
META-LEARNING APPROACHES FOR FEW-SHOT LEARNING: A SURVEY OF RECENT ADVANCES

A PREPRINT

Hassan Gharoun

Faculty of Engineering & IT
University of Technology Sydney
NSW, AU
Hassan.Gharoun@student.uts.edu.au

Fereshteh Momenifar

School of Business
Western Sydney University
NSW, AU
22046851@student.westernsydney.edu.au

Fang Chen

Faculty of Engineering & IT
University of Technology Sydney
NSW, AU
Fang.Chen@uts.edu.au

Amir H. Gandomi

Faculty of Engineering & IT, University of Technology Sydney, NSW, AU
University Research and Innovation Center (EKIK), Óbuda University, Budapest, HU
Gandomi@uts.edu.au

March 15, 2023

ABSTRACT

Despite its astounding success in learning deeper multi-dimensional data, the performance of deep learning declines on new unseen tasks mainly due to its focus on same-distribution prediction. Moreover, deep learning is notorious for poor generalization from few samples. Meta-learning is a promising approach that addresses these issues by adapting to new tasks with few-shot datasets. This survey first briefly introduces meta-learning and then investigates state-of-the-art meta-learning methods and recent advances in: (i) metric-based, (ii) memory-based, (iii), and learning-based methods. Finally, current challenges and insights for future researches are discussed.

Keywords Meta Learning · Few-shot Learning · Learning to learn · Representation learning

1 Introduction

Humans possess the extraordinary capability of learning a new concept even after minimal observation. To a greater extent, a child can distinguish a dog from a cat through a single picture [1]. This critical characteristic of human intelligence lies in the humans' ability to leverage obtained knowledge of prior experiences to unforeseen circumstances with small observation.

Unlike the human learning paradigm, traditional machine learning (ML) and deep learning (DL) models train a specific task from scratch through: (a) the training phase in which a model is initiated randomly and then updated, and (b) the test phase in which the model evaluates. While ML and DL have obtained remarkable success in a wide range of applications, they are notorious for requiring a huge number of samples to generalize. In many real-world problems, collecting more data is costly, time-consuming, and even might not feasible due to physical system constraints [2]. Moreover, most ML and DL models presume that training and testing datasets have the same distribution [3]. Thus, their performance suffers under data distribution shifts [4]. While the model's generalization ability significantly

depends on data distribution, training a model on a diverse dataset does not result in domain adaptation [5]. Domain adaptations aim to transfer obtained knowledge from prior experiences to learn new but relevant tasks [6]. Transfer learning, a well-established sub-field in machine learning, uses knowledge of related source domains with sufficient annotated data to improve the model’s performance on related target domains with inadequate or lack of labeled data [7]. Model parameters, feature representations, and instances are different forms of knowledge that can be transferred in transfer learning [8]. In addition, transfer learning has achieved tremendous success in a wide range of applications, such as computer vision [9] and natural language processing [10]. Arguably, transfer learning has obtained outstanding achievement for supervised pre-training of convolutional neural networks (CNNs) on ImageNet. However, the performance of the pre-trained network deteriorates during fine-tuning with few samples [11], or new samples become less similar [12]. One way to generalize unseen samples directly from few examples is by learning rich knowledge from similar events and re-using past knowledge to incrementally adapt to unseen examples. This idea, motivated by humans’ intelligence, is known as meta-learning.

Meta-learning refers to a variety of techniques that aim to improve adaptation through transferring generic and accumulated knowledge (meta-data) from prior experiences with few data points to adapt to new tasks quickly without requiring training from scratch. Meta-learning is most commonly known as “*Learning to learn*”. But, what does learning to learn mean? Mitchel [13] defined the term *learning* as the instance when a computer program’s performance at a given task improves with experience. In conjunction with Mitchel’s definition, Thrun and Pratt [14] defined “*Learning to learn*” as a capability of algorithms to improve their tasks’ performance with experience and with a number of observed tasks. In this context, an algorithm’s performance improvement depends on a family of tasks. In other words, if a learning algorithm’s performance does not improve with the experience of other tasks, it is not said to learn to learn [14]. Accordingly, the learning level is elevated from data to tasks in meta-learning. Figure 1 better demonstrates the learning to learn paradigm. Driven by the goal to scale AI to a real-world setting and approaches to the human intelligence paradigm, meta-learning has recently regained significant interest in few-shot problems using deep learning [15]. Currently, a number of surveys on meta-learning have been proposed [16, 17], however, the meta-learning methods progress rapidly. Accordingly, this paper examines meta-learning methods in detail and performs an extensive review of meta-learning to provide an up-to-date review. Additionally, a new organization from the algorithms’ mechanics point of view is introduced. A summary of the benchmark comparison of review methods is also presented, followed by a comprehensive discussion of challenges to establish future directions based on current limitations and strengths.

The remainder of this paper is organized as follows. Section 2 reviews the preliminary concepts of meta-learning. Section 3 introduces the category of meta-learning methodologies and presents an in-depth review of state-of-the-art methods and their advances. Finally, section 4 describes the challenges along with future research directions and brief conclusions.

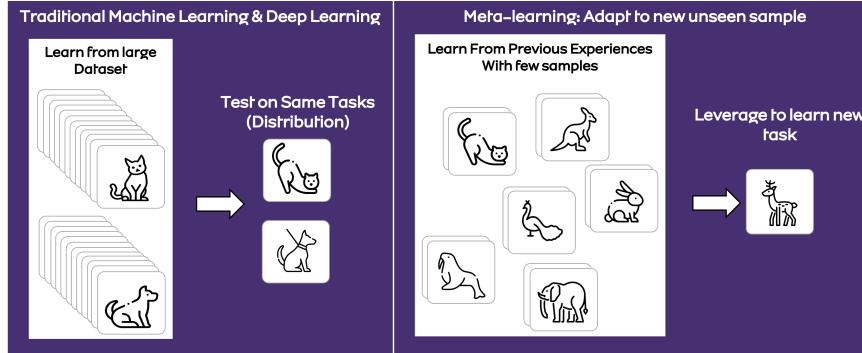


Figure 1: Standard Machine learning and deep learning vs. meta-learning paradigm

2 Meta-Learning Preliminary Terminology and Training Structure

2.1 Terminology and Training Structure

In the standard supervised machine learning setting, dataset D is divided into two separable and distinct datasets, namely train and test sets denoted by D_{train} and D_{test} , respectively. Conventional neural networks train during iterative cycles, termed “*Epoch*”. In every epoch, the entire D_{train} is feeding forward and backward. During each epoch, D_{train} may be partitioned into several packets termed “*Batch*”. Thus, each epoch may consist one or several “*iterations*”, which is the number of steps needed for one epoch that all batches (i.e., entire D_{train}) feed into the model.

In meta-learning, dataset D is divided with no overlapping into two meta sets, namely the meta train and meta test sets, denoted by D_{meta_train} and D_{meta_test} , respectively. Next, D_{meta_train} and D_{meta_test} are subdivided into train and test sets. To differentiate between these two forms of training and testing sets, the latter (inner) train and test sets are termed "*Support set*" and "*Query set*", respectively. Let's rewrite the meta train set and meta test set as $D_{meta_train} = \{(D_{meta_train}^{Support}, D_{meta_train}^{Query})\}$ and $D_{meta_test} = \{(D_{meta_test}^{Support}, D_{meta_test}^{Query})\}$. Support set is a N -way k -shot random samples and the query set is q random samples for each of the N categories of the support set. A meta-learner model trains in an "*Episodic*" mode. In each episode, support and query sets would be created. A meta-learner model takes $D_{meta_train}^{Support}$ to learn, then applies on the $D_{meta_train}^{Query}$ to make predictions. The meta-learner updates based on the prediction error over episodes. Over a series of episodes, the meta-learner learns to learn from the small dataset. This phase is called meta-learning. Afterward, it takes $D_{meta_test}^{Support}$ to build a classifier. Here, the model concentrates on learning task-specific parameters quickly. Then, the performance of the classifier is evaluated through $D_{meta_test}^{Query}$, which is known as the adaptation phase. In other words, the model uses the prior knowledge learned in the meta-learning phase to quickly adapt to a new task in the adaptation phase. From a conventional neural networks view, episodic training corresponds to training on a series of batches [18]. A visualization of data preparation and episodic training is demonstrated in Figure 2.



Figure 2: Meta training and test set, Episodic training schema

2.2 Datasets

Most state-of-the-art meta-learning models have been evaluated and compared on several well-known benchmarks. While there are several datasets available, the three most widely used benchmark datasets are:

- Omniglot: an image dataset of different handwritten characters of different alphabets [19]. This dataset is comprised of 1623 characters related to 50 different alphabets. In the few-shot learning setting, each character stands as classes/labels, and each character has 20 samples (or 20 shots) drawn by 20 different persons.
- miniImageNet: the mini version of the popular large ImageNet dataset published in 2009, which revolutionized the artificial intelligence realm. ImageNet is a large colored image dataset annotated by humans for visual recognition competition. Now, ImageNet contains 14 million images and covers 20,000 objects/classes. For few-shot learning problems, [1] published miniImageNet from ImageNet, a smaller dataset containing 100 classes with over 60,000 images.
- CUB-200-2011: the Caltech-USCD-Birds-200-2011 [20], in short CUB-200-2011, contains 11,788 images of 200 different bird species (200 classes). Each image has a textual description alongside labels describing various attributes, such as color and shape.

Oxford-102 (flowers dataset), CIFAR-10, and MNIST are other datasets that are less used in the literature. In this paper, the above three datasets are used to compare the performance of different methods.

3 Meta-Learning

This section provides a comprehensive overview of cutting-edge studies and their advances in meta-learning in three main categories (shown in Figure 3):

- Metric-based methods
- Memory-based methods
- learning-based methods
 - Learning the initialization
 - Learning the parameters
 - Learning the optimizer

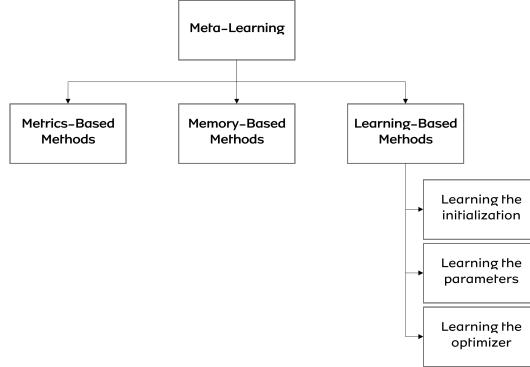


Figure 3: Meta-Learning methods categories

3.1 Metric-based methods

The key concept in "learning the metric space" methods is comparing the query set samples to support set samples. Then, the labels of query samples are predicted based on proximity to the support set. Principally, the success of these methods depends on embedding learning. In this context, the aim of embedding learning is to convert each high dimensional sample to a lower dimensional sample in the form of vector representations in such a way that similar samples reside closely in a new lower dimension space. Given an N-way k-shot problem, let's consider a support set $D_{support} = \{(x_1^s, y_1^s), \dots, (x_K^s, y_K^s)\}$, where x_i^s refers to input samples, $y_i^s \in N$ is predefined labels, and a query set $D_{query} = \{(x_1^q, y_1^q), \dots, (x_J^q, y_J^q)\}$, in which x_j^q is the test sample and $y_j^q \in N$ is the true label of test. In a generic form of learning the metric space methods, embedding function f and embedding function g apply on x_i^s and x_j^q , respectively, to generate embedding of the support set and query set. Next, the similarity between obtained embeddings, i.e., $f(x_i^s)$ and $g(x_j^q)$, is measured by similarity function $s(\cdot, \cdot)$. To predict the label of x_j^q , a decision is made based on the closest similarity between $f(x_i^s)$ and $g(x_j^q)$. In other words, the class $f(x_i^s)$ nearest to $g(x_j^q)$ is assigned to the sample test x_j^q . While the same embedding function can be used for both support and query sets, several researches have shown that different embedding functions lead to a better performance [1, 21]. Current state-of-the-art metric-based algorithms are:

3.1.1 Siamese Network

Siamese network is one of the simplest metric-based learners widely used in one-shot learning problems. The principal architecture of the Siamese network was proposed in 1993 by [22] for fingerprint similarity estimation. However, the term "Siamese" was first introduced by [23] for the signature verification problem. Principally, the Siamese network consists of two symmetrical neural networks joined with the similarity function called the "energy function." The architecture of the Siamese network is shown by Figure 4.

As demonstrated in Figure 4, a input pair x_1 and x_2 from the training set is fed to two embedding functions, $f_1(\cdot)$ and $f_2(\cdot)$, sharing the same parameters. Then, embeddings $f_1(x_1)$ and $f_2(x_2)$ feed into energy function. Basically, the energy function could be any similarity measure, such as cosine similarity or Euclidean distance. The output of the energy function is the similarity score of the two embeddings. Thus, rather than a classification task, the Siamese network principally aims to learn the network for matching tasks between similar pairs [24]. The proposed network

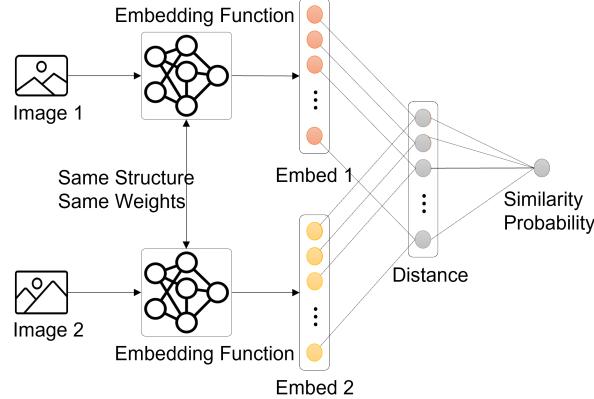


Figure 4: Siamese architecture for one-shot classification problem

utilizes contrastive loss as a loss function for optimization. The traditional contrastive loss used in the generic Siamese network is expressed as Eq.1 [25, 26]:

$$\mathcal{L}(E, Y) = \sum_{i=1}^P (1 - Y).L_G(E(Z_1, Z_2)_i) + Y.L_I(E(Z_1, Z_2)_i) \quad (1)$$

where $E(X_1, X_2)^i$ is the energy function that measures the similarity of i^{th} pairs of embeddings X_1 and X_2 . Y is the true label that denotes whether the given pair is similar $Y = 0$ or dissimilar $Y = 1$. L_G and L_I are partial loss functions applied on similar and dissimilar pairs, respectively, and designed in a way that monotonically increases the loss function for a similar pair and decreases loss function for a dissimilar pair. Siamese networks are widely used in various applications where data are limited. The face verification task is one of the extensive applications of the Siamese network. In the state-of-the-art research, [25] studied the task of face verification by refining the loss function as Eq. 2:

$$\mathcal{L}(E, Y) = (1 - Y).E^2 + Y.\max(0, margin - E)^2 \quad (2)$$

where the margin is a hyper-parameter that specifies the lower bound distance between dissimilar pairs. Similar to Eq.1, E is the energy function, and Y will be 0 for the similar pair and 1 otherwise. This study was one of the earliest works that used convolutional neural networks (CNNs) as an embedding function. Later, [27] studied Siamese networks for a one-shot image recognition problem, which is known as the first study of a metric-based method extended Siamese networks for few-shot learning. Additionally, Siamese networks have been extensively studied in various applications. For further details on Siamese network applications, interested readers can refer to comprehensive reviews by [28, 26].

