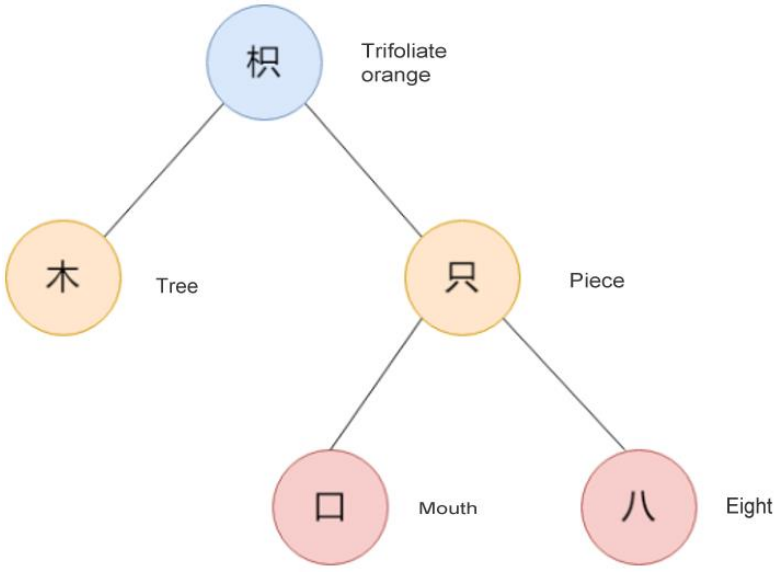


Figure 7. A schematic view of tree layout of the Chinese character.

In conclusion, it is crucial to find label correlations for label-based methods. There are two main approaches to obtaining their relationship: one uses their hierarchical structure, and the other uses a GCN. The correlations among labels are captured, and classification is performed with the help of label relationships.

© 2023 Conscientia Beam. All Rights Reserved.

3.3.3. Text-Based Methods

There are techniques for using a text corpus to mine the relationships between classes or categories. For processing text in FSL, there are primarily two categories: One involves utilizing existing models such as word2vec, BERT (Bidirectional Encoder Representations from Transformers), Text RCNN, Glove, and Reverb [53-58]. The other is an encoder-based method [59-61]. The processed text data is then fused with visual features or mapped onto the visual space (or vice versa) to enhance the single-modal visual features.

The use of the existing model will be discussed first. A multimodal diagnosis predictive model of Alzheimer's disease used reports of the disease and obtained a word vector by leveraging word2vec to address the FSL problem [53]. A sentence was pre-trained by BERT and CNN to obtain a low-dimensional vector for few-shot relation classification [54]. Few-shot vegetable disease recognition used Text RCNN to extract features from disease description texts [55]. Annotations were input into word2vec with a bag of words to build a text vector for image classification [56]. Any-shot sketch-based image retrieval problem was solved by using three different types of text-based side information: Word2Vec, GloVe, and FastText. This information was then used to judge if the generated information was real or fake [57]. English Wikipedia text was embedded into semantic vectors through word2vec, and the GAN was used to solve the ZSL problem [58].

A pre-trained text encoder processed textual descriptions to generate new data for FSL [59]. The knowledge Encoder processes textual descriptions and knowledge graphs to obtain global and local information for fine-grained classification for FSL [60]. Multimodal data-enhanced representation learning uses sentences from WN9-IMG-TXT (WordNet9-Image-Text), and image and text are input into a multimodal auto-encoder to learn entity representations or joint representations [61].

Text information cannot be leveraged directly. Therefore, it needs to be converted into word vectors by word2vec, BERT, TextRCNN, GloVe, FastText, or an encoder-based model. The extracted text features can be fused with visual features to generate new data, which is vital for FSL.

3.3.4. Lexicality-Based Methods

This method is based on WordNet and creates a taxonomy of classes. In one paper, hierarchy-based semantic embeddings were created based on the WordNet ontology and tree structure, and then a joint space of image and class embeddings was constructed to retrieve images [62]. Two other papers were based on a knowledge graph that was extracted from WordNet. A knowledge transfer module was introduced to address the data scarcity problem in FSL [63]. Knowledge-augmented networks (KANs) combine visual and semantic features and can obtain some discriminative features for FSL tasks [64].

3.4. Fine-Grained, Few-Shot Learning

Using the FSL model and semantic information to make fine-grained classifications is also challenging. The key challenge is capturing the subtle differences between fine-grained images [65]. The main approaches to this problem are divided into three categories: metric-based, meta-based, and data augmentation. Using semantic information is a type of data augmentation method.

