

Meta-Learning for Few-Shot Time Series Classification

Jyoti Narwariya, Pankaj Malhotra, Lovekesh Vig, Gautam Shroff, Vishnu TV
jyoti.narwariya@tcs.com, malhotra.pankaj@tcs.com, lovekesh.vig@tcs.com, gautam.shroff@tcs.com, vishnu.tv@tcs.com
TCS Research
New Delhi, India

ABSTRACT

Deep neural networks (DNNs) have achieved state-of-the-art results on time series classification (TSC) tasks. In this work, we focus on leveraging DNNs in the often-encountered practical scenario where access to labeled training data is difficult, and where DNNs would be prone to overfitting. We leverage recent advancements in gradient-based meta-learning, and propose an approach to train a residual neural network with convolutional layers as a meta-learning agent for few-shot TSC. The network is trained on a diverse set of few-shot tasks sampled from various domains (e.g. healthcare, activity recognition, etc.) such that it can solve a target task from another domain using only a small number of training samples from the target task. Most existing meta-learning approaches are limited in practice as they assume a fixed number of target classes across tasks. We overcome this limitation in order to train a common agent across domains with each domain having different number of target classes, we utilize a triplet-loss based learning procedure that does not require any constraints to be enforced on the number of classes for the few-shot TSC tasks. To the best of our knowledge, we are the first to use meta-learning based pre-training for TSC. Our approach sets a new benchmark for few-shot TSC, outperforming several strong baselines on few-shot tasks sampled from 41 datasets in UCR TSC Archive. We observe that pre-training under the meta-learning paradigm allows the network to quickly adapt to new unseen tasks with small number of labeled instances.

KEYWORDS

Time Series Classification, Meta-Learning, Few-Shot Learning, Convolutional Neural Networks

1 INTRODUCTION

Time series data is ubiquitous in the current digital era with several applications across domains such as forecasting, healthcare, equipment health monitoring, and meteorology among others. Time series classification (TSC) has several practical applications such as disease diagnosis from time series of physiological parameters [4], classifying heart arrhythmias from ECG signals [28], and human activity recognition [43]. Recently, deep neural networks (DNNs) such as those based on long short term memory networks (LSTMs) [17] and 1-dimensional convolution neural networks (CNNs) [9, 18, 40] have achieved state-of-the-art results on TSC tasks. However, it is well-known that DNNs are prone to overfitting, especially when access to a large labeled training dataset is not available. [10, 18].

Few recent attempts aim to address the issue of scarce labeled data for univariate TSC (UTSC) by leveraging transfer learning

[44] via DNNs, e.g. [10, 18, 22, 36]. These approaches consider pre-training a deep network in an unsupervised [22] or supervised [10, 18, 36] manner using a large number of time series from diverse domains, and then fine-tune the pre-trained model for the target task using labeled data from target domain. However, these transfer learning approaches for TSC based on pre-training a network on large number of diverse time series tasks do not necessarily guarantee a pre-trained model (or network initialization) that can be quickly fine-tuned with a very small number of labeled training instances, and rather rely on ad-hoc fine-tuning procedures.

Rather than learning a new task from scratch, humans leverage their pre-existing skills by fine-tuning and recombining them, and hence are highly data-efficient, i.e. can learn from as little as one example per category [27]. Meta-learning [34] approaches intend to take a similar approach for *few-shot learning*, i.e. learning a task from few examples. More recently, several approaches for few-shot learning for regression, image classification, and reinforcement learning domains have been proposed under the gradient-based meta-learning or the “learning to learn” framework, e.g. in [11, 24, 30]. A neural network-based meta-learning model is explicitly trained to quickly learn a new task from a small amount of data. The model learns to solve several tasks sampled from a given distribution where each task is, for example, an image classification problem with few labeled examples. Since each task corresponds to a learning problem, performing well on a task corresponds to learning quickly.

Despite the advent of aforementioned pre-trained models for time series, *few-shot learning* (i.e. learning from few, say five, examples per class) for TSC remains an important and unaddressed research problem. The goal of few-shot TSC is to train a model on large number of diverse few-shot TSC tasks such that it can leverage this experience through the learned parameters, and quickly generalize to new tasks with small number of labeled instances. More specifically, we train a residual network (ResNet) [9, 40] on several few-shot TSC tasks such that the ResNet thus obtained generalizes to solve new few-shot learning tasks. In contrast to existing methods for data-efficient transfer learning, our method provides a way to directly optimize the embedding itself for classification, rather than an intermediate bottleneck layer such as the ones proposed in [22, 36].

Key contributions of this work are:

- We define the problem of few-shot learning for univariate TSC (UTSC), and propose a training and evaluation protocol for the same.
- We propose a few-shot UTSC approach by training a ResNet to solve diverse few-shot UTSC tasks using a meta-learning procedure [24]. The ResNet thus obtained can be quickly adjusted (fine-tuned) on a new, previously unseen, UTSC task with few labeled examples per class.

- As opposed to fixed N -way classification setting in most existing few-shot methods, our approach can handle multi-way classification problems with varying number of classes without introducing any additional task-specific parameters to be trained from scratch such as those in the final classification layer [18, 36]: In order to generalize across few-shot tasks with varying number of classes, we leverage triplet loss [35, 42] for training the ResNet. This allows our approach to leverage the same neural network architecture across diverse applications without introducing any additional task-specific parameters to be trained from scratch.
- Since the proposed approach uses triplet loss to learn a Euclidean embedding for time series, it can also be seen as a data-efficient metric learning procedure for time series that can learn from very small number of labeled instances.

In few-shot setting, we demonstrate that a vanilla nearest-neighbor classifier over the embeddings obtained using our approach outperforms existing nearest-neighbor classifiers based on the highly effective dynamic time warping (DTW) classifier [2] and even state-of-the-art time series classifier BOSS [32]. The rest of the paper is organized as follows: we contrast our work to existing literature in Section 2. We define the problem of few-shot learning for UTSC in Section 3. We then provide details of the neural network architecture used for training the few-shot learner in Section 4 followed by the details of meta-learning based training algorithm for few-shot UTSC in Section 5. We provide details of empirical evaluation of proposed approach in Section 6 and conclude in Section 7.

2 RELATED WORK

Several approaches have been proposed to deal with scarce labeled data for TSC, via data augmentation, warping, simulation, transfer learning, etc. in e.g. [6, 8, 18, 20]. Regularization in DNNs, e.g. decorrelating convolutional filter weights [25] has been found to be effective for TSC and avoid overfitting in scarce data scenarios. Iterative semi-supervised learning [41] also addresses scarce labeled data scenario by iteratively increasing the labeled set but assumes availability of a relatively large amount of data albeit initially unlabeled. In this work, we take a different route to deal with scarce labeled data scenarios and leverage gradient-based meta-learning to explicitly train a network to quickly adapt and solve new few-shot TSC tasks.