3.1.2 Prototypical Networks

The prototypical network, proposed by [29], is basically learning an embedding function through clustering the embedding space. Principally, prototypical networks aim to generate a prototype of each class to use as an anchor for the classification task. Given the N-way k-task classification problem, the prototypical representation of each class is built based on the mean embedding of k data points in each N class. By using the same embedding function for both the support and query sets, the embedding of query points would be obtained. Mainly, Euclidean distance is used as a distance metric, measuring the distance of query embeddings and classes' prototypes. Each query point is classified based on the closest distance of the query embedding and classes' prototypes. Predictions probabilities are obtained by feeding query-prototypes distance into the Soft-max function. The prototypical network trains by minimizing the average negative log-probability of the correct prediction, as expressed by Eq.3. Figure 5 shows the general flow of the prototypical networks.

$$-\frac{1}{T} \sum_i \log p(y_i^* | x_i^*, p_c) \quad (3)$$

Two well-known variants of prototypical networks are semi-supervised and Gaussian prototypical networks. [30] revised the prototypical network to deal with the semi-supervised setting, where the dataset contains both labeled

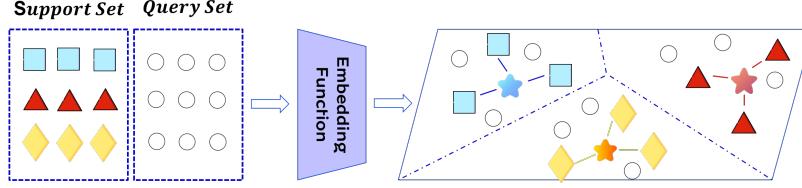


Figure 5: Architecture of Prototypical Networks.

and unlabeled data. In their model, similar to the generic prototypical network, class prototypes are first built with labeled data. Then, soft k-means algorithm is applied on the embeddings of both labeled and unlabeled data. However, inspired by seeded k-means, instead of initializing cluster centers randomly, the class prototypes are calculated from the labeled data, and are used as cluster centers. Then, unlabeled embeddings are assigned to each cluster based on their Euclidean distance to each cluster center, i.e., class prototypes. Finally, class prototypes are refined by updating cluster centers by the usual k-means algorithm. Since some points that do not belong to any existing classes may exist in unlabeled data, these data would interfere with prototypes refinements. Thus, the authors proposed an additional class for these data called the distractor class and considered the origin as the center of the distractor class. Since distractor class members may belong to different subjects, a single distractor class will have a high variance. To solve this issue, the authors proposed a soft-masking mechanism, in which a threshold distance is set for all original class prototypes. This threshold acts as an intra-cluster variance capturing distractors as exemplars outside the prototypical areas. This threshold is calculated by feeding statistics of prototypes, including min, max, skewness, and kurtosis, to a small multi-layer perceptron (MLP) network.

The Gaussian prototypical network is another novel architecture of the prototypical network proposed by [31]. In this model, a covariance matrix is constructed from the embedding function alongside the generated embedding vectors. The covariance matrix is used to build a confidence region around data points and adds weights to embedding vectors, then utilized to compute a total covariance matrix of each class. Afterward, the distance of each query from class prototypes is used to compute the linear Euclidean distance as follows (Eq.4):

$$d_c^2(i) = (\vec{x}_i - \vec{p}_c)^T S_c (\vec{x}_i - \vec{p}_c) \quad (4)$$

where p_c is the cluster center or the prototype of class c ; S_c is the inverse covariance matrix; d_c is the distance of query point i from class prototype c ; and \vec{x}_i is query point i . Figure 6 illustrates the operation of Gaussian prototypical networks. Variable p_c is calculated using variance-weighted embedding vectors in Eq.5:

$$\vec{p}_c = \frac{\sum_i \vec{s}_i^c \cdot \vec{x}_i^c}{\sum_i \vec{s}_i^c} \quad (5)$$

where \vec{x}_i^c is the embedding vector of support point i and class c ; and \vec{s}_i^c is the covariance matrix of related support point i and class c . Class covariance matrix is computed as Eq.6:

$$\vec{S}_C = \sum_i \vec{s}_i^c \quad (6)$$

[32] proposed the transferable prototypical network (TPN) by remoulding prototypical networks to learn unsupervised domain adaptation. TPN uses labeled data as a source and unlabeled data as target domains and tries to learn a neural network that minimizes embedding shifts between source and target domains. In other words, TPN constructs an embedding function that generates domain-invariant representations. Technically, TPN first builds class prototypes from the source domain (i.e., labeled data), then the target domain (i.e., unlabeled data) assigned to the nearest class prototypes. The authors proposed three different classifiers regarding three datasets, namely source only data, target only data, and combined source and target data. The authors argued that if the source and target domain distributions are the same, the class prototypes on both domains should be identical. Thus, to achieve the representation invariant over domains, they compute class level discrepancy over domains by calculating reproducing kernel Hilbert space (RKHS). Finally, TPN trains by minimizing combined classification loss and discrepancy loss.

With a similar belief and similar mechanics to the transferable prototypical network (TPN) [32], [33] introduced a new module called domain alignment component to prototypical networks to tackle the problem of domain shifts between test and train sets. Basically, the domain alignment component tries to minimize the maximum mean discrepancy

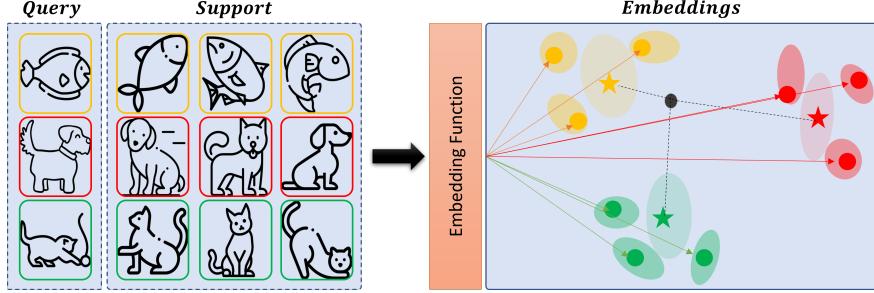


Figure 6: Gaussian prototypical networks workflow; Solid circles are embedding vectors, dark ellipses represent each embedding covariance matrix, and light halos depict the class's total covariance matrix. Finally, stars illustrate classes' prototypical.

(MMD) by calculating the similarity of two different domains, one from the train set and one from the test set, in kernel Hilbert space. Then, the networks are trained to minimize the combined negative log-probability loss (the generic prototypical networks loss) and MMD measure. Empirical analysis on benchmark datasets showed that their model outperforms the generic prototypical networks.

[34] argued that most metric-based meta-learning models forget intra-class distribution information. To tackle this problem, they proposed another tweak to prototypical networks called improved prototypical network (IPN), which has three main components: (i) feature extraction, (ii) weight distribution, and (iii) distance scaling. The first component basically extracts feature embeddings and constructs class prototypes. In generic prototypical networks, samples of the same class have the same importance (weight) in class prototypes. The authors refined this idea by proposing an attention-analogous strategy. Hence, the second component, weight distribution, is responsible for computing weighted class prototypes. The weight distribution component is designed as a three-layer neural network that takes concatenated embeddings of the same class as the input and provides samples' weights as the output. The last component, distance scaling is a three-layer neural network that maximizes inter-class differences and minimizes the intra-class differences. The results indicated that IPN outperforms the generic prototypical networks.

[35] expanded the prototypical network to multi-label few-shot problems. The authors argued that the main challenge with multi-label problems is that different labels can have the same support set leading to the same prototypes. To address this issue, they proposed a new prototypical network called label-enhanced prototypical network (LPN). For a multi-label few-shot problems, the authors targeted the aspect category detection in sentiment analysis. Accordingly, their data points and class labels consisted of sentences and aspect categories. The LPN is comprised of four main modules: (i) feature extraction, (ii) label-enhanced prototypical network, (iii) contrastive learning, and (iv) adaptive multi-label inference. The authors used a pre-trained embedding function for language, such as Bert [36] as the feature extraction module. The second module is responsible for computing the relation between data points, here sentences, and labels. For this purpose, after extracting support sentence representations from the previous module, the same process is repeated for labels that are verbal phrases to get label representations. The authors used the low-rank bilinear model [37] to assign weight to each sentence's representations. In simple words, the authors calculated the sentence representations' weights by element-wise multiplication of sentences' representations and their label representations, normalized with the Soft-max function. Then, class prototypes are built based on the weighted sentences' representations. The label-enhanced prototypical network module is trained to minimize the cross-entropy loss. The third module is designed in a way to maximize intra-class similarity and minimize inter-class similarity by employing contrastive learning loss. Since in their problem, every sentence can have a different number of labels as a multi-label problem, the next challenge was determining the number of labels that should predict for each new query point. This challenge was addressed by proposing the fourth module, which is basically a multi-layer perception network that predicts the number of labels for each sentence. This network is trained to minimize the cross-entropy loss of predicted label counts. By integrating all modules and combining all losses, LPN outperforms the generic prototypical network. [38] proposed another variant to prototypical named the multi-prototypes prototypical network. In contrast to the generic prototypical networks that consider only one prototype per class, the proposed network investigates the idea of multiple prototypes per class. For this purpose, the authors partitioned the embedding space of each class into k clusters by employing the k-means clustering algorithm. Class prototypes are simply cluster centers, and the proposed network is trained to minimize the COREL loss. The results showed that the number of class clusters, i.e., classes prototypes affects model performance; in other words, the difficulty of determining the optimal number of prototypes per class limits the performance.

[39] argued that most metric-learner models are not able to distinguish classes out of support set classes and it is necessary for meta-learner models to be able to detect unseen classes during prediction. The authors proposed dummy prototypical networks (D-ProtoNets) to address this issue. Conceptually, D-ProtoNets follows the same mechanics of the generic prototypical networks, but in every episode, a dummy prototype is built by introducing a dummy generator function. In simple words, the dummy generator is a Maxpool layer followed by non-linear fully connected layers performed on the support classes' prototypes. The dummy prototypes are responsible for handling unseen classes. The authors targeted the keyword spotting detection problem and created and published splitGSC, a benchmark dataset for key spotting detection problems based on the Google speech commands dataset. Through various numerical analyses, the authors demonstrated that D-ProtoNets achieved state-of-the-art performance on splitGSC.

[40] argued that a single prototype for each class, like the generic prototypical network, is not able to capture the class distribution effectively, so they proposed a local descriptor-based multi-prototype network (LMPNet). Technically, LMPNet utilizes a local descriptor to obtain feature embeddings. The local descriptor is basically the two-dimensional vector output from a convolutional layer [41]. LMPNet builds class prototypes based on the weighted average of the local descriptors. These prototypes were obtained through a novel mechanism by introducing channel squeeze and spatial excitation components, which flatten feature embeddings regarding their spatial position over channel dimensions (Figure 7). In summary, this attention module is designed to learn the output of multiple prototypes from local descriptors compressed by channel squeeze and spatial excitation. Using Resnet as a backbone, LMPNet outperforms the generic matching network.

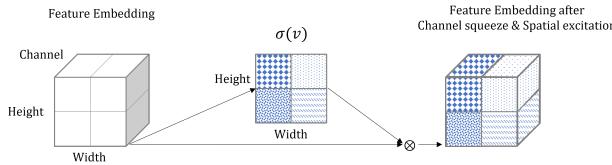


Figure 7: Schema of channel squeeze and spatial excitation components. $\sigma(v)$ corresponds to the relative positional importance of the information of embeddings.

[42] extended prototypical networks to a multi-modal prototypical network. The core novelty of this model is the use of textual descriptions attached to the image dataset to create more feature embeddings. For this purpose, the authors trained a text-conditional generative adversarial network (tcGAN), which creates image features from the given textual description. Then, two prototypes per class were calculated – one from the original images' embeddings and one from the generated images' embeddings. Next, combined class prototypes are computed based on the weighted average of both original and generated prototypes. Finally, the multi-modal prototypical network was trained in a similar fashion to generic prototypical networks.