The spatial attentive comparison network (SACN) is used to get multi-scale features to improve the performance of fine-grained few-shot (FGFS) recognition [66]. This method uses a powerful network that is based on meta-learning to obtain more discriminative features. Compared with traditional meta-learning, it extracts three scale features to capture the small differences between different classes. However, the author stated that the performance of this method was almost twice that of data augmentation. A teacher network extracts cross-modal information (text) and transfers it to a student network (only extracting visual features) to make predictions [65]. The teacher network and student work model can mitigate the semantic gap problem. The other algorithm for more discriminative features uses external cross-modal knowledge (text or graph) from global and local levels [60]. A mirror mapping network

© 2023 Conscientia Beam. All Rights Reserved.

(MMN) was introduced to map multimodal knowledge and visual features into a common space to bridge the semantic gap. It is obvious from the results presented in the studies above that a strong network can enhance FGFS recognition performance. However, this raises the calculation cost; for instance, SACN must handle three scale features. The enhancement is hardly noticeable. The alternative strategy relies on multimodal knowledge, which can solve the issue of data scarcity. It raises another issue, though, namely the semantic gap.

4. RECOMMENDATIONS

The present study focused on semantic-based few-shot image recognition, so the following recommendations for researchers are mainly on this topic. The most critical issue for semantic-based FSL is the domain shift problem [26, 35]. Visual and semantic features have different properties and cannot be combined directly. Therefore, transformation is required before their use. Simple semantic embedding or visual embedding causes serious domain shift problems.

Even though latent space embedding [37] can mitigate this problem, addressing this problem in the visual or semantic space remains a challenge. Because semantic-based FSL can always extract more discriminative features [37, 41, 42], it is suitable for FGFS learning. The features are in a lower-dimensional discriminative space [37] and are easily classified. Therefore, using semantic information for FGFS image classification has a promising future. Previous attempts at FGFS classification were based on metrics and meta-learning. Using semantic information, especially a semantic-guided attention mechanism, is worthy of further study. GAN can be used in FGFS classification [40, 67].

However, the models are extraordinarily complicated, and this method is time-consuming. The good news for the GAN process is that FSL does not require generating too many examples. How best to generate and use this data is the critical question. The distribution of the generated images is not similar to that of the original images [40]. Semantic-based generated methods can mitigate this problem using the semantic-guided method. Decomposing global features into parts through semantic information [43] is another direction for using GANs in FGFS image classification. This method can obtain more discriminative attributes for fine-grained classification.

5. CONCLUSION

A compelling subject within the field of deep learning is Few-Shot Learning (FSL). Nevertheless, semantic-based techniques lack definitive and structured forms or outlines. In this study, we have undertaken a comprehensive review of the methodologies that have been established in recent years in order to enhance our understanding of this particular topic.

Additionally, we have conducted an analysis of the prevailing trajectory of few-shot image recognition, with a specific focus on its reliance on semantic information. Initially, we presented the precise delineation of FSL and underscored its significance. In our analysis, we elucidated the utilization of cost-effective information in order to get a notable level of precision through the emulation of human cognitive processes. Subsequently, an examination was conducted on the primary methodologies employed in the field of Few-Shot Learning (FSL), including meta-learning, metric learning, and multimodal learning. In this paper, we divided multimodal learning into four categories based on the types of semantic information: attribute, text, label, and lexicality. Attribute data is the most convenient type to use.

We discussed the three types of attribute projection, semantic embedding, visual embedding, and latent space embedding, and their advantages and disadvantages. We explained that text, labels, and lexicality also require projection for visual features. Because they cannot be used directly, we mainly analyzed how to process this different information. Specifically, we examined the use of semantic information in fine-grained image classification. Finally, we introduced several promising directions for future research.

© 2023 Conscientia Beam. All Rights Reserved.

Funding: This study received no specific financial support.
Institutional Review Board Statement: Not applicable.
Transparency: The authors state that the manuscript is honest, truthful, and transparent, that no key aspects of the investigation have been omitted, and that any differences from the study as planned have been clarified. This study followed all writing ethics.
Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: Conceptualization, data collection, data analyze and writing, L.G.; formal analysis, E.M., H.I and N.M.; investigation, L.G. and E.M.; writing, review and editing, L.G. and E.M.; supervision, H.I. and N.M. All authors have read and agreed to the published version of the manuscript.

REFERENCES

[1] N. O'Mahony *et al.*, "Deep learning vs. traditional computer vision," in *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC)*, 2020, vol. 11, pp. 128-144.