Transfer learning using pre-trained DNNs has been shown to achieve better classification performance than training DNNs from scratch for TSC: a few instances of pre-trained DNNs for TSC have been recently proposed in e.g. [18, 22, 36]. However, none of these methods are explicitly trained to quickly adapt to a target task and tend to rely on ad-hoc fine-tuning procedures. Furthermore, they do not study the extreme case of few-shot TSC: while [22] relies on training an SVM classifier on top of unsupervised embeddings obtained via a deep LSTM, [18, 36] rely on introducing and training a new final softmax layer from scratch for each new task. Our approach explicitly pre-trains a DNN using triplet loss to optimize for quick adaptation to a few-shot task. Moreover, unlike existing methods, our approach directly optimizes for time series embeddings over which the similarity of time series can be defined, and hence can work in a kNN setting without requiring the training of

additional parameters like those of an SVM in [22], or those of a feedforward final layer in [18, 36].

Several approaches for few-shot learning have been recently introduced for image classification, regression, and reinforcement learning, e.g. [11, 24, 30, 38]. To the best of our knowledge, our work is the first attempt to study few-shot learning for TSC. We formulate the few-shot learning problem for UTSC, and build on top of the following recent advances in deep learning research to develop an effective few-shot approach for TSC: i) gradient-based meta-learning [11, 24], ii) residual network with convolutional layers for TSC [40], iii) leveraging multi-length filters to ensure generalizability of filters to tasks with varying time series length and temporal properties [18, 29], and iv) triplet loss [35] to ensure generalizability to tasks with varying number of classes without introducing any additional parameters.

Dynamic time warping (DTW) and its variants [16, 37] are known to be very robust and strong distance metric baselines for TSC over a diverse set of applications [2]. However, it is also well-known that no single distance metric works well across scenarios as they lack the ability to leverage the data-distribution and properties of the task at hand [2, 39]. It has been shown that k-nearest-neighbor (kNN) TSC can be significantly improved by learning a distance metric from labeled examples [1, 23]. Similarly, modeling time series similarity using Siamese recurrent networks based supervised learning has been proposed in [26]. CNNs trained using triplet loss for TSC have been very recently proposed for unsupervised learning in [13] and for supervised learning in [3]. However, to the best of our knowledge, none of the metric learning approaches consider pre-training a neural network that can be quickly fine-tuned for new TSC few-shot tasks.

3 PROBLEM DEFINITION

Consider a K -shot learning problem for UTSC sampled from a distribution $p(\mathcal{T})$ that requires learning a multi-way classifier for a test task given only K labeled time series instances per class. Rather than training a classifier from scratch for the test task, the goal is to obtain a neural network with parameters ϕ that is trained to efficiently (e.g. in a few iterations of updates of ϕ via gradient descent) solve several K -shot learning tasks sampled from $p(\mathcal{T})$. These K -shot tasks are divided into three sets: a *training meta-set* S^{tr} , a *validation meta-set* S^{va} , and a *testing meta-set* S^{te} . The training meta-set is used to obtain the parameters ϕ , the validation meta-set is used for model selection (hyperparameters for neural network training), and the testing meta-set is used only for final evaluation.

Each task instance $\mathcal{T}_j \sim p(\mathcal{T})$ in S^{tr} and S^{va} consists of a labeled training set of univariate time series $\mathcal{D}_j^{tr} = \{(x_j^{n,k}, y_j^{n,k}) \mid k = 1 \dots K; n = 1 \dots N_j\}$, where K is the number of univariate time series instances for each of the N_j classes. Ignoring the sub- and super-scripts, each univariate time series $\mathbf{x} = x_1, x_2, \dots, x_T$ with $x_t \in \mathbb{R}$ for $t = 1, \dots, T$, where T is the length of time series, and y is the class label. Unlike the tasks in S^{tr} and S^{va} , which only contain a training set, each task in S^{te} also contains a testing set $\mathcal{D}_j^{te} = \{(x_j^{n,k}, y_j^{n,k}) \mid k = 1 \dots K'; n = 1 \dots N_j\}$ apart from a training set \mathcal{D}_j^{tr} . The classes in \mathcal{D}_j^{tr} and \mathcal{D}_j^{te} are the same while

classes across tasks are, in general, different. For any $x_j^{n,k}$ from \mathcal{D}_j^{te} , the goal is to estimate the corresponding label $y_j^{n,k}$ by using an updated version $\hat{\phi}$ of ϕ obtained by fine-tuning the neural network using the $K \times N_j$ labeled samples from \mathcal{D}_j^{tr} . In other words, the training set \mathcal{D}_j^{tr} of a task $\mathcal{T}_j \in \mathcal{S}^{te}$ is used for fine-tuning the neural network parameters ϕ , while the corresponding testing set \mathcal{D}_j^{te} of the task \mathcal{T}_j is used for evaluation.

It is to be noted that the tasks in the three meta-sets correspond to time series from disjoint sets of classes, i.e. the classes in any task in training meta-set are different from those of any task in validation meta-set, and so on. In practice, we sample the tasks from diverse domains such as electric devices, motion capture, spectrographs, sensor readings, ECGs, simulated time series, etc. taken from the UCR TSC Archive [5]. Each dataset, and in turn tasks sampled from it, have a different notion of classes depending upon the domain, a different number of classes N , and a different T .

4 NEURAL NETWORK

As shown in Figure 1, we consider a ResNet consisting of multiple convolutional blocks with shortcut residual connections [14] between them, eventually followed by a global average pooling (GAP) layer such that the network does not have any feedforward layers at the end. Each convolutional block consists of a convolutional layer followed by a batch normalization (BN) layer [15] which acts as a regularizer. Each BN layer is in turn followed by a ReLU layer. We omit further architecture details and refer the reader to [18]. In order to quickly adapt to any unseen task, the neural network should be able to extract temporal features at multiple time scales and should ensure that the fine-tuned network can generalize to time series of varying lengths across tasks. We, therefore, use filters of multiple lengths in each convolutional block to capture temporal features at various time scales, as found to be useful in [3, 18, 29].

In a nutshell, ResNet takes a univariate time series x of any length T as input and converts it to a fixed-dimensional feature vector $z \in \mathbb{R}^m$, where m is the number of filters in the final convolutional layer. We denote the set of all the trainable parameters of the ResNet consisting of filter weights and biases across convolutional layers, and BN layer parameters by ϕ .