3.1.3 Matching Networks

Matching networks are another powerful metric-based meta-learner proposed by [1] that are able to generate labels for the unobserved class as well. In this work, a parametric neural network P was developed to produce a prediction for each query point from the support set $S = \{(x_i, y_i)\}_{i=1}^k$, (k number of sample per class), per Eq.7:

$$p(\hat{y}|\hat{x}, S) = \sum_{i=1}^k a(\hat{x}, x_i) y_i \quad (7)$$

where $p(\hat{y}|\hat{x}, S)$ is the predicted class for the query point \hat{x} given support set S ; x_i and y_i are the inputs and labels of samples from support set S ; and a is the attention mechanism between \hat{x} and x_i . The attention mechanism is simply the Soft-max function over the cosine distance between embeddings of \hat{x} and x_i , which are calculated via Eq.8:

$$a(\hat{x}, x_i) = \frac{e^{\text{cosine}(f(\hat{x}), g(x_i))}}{\sum_{j=1}^k e^{\text{cosine}(f(\hat{x}), g(x_j))}} \quad (8)$$

In Eq.7, the attention mechanism result is multiplied by the support set labels y_i . For this purpose, the support set labels are first converted into one-hot encoded values. The overall flow of the matching network is shown in Figure 8. As seen in Eq. 8, the authors used two different embedding functions, f and g , to map data points into embedding for the query \hat{x} and support input x_i , respectively. Finally, the proposed network is trained to minimize the prediction error. [43]

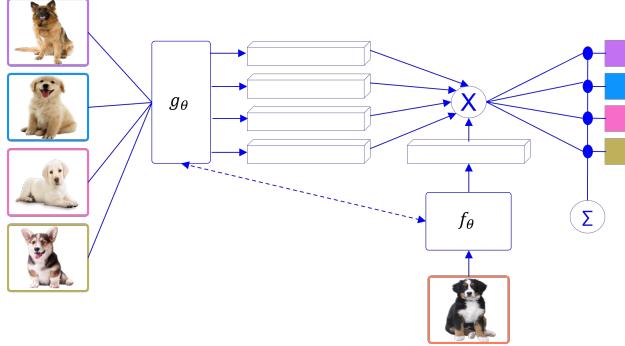


Figure 8: Architecture of Matching Networks.

argued that support and query images could differentiate in local features and shapes. They tackled this problem by proposing a cascaded feature matching network (CFMN), which conceptually maps features of the support and query sets to a latent space then reshapes them into a two-dimensional array. Next, it calculates the element-wise similarity of support and query features. Additionally, the support set features maps to a higher space based on element-wise correlation calculated in the previous step. In other words, the final support set features are constructed by taking into account the relations of feature positions, the so-called spatial attention mechanism. This approach allows relevant features to be considered and less important features to be ignored. Finally, CFMN predicts a label for a new sample based on the distance-based similarity score of the support set and a new sample point. [44] proposed another matching network variant called the attentive matching network (AMN), which exhibits two novelties. First, it incorporates a new loss function named complementary cosine loss (CCL) that is a combination of the hardest-category discernment loss (HDL) and cosine-distance metric, aiming to minimize the intra-class difference and maximize the inter-class differences. The second novelty is a new attention mechanism that focuses on the feature level. This module calculates the importance of features based on the standard deviation over classes. In other words, a higher standard deviation over classes means that features are more distinctive between classes. For this purpose, the authors first built class prototypes by class-wise mean of embeddings, then calculated the inter-class standard deviation from each class prototypes. In order for the feature attention module to be differentiable, they proposed a convolutional operation over the attention calculation.

[45] argued that a generic matching network randomly selects labels from the base dataset to create training samples, which leads to an over-easy label set. For example, the authors compared a label set cat, tree, water to another label set cat, dog, lion that is easy and low value in a classification task. To address this issue, they proposed schedule sampling for the matching network, which first samples a random label set, then calculates the difficulty value for the selected label set. The difficulty value is computed based on similarity metrics. The training of matching net started with a low difficulty label set and then continued into a more difficult label set. The authors showed that the performance of the classifier directly relates to the difficulty of the label set.

3.1.4 Relation Networks

Relation network is another simple yet effective metric-based meta-learner presented by [46]. This model has two main components: iembedding function denoted by f_ϕ and iirelation function g_ϕ . According to the flow of the relation network illustrated in Figure 9, the embedding function f_ϕ maps the support and query sets into the embedding environment. For the k -shot problem where $k > 1$, embeddings of the same class in the support set are aggregated through the element-wise summation to construct each class embedding map. The obtained consolidated classes embeddings are then combined with embedding of query samples and fed into the relation function g_ϕ . The output of the relation function g_ϕ is the N (number of classes) similarity score called the relation score that ranges between 0 and 1. While the proposed model is classification task, the authors trained their model using the mean square error (MSE) loss function since they aim to predict a relation score scale between 0 and 1.

With a similar belief to [34], [47] introduced the memory-augmented relation network (MRN) to investigate interactions of instances. MRN is composed of four components: (i) feature embedding function, (ii) information generation component, (iii) relation module, and (iv) metric-based classifier. The feature embedding function simply converts data points into latent space. The information generation component adds information to samples basically from its k -nearest samples' embeddings. This component is equipped with memory to control embedding propagation. The relation module, similar to a generic relation network, is a convolutional neural network that takes input samples and computes their distances. The last component, the metric-based classifier, predicts a label for an unlabeled query point

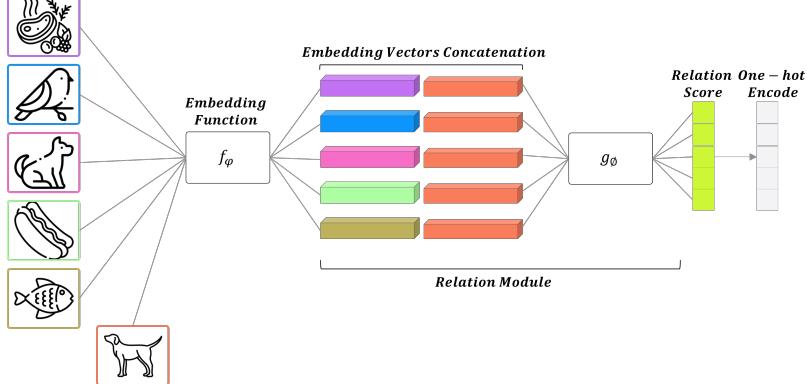


Figure 9: Relation network architecture.

based on the distance from class centroids, which are defined as the mean of the same class feature representations. This classifier trains to minimize cross-entropy loss. [48] proposed a prototype-relation network by basically combining prototypical networks and relation net. This model has two main components: 1) the prototype component that generates class prototypes, and 2) the relation component, similar to the generic relation net, that computes relations between query points and class prototypes. Another novelty of this work is the introduction of a new loss function that considers inter and intra-class distances. [49] developed another variant of relation networks called multi-scale relation networks (MSRN) that has two main components: a feature extractor and metric learner. In convolutional neural networks, low-level layers generate detailed information while the higher-level layers produce more semantic information. The authors argued that combining the feature outputs of different layers may improve representations. The second component, the metric learner, is basically a generic relation network; however, the authors subtracted these two components instead of concatenating two multi-scale features. They claimed that subtraction generates a better representation of feature differences. Their model outperforms the generic relation net. [50] proposed a multi-local feature relation network (MLFRNet) to improve relation networks. They used random cropping and random dropping to get local features, then mapped to embedding space. Finally, class prototypes are calculated by the mean of each class's local feature embeddings. The authors also introduced a local feature attention module to calculate the relation between local feature prototypes and query samples. Basically, the local feature attention module is responsible for computing the contribution of local features in query sample prediction. Additionally, the authors proposed a new strategy called support-support similarity to dynamically calculate the margin loss of local features. Support-support similarity was designed to compute pair-wise cosine similarity of classes that is used to generate a margin. The authors suggest that adding the margin is necessary for different classes that might have local features with high similarity in order to separate different classes effectively. [51] argued that the original relation network is not robust to spatial positions of feature maps and may perform poorly in scale variations of compared objects. To address this issue, they extended a new variation of the relation network called the multi-scale kronecker-product relation network (MsKPRN), which consists of three main modules. The first module is the encoding module that is responsible for converting different scales of support and query sets into feature space. The output of this module is feature maps with different scales that is organized in a way to feed the same scale support and query feature maps into the next module. The next module is the kronecker-product (in short KP) module, which calculates the spatial correlation from pair representations by operating the inner product on the same scale support and query feature maps. The obtained spatial correlation is concatenated with the representations to feed into the next module. The third module is the relation network module, which is responsible for learning the relation between the same scale support and query feature maps and producing similarity scores in the range between 0 and 1. Their results demonstrated that the kronecker-product module improves few-shot classification tasks by incorporating spatial correlation maps into feature maps.

3.1.5 Other Advanced Methods

The next advanced metric-based meta-learning algorithm reviewed in this section is the graph neural network (GNN) proposed by [52]. The authors argued that their proposed GNN for the few-shot classification problem generalizes Siamese networks, prototypical networks, and matching networks based on their architectures. Conceptually, in their model, each task T_j is approached as a fully-connected graph $G = (V, E)$, where V and E are denoted as nodes and edges, respectively. Nodes are associated with support set embeddings concatenated with one hot encoded label. Edges correspond to tasks' similarities, which are trainable. For a new query point, i.e. an unlabeled data point, a uniform distribution over all labels is used as label concatenation to query embedding. Finally, GNN was trained to

minimize the cross-entropy loss. Results show that GNN outperforms Siamese networks, prototypical networks, and matching networks on several benchmark problems. [53] proposed the global class representation (GCR) model, which achieved the best performance on benchmarks and outperformed many well-known meta-learning algorithms. The idea of GCR is to engage novel classes with base classes at the first stages of training, thus the reason for naming the model global class. Concretely, global class representation is first created from the whole class set, called C_{Total} , which is composed of both base and novel classes. Next, in each episode, N random classes and n_s random samples per each class are selected to form the support set. Then, episodic class representation is built based on the support set. Here, class representation is analogous to class prototypes in prototypical networks. As one novelty of this study, the authors proposed a new component called the registration module. Basically, the registration module is responsible for calculating the similarity between each episodic class representation and global class representations, or the so-called registration loss. GCR minimizes both registration loss and classification cross-entropy loss to achieve optimal global classes. GCR predicts a label of a new sample by comparing it to global class representations. [54] disputed that most metric-based meta-learners use standard similarity measures, such as cosine similarity and Euclidean distance. They support that those metrics could not capture non-linear relations between samples. To address this issue, they introduce a multi-scale feature network (MSFN), which has three main components. The first component is the feature extractor that extracts multi-scale features from the support set. The next component is the label feature that is responsible for learning classes representations (analogous to class prototypes). The label feature basically is a network that takes a concatenation of support features from the previous component as the input and generates class representations as the output. The last component is the metric-component, which principally is the relation network that computes the matching degree between a new query feature map and class representations. Through numerical analysis, the authors demonstrated that multi-scale features and non-metric similarity measures improve the overall model performance.

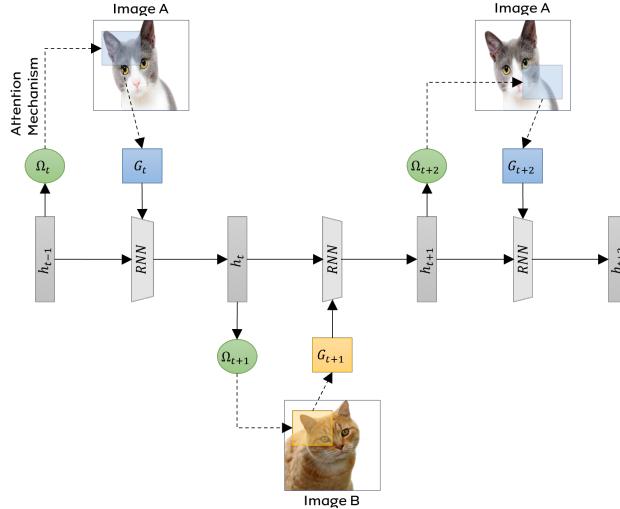


Figure 10: Attentive recurrent comparators (ARCs) scheme.

[55] argued that humans differentiate two objectives by recurrent multiple observations, and during each observation, a specific feature is inferred. This human-wise approach is quite different from Siamese networks. [55] proposed attentive recurrent comparators (ARCs), mimicking humans. ARCs consists of two components: (i) recurrent neural network (RNN) controller, and (ii) attention mechanism. The attention mechanism is designed in a way that concentrates on a specific region of the image. Figure 10 demonstrates the scheme of ARCs, in which the attention mechanism gets a glimpse G_t from the image I_t at each step (Eq.9):

$$G_t \leftarrow (I_t, \Omega_t) \quad (9)$$

where Ω_t is the attention glimpse parameter, denoting the location and size of the glimpse. Basically, Ω_t is computed based on the previous hidden state (h_{t-1}) of the RNN controller (Eq.10):

$$\Omega_t = w_t h_{t-1} \quad (10)$$

where w_t is the projection matrix necessary to map h_t to the number of attention parameters. Next, the hidden state is obtained via Eq.11:

$$h_t \leftarrow RNN(G_t, h_{t-1}) \quad (11)$$

where $RNN(\cdot)$ is the RNN controller update function. In short, ARCs recurrently observe different partitions of two images over a finite number of representations, so the model can focus better on the important context of each image. While the presented model outperforms Siamese networks, it can be quite computationally expensive due to sequential operations. Similarly, [56] believed that humans differentiate objects based on partitioning images and comparing them with the prototypical part of each image category. Accordingly, they proposed a new metric-based method called the region comparison network (RCN). The aim of RCN, as demonstrated in Figure 11, is to compare the whole query image with different partitions of support samples. RCN consists of three main components: (i) feature extractor, which is responsible for producing image representations; (ii) region matching, which is designed in a way to calculate the similarity scores between support and query samples; and (iii) explain network, which is used to obtain the final classification. To be more specific, the output of the first step is decomposed into several region vectors. Then, cosine similarity between support set region vectors and query set representation is computed. Subsequently, global max-pooling is applied to pick the most important information in the region similarity vector. The explain network takes region similarity maps from the previous step and region weight to make a classification. In each task, the important region may change. For example, in classifying dogs and birds, the important regions are the dog's head and bird's wings. The explain network uses task-specific region weights produced by the region meta-learner, which is designed to take a concatenation of support set representations and query set representations and produce task-specific region weights. The most salient aspects of RCN are its interpretability and quantitative explainability of classification tasks in the few-shot setting.