[2] X. Sun, H. Xv, J. Dong, H. Zhou, C. Chen, and Q. Li, "Few-shot learning for domain-specific fine-grained image classification," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 4, pp. 3588-3598, 2020. https://doi.org/10.1109/tie.2020.2977553

[3] E. Schwartz, L. Karlinsky, R. Feris, R. Giryes, and A. Bronstein, "Baby steps towards few-shot learning with multiple semantics," *Pattern Recognition Letters*, vol. 160, pp. 142-147, 2022. https://doi.org/10.1016/j.patrec.2022.06.012

[4] X.-S. Wei, P. Wang, L. Liu, C. Shen, and J. Wu, "Piecewise classifier mappings: Learning fine-grained learners for novel categories with few examples," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 6116-6125, 2019. https://doi.org/10.1109/tip.2019.2924811

[5] T. M. Mitchell, "Does machine learning really work?," *AI Magazine*, vol. 18, no. 3, pp. 11-11, 1997.

[6] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Computing Surveys (CSUR)*, vol. 53, no. 3, pp. 1-34, 2020. https://doi.org/10.1145/3386252

[7] F. Pourpanah, "A review of generalized zero-shot learning methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–24, 2022.

[8] F. Yang, R. Wang, and X. Chen, "SEGA: Semantic guided attention on visual prototype for few-shot learning," Proc," *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1586–1596, 2022.

[9] W. Wang, V. W. Zheng, H. Yu, and C. Miao, "A survey of zero-shot learning: Settings, methods, and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1-37, 2019. https://doi.org/10.1145/3293318

[10] Y. Song, T. Wang, S. K. Mondal, and J. P. Sahoo, "A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities," *ACM Computing Surveys*, vol. 55, no. 13s, p. 1–40, 2023.

[11] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A surve," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5149-5169, 2021.

[12] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine*, 2017, vol. 3, pp. 1856–1868.

[13] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," *arXiv preprint arXiv:1803.02999*, 2018.

[14] A. Raghu, M. Raghu, S. Bengio, and O. Vinyals, "Rapid learning or feature reuse?towards understanding the effectiveness of maml," *arXiv preprint arXiv:1909.09157*, 2019.

[15] A. Rajeswaran, S. M. Kakade, C. Finn, and S. Levine, "Meta-learning with implicit gradients," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[16] J. Rajasegaran, S. Khan, M. Hayat, F. S. Khan, and M. Shah, "Itaml: An incremental task-agnostic meta-learning approach," in *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13588-13597.

[17] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," presented at the In International Conference on Learning Representations, 2016.

© 2023 Conscientia Beam. All Rights Reserved.

[18] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning* 2016, vol. 4, pp. 2740–2751.

[19] X. Li, X. Yang, Z. Ma, and J.-H. Xue, "Deep metric learning for few-shot image classification: A selective review," *arXiv e-prints, arXiv-2105*, 2021.

[20] E. Van der Spoel, "Siamese neural networks for one-shot image recognition," *ICML - Deep Learn Work*, vol. 7, no. 11, pp. 956–963, 2015.

[21] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," *Advances in Neural Information Processing Systems*, pp. 3637–3645, 2016.

[22] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in Neural Information Processing Systems*, pp. 4078–4088, 2017.

[23] S. Rahman, S. Khan, and F. Porikli, "A unified approach for conventional zero-shot, generalized zero-shot, and few-shot learning," *IEEE Transactions on Image Processing*, vol. 27, no. 11, pp. 5652-5667, 2018. https://doi.org/10.1109/tip.2018.2861573

[24] G. Wang, N. Guan, H. Ye, X. Yi, H. Cheng, and J. Zhu, "Channel-wise mix-fusion deep neural networks for zero-shot learning," presented at the ICASSP, IEEE International Conference on Acoustics, Speech, and Signal Processing 2021.

[25] G. S. Xie *et al.*, "Attentive region embedding network for zero-shot learning," in *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9384-9393.

[26] X. Li, M. Fang, D. Feng, H. Li, and J. Wu, "Prototype adjustment for zero shot classification," *Signal Processing: Image Communication*, vol. 74, pp. 242-252, 2019. https://doi.org/10.1016/j.image.2019.02.011

[27] Y. Huo, J. Guan, J. Zhang, M. Zhang, J. R. Wen, and Z. Lu, "Zero-shot learning with few seen class samples," presented at the In 2019 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2019.

[28] A. Pambala, T. Dutta, and S. Biswas, "Generative model with semantic embedding and integrated classifier for generalized zero-shot learning," in *In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 1237-1246.