Most ResNet implementations for TSC [9, 18, 36, 40] use a feed-forward layer followed by a softmax layer to eventually map z to class probabilities, and use cross-entropy loss for training. Further, when training the ResNet for multiple tasks with varying number of classes across tasks, a multi-head output with different final feedforward layer for each task is typically used, e.g. as in [18, 36]. However, in our setting, this implies a different feedforward layer for each new few-shot task, introducing at least $m \times N_j$ additional task-specific parameters¹ that need to be trained from scratch for each new few-shot task. This is not desirable in a few-shot learning setting given only a small number of K samples per class, as this can lead to overfitting: this is one reason due to which most few-shot learning formulations, e.g. [11, 38], consider a fixed number of target classes across tasks. However, we intend to learn a few-shot learning algorithm that overcomes this limitation. We propose using triplet loss [3, 35, 42] as the training objective which allows for

¹when the GAP layer is followed by a single feedforward layer and a softmax layer

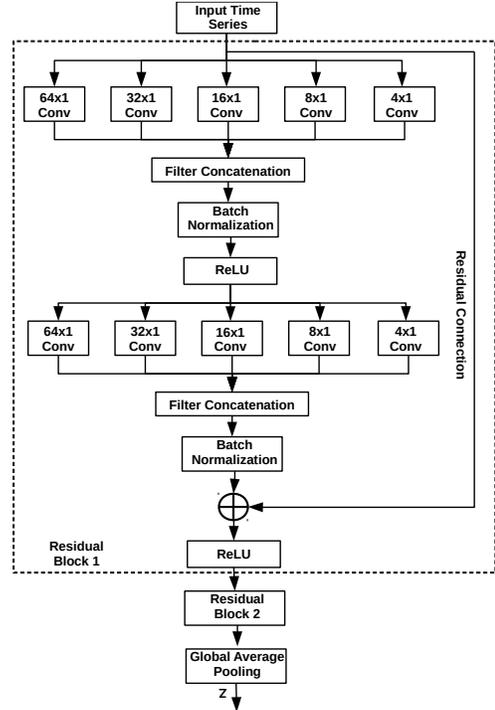


Figure 1: ResNet Architecture depicting two residual blocks each with two convolutional layers, and variable-length filters in each convolutional layer.

generalization to varying number of classes without introducing any additional task-specific parameters, as detailed next.

4.1 Loss Function

Triplet loss relies on pairwise distance between representations of time series samples from within and across classes, irrespective of the number of classes. Using triplet loss at time of fine-tuning for the test task, therefore, allows the network to adapt to a given few-shot classification task without introducing any additional task-specific parameters. Triplets consist of two matching time series and a non-matching time series such that the loss aims to separate the positive pair from the negative by a distance margin. Given the set \mathcal{S}_j of all valid triplets of time series for a training task \mathcal{T}_j of the form $(x_i^a, x_i^p, x_i^n) \in \mathcal{S}_j$ consisting of an anchor time series x_i^a , a positive time series x_i^p , and a negative time series x_i^n ; where the positive time series is another instance from same class as the anchor, while the negative is from a different class than the anchor. We aim to obtain corresponding representations (z_i^a, z_i^p, z_i^n) such that the distance between the representations of an anchor and any positive time series is lower than the distance between the representations of the anchor and any negative time series.

More specifically, we consider triplet loss based on Euclidean norm given by:

$$\|z_i^a - z_i^n\|_2^2 - \|z_i^a - z_i^p\|_2^2 > \alpha, \quad (1)$$

where $\alpha > 0$ is the distance-margin between the positive and negative pairs. The loss to be minimized is then given by:

$$\mathcal{L}_{\mathcal{T}_j} = \sum_{l=1}^{|\mathcal{S}_j|} \left[\|z_l^a - z_l^p\|_2^2 - \|z_l^a - z_l^n\|_2^2 + \alpha \right]_+, \quad (2)$$

where $[z]_+ = \max(z, 0)$, such that only those triplets violating the constraint in Eq. 1 contribute to the loss. Note that since we use triplet loss for training, the number of instances per class $K > 1$.

5 FEW-SHOT LEARNING FOR UTSC

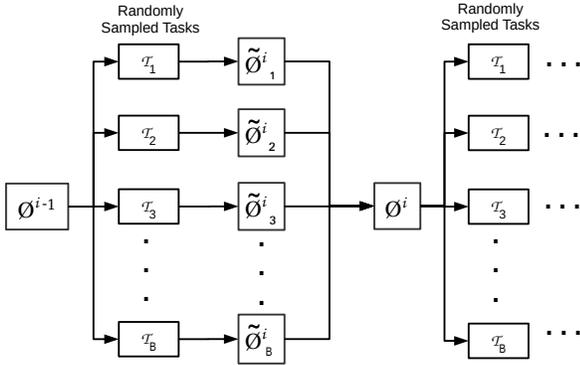


Figure 2: Few-Shot Training Approach.

We consider a meta-learning approach for few-shot UTSC based on Reptile [24], a first-order gradient descent based meta-learning algorithm, and refer to that as *FS-1*. We also consider a simpler variant of this approach and refer to that as *FS-2*: similar to the training procedure of *FS-1*, *FS-2* is also trained to solve multiple UTSC tasks but not explicitly trained in a manner that ensures quick adaptation to any new UTSC task. Except for the triplet loss, *FS-2* is similar to [18, 36] in the way data is sampled and used for training.

5.1 FS-1

5.1.1 Objective. *FS-1* learns an initialization for the parameters ϕ of the ResNet such that these parameters can be quickly optimized using gradient-based learning at test time to solve a new few-shot UTSC task—i.e., the model generalizes from a small number of examples from the test task. In order to learn the parameters ϕ , we train the ResNet on a diverse set of UTSC tasks in \mathcal{S}^{tr} with varying number of classes and time series lengths. As explained in Section 4, the same neural network parameters ϕ are shared across all tasks owing to the fact that: i. ResNet yields a fixed-dimensional representation for varying length time series, and ii. the nature of the loss function that does not require any changes due to the varying number of classes across tasks.

Similar to [11, 24], we consider the following optimization problem: find an initial set of parameters ϕ for the ResNet, such that for a randomly sampled task \mathcal{T}_j with corresponding loss $L_{\mathcal{T}_j}$ as given in Eq. 2, the learner will have low loss after k updates, such that:

$$\text{minimize}_{\phi} \mathbb{E}_{\mathcal{T}_j} \left[\mathcal{L}_{\mathcal{T}_j} \left(U_{\mathcal{T}_j}^k(\phi) \right) \right], \quad (3)$$

where $U_{\mathcal{T}_j}^k$ is the operator (e.g. corresponding to Adam optimizer or SGD) that updates ϕ using k mini-batches from \mathcal{D}^{tr} .

5.1.2 Implementation Details. *FS-1* sequentially samples few-shot tasks from the set of tasks \mathcal{S}^{tr} . As summarized in Algorithm 1 and depicted in Figure 2, the meta-learning procedure consists of M meta-iterations. Each meta-iteration involves sampling B K -shot tasks. Each task, in turn, is solved using k steps of gradient-based optimization, e.g. using stochastic gradient descent (SGD) or Adam [19] – this, in turn, involves randomly sampling mini-batches from the $K \times N$ instances in the task. Each task is associated with a triplet loss defined over the valid triplets as described in Section 4.1.

Given that each task has a varying number of instances owing to varying N , we set the number of iterations for each task to $k = \lfloor \frac{K \times N}{b} \rfloor \times e$, where b is the mini-batch size and e is the number of epochs. Therefore, instead of fixing the number of iterations k for each sampled task, we fix the number of epochs e across datasets, such that the network is trained to adapt quickly in a fixed number of epochs, as described later. Also note that the number of triplets in each batch is significantly more than the number of unique time series in a mini-batch.