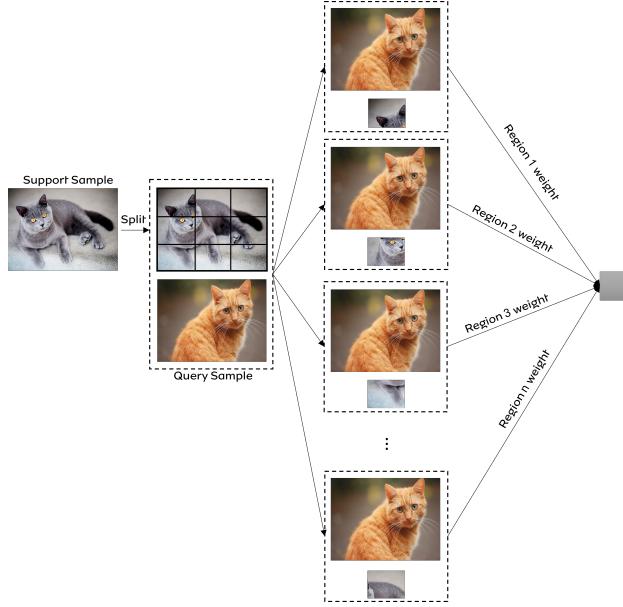


Figure 11: Comparison support sample's parts with query sample in Region comparison network (RCN).

[57] aimed to explore the effects of pre-defined similarity metrics, such as Euclidean and cosine distances, on existing metric-based methods and proposed a new method called metric-agnostic conditional embeddings (MACO). The architecture of MACO is demonstrated in Figure 12. Basically, MACO is composed of four stages: (i) feature stage, (ii) relation stage, (iii) conditioning stage, and (iv) classifier stage. The first stage is responsible for obtaining a feature vector from input images, which is simply a CNN-based network with a fixed structure for both the support and query sets. The next stage is designed to produce a single vector for each class. This class-specific single vector is analogous to the class prototype. The authors used the relational network (proposed by [58]) to generate the single vector class representations. The relational network is comprised of fully connected layers (instead of convolutional layers) followed by batch normalization and ELU activation function using pair-wise image comparisons within each class. The single vector class representation would be the average output across all comparisons (Eq.12):

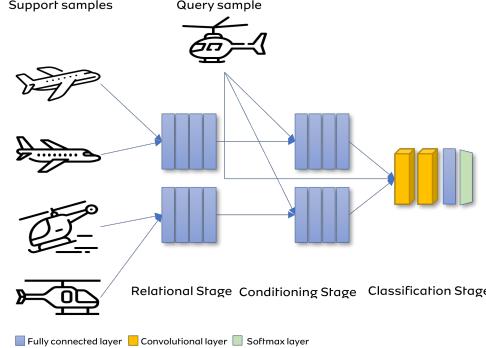


Figure 12: Metric-agnostic conditional embeddings (MACO) scheme.

$$R_\rho(s_k) = \frac{1}{\binom{n}{2}} \sum_{(x_i, x_j) \in s_k} g_\rho(x_i, x_j), i \neq j \quad (12)$$

where R_ρ is the single vector of support set class k ; n is the number of images in class k ; g_ρ is the relational network parameterized with ρ ; and x_i, x_j are images representations within support class k . The conditioning stage uses a network similar to a relational network that takes the concatenation of class vectors and query images. The output of this stage is the modified class representation in the context of the query image. The last stage is a neural network that takes the modified class vectors and query images as input and produces Softmax classification. This network is composed of convolutional layers followed by fully connected layers. In short, MACO is a novel model in which a learnable classifier supersedes the pre-defined metrics and modified class vectors (or prototypes) conditioned to the query image context. Similarly, [59] developed a learnable network instead of pre-defined similarity metrics termed relative position and map network (RPMN). Basically, RPMN consists of three components: (i) feature extractor, (ii) relation position network, and (iii) relative map network. The relative position network aims to produce different weights to the position of feature embeddings, and the relative map network is analogous to the original relation network but has been modified in such a way to compare two single maps from support and query feature representations. The relative map network is trained to learn task-specific embedding models. Moreover, the relative map network is equipped with a fully connected head, which is responsible for computing the final similarity score between the support and query sets. An important aspect of RPMN is the introduction of the relation position network as a new attention mechanism, allowing for more comparisons of effective feature maps, and the relative map network as a learnable similarity metric network.

3.2 Memory-based Methods

Memory-based meta-learning techniques have a more dynamic mechanism since they are adaptive to the presented tasks. Principally, memory-based meta-learners own the memory component (internally or externally) that enables them to retrieve information from previous inputs. The memory-augmented neural networks (MANN) is a well-known and effective memory-based meta-learner proposed by [60], which aims to use external memory to improve task adaptation utilizing the Neural Turing Machine (NTM) [61]. On a high level view, NTM is a neural network equipped with external memory to retrieve information. More specifically, NTM contains two main components: (i) controller and (ii) memory bank. Figure 13 shows the interaction of the controller and memory bank. Similar to a standard neural network, NTM receives inputs and predict outputs. In contrast to a standard neural network, NTM communicates with the memory bank through selective read and write operations, which the authors named as *head* analogous to the Turing machine. In order to NTM be trainable, including the way memory bank is utilized, the authors developed blurry read and write operations that should be differentiable to use backpropagation. For this purpose, read and write operations interact with distribution over memory locations instead of a single location. Specifically, the read operation simply reads values from the memory bank, where the value is basically the neural network parameters (i.e., weights). The memory bank is simply a two-dimension matrix. The write operation is more complex than read and includes two modules called erase and add, which erases information that is no longer required from a memory cell and adds a new operation to the memory, respectively.

To produce the parameters of the NTM neural network, the authors proposed two different addressing mechanisms. First, the content-based approach uses the similarity of the current value and controller's value. While the content-based is quite straightforward, the location-based approach is employed to iterate over different locations rather than content, which is essential for generalization.

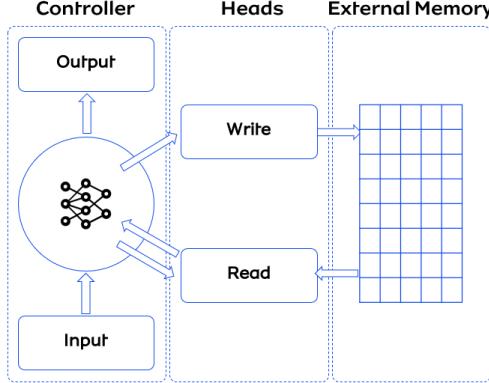


Figure 13: Neural Turing Machine scheme.

[60] proposed a new addressing mechanism called least recently used access (LRUA). In MANN, the read operation is performed through the content-based addressing mechanism, and the write operation uses LRUA to write the most recently used memory location. MANN accumulates information of a coupled representation-class label in external memory through the explained mechanism, which enables it to retrieve the information later for classification [specifically a sample from seen class is given - please reword this last part for clarity]. [62] proposed a new memory-based meta-learner model called simple neural attentive meta-learner (SNAIL) that combines temporal convolution and soft attention mechanism that acts as a memory compared to the external memory of MANN. The temporal convolution serves as high band-width memory access, and the soft attention allows access to a specific experience. The combination of these two components leads to better leveraging of information from past experiences. [63] proposed conditional neural processes (CNPs) that, in a similar fashion to SNAIL, do not employ external memory. On a high level, CNPs consists of a meta-learner and task-learner. Figure 14 shows the general architecture of CNPs. In this figure, h is an embedding function that generates representations r_i from the input and ground truth (labels) from the support set S . Operator a is the aggregation function that produces a single representation from r_i , and g is the neural network (task learner) that predicts labels for a new sample from the query set based on the output of a . As seen in Figure 14, the meta-learner produces a memory value by aggregating representations of the support set through operator a . Then, the label of new inputs is predicted by processing the aggregated representations. In other words, CNPs use concise representations of seen classes to make predictions.

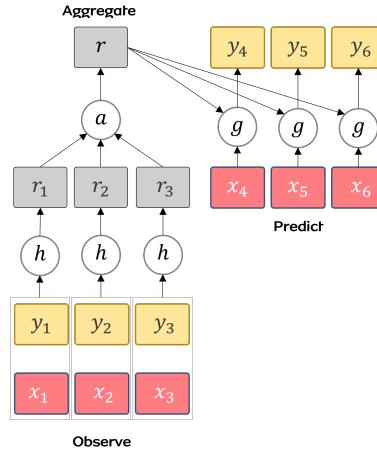


Figure 14: Conditional Neural Processes scheme.

[64] proposed a memory augmented matching network by combining MANN and matching network. The authors argued that if support data distribution and skewed, constructing a prototype by mean of embeddings leads to biased class prototypes. To support this argument, they developed weighted class prototypes by incorporating the distances of class-wise samples. The weight of sample i in each class is calculated as the inverse of the total distance of sample i to all other samples in that class. The results showed that weighted class prototypes lead to remarkable improvement.

3.3 Learning-based Methods

This work categorizes the three groups of *learning the initialization*, *learning the parameters*, and *learning the optimizer* as a subset of a larger group *learning-based methods*. In previous surveys, three groups were introduced as model-based methods or optimization-based methods, but these methods are mechanically different. This work refers to these three groups as learning-based methods because a network (meta-learner) is trained to learn shared global initializer (learning the initialization), to learn the learning rate (learning the optimizer), or to learn the network’s parameters for the base learner (i.e., the model that predicts for unseen tasks). The following subsections discuss each group in detail.

3.3.1 Learning the initialization

In a standard neural network, weights are initialized randomly, then after calculating loss, optimal weights are obtained by minimizing loss through the gradient descent process. The learning an initialization algorithms seeks to generate a better shared global initializer learned from the support set so that a new unseen task can learn quickly with fewer data points and less optimization steps. The model-agnostic meta-learning (MAML) is a pioneering learning the initialization algorithm proposed by [65]. Intuitively, MAML aims to learn a model’s parameters in such a way that a gradient-based learning rule on a new unseen task T progresses quickly. Figure 15 shows how MAML can lead to improvement on the direction of the gradient. In this figure, three different related tasks are represented with optimal parameters θ'_1 , θ'_2 and θ'_3 , where θ is the randomly initialized parameter. By shifting θ to a new position closer to θ'_1 , θ'_2 and θ'_3 , these optimal parameters can be achieved quickly with fewer steps. The general architecture of MAML consists of a model denoted by function f_θ with parameter θ . The model parameter θ is updated by Eq. 13 :

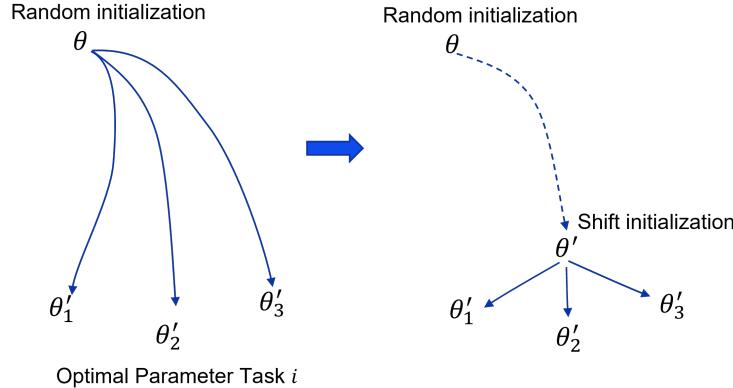


Figure 15: Schema learning initialization, Model agnostic meta-learning for fast adaptation.

$$\theta'_i = \theta - \alpha \nabla_{\theta} \mathcal{L}T_i(f_\theta) \quad (13)$$

where $\mathcal{L}T_i(f_\theta)$ is a gradient of loss for task T_i ; and α is a hyper-parameter. In this step (inner loop), the optimal parameter θ'_i for each task T_i is obtained. In the next step (outer loop), the randomly initialized parameter θ is updated on a new set of tasks by Eq.14. The flow of inner and outer loops update rule is shown in Figure 16. Thus, parameter θ , initialized randomly, shifts to an optimal position where fewer gradient steps are required for the next batch of tasks.

$$\theta = \theta - \beta \nabla \sum_{T_i \sim p(T)} \mathcal{L}T_i(f_\theta) \quad (14)$$

where β is a hyper-parameter; $p(T)$ is distribution over tasks T_i and $\mathcal{L}T_i(f_\theta)$ is the gradient for each new task T_i .

Continuing their previous work,[66] proposed probabilistic MAML, inspired by [67] who interpreted MAML as an approximate inference for the posterior distribution of parameters of the support set. The proposed model, termed probabilistic latent model for incorporating priors and uncertainty (PLATIPUS), was able to capture uncertainty via variational inference. Numerical experiments indicated that PLATIPUS can effectively be used for a diverse family of tasks with multi-modal task distributions. In another tweak to MAML, [68] proposed Bayesian Model-Agnostic Meta-Learning (BMAML) to learn uncertainty, which principally combines the stein variational gradient descent (SVGD) with the MAML algorithm. Their experiments demonstrated that this model is robust to overfitting during

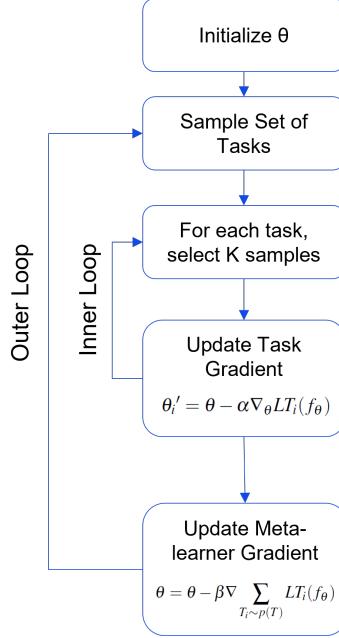


Figure 16: Model agnostic meta-learning workflow.

the meta-update phase. To alleviate the difficulties of learning in high-dimensional parameter space such as such as those faced by MAML, [69] proposed a novel algorithm called latent embedding optimization (LEO). Similar to MAML, LEO consists of an inner loop where the task-specific values are obtained and an outer loop where global shared initializations are updated. Theoretically, LEO improves generalization performance by lowering the dimension of embedding space by using a combination of an encoder and relation network. The workflow of LEO is shown by Figure 17. The encoder produces hidden codes from the support set. Then, hidden codes are concatenated pairwise and fed into a relation network, resulting in a probability distribution over latent codes in a lower dimension. Finally, the decoder generates task-specific initial parameters that are differentiable to backpropagate for adaptation.