[29] C. C. Lin, Y. C. F. Wang, C. L. Lei, and K. T. Chen, "Semantics-guided data hallucination for few-shot visual classification," presented at the In 2019 Ieee International Conference on Image Processing (icip). IEEE, 2019.

[30] Y. Liu, X. Gao, J. Han, L. Liu, and L. Shao, "Zero-shot learning via a specific rank-controlled semantic autoencoder," *Pattern Recognition*, vol. 122, p. 108237, 2022. https://doi.org/10.1016/j.patcog.2021.108237

[31] Z. Fu, T. Xiang, E. Kodirov, and S. Gong, "Zero-shot learning on semantic class prototype graph," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 8, pp. 2009-2022, 2017. https://doi.org/10.1109/tpami.2017.2737007

[32] Z. Liu, X. Zhang, Z. Zhu, S. Zheng, Y. Zhao, and J. Cheng, "Convolutional prototype learning for zero-shot recognition," *Image and Vision Computing*, vol. 98, p. 103924, 2020. https://doi.org/10.1016/j.imavis.2020.103924

[33] A. Mishra, V. K. Verma, M. S. K. Reddy, S. Arulkumar, P. Rai, and A. Mittal, "A generative approach to zero-shot and few-shot action recognition," presented at the In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018.

[34] Y. Long and L. Shao, "Learning to recognise unseen classes by a few similes," in *In Proceedings of the 25th ACM International Conference on Multimedia*, 2017, pp. 636-644.

[35] J. Liu, D. Tu, Z. Shi, and Y. Liu, "Zero-shot image classification based on a learnable deep metric," *Sensors*, vol. 21, no. 9, p. 3241, 2021.

[36] Y. Liu, J. Li, and X. Gao, "A simple discriminative dual semantic auto-encoder for zero-shot classification," in *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 940-941.

[37] A. Roy, B. Banerjee, and V. Murino, "Discriminative latent visual space for zero-shot object classification," in *In 2018 24th International Conference on Pattern Recognition (ICPR). IEEE*, 2018, pp. 2552-2557.

© 2023 Conscientia Beam. All Rights Reserved.

[38] Y. Liu, X. Gao, J. Han, and L. Shao, "A discriminative cross-aligned variational autoencoder for zero-shot learning," *IEEE Transactions on Cybernetics*, 2022.

[39] M. Gull and O. Arif, "Generalized zero-shot learning using identifiable variational autoencoders," *Expert Systems with Applications*, vol. 191, p. 116268, 2022. https://doi.org/10.1016/j.eswa.2021.116268

[40] R. Tao, Z. Li, R. Tao, and B. Li, "ResAttr-GAN: Unpaired deep residual attributes learning for multi-domain face image translation," *IEEE Access*, vol. 7, pp. 132594-132608, 2019. https://doi.org/10.1109/access.2019.2941272

[41] N. Xing, Y. Liu, H. Zhu, J. Wang, and J. Han, "Zero-shot learning via discriminative dual semantic auto-encoder," *IEEE Access*, vol. 9, pp. 733-742, 2020. https://doi.org/10.1109/access.2020.3046573

[42] H. Jiang, R. Wang, S. Shan, Y. Yang, and X. Chen, "Learning discriminative latent attributes for zero-shot classification," in *In Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4223-4232.

[43] P. Tokmakov, Y. X. Wang, and M. Hebert, "Learning compositional representations for few-shot recognition," in *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6372-6381.

[44] J. Liu, K. Song, Y. He, H. Dong, Y. Yan, and Q. Meng, "Learning object-centric complementary features for zero-shot learning," *Signal Processing: Image Communication*, vol. 89, p. 115974, 2020. https://doi.org/10.1016/j.image.2020.115974

[45] Y. Xu, Y. Bao, Y. Zhang, and H. Li, "Attribute-based structural damage identification by few-shot meta learning with inter-class knowledge transfer," *Structural Health Monitoring*, vol. 20, no. 4, pp. 1494-1517, 2021. https://doi.org/10.1177/1475921720921135

[46] W. Lin, Z. Gao, and B. Li, "Shoestring: Graph-based semi-supervised classification with severely limited labeled data," in *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4174-4182.

[47] G. Ou, G. Yu, C. Domeniconi, X. Lu, and X. Zhang, "Multi-label zero-shot learning with graph convolutional networks," *Neural Networks*, vol. 132, pp. 333–341, 2020.

[48] L. Bai, H. Wang, and Y. Guo, "Improving the generalised few-shot learning by semantic information," presented at the In 2020 6th International Conference on Big Data and Information Analytics (BigDIA). IEEE, 2020.