Algorithm 1 Few-Shot UTSC Approach-1 (FS-1)

ϕ^0 : initial parameters of the ResNet
for meta-iteration $i = 1, 2, \dots, M$ **do**
 for $j = 1, 2, \dots, B$ **do**
 Sample a K -shot task \mathcal{T}_j
 Get number of classes N_j for task \mathcal{T}_j
 Set $k = \lfloor \frac{K \times N_j}{b} \rfloor \times e$
 Compute $\tilde{\phi}_j^i = U_{\mathcal{T}_j}^k(\phi^{i-1})$ using k steps (mini-batches) of Adam to minimize loss $\mathcal{L}_{\mathcal{T}_j}$
 end for
 Update $\phi^i = \phi^{i-1} + \epsilon \frac{1}{B} \sum_{j=1}^B (\tilde{\phi}_j^i - \phi^{i-1})$
end for

Algorithm 2 Few-Shot UTSC Approach-2 (FS-2)

ϕ^0 : initial parameters of the ResNet
for iteration $i = 1, 2, \dots, M$ **do**
 for $j = 1, 2, \dots, B$ **do**
 Sample a K -shot task \mathcal{T}_j
 Get number of classes N_j for task \mathcal{T}_j
 Set $k = \lfloor \frac{K \times N_j}{b} \rfloor \times e$
 Compute $\phi^{i+j} = U_{\mathcal{T}_j}^k(\phi^{i+j-1})$ using k_j steps (mini-batches) of SGD or Adam to minimize loss $\mathcal{L}_{\mathcal{T}_j}$
 end for
end for

The filter weights of the ResNet are randomly initialized, e.g. via orthogonal initialization [31]. In the i th meta-iteration, ResNet for each of the B tasks is initialized with ϕ^{i-1} . Each task \mathcal{T}_j with labeled data \mathcal{D}_j^{tr} is solved by updating the parameters ϕ^{i-1} of the

network $k (= \lfloor \frac{K \times N_j}{b} \rfloor) \times e$, where N_j is number of classes in \mathcal{T}_j times to obtain

$$\tilde{\phi}_j^i = U_{\mathcal{T}_j}^k(\phi^{i-1}). \quad (4)$$

In practice, we use a batch version of the optimization problem in Equation 3 and use a meta-batch of B tasks to update ϕ as follows:

$$\phi^i = \phi^{i-1} + \epsilon \frac{1}{B} \sum_{j=1}^B (\tilde{\phi}_j^i - \phi^{i-1}). \quad (5)$$

Note that $\tilde{\phi}_j - \phi$ with $k > 1$ implies that ϕ is updated using the updated values $\tilde{\phi}_j$ obtained after solving B tasks for k iterations each. It is this particular way of updating ϕ by internally solving multiple tasks, that this algorithm is considered an example of gradient descent based meta-learning. As shown in [24], when performing multiple gradient updates as per Eqs. 4 and 5, i.e. having $k > 1$ while solving few-shot tasks, then the expected update $\mathbb{E}_{\mathcal{T}_j}[U_{\mathcal{T}_j}^k(\phi)]$ is very different from taking a gradient step on the expected loss $\mathbb{E}_{\mathcal{T}_j}[\mathcal{L}_{\mathcal{T}_j}(\phi)]$, i.e. having $k = 1$. In fact, it is easy to note that the update of ϕ consists of terms from the second-and-higher derivatives of $\mathcal{L}_{\mathcal{T}_j}$ due to the presence of derivatives of $\mathcal{L}_{\mathcal{T}_j}$ in $\tilde{\phi}_j$. Hence, the final solution using $k > 1$ is significantly different from the one obtained using $k = 1$.

5.1.3 Fine-tuning and inference in a test K -shot task. We denote the optimal parameters of ResNet after meta-training as ϕ^* , and use this as initialization of target task-specific ResNet. For any new K -shot N -way test task with labeled instances in \mathcal{D}^{tr} and any test time series \mathbf{x}^* taken from \mathcal{D}^{te} , first ϕ^* is updated to $\tilde{\phi}$ using \mathcal{D}^{tr} . The embeddings for all the $N \times K$ samples in \mathcal{D}^{tr} is compared to the embedding for \mathbf{x}^* using 1NN classifier to get the class estimate.

5.2 FS-2

As shown in Algorithm 2, FS-2 is a simpler variant of FS-1 where instead of updating the parameters ϕ by collectively using updated values from B tasks, ϕ is continuously updated at each mini-batch irrespective of the task. As a result, the network is trained for a few iterations on a task, and then the task is changed. Unlike FS-1, FS-2 uses only the first-order derivatives of $\mathcal{L}_{\mathcal{T}_j}$.

6 EXPERIMENTAL EVALUATION

6.1 Experimental Setup

6.1.1 Sampling few-shot UTSC tasks. We restrict the distribution of tasks to univariate TSC with a constraint on the maximum length of the time series such that $T \leq 512$. We sample tasks from the publicly available UCR Archive of UTSC datasets [5], where each dataset corresponds to a N -way multi-class classification task with number of classes N and the length of time series T varies across datasets. However, all the time series in any dataset are of same length. Each time series is z-normalized using the mean and standard deviation of all the points in the time series.

Out of the total of 65 datasets on UCR Archive with $T \leq 512$, we use 18 datasets to sample tasks for training meta-set \mathcal{S}^{tr} and 6 datasets to sample tasks for the validation meta-set \mathcal{S}^{va} (dataset level splits are same as in [22]). Any task in \mathcal{S}^{tr} or \mathcal{S}^{va} has K randomly sampled time series for each of the N classes in the

dataset. The remaining 41 datasets with length $T \leq 512$ as listed in Table 1 are used to create tasks for the testing meta-set. As a result of this way of creating the training, validation and testing meta-sets, the classes in each meta-set are disjoint. However, the classes in the train and test sets of a task in a testing meta-set is, of course, the same.

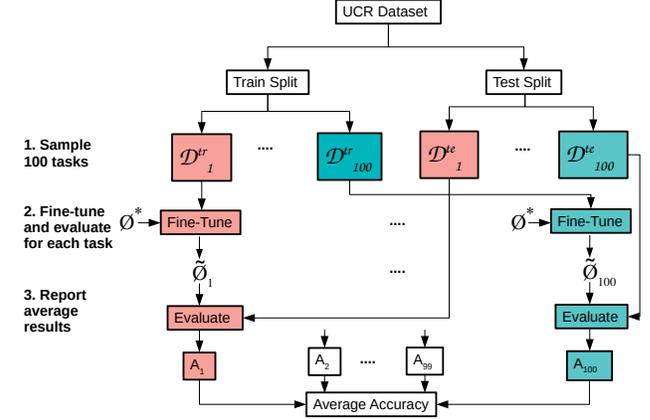


Figure 3: Evaluation protocol for FS-1 and FS-2 on a UCR dataset. For ResNet, ϕ is randomly initialized for each task. A_j is the accuracy on j -th task.