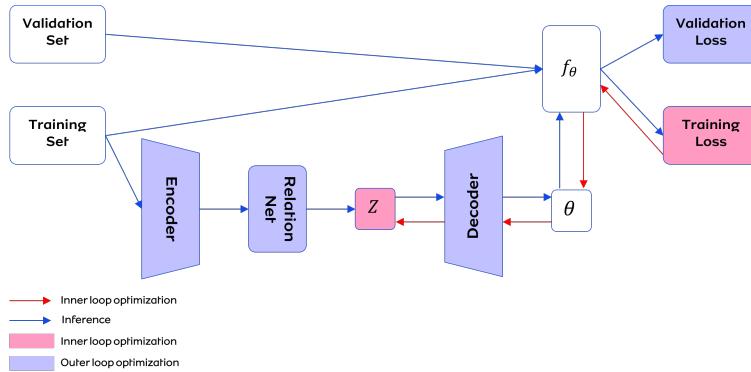


Figure 17: LEO scheme.

While meta-learning algorithms have a remarkable performance on benchmark few-shot problems, they are not robust to the adversarial samples [70]. [71] proposed a modified MAML model called context adaptation for meta-learning (CAML) by introducing context parameters. Context parameters essentially are task-specific parameters updated in an inner loop. In each iteration of the adaptation process, context parameters are initialized with zero value. This modification leads to better memory management and faster adaptation. Additionally, through the numerical examples, the authors demonstrated that CAML outperforms MAML in regression tasks. [72] proposed a network robust to adversarial samples by training a MAML model on both clean and adversarial samples. In the developed ADversarial Meta-Learner (ADML) model, inner and outer gradient updating is performed regarding correlations between clean and adversarial examples. This research opens up a new direction in meta-learning. However, the potential problem of meta-

learning models, such as MAML, with learning the initialization is that they only learn initializers, which can be biased toward some tasks and leads to poor generalization, specifically on unseen tasks. Addressing this issue, [73] proposed a task-agnostic meta-learner (TAML) model, in which two approaches are used to train a model unbiased over tasks. The first approach, "entropy maximization reduction," makes a random guess over predicted labels, so the initial model will have a large entropy. Then, entropy is combined with the meta-learner loss to minimize the entropy, or in other words make the model more confident about the predicted labels. The limitation of the entropy-based approach is that it can only apply to discrete labels. Thus, the authors proposed the second approach, "inequality-minimization," inspired by economic inequalities theory. They used a large group of statistics, including the Gini-coefficient, Theil-index, and variance of logarithm, to compute the tasks bias. They presumed the loss of the initial model of each task T_i as income inequality, which was minimized by the meta-learner.

As can be seen in Eq. 13 and Eq. 14, MAML has two hyperparameters, α and β , that are required to be tuned, which can be expensive. [74] alleviate the tuning burden of MAML hyperparameters by utilizing the hypergradient descent (HD) algorithm (proposed by [75]). Theoretically, HD adapts hyperparameters by backpropagating on the hyperparameters (here, α and β) in conjunction with the original gradient descent iterations. [76] believed that representation change plays an important role in the performance of MAML. Theoretically, the body of the embedding function in the MAML inner loop is almost fixed and the embedding function's head is responsible for updating the features. This mechanic led to only small changes in representations during MAML inner loop task updates, which authors termed as representation reuse. The authors proposed a new variant of the agnostic meta-learner model called body-only update in inner loop (BOIL). Contrary to the previous belief of MAML, in their model, the body of the embedding function updates features, leading to remarkable changes in representations, phrased representation change. Empirical results on benchmark datasets showed that the learning body of the embedding function in the inner loop improves MAML performance. [77] investigated the possible disagreement of the initialization location between tasks as well as layers of neural networks. They argued that sharing an initialization between conflicting tasks, i.e., tasks with unaligned gradient directions, degenerates the optimization landscape. To solve this issue, they proposed a new technique called learn to forget (L2F), which is designed to selectively forget prior knowledge for each task. For this purpose, they proposed a learnable parameter for each layer that is computed through gradients of the support set batch of tasks. Through numerical analysis, the authors showed that the proposed attenuation parameters lead to a stable performance improvement of MAML. [67] empirically showed that MAML can be cast within a hierarchical Bayesian model to estimate task-specific parameters and, thus, proposed another variation of MAML called Laplace approximation for meta-adaptation (LLAMA). More specifically, LLAMA is an extension of MAML that incorporates uncertainty about task-specific parameters, so instead of point estimation, Laplace approximation is applied to obtain the probability distribution over parameter tasks T_i . While LLAMA resulted in a more accurate estimate than MAML's point estimation, Laplace approximation might be quite inaccurate. [78] proposed another extension to MAML called implicit MAML (iMAML). Figure 18 illustrates the difference between MAML and iMAML. While MAML differentiates through the optimization path, iMAML estimates local curvature for the exact meta-gradient.

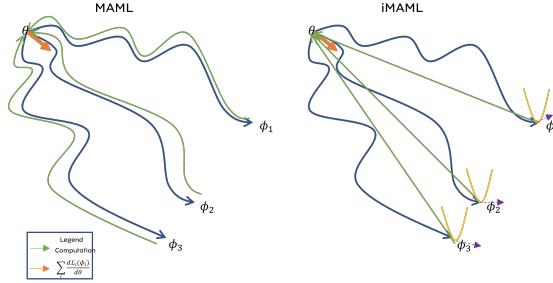


Figure 18: Implicit MAML (iMAML) approach. While MAML differentiates through the optimization path, iMAML estimates local curvature for the exact meta-gradient

3.3.2 Learning the parameters

In this group of algorithms, the parameters of task-specific networks (i.e., base learner) are generated by another learner (i.e., meta-learner) instead of being initialized randomly, leading to a faster adaptation for new tasks. One early work in this category was conducted by [79], who proposed neural statistician, a novel extension of the variational auto-encoder that summarizes dataset and learn dataset statistics. In simple words, a neural statistician generates summary statistics,

including mean and variance, to provide a generative model that can be trained unsupervised and is used to estimate the posterior distribution of task-specific parameters. However, the major limitation of the neural statistician is that its performance depends on a large dataset. [80] proposed a meta-learner model called MetaNet to quickly parameterize a predictive model (i.e., a neural network) by processing meta information. MetaNet consists of two main modules: the base learner operates at the level of tasks and provides the meta-learner with information including loss gradients; and the meta-learner principally generates weights quickly in a task-agnostic setting by processing received information. These generated weights are employed to leverage both the meta- and base learners. In other words, the meta-learner parameterizes itself and the base learner. Here, the meta-learner benefits an external memory to store the support set inputs, their representations, and generated example level weights. Prediction over the query set is made based on comparing the query sample's representation against stored representation in memory. In other words, weights of the base learner would compute based on the similar support set information in memory. Finally, the base learner makes predictions based on the computed weights and seen query points. [21] targeted the one-shot classification problem and proposed non-iterative feed-forward neural nets (called learnet) to predict the parameters of a task-specific predictor (named pupil network). While their model architect may be reminiscent of Siamese networks, the meta-learner (i.e., learnet) is dynamic and changes regarding the output of mappings. [81] proposed a novel model called LGM-Net consisting of two main components: (i) TargetNet, which operates in task-specific environment; and (ii) MetaNet, which is responsible for generating parameters for TargetNet. This MetaNet is different from the Meta-Net model presented by [80] and consists of two modules: (i) task context encoder and (ii) weight generator. The task context encoder is responsible for learning representations, while the weight generator is designed to compute the conditional distributions of weights. The authors used matching networks as TargetNet which use generated weights to compute matching probability score as well as classification loss. Another novelty of this study is the introduction of intertask normalization, a strategy that allows similar tasks in a batch of tasks to interact with each other.

[82] proposed a dynamic algorithm by introducing an attention-based classification weight generator and a cosine-similarity based ConvNet recognition model. The former is responsible for producing weights vectors for unseen classes by seeing a few examples from those classes. The authors argued that it is not feasible for a ConvNet model with a typical linear layer classifier to handle classification weights of both base and novel classes. They address this issue by proposing the second component that provides prediction by computing the cosine similarity between the generated weights and samples' representations. The combination of these two components allowed the authors to achieve fast adaptation for novel classes without forgetting the base classes. Continuing the previous work and with the same belief of developing the meta-learner without forgetting base classes, [83] proposed a denoising auto-encoder network (DAE) in the form of graph neural network (GNN), which is responsible for regularizing the weight generating meta-model. Basically, DAE takes contaminated classification weights with Gaussian noise and generates classification weights based on the conditional distribution of seen data from both novel and base classes. Results showed that this model outperforms their previous model in [82].

[84] argued that embedding vectors are comparable with the weights of the last linear layer of the ConvNet classifier. Accordingly, instead of using two different components of meta-learner and base learner, they proposed a new process called weight imprinting that basically sets final layer weights based on normalized embeddings of the novel class's training samples. Figure 19 shows the overall flow of their model, in which an embedding function first trains on base classes then tunes on novel classes where imprint weights are used in the final layers. With a similar belief, [85] addressed the issue of forgetting the base classes by proposing incremental learning. Formally, the classifier trains on base classes and learns a set of embeddings. Then, the meta-learner trains on novel classes. This study also introduced an additional regularizer that is optimized by combining both novel and base classes and is computed by an attention attractor network, which solves the catastrophic forgetting problem.

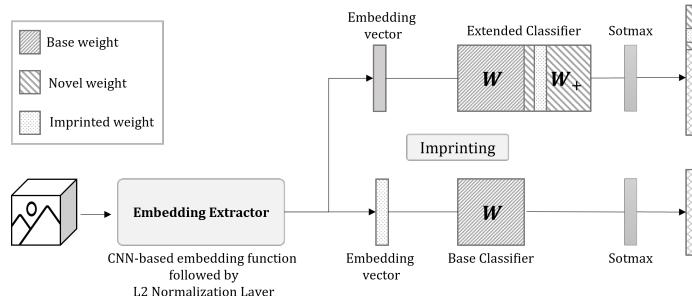


Figure 19: An overview of the weight imprinting [84] process. Feature vectors of latent space are utilized as imprinted weights in a fully connected layer of the extended classifier which is used for the prediction of new query point. Source: [84]

Until now, the reviewed studies in this section focused on generating parameters for a classifier network. [86] proposed a new model called task-aware feature embeddings for low-shot learning (TAFE-Net) that concentrates on generating parameters of feature layers to construct feature embeddings tuned for a given task. TAFE-Net consists of two components: (i) meta-learner that is designed to generate parameters for feature layers in the prediction network; and (ii) prediction network, which generates a binary label that indicates that the input image x is compliant to task description T . The authors used well-known pre-trained deep neural networks, such as ResNet, to get the features of input images. These generated features feed to the dynamic feature layers. The output of this layer is called task-aware feature embedding since feature embedding of each image may vary depending on the task description. Similarly, [87] introduced the meta-transfer learner (MTL) model, focusing on developing task-specific feature extractors by leveraging both transfer learning and meta-learning. Conceptually, they trained a deep neural network on a large dataset, such as miniImageNet dataset, and froze the feature extractor by scaling and shifting operations on pre-trained feature embeddings. This work also introduced a novel hard task meta-batch process, which increases the meta-learner focus on hard tasks through sampling cases that the classifier failed.

3.3.3 Learning the optimizer

A standard neural network trains by calculating loss and using gradient descent to minimize the loss. The key idea of learning the optimizer is to learn gradient descent by replacing it with the recurrent neural network (RNN). Figure 20 shows the high-level mechanics of learning the optimizer algorithms. The base learner's parameters are updated through an RNN network (meta-learner). Since RNN optimizes itself through gradient descent, these models are also known as learn gradient descent by gradient descent. [11] proposed an LSTM-based meta-learner to learn the optimization process used to train task-specific learner.

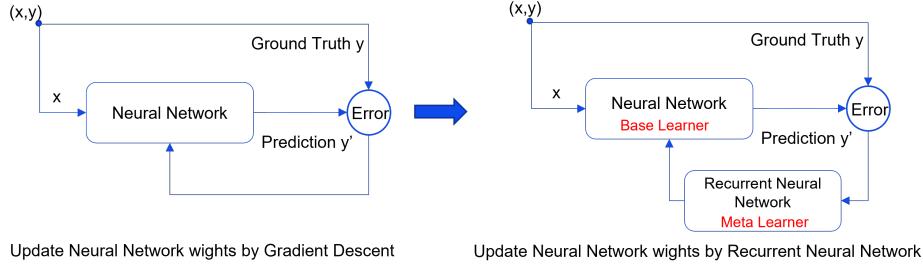


Figure 20: Learning the optimizer conceptual scheme.