[49] H. Zhao, K. H. Yap, and A. C. Kot, "Fusion learning using semantics and graph convolutional network for visual food recognition," *In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1711-1720, 2021.

[50] M. Li, R. Wang, J. Yang, L. Xue, and M. Hu, "Multi-domain few-shot image recognition with knowledge transfer," *Neurocomputing*, vol. 442, pp. 64-72, 2021. https://doi.org/10.1016/j.neucom.2021.01.123

[51] Z. Cao, J. Lu, S. Cui, and C. Zhang, "Zero-shot handwritten Chinese character recognition with hierarchical decomposition embedding," *Pattern Recognition*, vol. 107, p. 107488, 2020. https://doi.org/10.1016/j.patcog.2020.107488

[52] B. Zhang, K.-C. Leung, X. Li, and Y. Ye, "Learn to abstract via concept graph for weakly-supervised few-shot learning," *Pattern Recognition*, vol. 117, p. 107946, 2021. https://doi.org/10.1016/j.patcog.2021.107946

[53] D. Chen, L. Zhang, and C. Ma, "A multimodal diagnosis predictive model of alzheimer's disease with few-shot learning," presented at the In 2020 International Conference on Public Health and Data Science (ICPHDS). IEEE, 2020.

[54] B. Hui, L. Liu, J. Chen, X. Zhou, and Y. Nian, "Few-shot relation classification by context attention-based prototypical networks with BERT," *EURASIP Journal on Wireless Communications and Networking*, no. 1, 2020. https://doi.org/10.1186/s13638-020-01720-6

[55] C. Wang, J. Zhou, C. Zhao, J. Li, G. Teng, and H. Wu, "Few-shot vegetable disease recognition model based on image text collaborative representation learning," *Computers and Electronics in Agriculture*, vol. 184, p. 106098, 2021. https://doi.org/10.1016/j.compag.2021.106098

[56] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for image classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 7, pp. 1425-1438, 2015. https://doi.org/10.1109/tpami.2015.2487986

© 2023 Conscientia Beam. All Rights Reserved.

[57] A. Dutta and Z. Akata, "Semantically tied paired cycle consistency for any-shot sketch-based image retrieval," *International Journal of Computer Vision*, vol. 128, no. 10-11, pp. 2684-2703, 2020. https://doi.org/10.1007/s11263-020-01350-x

[58] T. Zeng, H. Xiang, C. Xie, Y. Yang, and Q. Liu, "Zero-shot learning based on knowledge sharing," in *In 2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD). IEEE*, 2021, pp. 643-648.

[59] F. Pahde, M. Nabi, T. Klein, and P. Jahnichen, "Discriminative hallucination for multi-modal few-shot learning," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018: IEEE, pp. 156-160.

[60] J. Zhao, X. Lin, J. Zhou, J. Yang, L. He, and Z. Yang, "Knowledge-based fine-grained classification for few-shot learning school of computer science and technology," *East China Normal University* , *Shanghai* , *China Shanghai Key Laboratory of Multidimensional Information Processing* , *ECNU Insigma Hengtian Softwa*, no. 1, pp. 1–6, 2020.

[61] Z. Wang, L. Li, Q. Li, and D. Zeng, "Multimodal data enhanced representation learning for knowledge graphs," in *2019 International Joint Conference on Neural Networks (IJCNN)*, 2019: IEEE, pp. 1-8.

[62] B. Barz and J. Denzler, "Hierarchy-based image embeddings for semantic image retrieval," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019, pp. 638-647.

[63] Z. Peng, Z. Li, J. Zhang, Y. Li, G.-J. Qi, and J. Tang, "Few-shot image recognition with knowledge transfer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 441-449.

[64] Z. Zhu and X. Lin, "Kan: Knowledge-augmented networks for few-shot learning," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 1735-1739.

[65] J. Zhao, X. Lin, Y. Yang, J. Yang, and L. He, "Cross-modal knowledge distillation for fine-grained one-shot classification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021: IEEE, pp. 4295-4299.

[66] X. Ruan, G. Lin, C. Long, and S. Lu, "Few-shot fine-grained classification with spatial attentive comparison," *Knowledge-Based Systems*, vol. 218, p. 106840, 2021. https://doi.org/10.1016/j.knosys.2021.106840

[67] S. Tsutsui, Y. Fu, and D. Crandall, "Reinforcing generated images via meta-learning for one-shot fine-grained visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Views and opinions expressed in this article are the views and opinions of the author(s), Review of Computer Engineering Research shall not be responsible or answerable for any loss, damage or liability etc. caused in relation to/arising out of the use of the content.

© 2023 Conscientia Beam. All Rights Reserved.