Each dataset in UCR Archive is a N -way classification problem with an original train and test split. As shown in Figure 3, we sample 100 K -shot tasks from each of the 41 datasets. Each task (out of the 100) sampled from a dataset contains K samples from each of the N classes for \mathcal{D}^{tr} and K' samples from each of the N classes for \mathcal{D}^{te} for each task are sampled from the respective original train and test split of the dataset². The K (or K') samples for each class in \mathcal{D}^{tr} (or \mathcal{D}^{te}) are sampled uniformly from the entire set of samples of the respective class. While \mathcal{D}^{tr} is used for fine-tuning ϕ^* to get $\tilde{\phi}$, \mathcal{D}^{te} is used to evaluate the updated task-specific model $\tilde{\phi}$. (Note that while the class distribution in the original dataset may not be uniform, each K -shot task consists of equal number, i.e. K , samples per class.)

6.1.2 Hyperparameters for FS-1 and FS-2. On the basis of initial experiments on a subset of the training meta-set, we use the ResNet architecture with $L = 4$ layers and $m = 165$ convolution filters per layer (33 filters each of length 4,8,16,32,64). We use Adam optimizer with a learning rate of 0.0001 for updating ϕ on each task while using $\epsilon = 1$ in the meta-update step in Equation 5. FS-1 and FS-2 are trained for a total of $M = 2000$ meta-iterations with meta-batch size of $B = 5$, and mini-batch size $b = 10$. We trained FS-1 and FS-2 using $K = 5$ and 10 for the tasks in training meta-set while $K = 5$ is used for validation and testing meta-sets. $K' = 5$ across all experiments unless stated otherwise. We found the model with $K = 10$ for tasks in training meta-set to be better based on average triplet loss on validation meta-set. We use epochs $e = 4$ for

²We also considered the original test split for each test task \mathcal{D}^{te} during evaluation. We obtained similar conclusions under this evaluation strategy as well, and hence, omit those results for brevity.

solving each task while training FS-1 and FS-2 models. The number of epochs e' to be used while fine-tuning for tasks in testing meta-set is chosen from the range 1-100 based on average triplet loss on tasks in validation meta-set. We found $e' = 16$ and 8 to be best for FS-1 and FS-2 models, respectively. Therefore, ϕ^* is fine-tuned for e' epochs for each task in testing meta-set. For the triplet loss, we use $\alpha = 0.5$.

6.1.3 Baselines Considered. For comparison, we consider following baseline classifiers each using 1NN as the final classifier over raw time series or extracted features³:

- (1) **ED:** 1NN based on Euclidean distance is the simplest baseline considered, where time series of length T is represented by a fixed-dimensional vector of the same length. (Note: For any given dataset and subsequent tasks sampled from it, the length T is same across samples, and hence 1NN based on ED is applicable.)
- (2) **DTW:** 1NN based on dynamic time warping (DTW) approach is one of the highly effective and strong baseline for UTSC [2]. We use leave-one-out cross-validation on \mathcal{D}^{tr} of each task to find the best warping window in the range $w = 0.02T, 0.04T, \dots, T$, where w is the window length and T is the time series length.
- (3) **BOSS:** Bag-of-SFA-Symbols [32] is a state-of-the-art time series feature extraction technique that provides time series representations while being tolerant to noise. BOSS provides a symbolic representation based on Symbolic Fourier Approximation (SFA) [33] on each fixed-length sliding window extracted from a time series while providing low pass filtering and quantization for noise reduction. The hyper-parameters, i.e. *wordLength* and *normalization* are chosen based on leave-one-out cross validation over the ranges $\{8, 10, 12, 14, 16\}$ and $\{True, False\}$ respectively, while default values of remaining hyper-parameters is used. 1NN is applied on the extracted features for final classification decision.
- (4) **ResNet:** Instead of using ϕ^* obtained via FS-1 or FS-2 as a starting point for fine-tuning, we consider a ResNet-based baseline where the model is trained from scratch for each task using triplet loss. The architecture is same as those used for FS-1 and FS-2 (also similar to state-of-the-art ResNet versions studied in [9, 18, 40]). Given that each task has a very small number of training samples and the parameters are to be trained from scratch, ResNet architectures are likely to be prone to overfitting despite batch normalization. To mitigate this issue, apart from the same network architecture as FS-1 and FS-2, we also consider smaller networks with smaller number of trainable parameters. More specifically, we considered four combinations resulting from number of layers = $\{\frac{L}{2}, L\}$ and number of filters per layer = $\{\lfloor \frac{m}{2} \rfloor, m\}$, where $L = 4$ and $m = 165$. We consider the model with best overall results amongst these four combinations as baseline, viz. number of layers = 2 and number of filters = 165. For fair comparison, each ResNet model is trained for 16 epochs⁴ as for FS-1.

³For DTW and BOSS, we use implementations as available at <http://www.timeseriesclassification.com/code.php>.

⁴We also tried training till 32 epochs for ResNet and found insignificant improvement in results.

6.1.4 Performance Metrics. Each task is evaluated using classification accuracy rate on the test set—*inference* is correct if the estimated label is same as the ground truth label. Each task consists of $K' \times N$ test samples: the performance results for each task equals the fraction of correctly classified test samples. Further, we follow the methodology from [2, 7] to compare the proposed approach with various baselines considered. For each dataset, we average the classification error results over 100 randomly sampled tasks (as described in Section 6.1.1). To study the relative performance of the approaches over multiple data sets, we compare classifiers by ranks using the Friedman test and a post-hoc pairwise Nemenyi test.

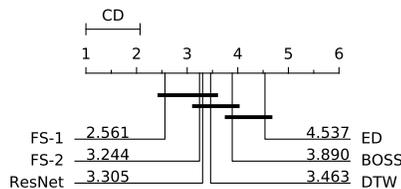


Figure 5: Critical Difference Diagram comparing ranks of few-shot learning approaches (FS-1 and FS-2) with other baselines for $K = 5$ samples per class used for fine-tuning.

Table 2: Comparison of various approaches in terms of ranks over classification accuracy rates on all the 4100 tasks from 41 datasets with varying K . Best approach is marked in bold and second-best is underlined.

K	ED	DTW	BOSS	ResNet	FS-2	FS-1
2	4.232	<u>2.976</u>	3.902	3.805	3.207	2.878
5	4.537	3.463	3.890	3.305	<u>3.244</u>	2.561
10	4.573	3.476	3.646	3.683	<u>3.427</u>	2.195
20	4.439	3.354	<u>2.927</u>	3.902	3.793	2.585

Table 3: Comparison of ranks across datasets with varying number of classes N in 5-shot task and n is the number of datasets.