Equation 15 denotes the standard gradient descent update rule:

$$\theta_t = \theta_{t-1} - \alpha_t \nabla_{\theta_{t-1}} \mathcal{L}_{\mathfrak{T}_j}(\theta_{t-1}) \quad (15)$$

where θ_{t-1} is the parameter of the learner at $t - 1$ step; α_t is the learning rate; and $\nabla_{\theta_{t-1}} \mathcal{L}_{\mathfrak{T}_j}$ is the gradient of loss for task T parameterized with θ at step $t - 1$. The authors argued that Eq. 15 resembles the update cell in LSTM and replaced term $-\alpha_t \nabla_{\theta_t} \mathcal{L}_{\mathfrak{T}_j}(\theta_t)$ by LSTM network denoted by g and parameterized with φ . Thus, Eq. 15 becomes Eq. 16:

$$\theta_{t+1} = \theta_t = g_\varphi(\nabla_{\theta_t} \mathcal{L}_{\mathfrak{T}_j}(\theta_t)) \quad (16)$$

With this new update rule (Eq. 16), the meta-learner (i.e., LSTM) is able to learn the learning rate θ , which leads to task adaptation quickly. Meta stochastic gradient descent (Meta-SGD) is another learning the optimizer algorithm proposed by [88], which is able to learn the learning rate on top of learning initialization too. Recall the standard gradient descent update rule (Eq. 15), Meta-SGD proposed a new update rule as Eq. 17:

$$\theta_t' = \theta_{t-1} - \alpha \odot \nabla_{\theta_{t-1}} \mathcal{L}_{\mathfrak{T}_j}(\theta_{t-1}) \quad (17)$$

where α is a vector instead of a scalar with the same shape as θ . The operator \odot denotes the element-wise product. In an inner loop, task-specific parameters are updated, and in an outer loop, meta-related optimization is performed on both θ and α by Eq.18 and 19.

$$\theta_t = \theta_{t-1} - \beta \nabla_{\theta, t-1} \sum_{T_i \sim p(T)} \mathcal{L}_{\mathfrak{T}_j}(f(\theta_{i, t-1}')) \quad (18)$$

$$\alpha_t = \alpha_{t-1} - \beta \nabla_{\alpha, t-1} \sum_{T_i \sim p(T)} \mathcal{L}_{\mathfrak{T}_j}(f(\theta_{i, t-1}')) \quad (19)$$

Another learning the optimizer algorithm called Reptile proposed by [89] tries to generate a task-specific learner close to optimal values. In this regard, the Reptile concept is similar to MAML but quite simpler. Figure 21 shows the mechanism of Reptile. Assume two tasks, T_1 and T_2 , sampled from task distribution $p(T_i)$ randomly, where each task is initialized randomly with parameter θ . For each task, T_1 and T_2 , optimal parameters θ_1' and θ_2' will be obtained through gradient descent after n steps. Then, a random parameter θ will shift in the direction closer to both obtained parameters, θ_1' and θ_2' , by minimizing the Euclidean distance of parameter θ and task-specific optimal parameters θ_1' and θ_2' , as shown in Eq. 20:

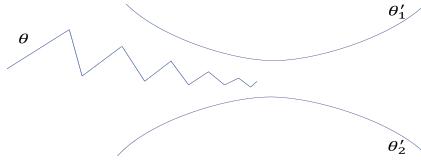


Figure 21: Reptile scheme shows the movement of the random parameter θ towards tasks optimal parameters.

$$\text{minimize } E[\frac{1}{2}D(\theta, \theta_i')] \quad (20)$$

Finally, gradient descent is performed on Eq. 20 to obtain the following update rule (Eq. 21):

$$\theta = \theta - \epsilon(\theta_i' - \theta) \quad (21)$$

where ϵ is a step size parameter. [90] introduced a new idea about the existing learning the optimizer methods, that is using a non-deep machine learning algorithm to adapt new tasks from few samples. Let's consider a pre-trained network (typically, a CNN-based network) $\phi(x) : \mathbb{R}^m \rightarrow \mathbb{R}^e$ (where \mathbb{R}^m and \mathbb{R}^e denote the input and embedding space, respectively); and episode-specific predictor $f(\phi(x), w_\varepsilon) : \mathbb{R}^e \times \mathbb{R}^p \rightarrow \mathbb{R}^o$ (where \mathbb{R}^p and \mathbb{R}^o refer to the predictor parameters and output spaces, respectively; and $w_\varepsilon \in \mathbb{R}^p$ is a set of task-specific predictor parameters). Given the support set of episode, $\varepsilon, z_\varepsilon = \{(x_i, y_i) \sim \varepsilon\}$ and query set of episode $\varepsilon, z'_\varepsilon = \{(x'_i, y'_i) \sim \varepsilon\}$, f can be considered as a final layer of embedding function denoted by $f(\phi(x)) = \phi(x)W, (w_\varepsilon \sim W)$. The authors argued that f can be trained to obtain parameter w_ε through a learning algorithm Λ , which can be considered as a typical machine learning algorithm applied in an episodically training fashion and incorporating meta-parameters w and ρ . These meta-parameters allow knowledge transfer between episodes. Accordingly, the authors proposed two algorithms, namely Ridge Regression Differentiable Discriminator (R2-D2) and Logistic Regression Differentiable Discriminator (LR-D2). In R2-D2, the optimal W is obtained by Eq.22:

$$\begin{aligned} \Lambda(z) &= \text{argmin} \|XW - Y\|^2 + \lambda \|W\|^2 \\ &= (X^T X + \lambda I)^{-1} X^T Y \end{aligned} \quad (22)$$

where X , Y , and λ are embeddings, outputs, and regularization terms, respectively. Considering the high dimensionality of embeddings, $X^T X \in \mathbb{R}^{e \times e}$ can be computationally expensive since the matrix $X^T X$ grows quadratically. The authors addressed this issue by using the Woodbury formula, as given in Eq.23:

$$W = \Lambda(z) = X^T (X X^T + \lambda I)^{-1} Y \quad (23)$$

With small number of n in the n -shot problem, the computation cost of Eq.23 is generally small. Although Eq.22 originally works for regression tasks, the authors modified it to Eq.24 for classification tasks:

$$\hat{Y} = \alpha X' W + \beta \quad (24)$$

The meta-learner learns α , β , and λ (hyper-parameters of base learner) along with CNN parameters W . The authors asserted that other similar solvers can be employed for differentiable operations. Thus, they proposed the LR-D2 method, which is basically logistic regression that uses Newton's method as an iterative solver. LR-D2, unlike R2-D2, is designed to apply to classification tasks directly. Considering inputs $X \in \mathbb{R}^{n \times e}$ and binary label $y \in \{-1, +1\}^n$, parameters $w_i \in \mathbb{R}^e$ are updated by Eq.25:

$$w_i = (X^T \text{diag}(s_i) X + \lambda I)^{-1} X^T \text{diag}(s_i) z_i \quad (25)$$

where w_i denotes the parameters at iteration i ; and I stands for the identity matrix. Moreover, $s_i = \mu_i(1 - \mu_i)$, $z_i = w_{i-1}^T X + (y - \mu_i)/s_i$ and $\mu_i = \sigma(w_{i-1}^T X)$, where σ is the sigmoid function. Similarly, the authors applied Woodbury's identity on $X^T \text{diag}(s_i) X$ and modified Eq.25 to Eq.26:

$$w_i = X^T (X X^T + \lambda \text{diag}(s_i)^{-1})^{-1} z_i \quad (26)$$

With the same idea of using a machine learning classifier as a base learner, [91] proposed a new method called MetaOptNet. In other words, a support vector machine (SVM) is applied as a base learner incorporated with a differentiable quadratic programming (QP) solver to learn parameters of the embedding function f_ϕ . This method showed to outperform both R2-D2 and LR-D2.

4 Overview, Challenges and Future Directions

Along with the considerable success of deep learning, its impotence in domain generalization and avaricious craving for a large dataset for training have increased the demand for few-shot meta-learning algorithms. Despite the fact that in recent years, meta-learning has attracted remarkable interest from researchers, meta-learning is still in its infancy. Although the Siamese network was originally proposed in 1993, the paper "Siamese neural networks for one-shot image recognition" in 2015 [27] triggered a new era of meta-learning and metric-based algorithms. While Siamese networks are limited to binary classification tasks, other metric-based methods resolve this limitation by training on tasks in an episodic training manner. Prototypical networks bring the idea of using class prototypes to compare every new query point instead of individual comparisons in Siamese networks. This idea was the starting point of many metric-based methods. Relation networks employ a recurrent neural network instead of a fixed similarity function, leading to a dynamic nature of learning domain-related similarity functions. A general view on recent metric-based meta-learners demonstrates that the concentration on inter-class and intra-class relationships and a proposed weighted prototype has been increased. Compared to other non-metric based models, metric-based methods are simpler but more innovative in proposing novel similarity measures. Given the pair-wise comparisons – whether a comparison with class prototypes or direct comparison with support data points – metric-based methods can be computationally expensive, especially with large tasks. The next group, memory-based meta-learning models, have received less attention than other groups, but show strong performance. The key advantage of memory-based models is their robustness to forget base classes as the fruit of external/internal memory. Finally, the last large group, the learning-based methods, comprises the major area of research. The LSTM-based meta-learner was the first to materialize the idea of learning gradient descent by gradient descent. It will not be an overstatement if MAML is considered the most inspiring method and commencement mark of learning the initialization methods. While the learning-based methods outperform other methods on a broader range of tasks [92], they can be computationally expensive since they optimize a base learner for every task [92]. Table 1 compares the performance of the discussed methods on benchmark datasets extracted from the papers. Figures 22, 23, 24 and 25 summarize the results for the omniglot and MiniImageNet 5-way problems. The first impression from the figures is that the 5-shot setting leads to better performance. Regarding the omniglot dataset, Meta-SGD (99.91%), Prototype-relation networks (99.90%), MAML (99.90%), and Global class representation (99.86%) achieve the highest accuracy in 5-way 5-shot problem. It can be concluded that when the tasks are closely related, most proposed methods perform well and closely regardless of their mechanics. However, in more complex datasets where tasks are more distant like miniImageNet, the circumstances change. The multi-local feature relation network (MLFRNet) and multi-prototype network (LMPNet) showed the best performance of metric-based methods on 5-way 5-shot miniImageNet problem with 80.23% and 80.00% accuracy, respectively. This result confirms the above statement that metric-based approaches, despite having less computational complexity, have shown to be very powerful since they have the ability to pay attention to local features and minimize the intra-classes' difference while maximizing the inter-classes' difference. Moreover, further improvement and development of these methods are still needed to achieve higher accuracy, yet it remains unclear whether the current methods are confident in their decisions or not.