N	n	ED	DTW	BOSS	ResNet	FS-2	FS-1
2-5	24	4.167	4.083	3.375	3.458	<u>3.042</u>	2.875
6-10	9	4.778	2.333	5.333	<u>2.389</u>	3.778	<u>2.389</u>
>10	8	5.375	2.875	3.812	3.902	3.875	1.812
Overall	41	4.537	3.463	3.890	3.305	3.244	2.561

6.2 Results and Observations

- As shown in Figure 5, we observe that FS-1 improves upon all the baselines considered for 5-shot tasks. The pairwise comparison of FS-1 with other baselines in Figure 4 show significant gains in accuracies across many datasets. FS-1 has Win/Tie/Loss (W/T/L) counts of 26/2/13 when compared to the best non-few-shot-learning model, i.e. ResNet. On 27/41 datasets, FS-1 is amongst the top-2 models. Refer Table 1 for dataset-wise detailed results. Our approach FS-2 with a simpler update rule than FS-1 is the second best model but is very closely followed by the ResNet models trained from scratch.

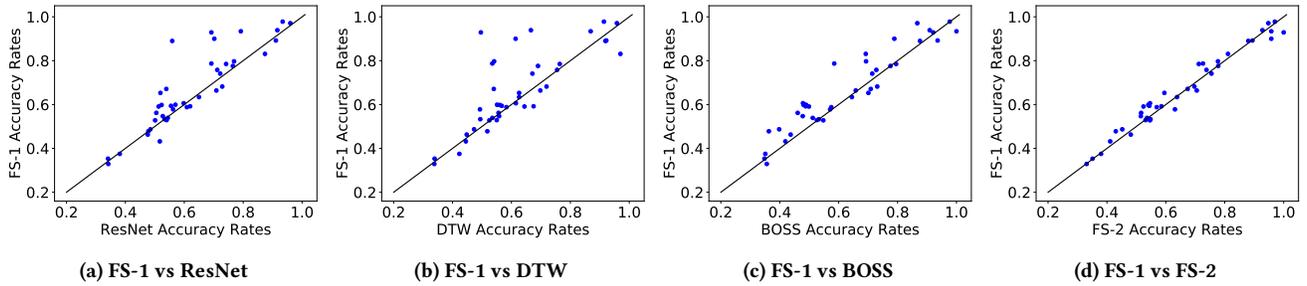


Figure 4: Classification accuracy rates comparison for 5-shot UTSC. Each point in a scatter plot corresponds to a dataset.

Table 1: Comparison of classification accuracy rates for 5-shot learning scenario. Best approach is marked in bold and second-best is underlined. N denotes the number of classes.

Dataset Name	N	ED	DTW	BOSS	ResNet	FS-2 (ours)	FS-1 (ours)	Dataset Name	N	ED	DTW	BOSS	ResNet	FS-2 (ours)	FS-1 (ours)
50words	50	0.483	0.644	0.499	0.513	0.524	<u>0.591</u>	InsectW.B.Sound	11	0.489	0.473	0.398	0.485	0.452	<u>0.487</u>
Adiac	37	0.538	0.540	0.709	0.539	<u>0.674</u>	0.671	Meat	3	0.919	0.919	0.876	0.559	0.880	<u>0.890</u>
Beef	5	0.618	0.626	0.701	0.519	0.595	<u>0.653</u>	MedicalImages	10	0.579	0.675	0.488	<u>0.620</u>	0.585	0.592
BeetleFly	2	0.667	0.614	0.789	0.702	0.958	<u>0.900</u>	Mid.Phal.O.A.G	3	0.529	0.558	0.478	0.527	0.515	0.547
BirdChicken	2	0.468	0.496	0.921	0.692	1.000	<u>0.929</u>	Mid.Phal.O.C	2	0.563	<u>0.550</u>	0.526	0.540	0.531	0.529
Chlor.Conc.	3	0.339	0.338	0.356	0.342	0.331	0.329	Mid.Phal.TW	6	0.338	0.339	0.348	0.341	<u>0.351</u>	0.353
Coffee	2	0.920	0.914	<u>0.977</u>	0.934	0.970	0.978	PhalangesO.C	2	0.532	0.535	0.512	0.544	0.536	<u>0.539</u>
Cricket_X	12	0.348	<u>0.567</u>	0.491	0.555	0.544	0.594	Prox.Phal.O.A.G	3	0.692	0.719	0.731	<u>0.729</u>	0.697	0.682
Cricket_Y	12	0.375	<u>0.556</u>	0.461	0.505	0.516	0.562	Prox.Phal.O.C	2	0.633	0.626	<u>0.645</u>	0.65	0.638	0.634
Cricket_Z	12	0.357	<u>0.560</u>	0.481	0.523	0.541	0.598	Prox.Phal.TW	6	0.427	<u>0.445</u>	0.419	0.517	0.411	0.432
Dist.Phal.O.A.G	3	0.710	0.698	0.658	<u>0.709</u>	0.705	0.664	Strawberry	2	0.682	0.671	0.714	0.722	0.755	<u>0.741</u>
Dist.Phal.O.C	2	0.571	0.583	0.575	0.609	0.569	<u>0.588</u>	SwedishLeaf	15	0.599	0.690	0.776	0.765	0.778	<u>0.776</u>
Dist.Phal.TW	6	0.444	0.448	0.437	<u>0.476</u>	0.481	0.463	synthetic_control	6	0.736	0.958	0.867	0.96	0.948	0.971
ECG200	2	0.771	0.755	0.728	0.712	0.738	<u>0.758</u>	Two_Patterns	4	0.361	0.970	0.692	<u>0.874</u>	0.811	0.831
ECG5000	5	0.524	0.494	<u>0.533</u>	0.533	0.548	<u>0.533</u>	uWave_X	8	0.591	0.615	0.479	0.598	0.546	0.606
ECGFiveDays	2	0.685	0.666	0.909	0.916	<u>0.928</u>	0.939	uWave_Y	8	0.504	0.518	0.363	<u>0.478</u>	0.430	<u>0.478</u>
ElectricDevices	7	0.239	0.423	0.351	<u>0.381</u>	0.380	0.375	uWave_Z	8	0.536	0.551	0.489	<u>0.57</u>	0.541	0.599
FaceAll	14	0.545	0.764	0.795	0.742	0.712	<u>0.785</u>	wafer	2	<u>0.922</u>	<u>0.922</u>	0.936	0.911	0.894	0.892
FaceFour	4	0.812	0.869	1.000	0.792	<u>0.958</u>	0.934	Wine	2	0.496	0.493	0.571	0.562	0.631	0.578
FordA	2	0.561	0.541	0.693	0.769	<u>0.777</u>	0.797	yoga	2	0.505	0.525	0.548	0.501	0.546	0.528
FordB	2	0.515	0.535	0.585	0.692	<u>0.726</u>	0.787	W/T/L of FS-1		32/0/9	27/0/14	30/2/9	26/2/13	24/0/17	-
								Mean Arithmetic Rank		4.537	3.463	3.890	3.305	3.244	2.561

- To study the effect of number of training samples per class available in end task, we consider $K = \{2, 5, 10, 20\}$ for \mathcal{D}^{tr} (while \mathcal{D}^{te} remains the same with $K' = 5$), and experiment under same protocol of 4100 tasks (with 100 tasks sampled from each of the 41 datasets). As observed by ranks comparison in Table 2,
 - FS-1 is the best performing model, especially for 5 and 10-shot scenarios with large gaps in ranks.
 - When considering very small number of training samples per class, i.e. for $K = 2$, we observe that FS-1 is still the best model although it is very closely followed by DTW. This is expected as given just two samples per class, it is very difficult to effectively learn any data distribution patterns, especially when the domain of the task is unseen while training. The fact that FS-1 and FS-2 still perform significantly better than ResNet models trained from scratch show the generic nature of filters learned in ϕ^* . As expected, data-intensive machine learning and deep learning models like BOSS and ResNet that are trained from scratch only on the target task data tend to overfit, and are even worse than DTW.