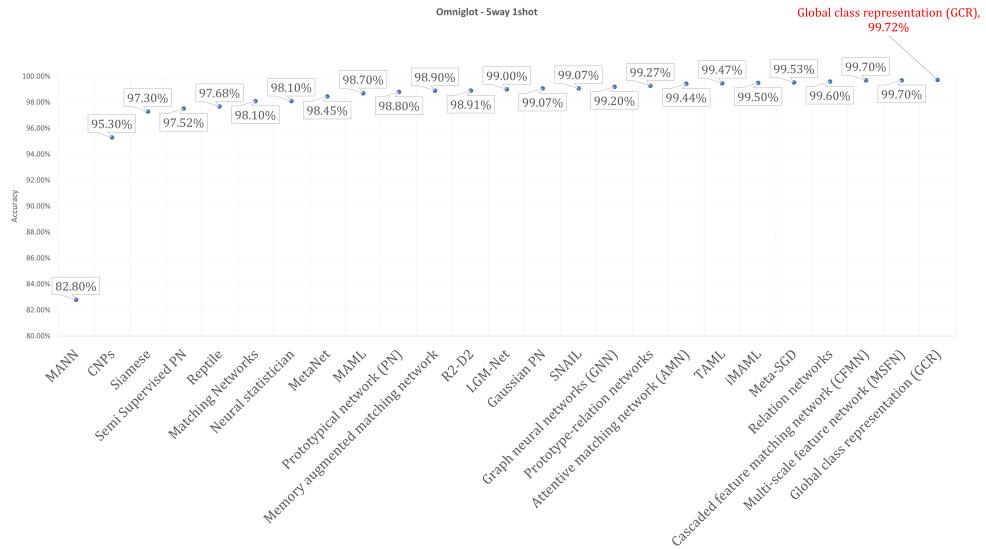


Figure 22: 5way-1shot Omniglot Accuracy results

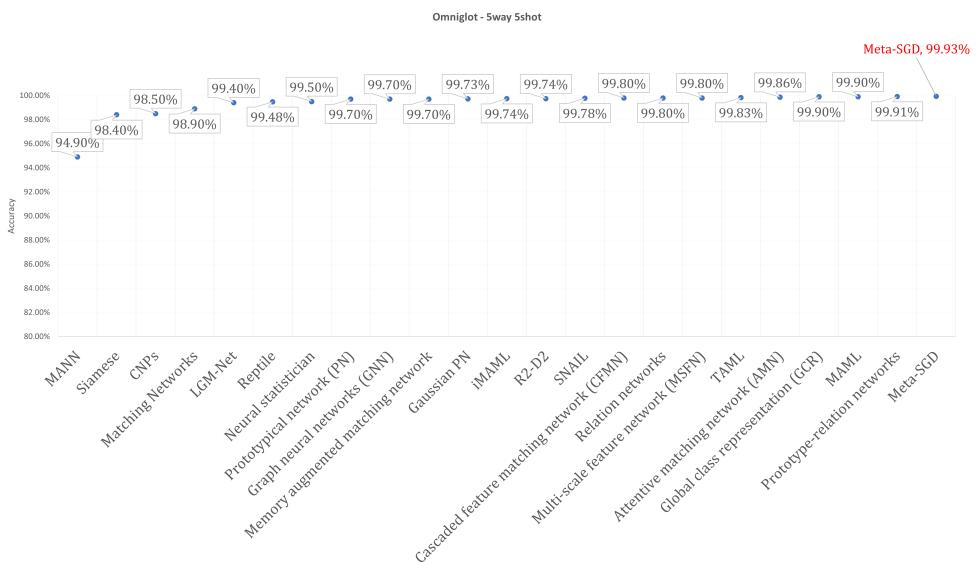


Figure 23: 5way-5shot Omniglot Accuracy results

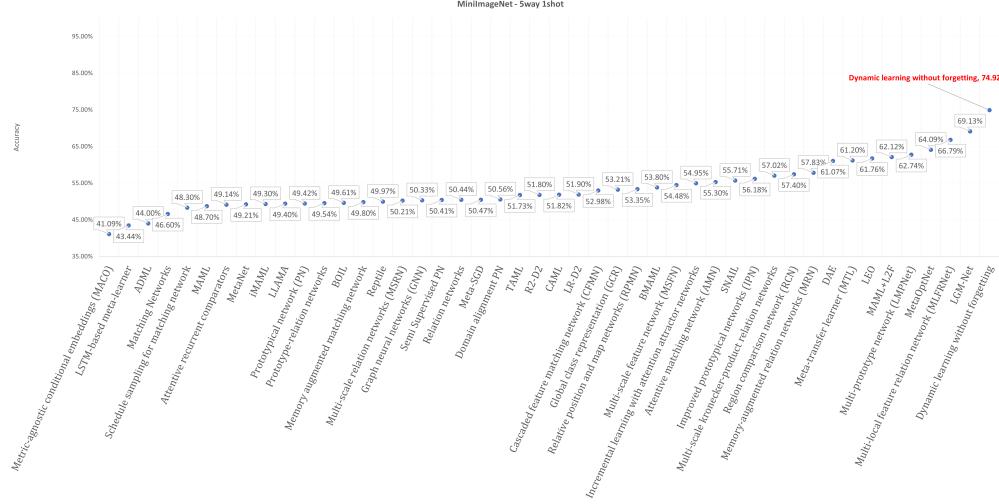


Figure 24: 5way-1shot MiniImageNet Accuracy results

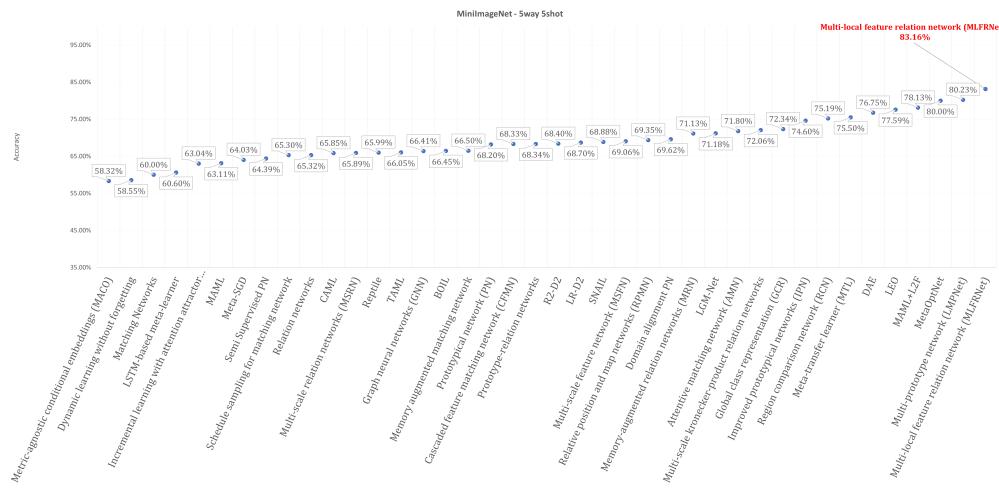


Figure 25: 5way-5shot MiniImageNet Accuracy results

Table 1: Accuracy results on benchmarks datasets reported in original papers, i: Metric-based; ii: Memory-based; iii: Learning the initialization; iv: Learning the parameter; v: Learning the optimizer; The \pm indicates the 95% confidence interval. “_” indicates not reported. The MANN[60] performance is extracted from [63].

Model	Category	Omniglot			Minilmaget			CUB-200-2011					
		1-shot	5-way	5-shot	1-shot	20-way	5-shot	1-shot	5-way	5-shot	1-shot	5-way	5-shot
Siamese[27]	i	97.3%	98.4%	88.2%	97%	98.9%	49.42 ± 0.78%	68.20 ± 0.66%	-	-	-	-	-
Prototypical network (PN)[29]	i	98.8%	99.7%	96%	-	-	50.41 ± 0.31	64.39 ± 0.24	-	-	-	-	-
Semi Supervised PN[30]	i	97.52 ± 0.07%	-	-	-	-	-	-	-	-	-	-	-
Gaussian PN[31]	i	99.07 ± 0.03%	99.73 ± 0.02%	96.94 ± 0.14%	99.29 ± 0.04%	-	-	-	-	-	-	-	-
Transferable prototypical network (TPN)[32]	i	-	-	-	-	-	-	-	-	-	-	-	-
Domain alignment PN[33]	i	-	-	-	-	-	-	-	-	-	-	-	-
Improved prototypical networks (IPN)[34]	i	-	-	-	-	-	-	-	-	-	-	-	-
Label enhanced prototypical network (LPN)[35]	i	-	-	-	-	-	-	-	-	-	-	-	-
Multi-prototypes PN[38]	i	-	-	-	-	-	-	-	-	-	-	-	-
Dummy prototypical networks[39]	i	-	-	-	-	-	-	-	-	-	-	-	-
Multi-prototype network (LMPNet)[40]	i	-	-	-	-	-	-	-	-	-	-	-	-
Multi-modal prototypical network[42]	i	-	-	-	-	-	-	-	-	-	-	-	-
Matching Networks[1]	i	98.1%	98.9%	93.8%	98.7%	-	-	-	-	-	-	-	-
Cascaded feature matching network (CFMN)[43]	i	99.7 ± 0.2%	99.8 ± 0.1%	98.0 ± 0.2%	99.2 ± 0.1%	52.98 ± 0.84%	68.33 ± 0.70%	-	-	-	-	-	-
Attentive matching network (AMN)[44]	i	99.44 ± 0.09%	99.86 ± 0.06%	98.06 ± 0.11%	99.50 ± 0.07%	55.30 ± 0.89%	71.80 ± 0.78%	-	-	-	-	-	-
Schedule sampling for matching network[45]	i	-	-	-	-	-	-	-	-	-	-	-	-
Relation networks[46]	i	99.6 ± 0.2%	99.8 ± 0.1%	97.6 ± 0.2%	99.1 ± 0.1%	50.44 ± 0.82%	65.32 ± 0.70%	-	-	-	-	-	-
Memory-augmented relation networks (MRN)[47]	i	99.27 ± 0.23%	99.91 ± 0.06%	95.97 ± 0.56%	99.32 ± 0.13%	49.54 ± 0.09%	68.34 ± 0.06%	-	-	-	-	-	-
Prototype-relation networks[48]	i	-	-	-	-	-	-	-	-	-	-	-	-
Multi-local feature relation network (MLFRNet)[50]	i	-	-	-	-	-	-	-	-	-	-	-	-
Graph neural networks (GNN)[93]	i	99.2%	99.7%	97.4%	99%	-	-	-	-	-	-	-	-
Global class representation (GCR)[53]	i	99.72 ± 0.06%	99.90 ± 0.10%	99.63 ± 0.09%	99.32 ± 0.04%	53.21 ± 0.40%	72.34 ± 0.32%	-	-	-	-	-	-
Multi-scale feature network (MSFN)[54]	i	99.7%	99.8%	98.1%	99.2%	54.48 ± 1.23%	69.06 ± 0.69%	62.4%	79.14%	-	-	-	-
Attentive recurrent comparators[55]	i	-	-	-	-	-	-	-	-	-	-	-	-
Relative position and map networks (RPMN)[59]	i	-	-	-	-	-	-	-	-	-	-	-	-
Region comparison network (RCN)[56]	i	-	-	-	-	-	-	-	-	-	-	-	-
Multi-scale kronecker-product relation networks[51]	i	-	-	-	-	-	-	-	-	-	-	-	-
Metric-agnostic conditional embeddings (MACO)[57]	i	-	-	-	-	-	-	-	-	-	-	-	-
MANN[60]	ii	82.8%	94.9%	-	-	-	-	-	-	-	-	-	-
SNAIL[62]	ii	99.07 ± 0.16%	99.78 ± 0.09%	97.64 ± 0.30%	99.36 ± 0.18%	55.71 ± 0.99%	68.88 ± 0.92%	-	-	-	-	-	-
CNPs[63]	ii	95.3%	98.5%	89.9%	96.8%	-	-	-	-	-	-	-	-
Memory augmented matching network[64]	ii	98.9%	99.7%	96.3%	98.9%	49.8%	66.5%	-	-	-	-	-	-
MAML[65]	iii	98.7 ± 0.4%	99.9 ± 0.1%	95.8 ± 0.3%	98.9 ± 0.2%	48.70 ± 1.84%	63.11 ± 0.92%	-	-	-	-	-	-
PLATIPUS[66]	iii	-	-	-	-	-	-	-	-	-	-	-	-
BMAML[68]	iii	-	-	-	-	-	-	-	-	-	-	-	-
LEO[69]	iii	-	-	-	-	-	-	-	-	-	-	-	-
CAML[71]	iii	-	-	-	-	-	-	-	-	-	-	-	-
ADML[72]	iii	-	-	-	-	-	-	-	-	-	-	-	-
TAML[73]	iii	99.47 ± 0.25%	99.83 ± 0.09%	95.62 ± 0.5%	98.64 ± 0.13%	51.73 ± 1.88%	66.05 ± 0.85%	-	-	-	-	-	-
Alpha MAML[74]	iii	-	-	-	-	-	-	-	-	-	-	-	-

Table 1: Accuracy results on benchmarks datasets reported in original papers, i: Metric-based; ii: Memory-based; iii: Learning the initialization; iv: Learning the parameter; v: Learning the optimizer; The \pm indicates the 95% confidence interval. “-” indicates not reported. The MANN[60] performance is extracted from [63].

Model	Category	Omniglot			Minimaget			CUB-200-2011				
		1-shot	5-way	5-shot	1-shot	20-way	5-shot	1-shot	5-way	5-shot	1-shot	5-way
BOIL[76]	iii	-	-	-	-	-	-	49.61 \pm 0.16%	66.45 \pm 0.37%	-	-	-
MAML+L2F[77]	iii	-	-	-	-	-	-	62.12 \pm 0.13%	78.13 \pm 0.15%	-	-	-
iMAML[78]	iii	99.50 \pm 0.26%	99.74 \pm 0.11%	96.18 \pm 0.36%	99.14 \pm 0.1%	49.30 \pm 1.88%	-	-	-	-	-	-
LLAMA[67]	iii	-	98.45%	-	-	95.92%	-	-	49.40 \pm 1.83%	-	-	-
MetaNet[80]	iv	-	99%	99.4%	96.5%	98.5%	-	-	49.21 \pm 0.96%	-	-	-
Pupil network[21]	iv	-	-	-	-	-	-	-	-	-	-	-
LGM-Net[81]	iv	-	-	-	-	-	-	74.92 \pm 0.36%	58.55 \pm 0.50%	-	-	-
Dynamic learning without forgetting[82]	iv	-	-	-	-	-	-	61.07 \pm 0.15%	76.75 \pm 0.11%	-	-	-
DAE[83]	iv	-	-	-	-	-	-	-	-	-	-	-
Weight imprinting[84]	iv	-	-	-	-	-	-	-	-	-	-	-
Incremental learning with attention attractor networks[85]	iv	-	-	-	-	-	-	-	-	-	-	-
TAFE-Net[86]	iv	-	-	-	-	-	-	-	-	-	-	-
Meta-transfer learner (MTL)[87]	iv	-	98.1%	99.5%	93.2%	98.1%	-	-	61.2 \pm 1.8%	75.5 \pm 0.8%	-	-
Neural statistician[79]	iv	-	-	-	-	-	-	-	-	-	-	-
LSTM-based meta-learner[11]	v	99.53 \pm 0.26%	99.93 \pm 0.09%	95.93 \pm 0.38%	98.97 \pm 0.19%	50.47 \pm 1.87%	60.60 \pm 0.71%	-	-	-	-	-
Meta-SGD[88]	v	97.68 \pm 0.04%	99.48 \pm 0.06%	89.43 \pm 0.14%	97.12 \pm 0.32%	49.97 \pm 0.32%	65.99 \pm 0.58%	-	-	-	-	-
Reptile[89]	v	-	-	-	-	-	-	64.09 \pm 0.62%	80.00 \pm 0.45%	-	-	-
MetaOptNet[91]	v	98.91 \pm 0.05%	99.74 \pm 0.02%	96.24 \pm 0.05%	99.20 \pm 0.02%	51.8 \pm 0.2%	68.4 \pm 0.2%	-	-	-	-	-
R2-D2[90]	v	-	-	-	-	-	-	51.9 \pm 0.2%	68.7 \pm 0.2%	-	-	-
LR-D2[90]	v	-	-	-	-	-	-	-	-	-	-	-

Despite the promising achievements of meta-learning, some challenges remain. These existing issues along with suggested future research directions are presented here.

- **Uncertainty of meta-learners' performance.** As the name of few-shot learning implies, the main challenge is learning from a handful of samples. This issue has been addressed by meta-learning via sharing knowledge or generating knowledge for each unseen task from previously experienced tasks. In this context, experience is interpreted as fitting to data. However, during episodic training, unrelated tasks or even less related tasks may result in divergent optimization directions as well as disparate performance, which eventually may lead to uncertainty in unseen class prediction [73]. Therefore, training a meta-learner with a confident performance in novel classes needs to be developed. One suggestion worth exploring is training a meta-learner on more diversified tasks and grouping more similar tasks to train in every episode.
- **Effective use of episodic training.** Episodic training is prone to catastrophic forgetting, which results in model underfitting in base classes. While several researches have addressed this issue [34, 82, 83], enhancing model performance on both base and unseen classes remains an important direction for future work. Combining memory-based methods with metric-based methods may be valuable exploration.
- **Improving stability.** Although GCR, a metric-based meta-learner, achieved the best performance on the Omniglot dataset, metric-based models cannot compete with other non-metric-based in miniImageNet, a colored dataset that is more complex than Omniglot. In other words, metric-based meta-learning algorithms are sensitive to the dataset. Moreover, it has been shown that other existing meta-leaning algorithms, such as MAML, are not robust to adversarial samples [70]. While ADML investigates the adversarial samples to improve meta-learner performance, more exploration will be valuable.
- **Training multi-domain, multi-modal and cross-domain meta-learning.** In theory, D_{Base} and D_{Novel} can be from separate domains. However, in practice, as the difference between D_{Base} and D_{Novel} increases, most meta-learning performance will decline. This can be seen by the poor performance of metric-based methods on the miniImageNet dataset. Additionally, the current methods are trained on a single domain D_{Base} . As a suggestion for future research exploration, meta-learning can be developed on multi-domain and cross-domain performance. While training on a multi-domain D_{Base} is expected to increase generalization ability, it can be quite challenging due to divergent optimization direction or the possibility of forgetting D_{Base} tasks. Therefore, using memory-based methods or grouping related tasks based on statistical information is worth exploring. Although [94] recently proposed a new benchmark dataset by combining Omniglot and ImageNet, which is useful for multi-domain and cross-domain evaluation, other benchmark datasets from a combination of real-world applications, such as remote sensing scene images and agricultural pest images, and other real-world domains should be investigated. Last, only a few researches have studied multi-modal domains [42, 66, 37]. Therefore, the development of multi-modal benchmark datasets and methods on multi-modal datasets will be valuable to study. Further research on the multi-modal datasets and real-world multi-domain and cross-domains datasets is pertinent to determine if the current methods are successful in learning from a few samples or simply re-using learned features.
- **Representation learning effects.** Different meta-learning methods use embedding functions with different architectures. While most methods use four-layer convolutional neural networks as a backbone, few studies have applied larger networks, such as resnet, and multi-scale networks as backbone [49, 54]. Accordingly, it may be insightful to analyze the effect of embedding functions, like multi-scale networks, or changing the architecture of the backbone or pre-trained neural networks on state-of-the-art methods such as prototypical networks performance as well as computational costs.
- **Improvement in computational cost.** As mentioned earlier in this section, learning-based methods can be quite computationally expensive in spite of learning from a few samples. Improving computational cost is another important direction for future research.

In conclusion, this study investigated state-of-the-art methods and advances in meta-learning. A clear description of meta-learning is also provided along with a thorough organization of the methods. Whereas deep learning methods focus on data, meta-learning concentrates more on tasks, thus it can be comparatively confusing with other similar approaches such as transfer learning or multi-task learning for researchers who are new to few-shot learning. It is hope that this study and the discussed research challenges and future directions encourage further development in this area.

References

- [1] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

- [2] Jiang Lu, Pinghua Gong, Jieping Ye, and Changshui Zhang. Learning from very few samples: A survey. *arXiv preprint arXiv:2009.02653*, 2020.
- [3] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022.
- [4] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.
- [5] Chelsea B Finn. *Learning to learn with gradients*. University of California, Berkeley, 2018.
- [6] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.
- [7] Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R Arabnia. A brief review of domain adaptation. *Advances in data science and information engineering*, pages 877–894, 2021.
- [8] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.
- [9] Manali Shaha and Meenakshi Pawar. Transfer learning for image classification. In *2018 second international conference on electronics, communication and aerospace technology (ICECA)*, pages 656–660. IEEE, 2018.
- [10] Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf. Transfer learning in natural language processing. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Tutorials*, pages 15–18, 2019.
- [11] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. 2016.
- [12] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.
- [13] Tom M Mitchell and Tom M Mitchell. *Machine learning*, volume 1. McGraw-hill New York, 1997.
- [14] Sebastian Thrun and Lorien Pratt. *Learning to learn*. Springer Science & Business Media, 2012.
- [15] Jane X Wang. Meta-learning in natural and artificial intelligence. *Current Opinion in Behavioral Sciences*, 38:90–95, 2021.
- [16] Mike Huisman, Jan N Van Rijn, and Aske Plaat. A survey of deep meta-learning. *Artificial Intelligence Review*, 54(6):4483–4541, 2021.
- [17] Xiaoxu Li, Zhuo Sun, Jing-Hao Xue, and Zhanyu Ma. A concise review of recent few-shot meta-learning methods. *Neurocomputing*, 456:463–468, 2021.
- [18] Steinar Laenen and Luca Bertinetto. On episodes, prototypical networks, and few-shot learning. *Advances in Neural Information Processing Systems*, 34:24581–24592, 2021.
- [19] Brenden Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua Tenenbaum. One shot learning of simple visual concepts. In *Proceedings of the annual meeting of the cognitive science society*, volume 33, 2011.
- [20] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.
- [21] Luca Bertinetto, João F Henriques, Jack Valmadre, Philip Torr, and Andrea Vedaldi. Learning feed-forward one-shot learners. *Advances in neural information processing systems*, 29, 2016.
- [22] Pierre Baldi and Yves Chauvin. Neural networks for fingerprint recognition. *neural computation*, 5(3):402–418, 1993.
- [23] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. *Advances in neural information processing systems*, 6, 1993.
- [24] Iaroslav Melekhov, Juho Kannala, and Esa Rahtu. Siamese network features for image matching. In *2016 23rd international conference on pattern recognition (ICPR)*, pages 378–383. IEEE, 2016.
- [25] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.
- [26] Olga Ilina, Vadim Ziyadinov, Nikolay Klenov, and Maxim Tereshonok. A survey on symmetrical neural network architectures and applications. *Symmetry*, 14(7):1391, 2022.
- [27] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, page 0. Lille, 2015.

- [28] Davide Chicco. Siamese neural networks: An overview. *Artificial Neural Networks*, pages 73–94, 2021.
- [29] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.
- [30] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018.
- [31] Stanislav Fort. Gaussian prototypical networks for few-shot learning on omniglot. *arXiv preprint arXiv:1708.02735*, 2017.
- [32] Yingwei Pan, Ting Yao, Yehao Li, Yu Wang, Chong-Wah Ngo, and Tao Mei. Transferrable prototypical networks for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2239–2247, 2019.
- [33] Jiang Lu, Zhong Cao, Kailun Wu, Gang Zhang, and Changshui Zhang. Boosting few-shot image recognition via domain alignment prototypical networks. In *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 260–264. IEEE, 2018.
- [34] Zhong Ji, Xingliang Chai, Yunlong Yu, Yanwei Pang, and Zhongfei Zhang. Improved prototypical networks for few-shot learning. *Pattern Recognition Letters*, 140:81–87, 2020.
- [35] Han Liu, Feng Zhang, Xiaotong Zhang, Siyang Zhao, Junjie Sun, Hong Yu, and Xianchao Zhang. Label-enhanced prototypical network with contrastive learning for multi-label few-shot aspect category detection. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1079–1087, 2022.
- [36] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [37] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks and learning systems*, 29(12):5947–5959, 2018.
- [38] Jessica Deuschel, Daniel Firnbach, Carol I Geppert, Markus Eckstein, Arndt Hartmann, Volker Bruns, Petr Kuricyn, Jakob Dexl, David Hartmann, Dominik Perrin, et al. Multi-prototype few-shot learning in histopathology. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 620–628, 2021.
- [39] Byeonggeun Kim, Seunghan Yang, Inseop Chung, and Simyung Chang. Dummy prototypical networks for few-shot open-set keyword spotting. *arXiv preprint arXiv:2206.13691*, 2022.
- [40] Hongwei Huang, Zhangkai Wu, Wenbin Li, Jing Huo, and Yang Gao. Local descriptor-based multi-prototype network for few-shot learning. *Pattern Recognition*, 116:107935, 2021.
- [41] Xiu-Shen Wei, Jian-Hao Luo, Jianxin Wu, and Zhi-Hua Zhou. Selective convolutional descriptor aggregation for fine-grained image retrieval. *IEEE Transactions on Image Processing*, 26(6):2868–2881, 2017.
- [42] Frederik Pahde, Mihai Puscas, Tassilo Klein, and Moin Nabi. Multimodal prototypical networks for few-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2644–2653, 2021.
- [43] Mengting Chen, Xinggang Wang, Heng Luo, Yifeng Geng, and Wenyu Liu. Learning to focus: cascaded feature matching network for few-shot image recognition. *Science China Information Sciences*, 64(9):1–13, 2021.
- [44] Sijie Mai, Haifeng Hu, and Jia Xu. Attentive matching network for few-shot learning. *Computer Vision and Image Understanding*, 187:102781, 2019.
- [45] Lingling Zhang, Jun Liu, Minnan Luo, Xiaojun Chang, Qinghua Zheng, and Alexander G Hauptmann. Scheduled sampling for one-shot learning via matching network. *Pattern Recognition*, 96:106962, 2019.
- [46] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018.
- [47] Jun He, Richang Hong, Xueliang Liu, Mingliang Xu, Zheng-Jun Zha, and Meng Wang. Memory-augmented relation network for few-shot learning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1236–1244, 2020.
- [48] Xiaoqian Liu, Fengyu Zhou, Jin Liu, and Lianjie Jiang. Meta-learning based prototype-relation network for few-shot classification. *Neurocomputing*, 383:224–234, 2020.

- [49] Yueming Ding, Xia Tian, Lirong Yin, Xiaobing Chen, Shan Liu, Bo Yang, and Wenfeng Zheng. Multi-scale relation network for few-shot learning based on meta-learning. In *International Conference on Computer Vision Systems*, pages 343–352. Springer, 2019.
- [50] Li Ren, Guiduo Duan, Tianxi Huang, and Zhao Kang. Multi-local feature relation network for few-shot learning. *Neural Computing and Applications*, 34(10):7393–7403, 2022.
- [51] Mounir Abdelaziz and Zuping Zhang. Multi-scale kronecker-product relation networks for few-shot learning. *Multimedia Tools and Applications*, 81(5):6703–6722, 2022.
- [52] Victor Garcia and Joan Bruna. Few-shot learning with graph neural networks. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.
- [53] Aoxue Li, Tiange Luo, Tao Xiang, Weiran Huang, and Liwei Wang. Few-shot learning with global class representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9715–9724, 2019.
- [54] Mengya Han, Ronggui Wang, Juan Yang, Lixia Xue, and Min Hu. Multi-scale feature network for few-shot learning. *Multimedia Tools and Applications*, 79(17):11617–11637, 2020.
- [55] Pranav Shyam, Shubham Gupta, and Ambedkar Dukkipati. Attentive recurrent comparators. In *International conference on machine learning*, pages 3173–3181. PMLR, 2017.
- [56] Zhiyu Xue, Lixin Duan, Wen Li, Lin Chen, and Jiebo Luo. Region comparison network for interpretable few-shot image classification. *arXiv preprint arXiv:2009.03558*, 2020.
- [57] Nathan Hilliard, Lawrence Phillips, Scott Howland, Artém Yankov, Courtney D Corley, and Nathan O Hodas. Few-shot learning with metric-agnostic conditional embeddings. *arXiv preprint arXiv:1802.04376*, 2018.
- [58] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. *Advances in neural information processing systems*, 30, 2017.
- [59] Zhiyu Xue, Zhenshan Xie, Zheng Xing, and Lixin Duan. Relative position and map networks in few-shot learning for image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 932–933, 2020.
- [60] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850. PMLR, 2016.
- [61] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.
- [62] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. *arXiv preprint arXiv:1707.03141*, 2017.
- [63] Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and SM Ali Eslami. Conditional neural processes. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2018.
- [64] Kien Tran, Hiroshi Sato, and Masao Kubo. Memory augmented matching networks for few-shot learnings. *International Journal of Machine Learning and Computing*, 9(6), 2019.
- [65] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.
- [66] Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. *Advances in neural information processing systems*, 31, 2018.
- [67] Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting gradient-based meta-learning as hierarchical bayes. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.
- [68] Jaesik Yoon, Taesup Kim, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn. Bayesian model-agnostic meta-learning. *Advances in neural information processing systems*, 31, 2018.
- [69] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960*, 2018.
- [70] Micah Goldblum, Liam Fowl, and Tom Goldstein. Adversarially robust few-shot learning: A meta-learning approach. *Advances in Neural Information Processing Systems*, 33:17886–17895, 2020.
- [71] Luisa M Zintgraf, Kyriacos Shiarlis, Vitaly Kurin, Katja Hofmann, and Shimon Whiteson. Caml: Fast context adaptation via meta-learning. 2018.

- [72] Chengxiang Yin, Jian Tang, Zhiyuan Xu, and Yanzhi Wang. Adversarial meta-learning. *arXiv preprint arXiv:1806.03316*, 2018.
- [73] Muhammad Abdullah Jamal and Guo-Jun Qi. Task agnostic meta-learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11719–11727, 2019.
- [74] Harkirat Singh Behl, Atilim Güneş Baydin, and Philip HS Torr. Alpha maml: Adaptive model-agnostic meta-learning. *arXiv preprint arXiv:1905.07435*, 2019.
- [75] Atilim Gunes Baydin, Robert Cornish, David Martinez Rubio, Mark Schmidt, and Frank Wood. Online learning rate adaptation with hypergradient descent. *arXiv preprint arXiv:1703.04782*, 2017.
- [76] Jaehoon Oh, Hyungjun Yoo, ChangHwan Kim, and Se-Young Yun. Boil: Towards representation change for few-shot learning. *arXiv preprint arXiv:2008.08882*, 2020.
- [77] Sungyong Baik, Seokil Hong, and Kyoung Mu Lee. Learning to forget for meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2379–2387, 2020.
- [78] Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. *Advances in neural information processing systems*, 32, 2019.
- [79] Harrison Edwards and Amos Storkey. Towards a neural statistician. *arXiv preprint arXiv:1606.02185*, 2016.
- [80] Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *International Conference on Machine Learning*, pages 2554–2563. PMLR, 2017.
- [81] Huaiyu Li, Weiming Dong, Xing Mei, Chongyang Ma, Feiyue Huang, and Bao-Gang Hu. Lgm-net: Learning to generate matching networks for few-shot learning. In *International conference on machine learning*, pages 3825–3834. PMLR, 2019.
- [82] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4367–4375, 2018.
- [83] Spyros Gidaris and Nikos Komodakis. Generating classification weights with gnn denoising autoencoders for few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21–30, 2019.
- [84] Hang Qi, Matthew Brown, and David G Lowe. Low-shot learning with imprinted weights. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5822–5830, 2018.
- [85] Mengye Ren, Renjie Liao, Ethan Fetaya, and Richard Zemel. Incremental few-shot learning with attention attractor networks. *Advances in Neural Information Processing Systems*, 32, 2019.
- [86] Xin Wang, Fisher Yu, Ruth Wang, Trevor Darrell, and Joseph E Gonzalez. Tafe-net: Task-aware feature embeddings for low shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1831–1840, 2019.
- [87] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 403–412, 2019.
- [88] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*, 2017.
- [89] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.
- [90] Luca Bertinetto, Joao F Henriques, Philip Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations*, 2018.
- [91] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10657–10665, 2019.
- [92] Chelsea Finn and Sergey Levine. Meta-learning and universality: Deep representations and gradient descent can approximate any learning algorithm. In *International Conference on Learning Representations*, 2018.
- [93] Victor Garcia and Joan Bruna. Few-shot learning with graph neural networks. *arXiv preprint arXiv:1711.04043*, 2017.
- [94] Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, et al. Meta-dataset: A dataset of datasets for learning to learn from few examples. *arXiv preprint arXiv:1903.03096*, 2019.