- For tasks with larger number of training samples per class, i.e. $K = 20$, FS-1 is still the best algorithm. As expected, machine learning based state-of-the-art model BOSS performs better than other baselines when sufficient training samples are available and is closer to FS-1.
- To study the generalizability of FS-1 to varying N as a result of leveraging triplet loss, we group the datasets based on N . As shown in Table 3, we observe that FS-1 is consistently amongst the top-2 models across values of N . While FS-1 is significantly better than other algorithms for $2 \leq N \leq 5$ and $N > 10$, it is as good as the best algorithm DTW for $6 \leq N \leq 9$.

6.2.1 Importance of fine-tuning different layers in deep ResNet. We also study the importance of fine-tuning different convolutional layers of FS-1. We consider four variants FS-1- l with $l = 1, 2, 3, 4$, where we freeze parameters of lowermost l convolutional layers of the pre-trained model, while fine-tuning top $L - l$ layers only. From Figure 6, we observe that FS-1-1, i.e. where the filter weights of only the first convolutional layer are frozen while those of all higher layers are fine-tuned, performs better than the default FS-1 model where all layers are fine-tuned. On the other hand, freezing higher layers as well (FS-1-2 and FS-1-3) or freezing all the layers

(FS-1-4, i.e. no fine-tuning on target task) leads to significant drop in classification performance. These results indicate that the first layer has learned generic features while being trained on diverse set of K -shot tasks and that the higher layers of the FS-1 model are important to quickly adapt to the target K -shot task.

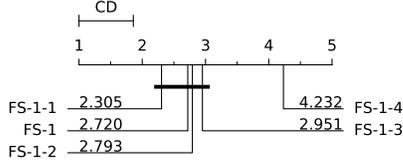


Figure 6: Effect of freezing parameters of different layers while fine-tuning for target few-shot task using FS-1.

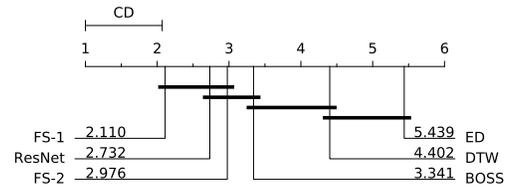
Table 4: Results on 5-shot 5-way classification tasks using dataset-specific pre-training.

Dataset	ED	DTW	BOSS	ResNet	FS-2	FS-1
50Words	0.614	0.812	0.713	0.733	0.719	0.784
Adiac	0.723	0.692	0.791	0.652	0.808	0.827
ShapesAll	0.854	0.897	<u>0.942</u>	0.915	0.924	0.958

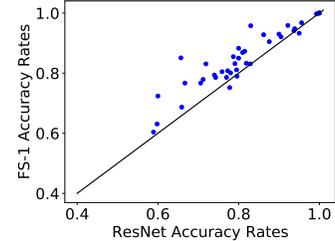
6.2.2 Few-shot learning to adapt to new classes for a given dataset. Apart from the above scenario where the UCR datasets used to sample tasks in training, validation and testing meta-sets are different, we also consider a scenario (similar to [38]) where there are a large number of classes within a TSC dataset, and the goal is to quickly adapt to a new set of classes given a model that has been pre-trained on another disjoint set of classes from the same dataset.

We consider three datasets with large number of classes from the UCR Archive, namely, 50Words, Adiac and ShapesAll, containing 50, 37, and 60 classes, respectively. We use half of the classes (randomly chosen) to form the training meta-set, 1/4th of the classes for validation meta-set, and remaining 1/4th of the classes for testing meta-set. We train the FS-1 and FS-2 models on 5-shot 5-way TSC tasks from training meta-set for $M = 50$ and $B = 5$. We chose the best meta-iteration based on average triplet loss on the validation meta-set (also containing 5-shot 5-way classification tasks). Note that ED, DTW and BOSS are trained on the respective task from the testing meta-set only. Also, whenever number of samples for a class is less than 5, we take all samples for that class in all tasks. The average classification accuracy rates on 100 5-shot 5-way tasks from the testing meta-set are shown in Table 4. We observe that FS-1 outperforms other approaches indicating the ability to quickly generalize to new classes for a given domain.

6.2.3 Non-few-shot learning scenario. We also evaluate FS-1 when sufficient labeled data is available for training, i.e. the standard non-few-shot learning scenario with original class distributions and train-test splits as provided in [5]. As shown in Figure 7a, we observe that the meta-learned FS-1 outperforms other approaches even in non-few-shot scenarios proving the benefit of meta-learning based initialization. Furthermore, when compared to the results in Figure 5, we observe increased performance gap between the deep learning approaches (FS-1, FS-2 and ResNet) and other approaches



(a) Critical Difference Diagram



(b) FS-1 vs second best method (ResNet)

Figure 7: Non-few-shot learning scenario using original train-test splits from UCR Archive.

(BOSS, DTW, ED) due to availability of sufficient training data. We provide scatter-plot comparison for FS-1 with second best approach ResNet in Figure 7b and omit other dataset-wise results for lack of space.

7 CONCLUSION AND FUTURE WORK

The ability to quickly adapt to any given time series classification task with a small number of labeled samples is an important task with several practical applications. We have proposed a meta-learning approach for few-shot time series classification (TSC). It can also be seen as a data-efficient metric learning mechanism that leverages a pre-trained model. We have shown that it is possible to train a model on few-shot tasks from diverse domains such that the model gathers an ability to quickly generalize and solve few-shot tasks from previously unseen domains. By leveraging the triplet loss, we are able to generalize across classification tasks with different number of classes.

We hope that this work opens a promising direction for future research in meta-learning for time series modeling. In this work, we have explored first-order meta-learning algorithms. In future, it would be interesting to explore more sophisticated meta-learning algorithms such as [11, 12, 30] for the same. A similar approach for time series forecasting will be interesting to explore as well.

REFERENCES

- [1] Abubakar Abid and James Y Zou. 2018. Learning a Warping Distance from Unlabeled Time Series Using Sequence Autoencoders. In *Advances in Neural Information Processing Systems*. 10547–10555.
- [2] Anthony Bagnall, Jason Lines, Aaron Bostrom, James Large, and Eamonn Keogh. 2017. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery* 31, 3 (2017), 606–660.
- [3] Anthony Brunel, Johanna Pasquet, Jérôme Pasquet, Nancy Rodriguez, Frédéric Comby, Dominique Fouchez, and Marc Chaumont. 2019. A CNN adapted to time series for the classification of Supernovae. *arXiv preprint arXiv:1901.00461* (2019).
- [4] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. 2018. Recurrent neural networks for multivariate time series with missing

- values. *Scientific reports* 8, 1 (2018), 6085.
- [5] Yanping Chen, Eamonn Keogh, Bing Hu, Nurjahan Begum, et al. 2015. The UCR Time Series Classification Archive. www.cs.ucr.edu/~eamonn/time_series_data/.
 - [6] Zhicheng Cui, Wenlin Chen, and Yixin Chen. 2016. Multi-scale convolutional neural networks for time series classification. *arXiv preprint arXiv:1603.06995* (2016).
 - [7] Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research* 7, Jan (2006), 1–30.
 - [8] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. 2018. Data augmentation using synthetic data for time series classification with deep residual networks. *arXiv preprint arXiv:1808.02455* (2018).
 - [9] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. 2018. Deep learning for time series classification: a review. *arXiv preprint arXiv:1811.01533* (2018).
 - [10] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. 2018. Transfer learning for time series classification. *arXiv preprint arXiv:1901.10738* (2019).
 - [11] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 1126–1135.
 - [12] Chelsea Finn, Kelvin Xu, and Sergey Levine. 2018. Probabilistic model-agnostic meta-learning. In *Advances in Neural Information Processing Systems*. 9516–9527.
 - [13] Jean-Yves Franceschi, Aymeric Dieuleveut, and Martin Jaggi. 2019. Unsupervised Scalable Representation Learning for Multivariate Time Series. *arXiv preprint arXiv:1901.10738* (2019).
 - [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
 - [15] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).
 - [16] Young-Seon Jeong, Myong K Jeong, and Olufemi A Omitaomu. 2011. Weighted dynamic time warping for time series classification. *Pattern Recognition* 44, 9 (2011), 2231–2240.
 - [17] Fazle Karim, Somshubra Majumdar, Houshang Darabi, and Shun Chen. 2018. LSTM fully convolutional networks for time series classification. *IEEE Access* 6 (2018), 1662–1669.
 - [18] Kathan Kashiparekh, Jyoti Narwariya, Pankaj Malhotra, Lovekesh Vig, and Gautam Shroff. 2019. ConvTimeNet: A Pre-trained Deep Convolutional Neural Network for Time Series Classification. In *Neural Networks (IJCNN), 2019 International Joint Conference on*. IEEE.
 - [19] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)* (2015).
 - [20] Arthur Le Guennec, Simon Malinowski, and Romain Tavenard. 2016. Data Augmentation for Time Series Classification using Convolutional Neural Networks. In *ECML/PKDD Workshop on Advanced Analytics and Learning on Temporal Data*.
 - [21] Min Lin, Qiang Chen, and Shuicheng Yan. 2013. Network in network. *arXiv preprint arXiv:1312.4400* (2013).
 - [22] Pankaj Malhotra, Vishnu TV, Lovekesh Vig, Puneet Agarwal, and Gautam Shroff. 2017. TimeNet: Pre-trained deep recurrent neural network for time series classification. In *25th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. 607–612.
 - [23] Jiangyuan Mei, Meizhu Liu, Yuan-Fang Wang, and Huijun Gao. 2015. Learning a mahalanobis distance-based dynamic time warping measure for multivariate time series classification. *IEEE transactions on Cybernetics* 46, 6 (2015), 1363–1374.
 - [24] Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999* (2018).
 - [25] Kaushal Paneri, Vishnu TV, Pankaj Malhotra, Lovekesh Vig, and Gautam Shroff. 2019. Regularizing Fully Convolutional Networks for Time Series Classification by Decorrelating Filters. *The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)* (2019).
 - [26] Wenjie Pei, David MJ Tax, and Laurens van der Maaten. 2016. Modeling time series similarity with siamese recurrent networks. *arXiv preprint arXiv:1603.04713* (2016).
 - [27] Amy Perfors and Joshua Tenenbaum. 2009. Learning to learn categories. Cognitive Science Society.
 - [28] Pranav Rajpurkar, Awni Y Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Y Ng. 2017. Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv preprint arXiv:1707.01836* (2017).
 - [29] Subhrajit Roy, Isabell Kiral-Kornek, and Stefan Harrer. 2018. ChronoNet: A Deep Recurrent Neural Network for Abnormal EEG Identification. *arXiv preprint arXiv:1802.00308* (2018).
 - [30] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. 2019. Meta-learning with latent embedding optimization. In *International Conference on Learning Representations (ICLR)*.
 - [31] Andrew M Saxe, James L McClelland, and Surya Ganguli. 2013. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120* (2013).
 - [32] Patrick Schäfer. 2015. The BOSS is concerned with time series classification in the presence of noise. *Data Mining and Knowledge Discovery* 29, 6 (2015), 1505–1530.
 - [33] Patrick Schäfer and Mikael Höggqvist. 2012. SFA: a symbolic fourier approximation and index for similarity search in high dimensional datasets. In *Proceedings of the 15th International Conference on Extending Database Technology*. ACM, 516–527.
 - [34] Jürgen Schmidhuber. 1987. *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook*. Ph.D. Dissertation. Technische Universität München.
 - [35] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.
 - [36] Joan Serra, Santiago Pascual, and Alexandros Karatzoglou. 2018. Towards a universal neural network encoder for time series. *arXiv preprint arXiv:1805.03908* (2018).
 - [37] Taras K Vintsyuk. 1968. Speech discrimination by dynamic programming. *Cybernetics and Systems Analysis* 4, 1 (1968), 52–57.
 - [38] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Tim Wierstra, et al. 2016. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*. 3630–3638.
 - [39] Haozhou Wang, Han Su, Kai Zheng, Shazia Sadiq, and Xiaofang Zhou. 2013. An effectiveness study on trajectory similarity measures. In *Proceedings of the Twenty-Fourth Australasian Database Conference-Volume 137*. 13–22.
 - [40] Zhiguang Wang, Weizhong Yan, and Tim Oates. 2017. Time series classification from scratch with deep neural networks: A strong baseline. In *Neural Networks (IJCNN), 2017 International Joint Conference on*. IEEE, 1578–1585.
 - [41] Li Wei and Eamonn Keogh. 2006. Semi-supervised time series classification. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 748–753.
 - [42] Kilian Q Weinberger, John Blitzer, and Lawrence K Saul. 2006. Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems*. 1473–1480.
 - [43] Jianbo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiao Li Li, and Shonali Krishnaswamy. 2015. Deep convolutional neural networks on multichannel time series for human activity recognition. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.
 - [44] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks?. In *Advances in neural information processing systems*. 3320–3328.